Einsatz ingenieurwissenschaftlicher Methoden in der Systembiologie

Habilitationsschrift

von Dr.-Ing. Andreas Kremling
geb. am 5. Mai 1965 in Lahr/Schwarzwald

zur Verleihung des akademischen Grades

Doktor-Ingenieur habilitatus
(Dr.-Ing. habil.)

genehmigt von der Fakultät für Verfahrens- und Systemtechnik
der Otto-von-Guericke-Universität Magdeburg am 7. April 2009

Gutachter:

Prof. Dr.-Ing. h.c. mult. E. D. Gilles
Max-Planck-Institut für Dynamik komplexer technischer Systeme, Magdeburg

Prof. Dr.-Ing. U. Reichl
Max-Planck-Institut für Dynamik komplexer technischer Systeme, Magdeburg, und Institut für Verfahrenstechnik, Otto-von-Guericke-Universität, Magdeburg

Prof. Dr. rer. nat. W. Wiechert
Institut für Biotechnologie, Forschungszentrum Jülich

Vorwort

Die vorliegende Arbeit wurde am Max-Planck-Institut für Dynamik komplexer technischer Systeme in Magdeburg angefertigt.

Ich danke Herrn Prof. Gilles für die Gelegenheit, die Arbeit in der Fachgruppe Systembiologie durchführen zu können und für die Übernahme des Gutachtens. Besonderen Dank auch an Herrn Prof. Reichl und an Herrn Prof. Wiechert für das Erstellen der Gutachten.

Systembiologische Forschung ist immer interdisziplinär. Die erstellten mathematischen Modelle, die in der Arbeit vorgestellt werden, basieren auf einer großen Anzahl von Experimenten die im molekularbiologischen Labor des Instituts durchgeführt wurden. Daher möchte ich mich besonders herzlich bei Frau Dr. Bettenbrock und ihren Mitarbeiterinnen Frau Focke und Frau Tietgens für die gute Zusammenarbeit bedanken. Mit Milind Joshi habe ich einen sehr guten und ehrgeizigen Wissenschaftler kennengelernt bei dem ich mich ebenfalls für die sehr intensive und erfolgreiche Zusammenarbeit bedanken möchte. Für die große Unterstützung während ihrer Zeit am MPI möchte ich mich auch bei meiner Frau Sophia bedanken. Ebenfalls ein herzliches Dankeschön an alle Mitarbeiter der Fachgruppe Systembiologie am MPI für anregende und interessante Gespräche.

Magdeburg im April 2009
Andreas Kremling

1 Einleitung

Die Systembiologie ist eine junge Forschungsrichtung, die besonders in den letzten Jahren einen rasanten Aufstieg erlebt hat (Stelling et al., 2001). Systembiologische Forschung zeichnet sich durch interdisziplinäre Ansätze aus den Bereichen Biologie/Medizin auf der einen Seite und Mathematik und Ingenieurwissenschaften auf der anderen Seite aus. Diese Forschung wird unterstützt durch Methoden aus den Informationswissenschaften, um die anfallenden Datenmengen sinnvoll zu strukturieren und in Datenbanken zur Verfügung zu stellen.

Die Zielsetzungen in der Systembiologie sind auf ein verbessertes Verständnis der in einer lebenden Zelle ablaufenden Prozesse ausgerichtet. Diese Prozesse lassen sich am besten durch biochemische Reaktionsnetzwerke beschreiben. Diese umfassen unterschiedliche Reaktionstypen wie enzymkatalysierte Reaktionen und Polymerisationsreaktionen. Im Gegensatz zu chemischen Reaktionsnetzwerken zeichnen sich die biochemischen Reaktionsnetzwerke durch eine große Anzahl von Rückkopplungschleifen aus. Das bedeutet, dass Komponenten in vielfältiger Art und Weise miteinander interagieren, so dass das zeitliche Verhalten des Gesamtsystems intuitiv schwer nachvollziehbar ist. Hier kommt das Hilfsmittel der mathematischen Modellierung zum Einsatz, das es erlaubt, die Vorgänge in der einzelnen Zelle zu abstrahieren und damit einer theoretischen Analyse zugänglich zu machen.

In der Systembiologie lassen sich nun zwei unterschiedliche und sich ergänzende Ansätze finden, die zu mathematischen Modellen führen. Im Bottom-up-Ansatz geht man von einem kleinen Teilsystem aus, für welches biologisches Wissen aus der Literatur bekannt ist, das ausreicht, ein mathematisches Modell aufzustellen. Wie in Abbildung 1.1 links gezeigt, kann dieses Modell dann durch experimentelle Daten verifiziert und anschließend analysiert werden. Weiterhin kann es mit anderen Teilmodellen zu einer größeren Einheit verschaltet werden. Im experimentell orientierten Top-down-Ansatz liegt ein Gesamtbild der zellulären Aktivität bspw. in Form von cDNA Array Daten vor (Abbildung 1.1 rechts). Die Daten beschreiben also die Gesamtheit aller in den Zellen gebildeten mRNA und machen daher eine Aussage über die

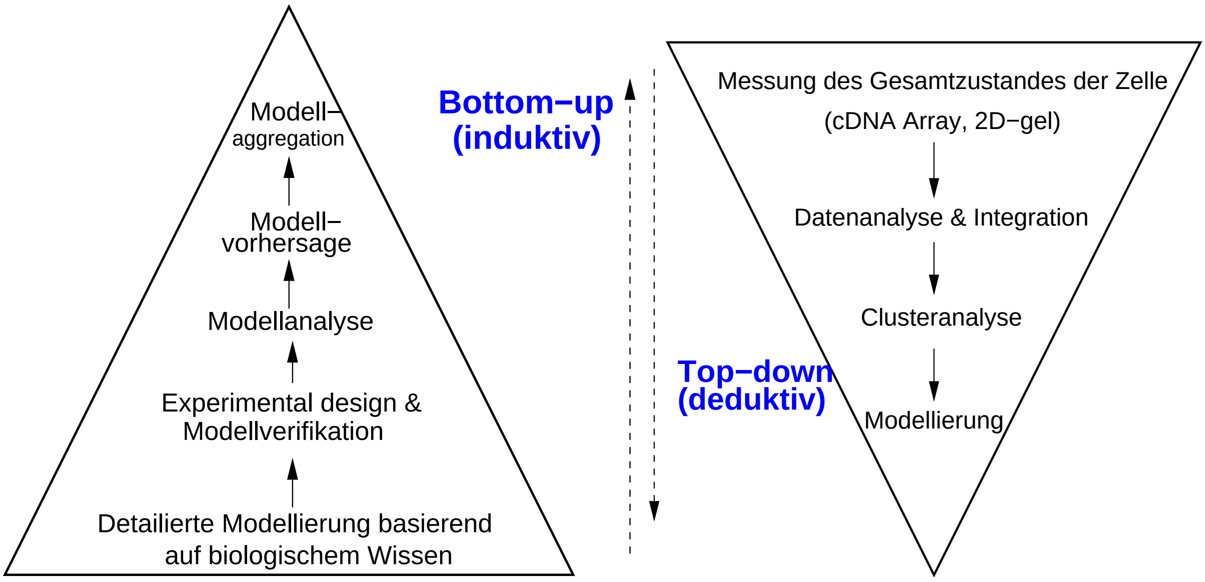


Abbildung 1.1: Zwei Vorgehensweisen in der Systembiologie. Links: Bottom-up, rechts: Top-down. In dieser Arbeit werden Methoden aus dem Bottom-up-Ansatz vorgestellt.

gesamte Transkriptionsaktivität der Zellen. Diese Daten werden mit geeigneten Tools analysiert und mit anderen Datentypen zu einem Gesamtbild der Zelle integriert. Basierend auf verschiedenen Techniken lassen sich dann aus den Daten ebenfalls mathematische Modelle ableiten.

Die vorliegende Arbeit stellt Methoden aus dem Bereich des Bottom-up-Ansatzes vor, die Aspekte der Modellerstellung, Verifizierung, Analyse und Versuchsplanung umfassen. Da die Modelle sehr umfangreich sind, erfolgt die Modellierung rechnergestützt, d.h., eine enstprechende Software kommt zum Einsatz, die den Modellierer bei der Eingabe des Netzwerkes unterstützt und gleichzeitig Schnittstellen zu Simulationswerkzeugen bereitstellt (Ginkel et al., 2003; Hucka et al., 2003). Die Modellverfikation erfolgt in enger Zusammenarbeit mit biologisch arbeitenden Gruppen, mit denen die Experimente gemeinsam geplant werden. Zielsetzung war in einem ersten Schritt zu zeigen, dass umfangreiche Modelle aufgestellt werden können, die in der Lage sind, mit einem einzigen Satz von kinetischen Parametern eine Vielzahl von experimentellen Bedingungen zu beschreiben (Kremling et al., 2003; Bettenbrock et al., 2006; Kremling, 2007). Die Vorgehensweise zeigt, dass die abgebildeten Prozesse mit einer hohen Güte der Modelle beschrieben werden können. Da im Prokaryontenbereich die ablaufenden Prozesse ähnlich strukturiert sind, kann davon ausgegangen werden, dass die Modelle als Grundlage einer Modellerstellung der ganzen Gruppe dienen kann.

In einem zweiten Schritt wurden dann Eigenschaften der Signalumwandlung und -verarbeitung für ein ausgewähltes System näher untersucht (Kremling et al., 2004b). Bei Bakterien unterscheidet man zwischen der für einen spezifischen Reiz beobachteten Antwort (lokale Kontrolle) und der bei einer allgemeinen Stimulation beobachteten Antwort (globale Kontrolle). Für das hier betrachtete Teilnetzwerk der Kohlenhydrataufnahme stellt eine Hungersituation eine allgemeine Stimulation dar, was sich durch eine veränderte Wachstumsrate der Zellen bemerkbar macht. Zur Analyse der Eigenschaften des Sensors und des Signalverarbeitungsweges wurde eine vereinfachte Modellstruktur verwendet und analysiert (Kremling et al., 2007, 2008). Für ein zweites Beispielsystem – ein Zwei-Komponenten-System – wurde ebenfalls eine umfassende Analyse des erstellten mathematischen Modells durchgeführt. Mit den Ergebnissen konnten zwei bisher noch nicht charakterisierte Rückkopplungsschleifen hinsichtlich ihrer kinetischen Eigenschaften analysiert werden (Kremling et al., 2004d; Saez-Rodriguez et al., 2004, 2005).

In vielen Fällen reichen vorliegende experimentelle Daten nicht aus, um Modelle ausreichend gut zu verifizieren. Daher können Parameter oft nur sehr ungenau, d.h., mit einer hohen Unsicherheit ermittelt werden. In anderen Fällen liegen Modellvarianten vor, die ein bestimmtes Experiment gleich gut wiedergeben. In beiden Fällen ist es notwendig, ein neues Experiment vorzuschlagen. In der vorliegenden Arbeit liegt der Schwerpunkt der Methodenentwicklung zunächst auf der Planung neuer Experimente, wenn zwischen zwei Modellvarianten unterschieden werden muß (Kremling et al., 2004a). Ausgehend von in der Literatur beschriebenen Ansätzen wird eine Methode vorgeschlagen, die sowohl Unsicherheiten der Messgrößen, als auch aus dem vorliegenden Experiment ermittelte Parametervarianzen berücksichtigt. Bei der Ermittlung der Parametervarianzen zeigte sich, dass nichtlineare Effekte (Nichtlinearität bezüglich der Parameter) eine wichtige Rolle spielen. Da klassische Verfahren die Parametervarianzen nur unzureichend abschätzen, kommt hier ein neuer Ansatz, basierend auf einer statistischen Methode zum Einsatz, der wesentlich bessere Resultate liefert, als die klassische Vorgehensweise (Joshi et al., 2006a,b).

Methoden aus den Ingenieurwissenschaften eröffnen in der Systembiologie zahlreiche Möglichkeiten, mathematische Modelle schnell und effizient zu erstellen, zu verifizieren und zu analysieren. Die vorliegende Arbeit betrachtet dazu einige ausgewählte Beispiele. Eine ausführliche Darstellung von ingenieurwissenschaftlichen Methoden in der Systembiologie ist in einem

Übersichtsartikel zu finden (Kremling and Saez-Rodriguez, 2007).

Die Arbeit gliedert sich in drei Teile. In einem ersten Teil wird ein umfassendes Modell der Kohlenhydrataufnahmesysteme für *Escherichia coli* vorgestellt und diskutiert. Im zweiten Teil steht die Analyse eines speziellen Signalweges im Zentrum, der zeigt, dass eine bestimmte Verschaltung im Netzwerk zu einem robusten Verhalten des Systems führt. Im dritten Teil stehen Methoden zur Versuchsplanung und zur Ermittlung von Parameterunsicherheiten im Mittelpunkt.

2 Rechnergestützte Modellierung von Kohlenhydrataufnahmesystemen bei *Escherichia coli*

Das Bakterium *Escherichia coli* besitzt eine ganze Reihe von Transportsystemen, die es erlauben, Substrate aus dem Medium heraus aufzunehmen. Für die Klasse der Kohlenhydrate umfassen diese Transportsysteme ein membranständiges Protein, welches für die eigentliche Aufnahme verantwortlich ist sowie weitere Proteine, die bei einigen Substraten für eine Modifikation des Substrates, beispielsweise eine Phosphorylierung, sorgen (eine Übersicht über Transportsysteme ist in Postma et al. (1996) zu finden). Die Transportsysteme sind in der Regel spezifisch und besitzen daher nur ein kleines Substratspektrum. Damit die Zelle nun nicht alle Systeme vorhalten muß, was ökonomisch betrachtet auch nicht sinnvoll wäre, werden diese Systeme erst bereitgestellt, wenn das betreffende Substrat im Medium vorliegt. Legt man nun in einer Batch-Kultur ein Substrat vor, so stellt sich nach kurzer Zeit eine konstante Wachstumsrate ein. Die Wachstumsrate für verschiedene Kohlenhydrate variiert sehr stark. Dies bedeutet, dass die Zelle bestimmte Kohlenhydrate besser verwerten kann als andere. Aus molekularbiologischen Untersuchungen ist nun bekannt, dass bei der Synthese fast aller Kohlenhydrat-Transportsysteme neben einer spezifischen Kontrolle auch das Regulatorprotein Crp beteiligt ist. Es sorgt dafür, dass die entsprechenden Proteine bei Bedarf exprimiert werden. Da dieser Regulator bei der Genexpression einer großen Anzahl von Genen beteilig ist, spricht man von einem globalen Regulator (Neidhardt et al., 1990). Die einzelnen Transportsysteme sind hinsichtlich ihrer qualitativen Eigenschaften recht gut untersucht. Dies gilt auch für die übergeordnete Koordination der Transportsysteme. Es fehlt allerdings eine Beschreibung auf einer quantitativen Ebene.

Interessante Beobachtungen werden auch gemacht, wenn zwei Substrate als Mischung in der Kultur vorgelegt werden. Für die Kombination der beiden Zucker Laktose und Glukose wird beispielsweise beobachtet, dass die Aufnahme nicht gleichzeitig erfolgt, sondern zuerst die Glukose verstoffwechselt wird und erst nach fast vollständigem Verbrauch die Laktose aufgenommen wird. Das bedeutet, dass die einzelnen Systeme auch untereinander vernetzt sind.

Von einer quantitativen Beschreibung der übergeordneten Koordination der Transporter und der Interaktionen zwischen den (lokalen) Transportern kann erwartet werden, dass sie zu einem besseren Verständnis von zellulären Regulationsprinzipien führen. Damit ergeben sich zukünftig neue Möglichkeiten, das Potential von Mikroorganismen, beispielsweise in der Biotechnologie, effizienter zu nutzen als bisher. Im folgenden soll zunächst ein umfangreiches mathematisches Modell vorgestellt werden, welches Aufnahme und Stoffwechsel mehrerer Kohlenhydrate beschreibt und sowohl die Aktivitätsregulation einzelner enzymkatalysierte Reaktionen als auch die Regulation der Genexpression der Enzyme umfaßt (Kremling et al., 2003; Bettenbrock et al., 2006). Zur Verifikation des Modells liegen eine ganze Reihe von Experimenten vor, bei denen intra- und extrazellulären Messgrößen erfasst wurden. Das Modell wurde mit dem Modellierungswerkzeug ProMoT erstellt, das eine graphische Benutzeroberfläche besitzt und es erlaubt, einzelne Teilmodelle auszuwählen, zunächst zu parametrieren und dann mit anderen Teilmodellen zum Gesamtmodell zu verschalten (Ginkel et al., 2003). In einem zweiten Schritt wurde ein vereinfachtes Modell herangezogen, um spezielle Eigenschaften der Signalübertragung zu analysieren (Kaptiel 3).

2.1 Modell der Kohlenhydrataufnahme

Der Aufbau des Gesamtmodells zur Beschreibung der Kohlenhydrataufnahme ist den Abbildungen 2.1 und 2.2 zu entnehmen. Zentrales Element ist das bakterielle Phosphotransferase System (PTS), welches Sensor- und Transportfunktion gleichzeitig wahrnimmt (Abbildung 2.1). Zum einen ist es für den Transport einer ganzen Reihe von Kohlenhydraten verantwortlich, zum anderen sind die Proteine an Signaltransduktionswegen beteiligt. Dies gilt für die Aktivierung des globalen Regulators Crp und für die Chemotaxis. Im Falle der Glukoseaufnahme sind bei der Übertragung der Phosphatgruppen die zwei allgemeinen PTS-Proteine, EI und HPr sowie die Proteine EIIA^{Crr} (im folgenden EIIA genannt) und PtsG (EIICB) beteiligt. Die Proteine übertragen Phosphatgruppen vom Metaboliten PEP auf den Zucker. Das bedeutet, dass die Proteine in phosphorylierter oder nicht phosphorylierter Form in der Zelle vorliegen. Allerdings verändert sich je nach Umgebungsbedingung auch die Gesamtmengen der Proteine. Sie unterliegen ebenfalls der Kontrolle des globalen Regulators Crp. Die Aktivierung von Crp erfolgt durch das Alarmon cAMP, ein Molekül, welches aus ATP durch die Adenylatzyklase (Cya) entsteht. Die Adenylatzyklase wird durch die phophorylierte Form des PTS-Proteins EIIA aktiviert und steht ebenfalls unter der Kontrolle von Crp, was deutlich macht, dass es sich nicht um einen einfachen linearen Signalweg handelt, sondern um eine komplexe Signaltransduktionseinheit, mit einer ganzen Reihe von Rückkopplungsschleifen (Lengeler and Jahreis, 1996). Im Modell ist auch eine Regulation des Glukose-

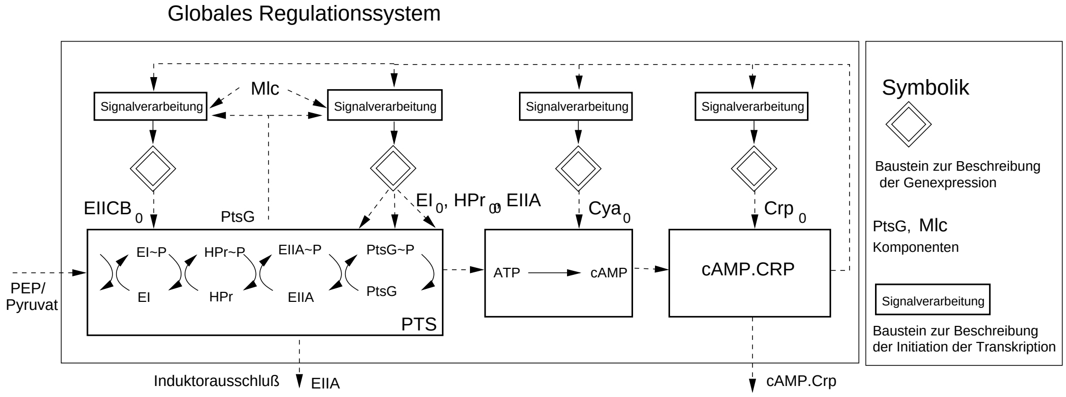


Abbildung 2.1: Übergeordnete globale Signalverarbeitung und Regulation der Kohlenhydrataufnahme. Das PTS übt eine Sensorfunktion aus, die es erlaubt, den Fluss durch die Glycolyse zu erfassen. Ein niedriger Fluss fürht zu einer Aktivierung des globalen Transkriptionsfaktors Crp. Dieser ist in die Genexpression einer großen Anzahl von Genen involviert.

Transporters PtsG berücksichtigt (Plumbridge, 1998). Wie die meisten Transporter wird er über das Regulationsprinzip der Induktion reguliert. Das bedeutet, dass bei Vorliegen von Glukose im Medium durch einen autokatalytischen Prozess das Transportprotein erst gebildet wird. Allerdings unterscheidet sich die Regulation von PtsG von den bisher bekannten Induktionsmustern dadurch, dass nicht ein Metabolit des Stoffwechsels für eine Deaktivierung des Regulatorproteins sorgt, sondern eine Konformation des Transporters selbst mit dem Regulator (Mlc) interagiert.

Abbildung 2.2 macht deutlich, wie sich der Kreis, ausgehend vom Regulator Crp, wieder schließt. Die einzelnen Aufnahmesysteme werden individuell von einem Regulatorprotein reguliert, die Aktivität des Regulators Crp wird dieser Regulation überlagert, wobei die Verrechnung der Signale der verschiedenen Ebenen nach einem in der Literatur vorgeschlagenen

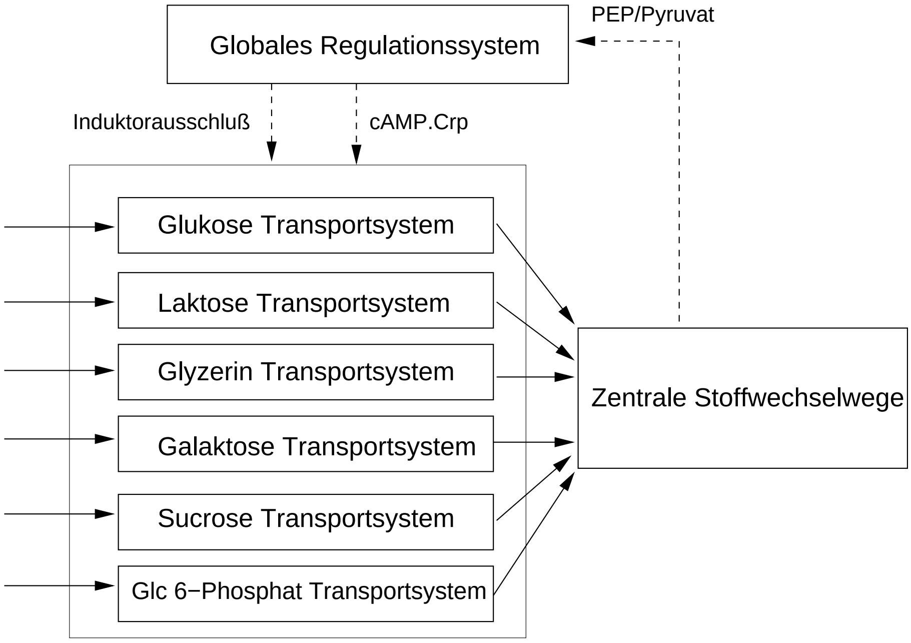


Abbildung 2.2: Individuelle Stoffwechselwege der Kohlenhydrataufnahme. Die Elemente aus Abbildung 2.1 sind im Block "Globales Regulationssystem" zusammengefaßt.

Verfahren erfolgt (Kremling and Gilles, 2001).

Die Modellerstellung erfolgt unter Betrachtung eines gemittelten Verhaltens der Zellpopulation. Intrazelluläre Komponenten können dann formal über folgende Differentialgleichungen beschrieben werden:

$$\dot{c}_i \; = \; \sum_j \gamma_{ij} \, r_j \; - \; \mu \, c_i \, , \tag{2.1}$$

wobei c_i die intrazelluläre Konzentration der Komponente darstellt, γ_{ij} stöchiometrische Koeffizienten sind, r_j intrazelluläre Raten und μ die spezifische Wachstumsrate bedeuten. Die spezifische Wachstumsrate repräsentiert einen Verdünnungsterm, der berücksichtigt, dass sich bei der Zellteilung die Komponenten gleichmäßig auf die beiden Tochterzellen verteilen. Für die Substrate im Medium (Konzentration c_{Si}), die in einem Reaktorsystem (Volumen V_R) durch jeweils einen separaten Zufluss (Rate q_i^{in}, Konzentration c_{Si}^{in}) zudosiert werden können, ergeben sich dann folgende Gleichungen:

$$\dot{c}_{Si} \; = \frac{1}{V_R} \left(q_i^{in} c_{Si}^{in} \; - \; \sum_k q_k^{in} \, c_{Si} \right) \; - \; \sum_j \gamma_{ij} \, {r_j}^t \, g_i \, , \tag{2.2}$$

wobei die Transportraten ${r_j}^t$ in die Zelle hinein oder aus der Zelle heraus berücksichtigt werden. Da diese in der Einheit $[\mu mol/gTM\, h]$[1] angegeben sind, die Substrate aber in $[g/l]$ eingewogen werden, erfolgt noch eine Umrechung mit dem Molekulargewicht g_i der Komponente.

Zur Beschreibung der kinetischen Raten für den Bereich des Stoffwechsels werden klassische Ansätze in Form der Michaelis-Menten-Kinetik verwendet. Diese basieren auf der Annahme, dass das Enzym in mehreren Konformationen vorliegt und einige dieser Konformationen zur Produktbildung beitragen. Nimmt man an, dass die einzelnen Konformationen sich im Gleichgewicht befinden, ergibt sich ein algebraisches System von Gleichungen zur Bestimmung der Reaktionsrate r. Werden in der Literatur keine Angaben über mögliche Mechanismen gefunden, so wird die Michaelis-Menten-Kinetik in der einfachsten Form angenommen. Diese lautet:

$$r \;=\; k_e\, c_{E0}\, \frac{c}{c + K_M}\,, \tag{2.3}$$

wobei k_e die Produktbildungsgeschwindigkeitskonstante des Enzyms ist, c_{E0} die Gesamtenzymmenge, c die Metabolitkonzentration und K_M der Halbsättigungswert. Bei *E. coli* sind eine ganze Reihe von Enzymen des Zentralstoffwechsels gut untersucht, so dass hier auf Ansätze zurückgegriffen werden konnte, die bereits experimentell (*in vitro*) verfiziert sind und damit eine gute Ausgangsbasis darstellen.

Aufwändiger gestaltet sich die Ermittlung der Raten zur Beschreibung der Synthese der Enzyme. Der Prozess der Proteinsynthese besteht aus zwei gekoppelten Polymerisationsprozessen, der Transkription (Abschreiben der auf der DNA gespreicherten Information in mRNA) und der Translation (Umschreiben der mRNA). Formal läst sich folgender Ansatz formulieren, der analog zu oben (Gleichung (2.3)) die Rate r proportional zur Anzahl der Genkopien c_{DNA_0} annimmt:

$$r \;=\; k_s\, c_{DNA_0}\, \eta\,. \tag{2.4}$$

Die Expressionseffizienz η berücksichtigt nun den Einfluss der Regulatorproteine, der RNA-Polymerase und der Ribosomen. Eine ausführliche Darstellung der Berechnungsmöglichkeiten für η ist in Kremling (2007) zu finden. Im Modell sind im wesentlichen nur die Abhängigkeiten von den Regulatorproteinen und der RNA-Polymerase berücksichtigt, da davon ausgegangen wird, dass die Ribosomen keinen limitierenden Einfluss ausüben. In diesem Fall ist η definiert

---

[1] gTM ≡ g Trockenmasse

als das Verhältnis der mit RNA Polymerase belegten Promotoren c_{PD} zur Gesamtzahl der zur Verfügung stehenden Promotoren c_{DNA_0}:

$$\eta \; = \; \frac{c_{PD}}{c_{DNA_0}} \, . \tag{2.5}$$

Aufgrund der vielfältigen Wechselwirkungen zwischen Regulatorproteinen, DNA-Bindestellen, RNA-Polymerase und Promotor ergibt sich ein umfangreiches Reaktionssystem, welches analog der Vorgehensweise bei der Ableitung von Enzymkinetiken in Gleichungen umgesetzt werden muß. Da die Wechselwirkungen für diese Teilnetzwerke sehr schnell im Vergleich zu der Proteinsynthese ablaufen, kann man davon ausgehen, dass in diesem Falle alle Reaktionen im System im Gleichgewicht sind.

Die Berechnung der Genexpressionseffizienz soll an einem Beispiel illustriert werden (Abbildung 2.3). Die oben gezeigte Ebene des globalen Regulationssystems (Abbildung 2.1) beschreibt die Interaktion der RNA-Polymerase und des cAMP·Crp-Komplexes mit den Promotorbindestellen. Die Ausgänge dieser Einheit sind die Größen K_s und K_{ss}, die die Anteile der mit Polymerase allein und der mit Polymerase und dem cAMP·Crp-Komplex belegten Bindestellen beschreiben. Auf der Ebene der individuellen Aufnahmesysteme (Regulonebene) wird diese Information weiter verarbeitet. In einer früheren Arbeit (Kremling and Gilles, 2001) konnte gezeigt werden, dass die Informationsübertragung als einseitig, d.h. rückwirkungsfrei angenommen werden kann.

Die Regulonebene beschreibt im Falle der Glukoseaufnahme die Interaktion von Mlc (c_{Mlc}) und PtsG (c_{PtsG}) mit der Operatorbindestelle ($c_{D_{PtsG}}$). Sobald extrazelluläre Glukose im Medium vorliegt oder intrazelluläre Glukose vorhanden ist, bindet Mlc an den PtsG-Komplex, wie in der Abbildung gezeigt. Damit kann Mlc nicht mehr an den Operator binden, und es kann zum Ablesen der auf der DNA gespeicherten Information kommen. Betrachtet man die Regulonebene, so ergibt sich folgende Erhaltungsgleichung für die DNA Bindestelle:

$$c_{D_{PtsG0}} \; = \; c_{D_{PtsG}} + \frac{c_{D_{PtsG}}}{K_s} + \frac{c_{D_{PtsG}}}{K_{ss}} + \frac{c_{Mlc}\, c_{D_{PtsG}}}{K_b} + \frac{c_{Mlc}\, c_{D_{PtsG}}}{K_b K_{ss}} \tag{2.6}$$

Berücksichtigt werden die Signale aus der Modulonebene und die Anbindung von Mlc. Mlc ist nun in verschiedenen Komplexen gebunden, die bei der Erhaltungsgleichung berücksichtigt werden müssen. Zu beachten ist, dass Mlc an die mit dem cAMP·Crp-Komplex belegte Bindestelle binden kann, jedoch nicht an die mit RNA-Polymerase belegte Bindestelle. Es ergibt

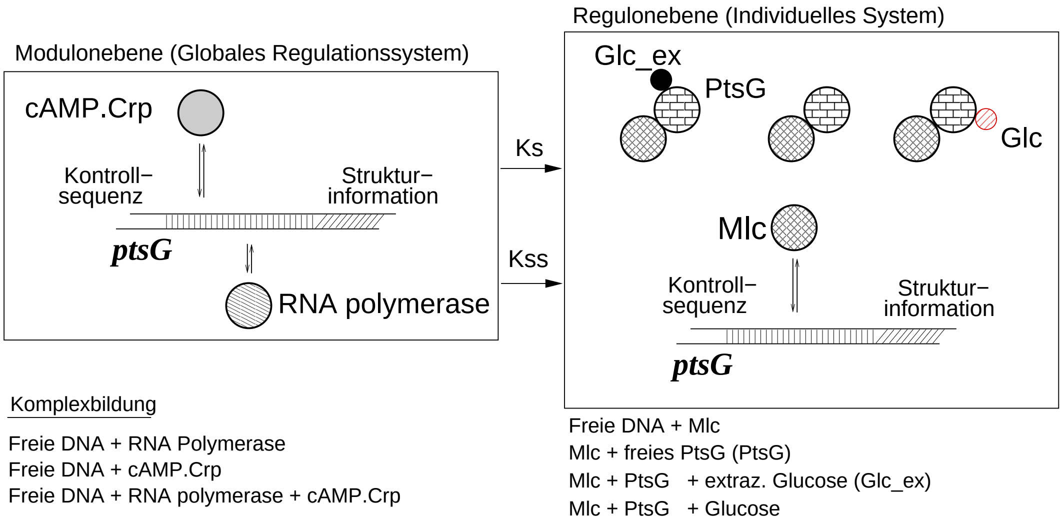


Abbildung 2.3: Modulon- und Regulonebene. Die Modulon-Ebene beschreibt die Interaktion der DNA-Bindestellen mit dem globalen Regulator Crp. Die entsprechenden Belegungsgrade (K_s, K_{ss}) werden an die Regulonebene weitergeleitet. Die Regulonebene beschreibt nun die Interaktionen des spezifischen Regulators Mlc mit der Bindestelle und anderen beteiligten Proteinen (hier PtsG) und Effektoren (hier intra- und extrazelluläre Glukose).

sich:

$$\begin{aligned} c_{Mlc0} &= c_{Mlc} + \frac{c_{PtsG}\; c_{Mlc}}{\alpha\, K_d} + \frac{c_{PtsG}\; c_{Mlc}}{K_d}\left(\frac{c_{Glc}}{K_{Glc}} + \frac{c_{Glc_{ex}}}{K_{Glc_{ex}}}\right) \\ &\quad + \frac{c_{Mlc}\; c_{D_{PtsG}}}{K_b}\left(1 + \frac{1}{K_{ss}}\right), \end{aligned} \tag{2.7}$$

wobei α und die K_i Wechselwirkungsparameter sind. Summiert man nun alle Konformationen für PtsG auf, erhält man:

$$\begin{aligned} c_{PtsG0} &= c_{PtsG} + \frac{c_{PtsG}\; c_{Mlc}}{\alpha\, K_d} + \frac{c_{PtsG}\; c_{Mlc}}{K_d}\left(\frac{c_{Glc}}{K_{Glc}} + \frac{c_{Glc_{ex}}}{K_{Glc_{ex}}}\right) \\ &\quad + \frac{c_{PtsG}\; c_{Glc}}{K_{Glc}}\left(1 + \frac{c_{P\sim EIIA}}{K_{P\sim EIIA}}\right) + \frac{c_{PtsG}\; c_{Glc_{ex}}}{K_{Glc_{ex}}}\left(1 + \frac{c_{P\sim EIIA}}{K_{P\sim EIIA}}\right) \\ &\quad + \frac{c_{PtsG}\; c_{P\sim EIIA}}{K_{P\sim EIIA}}\,. \end{aligned} \tag{2.8}$$

Bei dieser Gleichung ist berücksichtigt, dass PtsG sowohl mit extrazellulärer als auch mit intrazellulärer Glukose sowie mit EIIA interagiert.

Tabelle 2.1 gibt einen Überblick über die Anzahl der Zustandsgrößen, der Reaktionsraten und die Anzahl der kinetischen Parameter für das Gesamtmodell.

Tabelle 2.1: Übersicht Funktionseinheiten (FE). Neben den 6 Zuckern werden in der Flüssigphase die Biomasse, Acetat sowie extrazelluläres cAMP bilanziert (DA: Differential-Algebra System; Dgl.: Differentialgleichungen). Parameter, die nicht geschätzt werden konnten, besitzen nur eine sehr geringe Sensitivität.

Name FE	Parameter		Gleichungen	
	Gesamt	Geschätzt	Gesamt	Typ
Crp Modulon	28	12	14	DA
PTS	22	10	11	DA
Glukose Transport	9	6	2	Dgl.
Sukrose Transport	18	3	5	Dgl.
Laktose Transport	13	8	6	DA
Glyzerin Transport	14	7	5	Dgl.
Galactose Transport	22	4	10	DA
Glc 6-Phosphat Transport	10	4	4	DA
Katabole Reaktionen	45	7	8	Dgl.
Flüssigphase			9	Dgl.

2.2 Modellverifikation

Die Bestimmung der kinetischen Parameter erfolgt durch einen Vergleich von Messdaten mit den simulierten Modellausgängen. Dabei wird im Ingenieurbereich oft eine Eingangsfunktion (beispielsweise Sprung, Impuls oder PRBS (Pseudo Random Binary Signal)) auf das System aufgegeben und die Systemantwort vermessen. Im Falle von PRBS Signalen erfordert das Experiment allerdings eine lange Messzeit, was für zelluläre Systeme immer mit großen Schwierigkeiten verbunden ist (unzureichende Sterilität der Anlage, zufällige Mutation bei der

Zellteilung, Erzeugung von Stressbedingungen, die nicht im Modell vorgesehen sind). Aus diesem Grund wurde bei der vorliegenden Arbeit ein anderer Ansatz gewählt. Die Anregung des Systems erfolgte durch Einstellen von vier Einflussgrößen:

i) Vorkultur. Die *E. coli*-Bakterien werden vor Beginn des eigentlichen Experimentes in einem Schüttelkolben angezogen. Die Auswahl der Vorkultur hat eine Auswirkung auf die Menge des spezifischen Transporters, der dann in der Regel bereits in hoher Konzentration in der Zelle vorliegt.

ii) Hauptkultur. Zu Beginn des eigentlichen Experimentes werden ein oder zwei Kohlenhydrate im Medium vorgelegt. Durch Vergleich mit Experimenten mit Einzelsubstraten kann analysiert werden, ob im Falle von zwei Substraten die beiden Transportsysteme miteinander in Wechselwirkung stehen.

iii) Stammvariante/Mutantenstämme. Durch Vergleich und Analyse des Verhaltens von Stammvarianten wird überprüft, ob Regelkreisstrukturen oder Stoffwechselwege im Modell ausreichend gut abgebildet werden. Die verwendeten Stämme sind alle isogen, was bedeutet, dass sie ausgehend vom gleichen Wildtyp konstruiert sind.

iv) Prozeßführung. Die meisten Experimente wurden im Batch-Betrieb durchgeführt. Durch Zufütterung von Substraten kann das Zeitfenster mit niedrigen Substratkonzentrationen verlängert werden. Eine besondere Variante stellt das "Disturbed"-Batch-Experiment dar. Hier wird die Kultur mit einem Hauptsubstrat bis in die exponentielle Phase hinein angezogen. Dann wird ein zweites Substrat pulsförmig dazugegeben. Wird eine Kultur aus der Ruhelage durch einen Substratpuls ausgelenkt, wird das Experiment als Pulsexperiment bezeichnet.

Tabelle 2.2 stellt die gesamte Datenbasis zusammen. Insgesamt stehen 18 Experimente zur Verfügung. Messtechnisch erfasst wurden sowohl extra- als auch intrazelluläre Größen. Für eine ausführliche Beschreibung der Messmethodik wird auf Bettenbrock et al. (2006) verwiesen. Die Schätzung der kinetischen Parameter erfolgte iterativ, da es nicht sinnvoll erschien und auch technisch nicht durchfürbar war, die Parameter durch Vorgabe aller Experimente mit einer einzigen Optimierungsrechnung zu bestimmen. Iterativ bedeutet, dass für eine Gruppe von Experimenten Parameter geschätzt und dann, bei einer weiteren Optimierung einer zweiten Gruppe von Experimenten, konstant gehalten werden. So sind alle Experimente, bei den Laktose im Medium vorgelegen hat, zusammengefaßt und die entsprechenden

Tabelle 2.2: Übersicht Datenbasis.

	Vorkultur	Hauptkultur	Stammvariante	Prozeßführung
1.	Glukose	Glukose, Laktose	Wildstamm	Batch
2.	Laktose	Glukose, Laktose	Wildstamm	Batch
3.	Laktose	Laktose	Wildstamm	Batch
4.	Glyzerin	Laktose	Wildstamm	Batch
5.	Laktose	Galaktose, Laktose	Wildstamm	Batch
6.	Laktose	Laktose, Puls: Glukose	Wildstamm	"Disturbed"-Batch
7.	Glukose	Glyzerin, Puls: Glukose	Wildstamm	"Disturbed"-Batch
8.	Glukose	Glukose, Laktose	Glk Mutante	Batch
9.	Glukose	Glukose, Glyzerin	Wildstamm	Batch
10.	Glyzerin	Glukose	Wildstamm	Batch
11.	Glyzerin	Glyzerin	Wildstamm	Batch
12.	Glukose	Glukose, Glyzerin	Mlc Mutante	Batch
13.	Glyzerin	Glyzerin	Mlc Mutante	Batch
14.	Glukose	Glukose, Laktose	Mlc Mutante	Batch
15.	Glukose	Glukose, Laktose	LacI Mutante	Batch
16.	Laktose	Glukose, Laktose	PtsG Mutante	Batch
17.	Sukrose	Sukrose	Sucrose$^+$ Stamm	Pulsexperiment
18.	Glukose	Glukose	Wildstamm	Kontin. Kultur

Parameter geschätzt worden. Zu Beginn jeder Optimierung wird, wie an anderer Stelle beschrieben (Posten and Munack, 1990), eine Vorauswahl an Parametern ermittelt, die für die Experimente eine hohe Sensitivität aufweisen. Durch diese Vorgehensweise konnte die Anzahl der relevanten Parameter stark eingeschränkt werden. Insgesamt konnte rund ein Drittel der kinetischen Parameter mit Hilfe der Simulationsumgebung DIVA (Mangold et al., 2000) aus den Experimenten geschätzt werden.

Zur Vorstellung der Ergebnisse soll nur auf zwei Experimente eingegangen werden (eine detaillierte Darstellung findet sich im Supplement der Publikation Bettenbrock et al. (2006)). Abbildung 2.4 zeigt den Verlauf verschiedener Größen bei Wachstum auf Glukose und Glyzerin. In Plot A ist der Verlauf von Biomasse und Glukose zu sehen. Es sind zwei unterschiedliche Wachstumsphasen zu erkennen, wobei die Umschaltung auf Glyzerin erst nach Verbrauch der

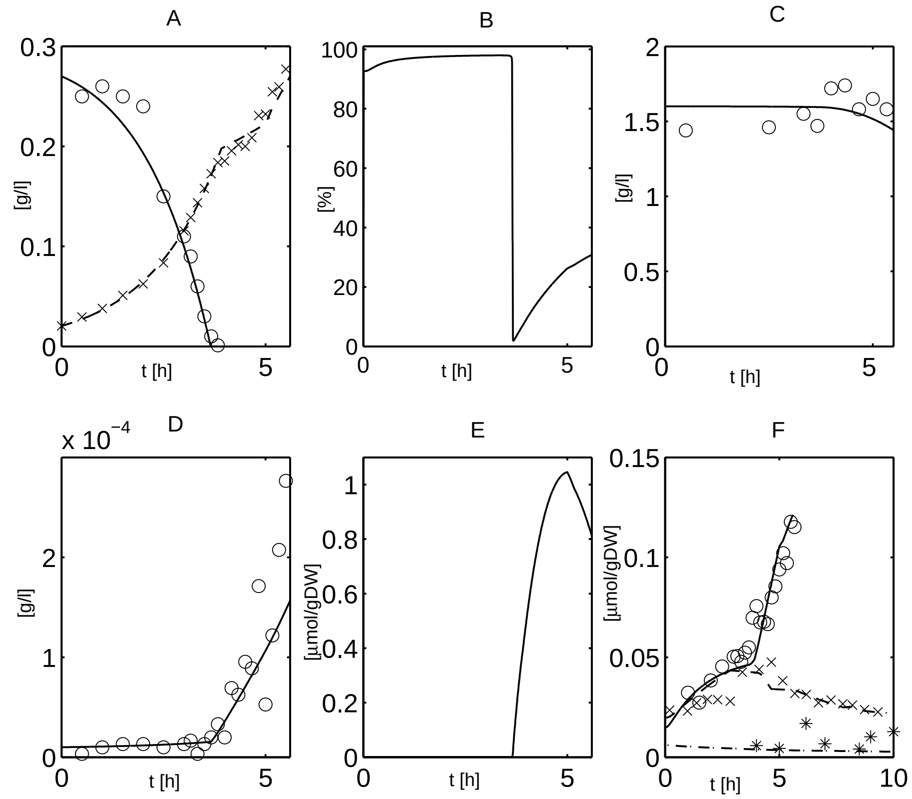


Abbildung 2.4: Verlauf von Zustandsgrößen für ein Diauxie-Experiment der Mlc-Mutante mit Glukose und Glyzerin. Glukose und Biomasse (Plot A), EIIA (Plot B), Glyzerin (Plot C), extrazelluläres cAMP (Plot D), intrazelluläres cAMP (Plot E). PtsG (Plot F), Siehe auch Text.

Glukose erfolgte (Plot C). Zielsetzung bei dieser Versuchsreihe war die Analyse der Genexpression des Glukosetransporters PtsG unter verschiedenen Bedingungen. Plot F zeigt den Verlauf von PtsG für das Diauxie-Experiment, sowie für ein Experiment mit dem Wildtyp und für ein Experiment, bei dem keine Glukose im Medium vorlag. Wie oben bereits gezeigt, erfolgt die Regulation des Transporters durch eine komplexe Interaktion zwischen dem Regulator Mlc und den DNA-Bindestellen. Das Experiment brachte Aufschluß über drei charakteristische Merkmale der Genexpression dieses Proteins, die ausreichend gut abgebildet wurden:

- Das Basallevel (Plot F, strichpunktiert, Experiment $'*'$) von PtsG wurde gebildet, wenn keine Glukose im Medium vorlag. Beim gezeigten Versuch wurde Glyzerin vorgelegt und das Basalniveau gemessen.

- Regulation durch Mlc (Plot F, gestrichelt, Experiment 'x'): Geht die Glukose bei dem Diauxieexperiment aus, sorgt die Regulation durch Mlc dafür, dass keine Neusynthese an Protein mehr stattfindet. Das Protein wird dann durch Wachstum ausgedünnt.

- Einfluss des Transkriptionsfaktors Crp (Plot F, durchgezogen, Experiment $'o'$): Wird eine Mlc-Mutante beim gleichen Diauxieexperiment eingesetzt, so macht sich die Mutation nur in der zweiten Wachstumsphase auf Glyzerin bemerkbar. Durch das Fehlen des Regulators kommt es durch den Einfluss von Crp zu einer verstärkten Neusynthese des Proteins.

Beim zweiten Experiment wird ein Batch-Versuch mit Glyzerin gestartet. In der exponentiellen Phase (Stunde 4) wird Glukose pulsförmig dazugegeben und der Verlauf der Messgrößen verfolgt. Das Experiment dient dazu, die Glukoseaufnahme zu beschreiben, wenn das System sehr schnell ausgelenkt wird. In einer der ersten Modellvarianten wurde angenommen, dass die beiden allgemeinen PTS-Proteine HPr und EI in konstanter Konzentration vorliegen. Aus der Literatur war bekannt, dass EI und HPr zwar der Kontrolle durch Crp und Mlc unterliegen, allerdings wurde festgestellt, dass der Bereich der Konzentrationen unter verschiedenen Bedingungen nur kleinen Schwankungen aufweist. Abbildung 2.5 vergleicht einige Modellvarianten bezüglich der Glukoseaufnahme (Plot B). Simuliert wurde mit (i) konstanter Proteinkonzentration für die Proteine PtsG, EI und HPr, (ii) konstanter Proteinkonzentration für PtsG und variabler Proteinkonzentration für EI und HPr, (iii) variabler Proteinkonzentration für PtsG und konstanter Proteinkonzentration für EI und HPr sowie (iv) variabler Proteinkonzentration für die Proteine PtsG, EI und HPr. Eine gute Anpassung der Messdaten (aller 18 Experimente) ist nur möglich, wenn die Regulation durch Crp und Mlc auch für HPr und EI berücksichtigt wird (Fall (iv)). Alle anderen Modellvarianten zeigten eine zu schnelle Aufnahme der Glukose. Plot E zeigt den zeitlichen Verlauf von PtsG und HPr, Plot F den Verlauf von EI. Die Verläufe machen deutlich, dass sich die Proteinkonzentrationen nur geringfügig ändern, aber diese Änderungen einen starken Einfluss auf den Verlauf der Glukoseaufnahme haben.

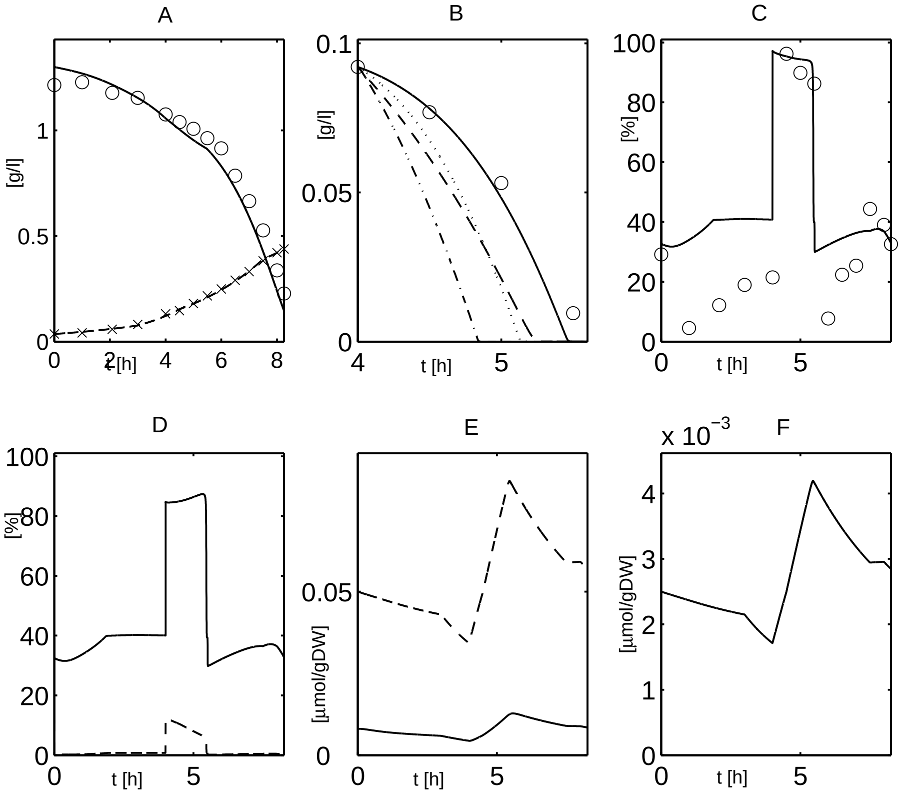


Abbildung 2.5: Verlauf von Zustandsgrößen für ein "Disturbed"-Batch-Experiment. Experimentelle Daten sind mit Symbolen dargestellt. Die Kultur wächst zunächst auf Glyzerin. In der exponentiellen Phase wird Glukose gepulst und die Systemantwort beobachtet. Gezeigt ist der Verlauf von Glyzerin und Biomasse (Plot A); Plot B zeigt den Verlauf von Glukose für Simulationsrechnungen mit verschiedenen Modellvarianten (siehe Text); EIIA (Plot C), in Plot D wird EIIA nochmal aufgeschlüsselt nach verschiedenen Konformationen gezeigt. Durchgezogen ist freies EIIA, gestrichelt ist EIIA mit GlpK komplexiert. PtsG (durchgezogen) und HPr (Plot E) und EI (Plot F).

2.3 Implementierung in ProMoT

ProMoT (Ginkel et al., 2003; Kremling et al., 2004c) ist ein objektorientiertes, gleichungsbasiertes Modellierungswerkzeug. Es kann kontinuierliche und gemischt kontinuierlich-ereignisdiskrete Modelle für die Simulationsumgebung Diva erstellen. Die Modelle werden dabei aus einer abstrakten symbolischen Repräsentation im Modellierungswerkzeug in Unterprogramme überführt, wodurch eine sehr effiziente Simulation ermöglicht wird. Basierend auf der Netzwerktheorie (Gilles, 1998) werden strukturelle und verhaltensbeschreibende Modellbausteine unterschieden. Strukturell wird das Gesamtmodell in Module unterteilt, die bestimmten biologischen Einheiten auf verschiedenen Ebenen des Gesamtsystems zugeordnet werden. Entsprechend dem Konzept werden auf der untersten Ebene molekularbiologische Spezies und Reaktionen beschrieben. Höher strukturierte Funktionseinheiten wie Stoffwechselwege, Signaltransduktionswege oder auch Bioreaktoren werden dann durch Module repräsentiert, die sich aus mehreren Teilmodellen zusammensetzen. Das lokale Verhalten eines Moduls wird durch Variablen, algebraische Gleichungen und gewöhnliche Differentialgleichungen beschrieben. Das Gesamtgleichungssystem kann im allgemeinen Fall ein differentialalgebraisches System sein. Einige Variablen der Module werden Schnittstellen zugeordnet und stehen damit zur Verknüpfung mit anderen Modulen zur Verfügung.

Modellbausteine in ProMoT sind in einer objektorientierten Klassenhierarchie mit multipler Vererbung organisiert. Dieses Konzept aus der Informatik wurde aufgegriffen, um komplexe Bibliotheken von Modellbausteinen flexibel gestalten und besser organisieren zu können. Für systembiologische Projekte wurde eine Bibliothek von Modulen erstellt, die sowohl elementare Modellelemente wie Stoffspeicher und Stoffwandler aber auch wiederholt vorkommende Funktionseinheiten zur Beschreibung der Genexpression und der Signaltransduktion umfassen. Die Module in ProMoT besitzen standardisierte Terminals, die eine universelle Verknüpfung der Bausteine ermöglichen. Die Terminals repräsentieren dabei Signale (Konzentrationen oder Konzentrationsverhältnisse) oder Stoffflüsse (bidirektionaler Austausch einer Konzentration und einer Flussrate). Benutzer können neue Module mit Hilfe eines graphischen Editors oder textuell in der “Model Definition Language” (MDL) von ProMoT eingeben. Im graphischen Editor können Module aus der geladenen Modellbibliothek durch “Drag’n Drop” aggregiert und miteinander verbunden werden. Die Modelliersprache erlaubt es, spezielle eigene Gleichungsmodelle in elementaren Modulen zu implementieren. Dabei kann der Modellierer auf abstrakte Superklassen aus der Modellbibliothek zurückgreifen. Die Modelliersprache MDL ist eine deklarative, objektorientierte Sprache, die eine Beschreibung der Modellelemente enthält. Sie wird vom Modellierungswerkzeug gelesen und geschrieben und auch als Datenformat zur Speicherung der Modellbibliotheken genutzt.

Alle oben vorgestllleten Funktionseinheiten des Modells sind in ProMoT implementiert und getestet worden. Damit steht das gesamte Modell auf der graphischen Benutzeroberfläche zur Verfügung. Abbildung 2.6 zeigt links den Browser, der die Auswahl der Modellbausteine erlaubt. Rechts ist beispielhaft für Stoffwechselwege die Funktionseinheit zur Beschreibung der Laktoseaufnahme gezeigt.

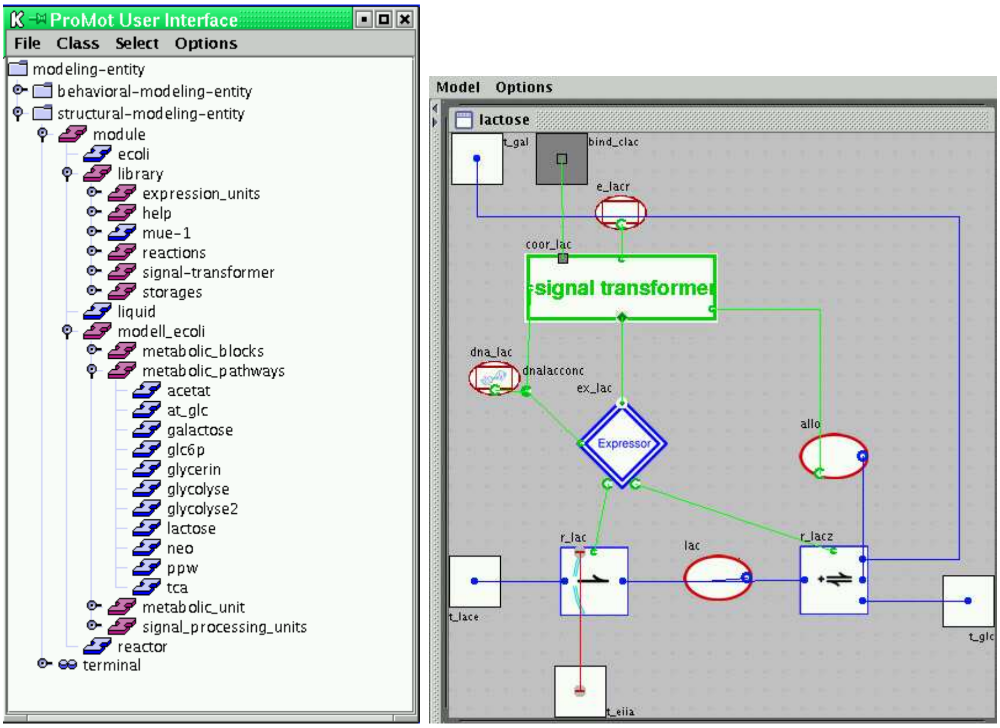


Abbildung 2.6: Zwei Bildschirmabzüge aus ProMoT. Links der “Modellbrowser”, der die Auswahl der Modellbausteine erlaubt, rechts der Stoffwechselweg des Laktoseabbaus.

3 Analyse globaler Regulationsnetzwerke

Zelluläre Systeme zeichnen sich auf der einen Seite durch eine große Flexibilität bezüglich der Umgebungsbedingungen aus, auf der anderen Seite zeigen sie sich auch äußerst robust, wenn sie durch Mutationen verändert werden. Verantwortlich für dieses Verhalten ist ein Netzwerk von globalen und spezifischen Transkriptionsfaktoren, die entsprechende Stoffwechselwege an- oder abschalten können. Wie der Name schon andeutet, reagieren spezifische Regulatoren, wenn sich bspw. eine bestimmte Nährstoffquelle im Medium befindet oder Spurenelemente aufgenommen werden müssen. Globale Regulatoren werden aktiv, wenn eine allgemeinere Situation, bspw. ein Mangel an Kohlenhydraten vorliegt. In dieser Situation, die dann "Hunger" signalisiert, müssen eine ganze Reihe von Stoffwechselwegen angepasst werden, um die neue Situation zu meistern. Es ist daher nicht verwunderlich, wenn diese wichtigen Funktionen in der Zelle besonders abgesichert sind, damit sie unter allen Bedingungen auch funktionieren.

In den letzten Jahren hat die Analyse von Robustheitseigenschaften einen breiten Raum in den Publikationen zur Systembiologie eingenommen (Barkai and Leibler, 1997; Stelling et al., 2004). Das Paradebeispiel für strukturelle Robustheit ist die bakterielle Chemotaxis. Bakterien wie *Escherichia coli* bewegen sich dabei in zwei Bewegungsformen, "Taumeln" und "Lauf". Interessant ist, dass sich die Zellen bei einem Reiz zunächst fast nur in der "Lauf"-Bewegung auf den Lockstoff zubewegen, nach einiger Zeit allerdings wieder zum ursprünglichen Bewegungsmuster zurückkehren. Das bedeutet, dass das Signal, das die Bewegung auslöst, zu einer adaptiven Antwort der Zelle führt: Die Zelle reagiert auf das Signal und kehrt dann wieder zum Ausgangswert zurück. Untersuchungen haben gezeigt, dass sich Mutationen, also das gezielte Einbringen oder Entfernen von bestimmten Genen nicht auf die Genauigkeit der Adaption des Systems auswirkt, sondern nur die Zeitkonstanten der Systemantwort verändern (Alon et al., 1999). Eine molekularbiologische Analyse hat dann gezeigt, dass eine integrale Rückführung dafür sorgt, dass das System immer zur Ausgangslage zurückkehrt, selbst bei einer 10-fachen Erhöhung eines Regulatorproteins (Alon et al., 1999). Neben diesen Analysen zur strukturellen Robustheit wird in vielen Publikationen auch der Einfluss von kinetischen Parametern untersucht, um das dynamische Verhalten des Systems zu charakterisieren. Bei vielen Systemen machen sich Veränderungen von Parametern kaum bemerkbar, die Systeme werden als robust bezeichnet. Bei einigen Systemen stellt man jedoch fest, dass sich das dynamische Verhalten stark verändert, wenn die Parameter variiert werden.Das zeigt sich besonders bei multistabilen Systemen, bei denen bei der Parametervariation ein kritischer Punkt (Bifurkationspunkt) überschritten wird.

Der globale Transkriptionsfaktor Crp wurde oben bereits eingeführt. Er ist bei der Initiation der Transkription bei einer Vielzahl von Genen des katabolen Stoffwechsels beteiligt. Die Aktivierung des Transkriptionsfaktors hängt vom Phosphorylierungsgrad der Komponente EIIA des PTS ab. Experimentelle Untersuchungen am MPI Magdeburg haben nun gezeigt, dass sich ein Zusammenhang zwischen der Wachstumsrate von *E. coli* und dem Phosphorylierungsgrad von EIIA ergibt. Der Zusammenhang kann in Form einer Kennlinie aufgetragen werden: bei hohen Wachstumsraten ist der Phosphorylierungsgrad niedrig, bei niedrigen Wachstumsraten ist er hoch. Die folgenden Ausführungen sollen den Zusammenhang mit einem mathematischen Modell beschreiben, wobei gezeigt wird, dass hier eine strukturelle Robustheit vorliegt: Veränderungen der Parameter oder der Modellstruktur haben nur einen geringen Einfluss auf das Systemverhalten.

3.1 Struktureigenschaften des Netzwerkes

Der Zusammenhang zwischen der spezifischen Wachstumsrate μ und dem Phosphorylierungsgrad von EIIA soll durch Betrachtung eines vereinfachten Stoffwechselschemas, wie in Abbildung 3.1 gezeigt, deutlich werden. Im Fall, dass das PTS nicht aktiv ist, wie bei Wachstum auf Laktose, Glyzerin oder Glukose-6-Phosphat, muß die Rate der reversiblen PTS Reaktion r_{PTS} in Abbildung 3.1 gleich Null sein. Die Rate r_{PTS} faßt dabei alle Reaktionen des PTS ausgehend von PEP bis zu EIIA zusammen. Nimmt man eine reversible Reaktion zweiter Ordnung für r_{PTS} an, ergibt sich eine Beziehung zwischen dem PEP/Pyruvat Verhältnis und dem Phosphorylierungsgrad der Komponente EIIA:

$$c_{EIIA\sim P} \quad = \quad c_{EIIA_0} \, \frac{\frac{c_{PEP}}{c_{Prv}}}{K_{PTS} \, + \, \frac{c_{PEP}}{c_{Prv}}} \, . \tag{3.1}$$

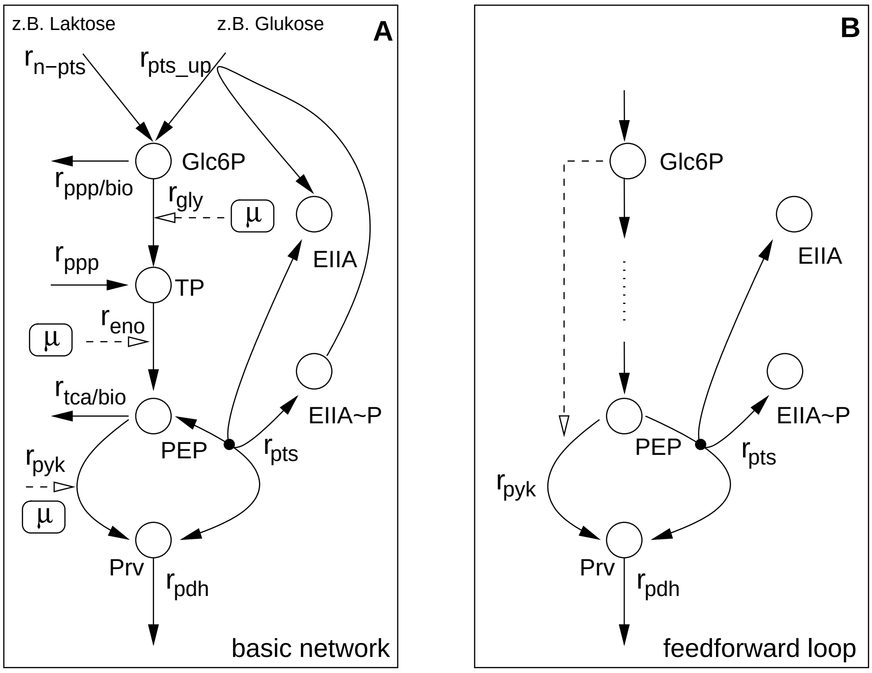


Abbildung 3.1: **A**: Vereinfachtes Stoffwechselschema. Die PTS-Proteine werden nur durch die Komponente EIIA repräsentiert. Reaktionen der Glykolyse sind zusammengefaßt, da der Abfluss in die Biosynthese nur marginal ist. Wachstumsabhängige Terme der Enzymmengen der Glykolyse sind berücksichtigt (Symbol μ). **B** Im Feedforward-Loop (in A aus Gründen der Übersichtlichkeit nicht eingezeichnet) wird die Pyruvatkinase r_{pyk} durch einen Metaboliten der Glykolyse aktiviert.

Im Falle eines aktiven PTS ist die Aufnahmerate des Zuckers r_{pts_up} gleich der Rate durch das PTS und man erhält:

$$c_{EIIA\sim P} \quad = \quad \frac{c_{EIIA_0} \, \frac{c_{PEP}}{c_{Prv}} \; - \; \frac{r_{up/pts}}{k_{pts}}}{K_{PTS} \; + \; \frac{c_{PEP}}{c_{Prv}}} \, . \tag{3.2}$$

Beide Gleichungen zeigen einen ähnlichen Aufbau, der zentral für die weiteren Betrachtungen ist: Um den für hohe Wachstumsraten und damit hohe Aufnahmeraten experimentell beobachteten niedrigen Phosphorylierungsgrad zu erreichen, muß das PEP/Pyruvat-Verhältnis mit steigender Wachstumsrate kleiner werden (Abbildung 3.2, Plot A). Fragt man sich nun, wie PEP und Pyruvat, über der Wachstumsrate aufgetragen, verlaufen müssen, um das gewünschte Verhältnis zu realisieren, so ergeben sich die in Abbildung 3.2, Plot B & C gezeigten Möglichkeiten. Allerdings ist schnell klar, dass, wenn beide Funktionen PEP(μ) und Prv(μ) monoton

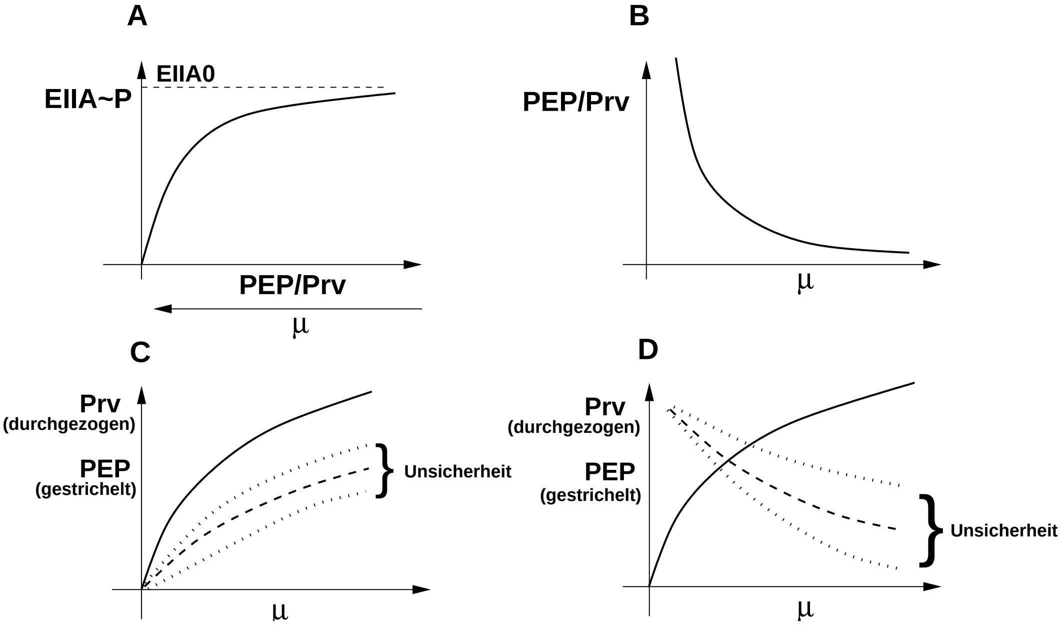


Abbildung 3.2: **A**: Nach Gleichung (3.1) ergibt ein hohes PEP/Pyruvat-Verhältnis einen hohen Phosphorylierungsgrad von EIIA. Ein hoher Phosphorylierungsgrad wird für eine niedrige Wachstumsrate gemessen. **B** Das PEP/Pyruvat-Verhältnis muss dann über der Wachstumsrate aufgetragen, abfallen. **C**: PEP und Pyruvat als Funktion der Wachstumsrate. Sensitive Struktur, da Unsicherheiten leicht zu starken Verschiebungen des Verhältnisses führen können. **D**: PEP und Pyruvat als Funktion der Wachstumsrate. Robuste Struktur, da Unsicherheiten kaum zu Verschiebungen des Verhältnisses führen.

steigende Abhängigkeiten von der Wachstumsrate zeigen (Plot B), dies auf ein sensitives Verhalten hindeutet: Leichte Änderungen oder Störungen auf das System (in der Abbildung für PEP gezeigt) können dazu führen, dass das Verhältnis stark verändert wird, sich im Extremfall sogar umdrehen kann. Der in Abbildung 3.2 in Plot C gezeigte Fall stellt dagegen einen robusten Verlauf dar: Veränderungen oder Störungen verändern das PEP/Pyruvat-Verhältnis kaum.

Geht man der Frage nach, wie das in Plot D gezeigte robuste Verhalten von der Zelle realisiert werden kann, so stellt eine Möglichkeit die in Abbildung 3.1 gezeigte Feedforward-Steuerung dar. Eine hohe Wachstumsrate und damit verbunden ein hoher Fluss durch die Glycolyse erfordert auch einen hohen Fluss durch die Pyruvatkinase. Wenn nun allerdings die PEP Konzentration mit steigender Wachstumsrate kleiner werden soll, muss dies kompensiert werden, da sonst die Rate nicht erhalten werden kann. Die Komponenten in der oberen Hälfte der Glykolyse können nun als Signal eingreifen und durch eine Aktivierung der Pyruvatkinase die hohe Rate bewerkstelligen, da die Konzentrationen dieser Metabolite mit steigender Wachstumsrate ebenfalls steigen. In der Tat ist schon lange aus Versuchen mit isolierter Pyruvatkinase bekannt, dass diese stark von Fruktose 1,6 Bis-Phosphat aktiviert wird (Waygood and Sanwal, 1974) (in Abbildung 3.1 ist Fruktose-1,6-Bis-Phosphat nicht explizit im Modell enthalten und durch Glukose 6-Phosphat ersetzt).

Der Feedforward-Loop kann hier als eine robuste Struktur aufgefaßt werden, die im regelungstechnischen Sinne keinen geschlossenen Kreis aufweist, wie bei der robusten Struktur bei der Chemotaxis zu beobachten ist. Trotzdem kann aber erwarten werden, dass sie unempfindlich gegenüber Störungen ist. Im folgenden Kapitel sollen nun für unterschiedliche Modellvarianten Parameter ermittelt werden, die experimentelle Daten beschreiben. Es wird gezeigt, dass die Modellvarianten mit Feedforward-Loop deutlich besser die experimentellen Daten beschreiben als die Modelle ohne Feedforward-Loop.

3.2 Parameterschätzung und Modellbewertung

Mathematische Modelle in der Systembiologie müssen auf der einen Seite versuchen, das bekannte Wissen über die biologischen Netzwerke so genau wie möglich zu erfassen, auf der anderen Seite aber auch die wesentlichen strukturellen Elemente, die für ein beobachtetes Verhalten verantwortlich sind, herausstellen. Oben wurde ein robustes Verhalten durch den Feedforward-Loop vorgestellt. Diese Arbeitshypothese soll nun weiter verfolgt werden. Dabei wird gezeigt, dass mit unterschiedlichen Modellvarianten, die sich durch den Detailliertheitsgrad und damit das biologische Wissen, welches repräsentiert wird, unterscheiden, gleich gute quantitative Ergebnisse erreicht werden können, wenn das entscheidende strukturelle Element im Modell berücksichtigt ist. Dazu wurden in einer Studie eine ganze Reihe von Modellvarianten zunächst hinsichtlich ihres stationären Verhaltens untersucht. Dabei wird die geschätzte Streuung σ (residual mean square) der Messgrösse EIIA$\sim$P in % bezogen auf die Gesamtkonzentration nach der Beziehung

$$\sigma \;=\; \frac{100}{c_{EIIA_0}} \sqrt{\frac{1}{N-n} \sum^{i} \epsilon_i^2} \tag{3.3}$$

zur Beurteilung herangezogen. Hierbei sind ϵ_i die Residuen, N ist die Anzahl der Datenpunkte und n ist die Anzahl der Parameter, also der Freiheitsgrade des Systems. Es sind nur diejenigen Parameter, die geschätzt werden, berücksichtigt. Tabelle 3.1 faßt die Modellvarianten zusammen. Da mit den Modellen ein großer Bereich der Wachstumsrate abgedeckt werden soll, müssen auch wachstumsratenabhängige Vorgänge einbezogen werden. Dies ist im wesentlichen die Aktivität der RNA Polymerase, die sich im Bereich der Wachstumsrate verdoppelt. Da bekannt ist, dass die Synthese der Enzyme der Glykolyse sonst nicht reguliert ist, wird folgender Ansatz für eine auf den Maximalwert skalierte Enzymmenge $\frac{e}{e_0}$ verwendet:

$$\frac{e}{e_0} = \frac{0.5\ \mu}{\mu + K_\mu} + 0.5\,. \tag{3.4}$$

Anschaulich bedeutet die Gleichung, dass es ein recht hohes Basalniveau für kleine Wachstumsraten gibt und dass die Enzymmengen dann in Form einer Michaelis-Menten-Kurve auf den Maximalwert ansteigen. Neben dieser von der Wachstumsrate abhängigen Enzymmenge sind auch Abflussraten in die Monomersynthese berücksichtigt. Die Werte für die Raten sind einer stationären Flussverteilung entnommen (Kremling, 2002).

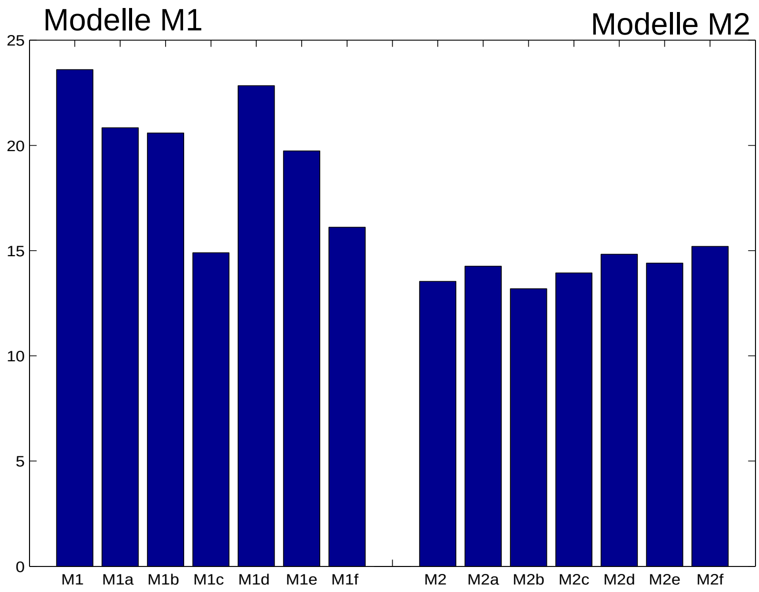


Abbildung 3.3: Vergleich der σ-Werte nach Gleichung (3.3) für die verschiedenen Modellvarianten. Die Modelle der M2-Familie zeigen niedrigere Werte als die Modelle der Variante M1.

Tabelle 3.1: Zusammenfassung der Modellvarianten. Die Spalten geben die Kinetiken von r_{gly}, r_{pyk} und r_{pdh} an. E bedeutet wachstumsabhängige Enzymkonzentration, A bedeutet Abflüsse in die Biosynthese. †Monod-Wyman-Changeux-Modell (Blangy et al., 1968): $r = \frac{k_{pyk}\, c_{PEP}/K_{PEP}\, (1+c_{PEP}/K_{PEP})^3\, (1+c_{G6p}/K_{G6p})^4}{(1+c_{PEP}/K_{PEP})^4\, (1+c_{G6p}/K_{G6p})^4 + L}$; ‡MM: Michaelis Menten.

Variante	r_{gly}	r_{pyk}	r_{pdh}	E	A
M1	g1	p1	d1	ja	ja
M1a	g1	p2	d1	ja	ja
M1b	g2	p2	d2	ja	ja
M1c	g2	p1	d1	ja	ja
M1d	g1	p1	d1	nein	ja
M1e	g1	p2	d1	nein	ja
M1f	g2	p2	d1	nein	ja
M2	g1	p1*	d1	ja	ja
M2a	g1	p2*	d2	ja	ja
M2b	g1	p3	d1	ja	ja
M2c	g2	p1*	d1	ja	ja
M2d	g1	p1*	d1	nein	ja
M2e	g1	p3	d1	nein	ja
M2f	g1	p1*	d1	nein	nein

Kinetik		Formel
g1	Massenwirkung	$k_{gly}\, c_{Glc6P}$
g2	MM‡	$k_{gly}\, \frac{c_{Glc6P}}{c_{Glc6P}+K_{G6p}}$
p1	Potenzansatz	$k_{pyk}\, {c_{PEP}}^2$
p2	Potenzansatz	$k_{pyk}\, {c_{PEP}}^4$
p1*	Potenzansatz	$k_{pyk}\, {c_{PEP}}^2\, {c_{Glc6P}}^2$
p2*	Potenzansatz	$k_{pyk}\, {c_{PEP}}^4\, {c_{Glc6P}}^4$
p3	MWC†	s.o.
d1	Mass action	$k_{pyk}\, c_{Prv}$
d2	Potenzansatz	$k_{pyk}\, {c_{Prv}}^2$

Die Modellvarianten M1 und M2 unterscheiden sich in der Regulationsstruktur der Glykolyse. Die M1-Modelle besitzen keinen Feedforward-Loop während bei den Modellen M2 Glukose-6-Phosphat die Pyruvatkinase aktiviert. Für den kinetischen Ansatz der Pyruvatkinasereaktion sind unterschiedliche Varianten vorgeschlagen. Beispielsweise berücksichtigt der Monod-Wyman-Changeux-Ansatz, dass das Enzym in einer aktiven und in einer inaktiven Form vorliegt und dass der Aktivator Glukose-6-Phosphat den Übergang in die aktive Konformation erleichtert.

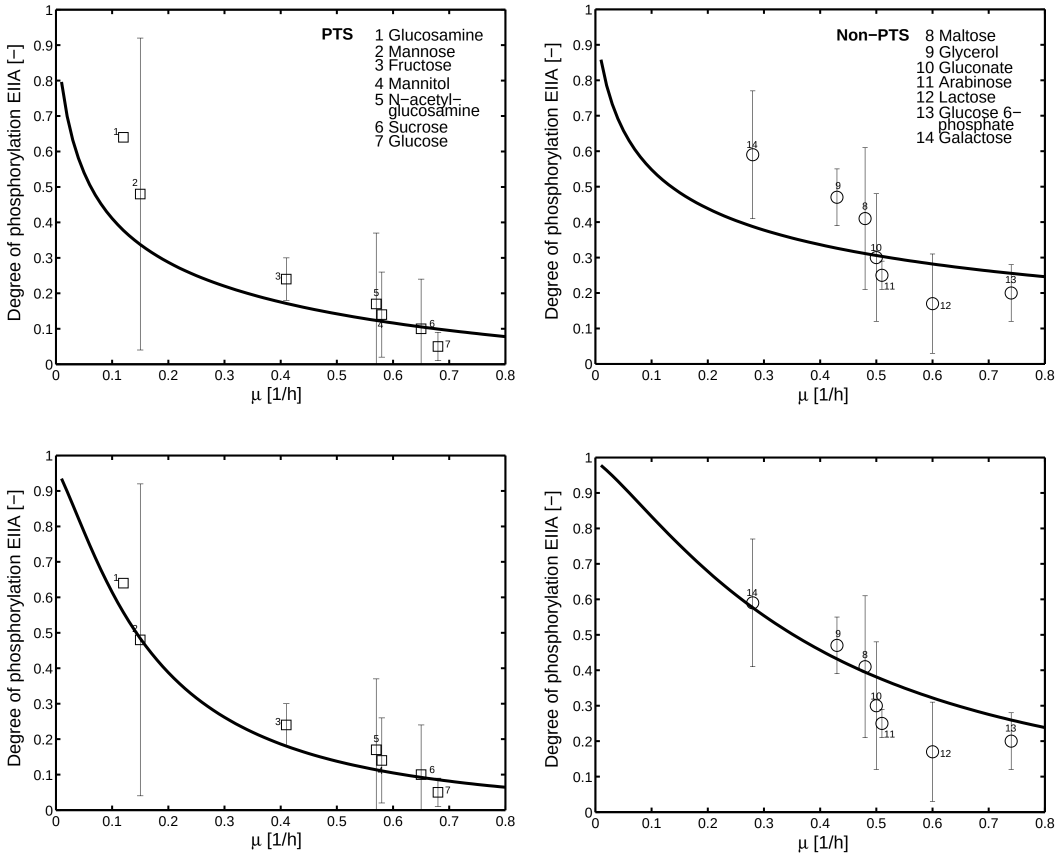


Abbildung 3.4: Vergleich der Modellvarianten M1a (oben) und M2b (unten). Die Legende gibt Aufschluß über die verwendeten Substrate. Die linke Spalte stellt Messungen für PTS Kohlenhydrate dar (Symbol □). Die rechte Spalte stellt Messungen für Nicht-PTS Kohlenhydrate dar (Symbol ◯). Die Messungen wurden mit Einzelzuckern durchgeführt wie in der Legende beschrieben. Die Messung erfolgte in der exponentiellen Phase (Bettenbrock et al., 2007).

Alle Modelle wurden mit dem gleichen Verfahren (MATLAB Optimierungsroutinen) an die vorliegenden experimentellen Daten angepaßt. Die Daten beschreiben den Zusammenhang zwischen spezifischer Wachstumsrate μ und dem Phosphorylierungsgrad des PTS-Proteins EIIA. Sie wurden aus Experimenten bei Wachstum auf Glukose und Laktose sowie aus Experimenten mit nur einer einzigen Nährstoffquelle gewonnen (Bettenbrock et al., 2006, 2007).

Abbildung 3.3 zeigt die Ergebnisse der Parameterschätzung für alle Modelle. Die beste Anpassung gelingt mit Modell M2b, welches auch die größte Komplexität aufweist. Das Modell verwendet die Monod-Wyman-Changeux-Kinetik für die Pyruvatkinasereaktion und berücksichtigt sowohl die Abflüsse in die Monomersynthese als auch die wachstumsratenabhängigen Enzymmengen. Im Vergleich dazu kann aber ebenfalls ein gutes Ergebnis erzielt werden, wenn ein Minimalmodell wie Variante M2f verwendet wird. Abbildung 3.4 zeigt die experimentellen Daten der Messungen mit Einzelzuckern in der Gegenüberstellung der Varianten M1a und M2b. Die simulierten Verläufe machen deutlich, dass mit der Modellvariante ohne Feedforward-Loop nur qualitativ das richtige Verhalten erzielt werden kann, die quantitative Beschreibung aber deutlich schlechter ist. Abbildung 3.5 zeigt die zu Abbildung 3.4 passenden Verläufe von PEP und Pyruvat. In der rechten Abbildung ist gut zu sehen, wie sich der Feedforward-Loop bemerkbar macht. Der Verlauf der Konzentration von PEP besitzt ein Maximum bei einer sehr kleinen Wachstumsrate und wird für höhere Wachstumsraten dann immer kleiner. In der Variante ohne den Loop sind sowohl PEP als auch Pyruvat monoton steigend über den

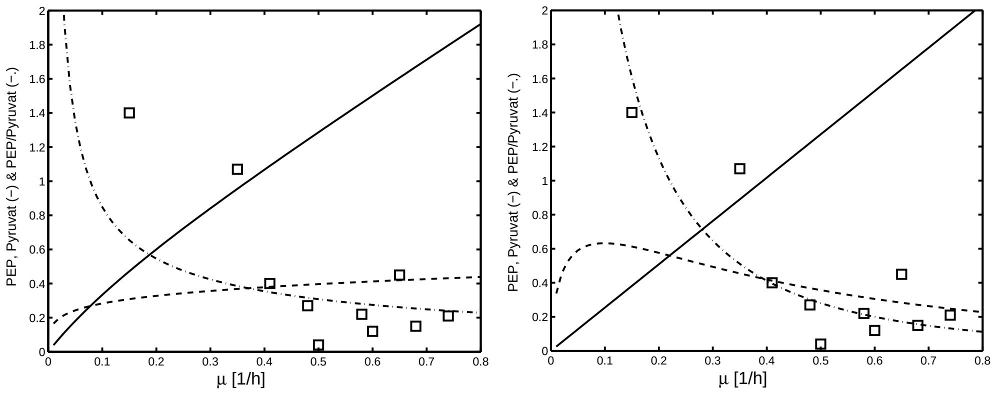


Abbildung 3.5: Abhängigkeit von PEP und Pyruvat von der Wachstumsrate μ für die Modellvarianten M1a (links) und M2e (rechts). Das PEP/Pyruvat Verhältnis ist strich-punktiert gezeichnet, experimentelle Daten des PEP/Pyruvat Verhältnises mit Symbol gekennzeichnet (Bettenbrock et al., 2007).

Verlauf der Wachstumsrate, allerdings mit unterschiedlichem Anstieg, was zu einem fallenden PEP/Pyruvat-Verhältnis führt. Die Messung des PEP/Puryvat-Verhältnisses *in vivo* ist immer noch sehr aufwändig und fehlerbehaftet. In der Abbildung sind Messungen aus verschiedenen Experimenten eingetragen, die deutlich machen, dass die Tendenz im Modell richtig wiedergegeben wird.

3.3 Dynamische Simulation verschiedener Kohlenhydrataufnahmesysteme

Das in Abbildung 3.1 gezeigte Schema stellt ein Grundgerüst für ein erweitertes Modell dar, welches nun eine dynamische Simulation und Analyse von verschiedenen Kohlenhydrataufnahmesystemen erlaubt. Dabei werden Teilnetzwerke zur Beschreibung der Aufnahmekinetik

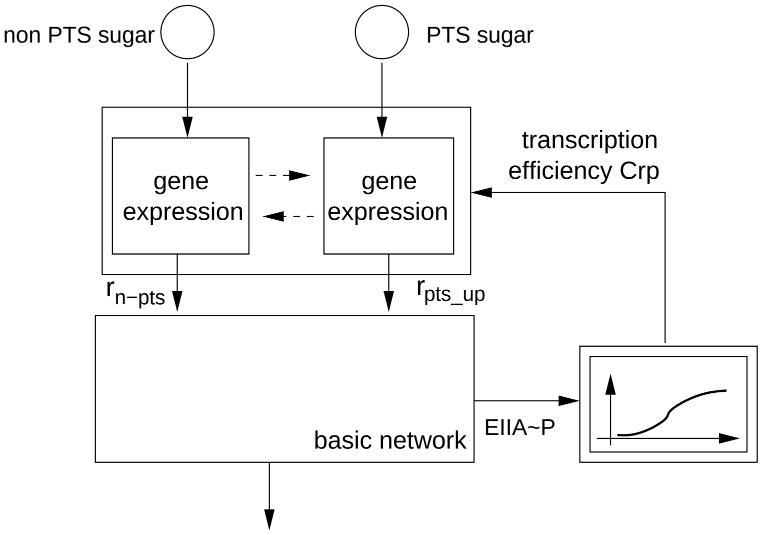


Abbildung 3.6: Erweitertes Modell zur Beschreibung der Genexpression. Das Ausgangssignal des Sensors, $EIIA{\sim}P$, wird weiterverarbeitet und es wird eine von Crp abhängige Transkriptionseffizienz ermittelt (nichlinear Kennlinie). Dieses Signal wird dann verwendet, um für individuelle Transporter für PTS und Nicht-PTS Zucker die Genexpression zu beschreiben. Die Pfeile zwischen den Transporteinheiten deuten an, dass sich solche Systeme auch gegenseitig beeinflussen können. Die Aufnahmeraten dienen dann als Eingang in das Grundmodell.

und der Genexpression entsprechend Abbildung 3.6 in modularer Art und Weise ergänzt. Das bedeutet, dass die Struktur und die kinetischen Parameter des Grundmodells unverändert bleiben und nur die im Bild gezeigten Verbindungen berücksichtigt werden. Der Ausgang des Grundmodells, der Phosphorylierungsgrad von EIIA wird als Eingang in einen weiteren Block zur Beschreibung der Crp-abhängigen Transkriptionseffizienz geführt. Experimentelle Daten, die für einen Crp-abhängigen und einen Crp-unabhängigen Promotor aufgenommen wurden (Bettenbrock et al., 2007), erlauben die Ermittlung einer nichtlinearen Kennlinie, die die Syntheserate der Proteine beschreibt (Abbildung 3.7). Dazu wurden Experimente mit verschiedenen Kohlenhydraten verwendet und die Daten in der exponentiellen Phase aufgenommen. Aus den Daten kann folgender Zusammenhang ermittelt werden:

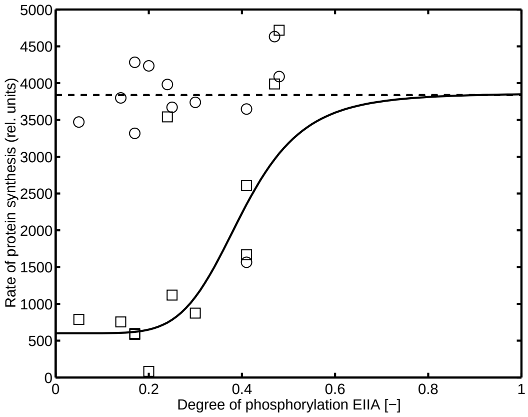


Abbildung 3.7: Zusammenhang zwischen dem Phosphorylierungsgrad von EIIA und der Rate der Proteinsynthese. Die Messungen wurden mit einem Crp-unabhängigen Gen (◯) durchgeführt, welches konstitutiv exprimiert wird. Das Crp-abhängige Gen (□) hat ein niedriges Basalniveau (k_b) und zeigt ein sigmoides Verhalten ($h = 6$) in Abhängigkeit vom Phosphorylierungsgrade.

$$r_{syn} \;=\; k_b + k_{syn}\, \frac{{c_{EIIAP}}^6}{{c_{EIIAP}}^6 + K^6}\,. \tag{3.5}$$

Die Gleichung weist einen recht hohen Hillkoeffizienten h von 6 auf. Dies deutet darauf hin, dass durch die Kennlinie mehrere stark verkoppelte Prozesse zusammengefaßt sind. Wie in Abbildung 2.1 zu sehen ist, faßt die Kennlinie die beiden rechten Blöcke zusammen, die durch die Aktivität des cAMP·Crp Komplexes auch rückgekoppelt sind. Im folgenden sollen am Beispiel der Glukose-Glukose 6-Phosphat-Diauxie die Ergebnisse vorgestellt werden. Es ist bekannt, dass beide Transporter von Crp abhänigig sind.

3.3.1 Glukose Glukose 6-Phosphat Diauxie

Das Grundmodell wurde um Gleichungen für Biomasse, Substrate und Transportenzyme ergänzt. Hierzu wurde Modell M2f ausgewählt, da es die einfachste Struktur besitzt. Messtechnisch liegen nur Messungen des Glukosetransporters vor. Für die Simulation sind nur die kinetischen Parameter der Transporter angepasst worden (r_{max}- und K_M-Werte). Die Parameter, die oben ermittelt wurden (Kapitel 3.2), sind unverändert beibehalten. Im Experiment wird zunächst Glukose 6-Phosphat aufgenommen, während Glukose nicht verstoffwechselt wird. Aus dem zeitlichen Verlauf des Glukose-Transporters ist zu entnehmen, dass die Glukose-6-

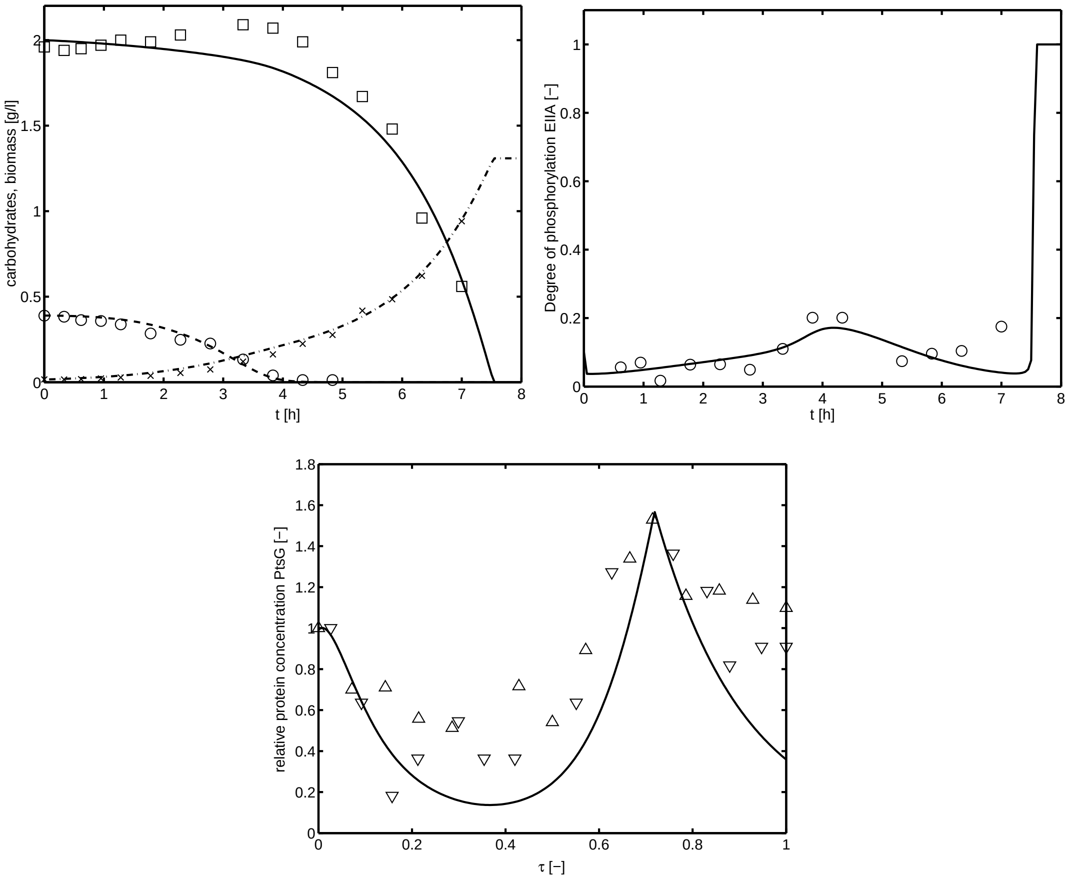


Abbildung 3.8: Glukose Glukose 6-Phosphat Diauxie. Links oben: Zeitverlauf von Glukose 6-Phosphat, Glukose und Biomasse. Rechts oben: Zeitverlauf des Phosphorylierungsgrades von EIIA. Unten: Zeitverlauf von PtsG (aus verschiedenen Experimenten aufgenommen, daher mit skalierter Zeit).

Phosphat-Aufnahme zu einer reduzierten Genexpression an PtsG führt. Da die molekularen Details dieser Interaktion noch nicht bekannt sind, ist ein einfacher Ansatz in den entsprechenden Kinetiken berücksichtigt. Das mittlere Bild in Abbildung 3.8 zeigt den Verlauf des Phosphorylierungsgrades von EIIA. Es ist zu sehen, dass sich bei einer hohen Wachstumsrate nur ein niedriger Wert einstellt. In der Übergangsphase kommt es zu einer leichten Erhöhung des Phosphorylierungsgrades, was durch das Modell sehr gut wiedergegeben wird. Das Beispiel macht deutlich, dass das Modell auch eine dynamische Simulation erlaubt und zu guten Ergebnissen führt.

Zusammenfassend läst sich festhalten, dass *E. coli* durch eine einfache Steuerung mittels eines Feedforward-Loops eine robuste Struktur besitzt, die es ermöglicht, die Flüsse durch die Glykolyse zu messen und entsprechend darauf zu reagieren. Der Vergleich von Modellvarianten mit unterschiedlichem Detailliertheitsgrad zeigte deutlich, dass sich die Messdaten nur ausreichend gut anpassen lassen, wenn der Feedforward-Loop berücksichtigt wurde. Eine modulare Erweiterung des Modells um die Beschreibung der Genexpression und der Kinetiken der Substrataufnahme zeigt gute Ergebnisse bei der Beschreibung des dynamischen Verhaltens des Gesamtsystems.

4 Versuchsplanung und Ermittlung von Parameterunsicherheiten

Die Planung neuer Experimente spielt in der chemischen Verfahrenstechnik generell eine wichtige Rolle. Auch in der Systembiologie werden Methoden und Verfahren zur Versuchsplanung immer wichtiger, da oft experimentelle Daten anfallen, die nicht aussagekräftig genug sind, um mathematische Modelle zu validieren. Bei der modellgestützten Versuchsplanung stehen die Unterscheidung verschiedener Modellhypothesen und die Verbesserung der Parametergüte im Zentrum der Untersuchungen.

4.1 Modelldifferenzierung

Liegen experimentelle Daten für die Parameterschätzung vor, so lassen sich oft zwei oder mehrere Modellstrukturen damit anpassen. Damit ist zunächst keine Aussage darüber möglich, welches Modell am besten geeignet ist. Zur Lösung des Problems wird ein weiteres Experiment gemacht, in der Hoffung, durch zusätzliche Messinformation eine Modelldiskriminierung durchführen zu können. Zur systematischen Planung eines solchen Experimentes wird in der Regel ein Kriterium der Art

$$\max_{u_{t_k}} \sum_{t_k} (\underline{y}_1 - \underline{y}_2)^T \; Q \; (\underline{y}_1 - \underline{y}_2) \tag{4.1}$$

definiert, welches Werte für die Eingangsgröße u zu Zeitpunkten t_k ermittelt, die die Differenzen der beiden Modellausgänge $\underline{y}_1$ und $\underline{y}_2$ maximieren. Die Differenzen werden mit einer geeigneten Matrix Q gewichtet.

In der Literatur werden eine ganze Reihe von Ansätzen für die Matrix Q vorgeschlagen (Asprey and Macchi 2000; Chen and Asprey, 2003; Munack, 1992). Allerdings fehlt in den meisten Fällen eine ausführliche Analyse von Beispielsystemen und Anwendungen. Folgende Punkte beschreiben den Ablauf bei der Versuchplanung:

- Definition der Modellvarianten und Durchführung eines ersten Experimentes
- Validierung der einzelnen Modellvarianten (Parameterschätzung und Ermittlung der Parameterunsicherheiten).
- Durchführung des neuen Experimentes und Aufnahme von Messdaten.
- Erneute Anpassung der Experimente an die Modellstrukturen. Dabei werden die Daten des alten und des neuen Experimentes verwendet.

Besonders die letzten beiden Punkte finden in der Literatur kaum Beachtung; in der Regel beschränkt man sich nur auf den Entwurf des Experimentes. Im folgenden sollen an einem kleinen zellulären Netzwerk alle Schritte beschrieben werden, die die Modelldiskriminierung umfaßt. Betrachtet wird dabei die in Abbildung 4.1 gezeigte Anlage mit einem Bio-Reaktor, der eine Zudosierung eines Substrates mit der Rate q und der Konzentration c^{zu} erlaubt. Das zelluläre Netzwerk beschreibt die Aufnahme des Substrates und den weiteren Stoffwechsel. Dabei stehen für den zweiten Reaktionsschritt zwei Modellalternativen zur Verfügung (Modell A und Modell B). Der Stoffwechselweg soll die Umsetzung des Stoffes $M1$ in $M2$ mittels Enzym $M3$ beschreiben. Das Enzym wird mit $M2$ entweder über die Aktivität (Modell A)

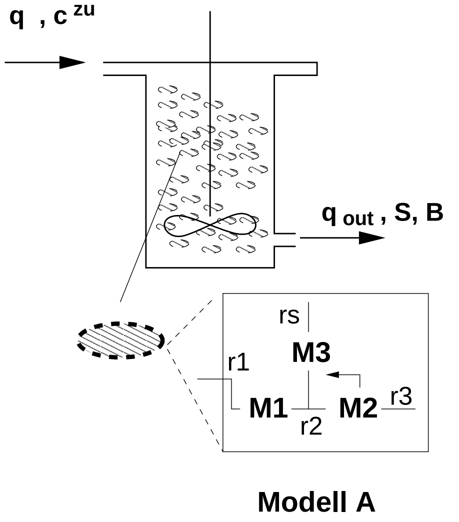


Abbildung 4.1: Reaktorsystem und biochemisches Reaktionnetzwerk. Substrat S seht zum Wachstum der Biomasse B zur Verfügung. Das intrazelluläre Netzwerk beschreibt die Interaktionen der Metabolite $M1$, $M2$ und $M3$, wobei die Ansätze für die Raten in den Varianten A und B unterschiedlich sind.

oder über die Synthese (Modell B) variiert. Damit liegen in den Raten r_2 und r_s die Modellunterschiede:

$$\begin{aligned} \text{Modell A} \quad r_2 &= f(M1, M2, M3) \\ r_s &= const. \\ \text{Modell B} \quad r_2 &= f(M1, M3) \\ r_s &= f(M2)\,. \end{aligned} \tag{4.2}$$

Zur Überprüfung der Methode wurden keine realen Daten aus dem Labor verwendet, sondern simulierte Daten, die anschließend verrauscht wurden. Abbildung 4.2 zeigt zunächst das Ausgangsexperiment mit den angepassten zwei Modellvarianten. Eine Analyse der Residuen zeigt, dass beide Modelle die Experimente in etwa gleich gut beschreiben. Gesucht sind nun Eingangsfunktionen für die Feedrate q und die Feedkonzentration c^{zu}, die das Kriterium Gleichung (4.1) maximal machen. In einer Fallstudie wurden unterschiedliche Kombinationen an Ausgangs-/Messgrößen und Wahl der Matrix Q untersucht, wobei neben einem Standard-Gradienten-Verfahren auch eine stochastische Optimierungsmethode verwendet wurde (Kremling et al.,

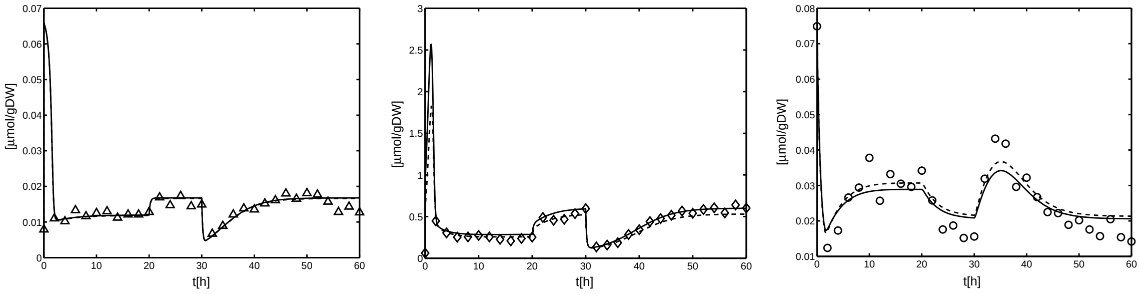


Abbildung 4.2: Messpunkte (△, ○) und simulierte Verläufe der beiden Modelle (Modell A durchgezogen, Modell B gestrichelt). Oben links: Biomasse (durchgezogen) und Substrat (in beiden Modellen gleich), oben rechts: M_1; unten links: M_2; unten rechts: M_3.

2004a). Folgende Gewichtsfunktionen Q werden untersucht:

$$Q = I \qquad \text{(Einheitsmatrix)} \tag{4.3}$$

$$Q_{ii} = \frac{1}{(\frac{x_{iA}+x_{iB}}{2})^2} \tag{4.4}$$

$$Q = (C_M + SC_A + SC_B)^{-1} \tag{4.5}$$

$$Q = (SC_A + SC_B)^{-1} . \tag{4.6}$$

Gleichung (4.4) zeigt die Gewichtung mit dem Mittelwert der Zustände aus beiden Modellen. Damit ist sicher gestellt, dass große Unterschiede im Verlauf verschiedener Zustände auf ein gleiches Maß gebracht werden. In Gleichung (4.5) bedeutet C_M die Varianz-Kovarianz-Matrix der Messfehler, wobei angenommen wird, dass die einzelnen Messgrößen nicht korreliert sind. Damit hat C_M nur Einträge in der Hauptdiagonalen. Die Matrizen SC_A und SC_B sind ebenfalls eine Diagonalmatrix mit den Elementen

$$SC_{A_ii} = \sum_j s_{A_ij}^2 \; \sigma_{p_j}^2, \tag{4.7}$$

und

$$SC_{B_ii} = \sum_j s_{B_ij}^2 \; \sigma_{p_j}^2, \tag{4.8}$$

wobei s Sensitivitäten sind, die wie folgt angeschrieben werden:

$$s_{ij} = \frac{\partial y_i}{\partial p_j}|_{\underline{p}} \tag{4.9}$$

und $\sigma_{p_j}^2$ die Varianz des Parameters p_j aus dem ersten Experiment ist. Die Sensitivitäten s_{ij} sind dynamisch zu rechnen und sind abhängig vom gewählten Arbeitspunkt (hier der aktuelle Parametersatz). Die Bedeutung der Gewichtsfunktion nach Gleichung (4.5) kann man sich wie folgt veranschaulichen. Es erfolgt eine geringe Gewichtung der Zustandsgrößen in der Zielfunktion, wenn

- die Messunsicherheit σ_{Mi}^2 der Zustandsgröße y_i hoch ist und wenn
- die Zuständsgröße y_i eine hohe Sensitivität s_{ij} gegenüber einem Parameter p_j aufweist, für den im ersten Experiment eine hohe Unsicherheit ermittelt wurde.

Tabelle 4.1: Zusammenfassung der Fallstudie (Teil 1).

			Optimierungsmethode	
Ansatz	Fall	Ausgang	Stochastisch	Gradientenbasiert
Gewichtung mit	1	***M1***	$\mathbf{3.7195 \cdot 10^6}$	$\mathbf{3.7342 \cdot 10^6}$
	2	$M2$	$9.0224 \cdot 10^{-5}$	$7.8716 \cdot 10^{-5}$
	3	E	$3.6658 \cdot 10^{-4}$	$8.015 \cdot 10^{-7}$
	4	$M1$, $M2$	$3.7198 \cdot 10^6$	$3.7342 \cdot 10^6$
Einheitsmatrix	5	$M1$, E	$3.7198 \cdot 10^6$	$3.7342 \cdot 10^6$
	6	$M2$, E	$3.6658 \cdot 10^{-4}$	$7.9476 \cdot 10^{-5}$
	7	$M1$, $M2$, E	$3.7195 \cdot 10^6$	$3.7367 \cdot 10^6$
Gewichtung mit	8	***M1***	**2.8287**	**2.8378**
Quadrat d. Mittelwerte	9	$M2$	0.0222	0.1201
	10	E	0.0476	0.0493
Q wie in Gleichung (4.4)	11	$M1, M2$	2.8394	2.8428
	12	$M1$, E	2.8685	2.8806
	13	$M2$, E	0.0686	0.0692
	14	$M1$, $M2$, E	2.8739	2.8857

Ein Vorteil des vorgestellten Ansatzes ist, dass hier die Gewichtsfunktion in Gleichung (4.5) sehr gut zu veranschaulichen ist. Tabelle 4.1 faßt die Ergebnisse der Fallstudie zusammen.

Einige Fälle sind in den Tabellen fett dargestellt. Sie stellen für die verwendeten Gewichte die optimalen Fälle dar, wobei bei annährend gleichen Zahlenwerten der Fall mit der geringeren Anzahl von Messgrößen herangezogen wurde. Interessanterweise ist in drei von vier Fällen die Verwendung einer einzelnen Größe ausreichend gewesen, um die Zielfunktion zu maximieren. Da die Gewichte in unterschiedlichen Zahlenbereichen liegen, sind auch die Werte der Zielfunktion stark unterschiedlich und können für die vier Fälle nicht verglichen werden. Abbildung 4.3 zeigt, dass mit dem Entwurf des Experimentes und der Durchführung allein noch kein befriedigendes Ergebnis erzielt werden kann.

Gezeigt sind die Eingangsfunktionen für die Feedrate und die Feedkonzentration (obere Reihe), die simulierten Vorhersagen der Modelle und die Messdaten, die sich ergeben, wenn mit dem "richtigen" Modell Daten erhoben werden. Da sich große Unterschiede zwischen den Vorher-

Tabelle 4.1: Zusammenfassung der Fallstudie (Teil 2).

			Optimierungsmethode	
Ansatz	Fall	Ausgang	Stochastisch	Gradientenbasiert
Messvarianzen &	15	$M1$	1.356	3.6054
Parametervarianzen	16	$M2$	36.3359	2.1768
	17	E	2.6896	0.099
	18	M1, M2	36.4986	7.6538
Q wie Gleichung (4.5)	19	$M1$, E	2.6897	2.5091
	20	$M2$, E	36.9019	2.4782
	21	$\boldsymbol{M1}$**,** $\boldsymbol{M2}$**,** $\boldsymbol{E}$	**37.0645**	**7.763**
nur Parameter-	22	$M1$	24.88	22.9863
varianzen	23	$\boldsymbol{M2}$	$\mathbf{1.5598 \cdot 10^{11}}$	$\mathbf{1.4355 \cdot 10^{11}}$
	24	E	3.6623	2.6398
	25	$M1$, $M2$	$1.5598 \cdot 10^{11}$	$1.4355 \cdot 10^{11}$
Q wie in Gleichung (4.6)	26	$M1$, E	24.88	3.6468
	27	$M2$, E	$1.5598 \cdot 10^{11}$	3.6207
	28	$M1$, $M2$, E	$1.5598 \cdot 10^{11}$	$1.4355 \cdot 10^{11}$

sagen und dem daraufhin durchgeführten Experiment zeigen, ist eine wiederholte Anpassung der Messdaten erforderlich. Für das vorgeschlagene (neue) Experiment lassen sich allerdings die beiden Zustandsgrößen $M1$ und $M2$ recht gut an beide Varianten anpassen. Nur für den Verlauf $M3$ findet sich kein Parametersatz von Modell A, der beide Experimente gleich gut beschreibt (Abbildung 4.4). Nur Modell B ist in der Lage, beide Experimente wiederzugeben.

Die vorgelegte Studie macht am Bespiel eines kleines biochemischen Netzwerkes deutlich, dass allein der Entwurf und die Durchführung eines neuen Experiments nicht ausreichen, um eine Modelldiskriminierung durchzuführen. Für das obige Beispiel wurde nur die Größe $M2$ für die Optimierung herangezogen. Jedoch konnte $M2$ mit beiden Modellvarianten gut angepaßt werden. Die Vorhersage des neuen Experimentes für $M3$ zeigte kaum Unterschiede. Hier war es dann allerdings nicht möglich beide Experimente anzupassen. In Kremling et al. (2004a) wurden auch die Vertrauensbereiche der Parameter mit dem ersten und mit beiden Experimenten berechnet. Es konnte gezeigt werden, dass durch das neue Experiment auch

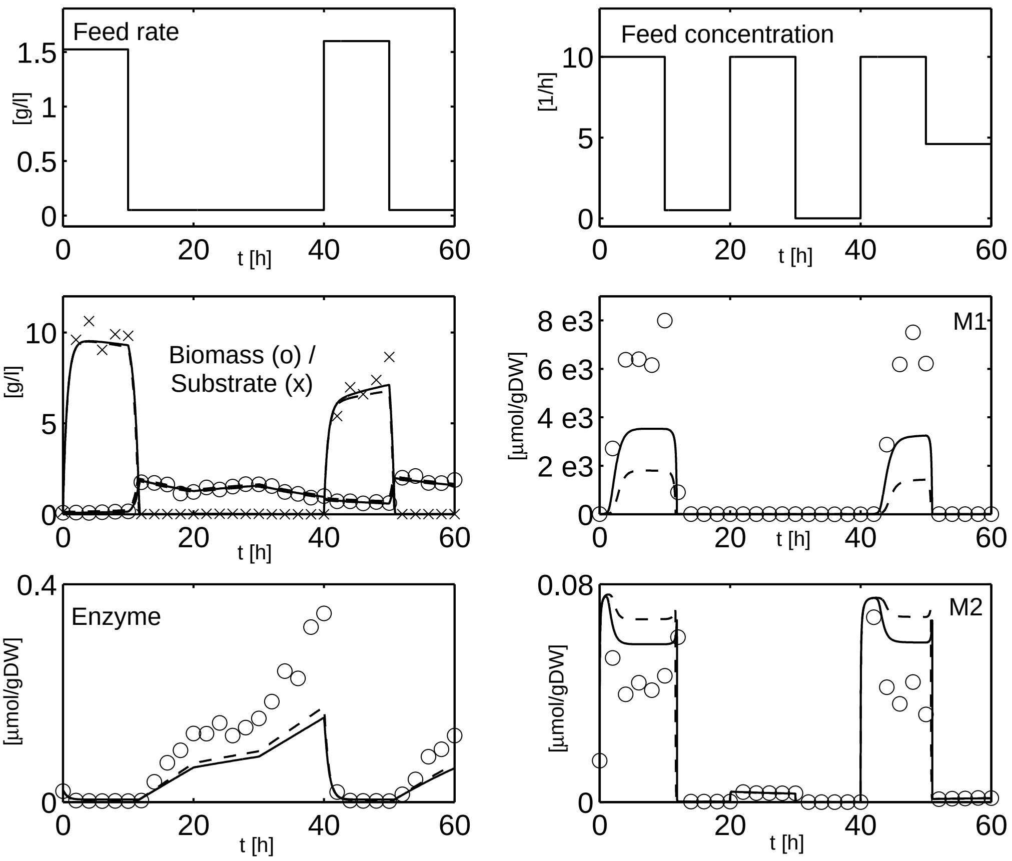


Abbildung 4.3: Optimales Experiment, entworfen mit **W** nach Fall 23. Obere Reihe: Optimale Eingangsgrößen q und c^{zu}. Zweite und dritte Reihe: Zustandsgrößen wie gekennzeichnet Modell A (durchgezogen) und Modell B (gestrichelt). Größere Unterschiede in den Modellprädiktionen finden sich im Verlauf von $M1$.

eine erhebliche Verbesserung der Konfidenzintervalle erreicht wurde.

4.2 Ermittlung von Parameterunsicherheiten mit dem Bootstrap-Verfahren

Die Ermittlung der Parameterunsicherheiten hatte sich im letzten Kapitel als ein wichtiges Element in der Versuchsplanung herausgestellt. Für Modelle, die zeitliche Verläufe aufweisen, die linear in den Parametern sind, lassen sich die Varianzen und Kovarianzen der Parameter leicht

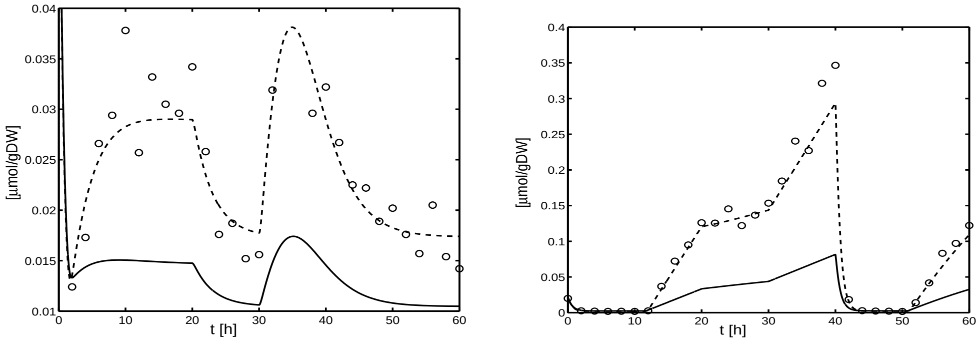


Abbildung 4.4: Links: Verlauf der Enzymkonzentration beim Ausgangsexperiment nach Anpassung beider Experimente. Rechts: Verlauf der Enzymkonzentration des neuen Experiments nach Anpassung. Modell A (durchgezogen), Modell B (gestrichelt) und Messdaten (Symbol ○).

angeben. In der Regel ist dieser Fall aber nicht gegeben und man greift wie oben gezeigt auf eine Linearisierung um einen gewählten Parameter zurück, um die Varianz-Kovarianz-Matrix zu ermitteln (die Sensitivitätsmatrix stellt diese Linearisierung dar).

Die Varianz-Kovarianz-Matrix C_P der Parameter erhält man mit der Matrix der Sensitivitäten S und der Varianz-Kovarianz Matrix der Messfehler C_M nach folgender Gleichung:

$$C_P \; = \; \left(\sum^{N} S^T \cdot C_M^{-1} \cdot S \right)^{-1} \; = \; F^{-1} \tag{4.10}$$

wobei über N Messpunkte aufsummiert wird. Der Kehrwert von C_P wird auch als Fisher-Informations-Matrix (FIM) F bezeichnet. Allerdings gilt auf Grund der Cramer-Rao-Ungleichung die folgende Beziehung für Modelle, die nichtlinear in den Parametern sind:

$$\sigma_{p_j} \; \geq \; \sqrt{\left(F^{-1}\right)_{jj}}\,, \tag{4.11}$$

d.h., es kann nur eine untere Grenze für die Parameterkonfidenzintervalle angegeben werden (Ljung, 1999). Über die obere Grenze kann keine Aussage getroffen werden, da sie stark von der Anzahl der Messpunkte und der Parameter-Nichtlinearität abhängt. Die untere Grenze wird für lineare Fälle und für eine (theoretisch) unendliche Anzahl von Messpunkten erreicht.

Eine Alternative zur Ermittlung der Konfidenzintervalle stellt das sogenannte Bootstrap-Verfahren dar (Efron and Tibshirani, 1993; DiCiccio and Efron, 1996). Dies ist ein statistisches Verfahren, um Parameterunsicherheiten zu bestimmen. Die Idee besteht darin, basierend auf dem Satz von Messdaten $\mathbf{D}$ nicht nur eine einzige Schätzung der Parameter $\underline{\hat{p}}$ durchzuführen, sondern die Messdaten im Bereich ihrer Messgenauigkeit stochastisch zu verändern und damit eine Reihe neuer Datensätze $\mathbf{D}_1^*$, $\mathbf{D}_2^*$ etc. zu generieren. Für jeden Datensatz erhält man dann einen anderen Satz an geschätzten Paramtern $\underline{\hat{p}_1^*}$, $\underline{\hat{p}_2^*}$ etc. Die Ermittlung der Parameterunsicherheiten besteht dann in der Auswertung eines Histogrammes für jeden einzelnen Parameter. Dabei werden die Parameterwerte sortiert und dann nach ihrer Häufigkeit aufgetragen. Ein Konfidenzintervall von 95% läßt sich dann direkt aus dem Histogramm ablesen.

Um zu zeigen, dass sich schon bei recht einfachen Modellen große Unterschiede ergeben können, soll eine algebraische Gleichung der unabhängigen Variablen x mit einem Parameter b betrachtet werden:

$$Y = \frac{1}{b+x} \tag{4.12}$$

Abbildung 4.5 zeigt beispielhaft die Histogramme, wenn der Bereich von x ($0 \leq x \leq 50$) 1000 Messpunkte (links) oder 25 (rechts) Messpunkte enthält. Die Daten wurden dabei mit weißem Rauschen mit einer Varianz von 20% verrauscht. Für eine große Anzahl von Messpunkten wird fast eine Normalverteilung erreicht, während bei einer kleinen Anzahl von Messpunkten ein deutlicher Bias im Histogramm zu sehen ist. Allerdings ist der Mittelwert beider Verteilungen fast identisch. Bei der Ermittlung der Vertrauensbereiche von Parameter b ergeben

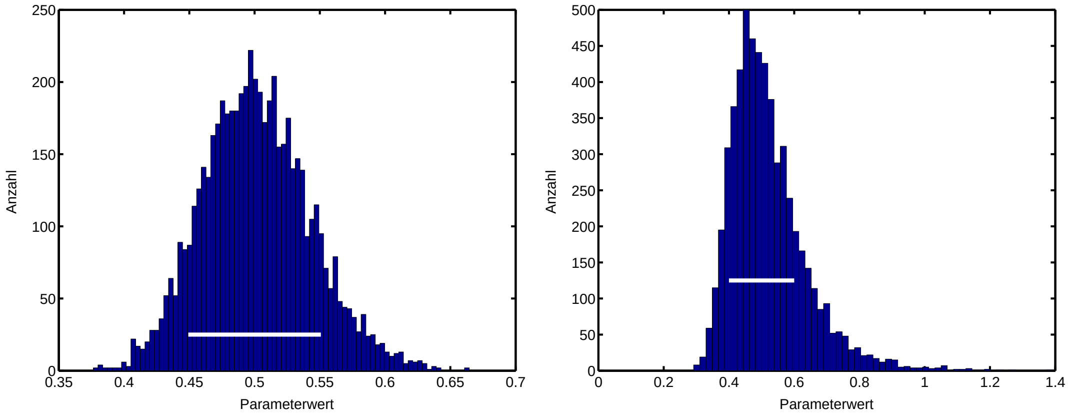


Abbildung 4.5: Histogramm für 6000 Bootstrap-Durchläufe. Für 1000 Messpunkte (links) erhält man fast eine Normalverteilung während bei 25 Messpunkten (rechts) ein deutlicher Bias zu sehen ist. Die weißen Balken geben die entsprechenden Werte des Konfidenzbereiches berechnet mit der FIM wieder.

sich Unterschiede zur FIM (weißer Balken). Die Auswertung der Histogramme liefert mit der Bootstrap-Methode in beiden Fällen größere Werte für die Parameterunsicherheit als mit der FIM.

Um Parameterunsicherheiten zu ermitteln wurde eine umfangreiche Studie an einem kleinen biochemischen Netzwerk durchgeführt (siehe Abbildung 4.6). Ausgehend von einem Standardexperiment werden Messdaten durch Simulation erzeugt, anschließend verrauscht und dann die Parameter gefittet. Um den Einfluss der Wahl verschiedener kinetischer Ausdrücke für die einzelnen Raten zu ermitteln, wurde eine ganze Reihe von Varianten untersucht, die in Tabelle 4.2 zusammengestellt sind. Abbildung 4.7 links zeigt für das Beispielsystem den Vergleich für den Parameter r_{max2} und macht den Unterschied zwischen dem Bootstrapverfahren und dem klassischen Ansatz über die Fisher-Informations-Matrix deutlich. Abbildung 4.7 rechts zeigt den zeitlichen Verlauf der Vertrauensbereiche von M_1, berechnet einmal mit dem Bootstrap-Verfahren (durchgezogene Linien) und einmal mit der FIM (gestrichelt). Der Vertrauensbereich berechnet mit dem Bootstrap-Verfahren ist deutlich größer und umfaßt ca. 95% aller Messdaten. Die Berechnung mit der FIM schneidet hierbei deutlich schlechter ab. Die Überschneidung der Kurven ergibt sich dadurch, dass für die kinetischen Parameter die

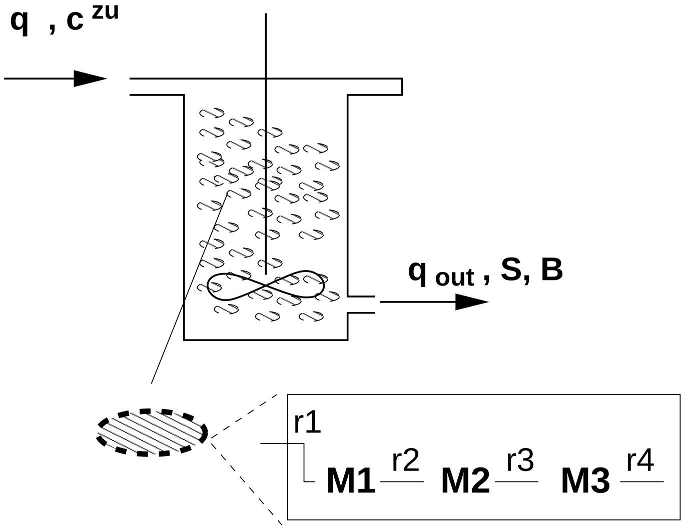


Abbildung 4.6: Reaktorsystem und biochemisches Reaktionnetzwerk. Substrat S seht zum Wachstum der Biomasse B zur Verfügung. Das intrazelluläre Netzwerk beschreibt die Interaktionen der Metabolite $M1$, $M2$ und $M3$, wobei die Ansätze für die Raten in den Varianten A und B unterschiedlich sind.

Tabelle 4.2: Übersicht kinetische Ansätze.

	Modell 1	Modell 2	Modell 3
r_1	$r_{max1}\,\frac{c_S}{c_S+K_S}$	$r_{max1}\,\frac{c_S}{c_S+K_S}$	$r_{max1}\,\frac{c_S}{c_S+K_S}$
r_2	$r_{max2}\,c_{M1}$	$r_{max2}\,c_{M1}$	$r_{max2}\,c_{M1}^m$
r_3	$r_{max3}\,\frac{c_{M2}^n}{c_{M2}^n+K_{M2}^n}$	$r_{max3}\,c_{M2}^n$	$r_{max3}\,c_{M2}^n$
r_4	$r_{max4}\,c_{M3}$	$r_{max4}\,c_{M3}$	$r_{max4}\,c_{M3}$

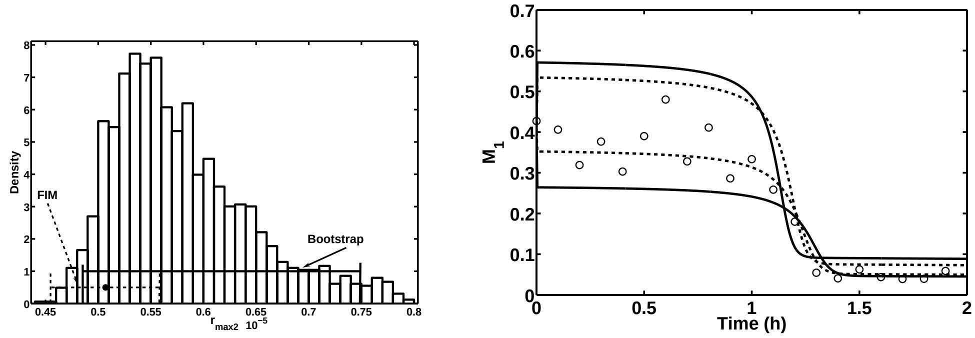


Abbildung 4.7: Links: Parametrisches Histogramm eines Parameters, ermittelt mit der Bootstrap-Methode. Zum Vergleich ist das Konfidenzintervall berechnet mit der FIM ebenfalls eingetragen. Rechts: Vergleich der beiden Verfahren bei der Vorhersage von Konfidenzbereichen der Zustände (gestrichelte Linien geben das Intervall berechnet mit der FIM wieder, durchgezogene Linie geben das Intervall berechnet mit dem Bootstrap-Verfahren wieder).

Maximal- bzw. Minimalwerte eingesetzt wurden.

Zusammengefaßt ergibt die Studie, dass für alle Varianten die Konfidenzbereiche mit der FIM kleiner sind, als die mit dem Bootstrap-Verfahren berechneten. Die Unterschiede zum Bootstrap-Verfahren lagen bei einem Fall (r_{max1} in Modell 2) bei Faktor 4. Ausserdem konnte gezeigt werden, dass mit der verbesserten Berechnung der Vertrauensbereiche der Parameter auch die Vertrauensbereiche der Zustandsgrößen genauer angegeben werden können. Im vorliegenden Beispielsystem ließen sich die Vertrauensbereiche der Zustandsgrößen einfach ermitteln, da eine lineare Reaktionssequenz vorlag. Für komplexere Systeme ist eine Bestimmung sicher nicht einfach möglich und kann u.U. nur über eine Monte-Carlo-Simulation erfolgen.

5 Zusammenfassung

Die vorliegende Schrift beschäftigt sich mit dem Einsatz systemtheoretischer Konzepte in der Systembiologie. Die Systembiologie ist eine noch recht junge Disziplin, bei der Methoden und Werkzeuge unterschiedlicher Fachrichtungen zum Einsatz kommen und die interdisziplinäres Denken für eine erfolgreiche Bearbeitung der Problemstellungen verlangt.

Zentrale Zielstellungen in der Systembiologie sind die Entwicklung leistungsfähiger mathematischer Modelle, die Aufklärung und das Verständnis von Regulationsstrukturen sowie das modellgestützte Entwerfen neuer Experimente zur Überprüfung molekularbiologischer Hypothesen. Die vorliegende Arbeit leistet zu diesen drei Aspekten Beiträge, die in den Kapiteln 2-4 zusammengefaßt und in den beiligenden Publikationen in internationalen Zeitschriften detailliert ausgeführt sind. Bei der Modellierung kommt es dabei auf eine sehr enge Verbindung mit experimentell arbeitenden Gruppen an, damit aussagekräftige Modelle gewonnen werden können. Eine optimale Kooperation konnte am MPI in Magdeburg mit der Arbeitsgruppe von Frau Dr. Bettenbrock gestaltet werden.

Zur Beschreibung der Kohlenhydrataufnahme des Bakteriums *E. coli* wurde ein sehr detailliertes Modell erstellt, welches mit einer umfangreichen Basis an experimentellen Daten verifiziert wurde (Bettenbrock et al., 2006). Dabei gelang es, alle Experimte mit einem einzigen Parametersatz zu beschreiben und so eine hohe Güte bei der Anpassung zu erreichen. Das Modell ist in der Lage, das dynamische Verhalten sowohl des Wildstammes, als auch einiger Mutantenstämme bei verschiedenen Betriebsführungen wiederzugeben. Mit dem Modell ist es in Zukunft möglich, weitere Studien in Richtung Modellreduktion oder Softwaresensorik vorzunehmen. Dabei kann das Modell als Referenz für das reale Geschehen in der Zelle eingesetzt werden. Dies erlaubt es, neue Methoden schneller und effizienter zu entwicken.

Die Analyse globaler Regulationsstrukturen ist von besonderem Interesse in der Systembiologie. Im Gegensatz zu ganz spezifischen Reizen, bspw. der Zugabe eines einzigen Zuckers, die von Seiten der Molekularbiologie bereits auch gut verstanden sind, verspricht die Aufklärung von globalen Regulationsstrukturen einen besonderen Erkenntnisgewinn. Globale Regulatoren beeinflussen eine große Anzahl von Genen und Operons und damit große Teile des Stoffwechsels. Mit einer einfachen Modellstruktur, die sich aus dem obigen Modell ableiten läßt, konnte gezeigt werden, dass ein Feedforward-Loop in der Glykolyse, der schon lange bekannt ist, zu einem robusten Verhalten der Zelle führt (Kremling et al., 2007, 2008). Die robuste Netzwerkstruktur sorgt dafür, dass sich Unsicherheiten oder Störungen nicht oder nur kaum auf die Funktion des Systems auswirken.

Die Planung von neuen Experimenten und die Bestimmung von Parameterunsicherheiten sind wichtige Methoden in der Systembiologie, um die Modelle aussagekräftig zu machen. In zwei Studien (Joshi et al., 2006b,a) wurden anhand von kleinen biochemischen Netzwerken Methoden entwickelt bzw. verbessert, die es erlauben Modelle zu diskriminieren und die Parameterunsicherheiten besser anzugeben als mit klassischen Methoden. Zur Verbesserung der Parameterschätzgüte wurde dabei ein statistisches Verfahren eingesetzt, welches es erlaubt, unter Einsatz von Monte-Carlo Verfahren und wiederholter Parameterschätzung für jeden Parameter ein parametrisches Histogramm zu erstellen, welches dann ausgewertet wird. Mit der Methode konnte gezeigt werden, dass obere und untere Schranken der Parameter so angegeben werden können, dass entsprechende Messdaten ausreichend gut wiedergegeben werden können.

6 Symbol- und Abkürzungsverzeichnis

Symbole, die in den Formeln verwendet werden:

Symbol		**Einheit**
c	Konzentration	$\mu mol/gTM$
γ	Stöchiometrischer Koeffizient	
r	Rate	$\mu mol/gTM\, h$
V_R	Reaktorvolumen	l
q_i	Zuflussrate für Substrat i	l/h
g	Molekulargewicht	g/mol
k	Geschwindigkeitskonstante	$1/h$
K	Halbsättigungswert, Bindungskonstante	$\mu mol/gTM$
η	Expressionseffizienz	$-$
α	Kinetischer Parameter für Kooperativität	$-$
h	Hillkoeffizient	$-$
μ	Spezifische Wachstumsrate	$1/h$
ϵ	Residuum	
σ^2	Varianz	
C	Varianz-Kovarianz-Matrix	
F	Fisher-Informations-Matrix	
Q	Gewichtungsmatix	
s_{ij}	Sensitivität der Zustandsgröße i bzgl. Parameter j	
S	Matrix der Sensitivitäten	

Abkürzungen von Komponenten des biochemischen Netzwerkes von *E. coli*:

Abkürzungszeichen	
PtsG	Glukose-Transport-Protein
Mlc	Repressor des *ptsG* Gens
Crp	Globaler Transkriptionsfaktor
EI	Protein des Phosphotransferasesystems
HPr	Protein des Phosphotransferasesystems
EIIA	Protein des Phosphotransferasesystems
EIIA$\sim$P	Phosphorylierte Form von EIIA
PEP	Phosphoenolpyruvat
Prv	Pyruvat
Glc6P	Glukose 6-Phosphat
TP	Triosephosphat

Literaturverzeichnis der eigenen Arbeiten der Jahre 2003 - 2008

K. Bettenbrock, S. Fischer, **A. Kremling**, K. Jahreis, T. Sauter, and E. D. Gilles. A quantitative approach to catabolite repression in *Escherichia coli*. *J. Biol.Chem.*, 281:2578–2584, 2006.

K. Bettenbrock, T. Sauter, K. Jahreis, **A. Kremling**, J. W. Lengeler, and E. D. Gilles. Analysis of the correlation between growth rate, EIIA^{Crr} (EIIA^{Glc}) phosphorylation levels and intracellular cAMP levels in *Escherichia coli* K-12. *J. Bacteriology*, 189:6891–6900, 2007.

M. Ginkel, **A. Kremling**, T. Nutsch, R. Rehner, and E. D. Gilles. Modular modeling of cellular systems with ProMoT/Diva. *Bioinformatics*, 19(9):1169–1176, 2003.

M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, the rest of the SBML Forum: A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, **A. Kremling**, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19:524–531, 2003.

M. Joshi, **A. Kremling**, and A. Seidel-Morgenstern. Model based statistical analysis of adsorption equilibrium data. *Chem. Eng. Sci.*, 61:7805–7818, 2006a.

M. Joshi, A. Seidel-Morgenstern, and **A. Kremling**. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metab. Eng.*, 8(5):447–455, 2006b.

A. Kremling. Comment on mathematical models which describe transcription and calculate the relationship between mRNA and protein expression ratio. *Biotech. Bioeng.*, 96(4):815–819, 2007.

A. Kremling and J. Saez-Rodriguez. Systems biology – an engineering perspective. *J. Biotech.*, 129(2):329–351, 2007.

A. Kremling, K. Bettenbrock, T. Sauter, S. Fischer, M. Ginkel, and E.D.Gilles. Towards whole cell "in silico" models for cellular systems: model set-up and model validation. In A. de Santis L. Benvenuti and L. Farina, editors, *Positive Systems. Proceedings of the first multidisciplinary international symposium on Positive Systems: theory and application*, pages 95–102. Springer, 2003.

A. Kremling, S. Fischer, K. Gadkar, F. J. Doyle, T. Sauter, E. Bullinger, F. Allgöwer, and E.D. Gilles. A benchmark for methods in reverse engineering and model discrimination: problem formulation and solutions. *Genome Research*, 14(9):1773–1785, 2004a.

A. Kremling, S. Fischer, T. Sauter, K. Bettenbrock, and E. D. Gilles. Time hierarchies in the *Escherichia coli* carbohydrate uptake and metabolism. *BioSystems*, 73(1):57–71, 2004b.

A. Kremling, M. Ginkel, S. Klamt, and E.D. Gilles. Workbench zur Modellbildung, Simulation und Analyse zellulärer Systeme. *it-Information Technology*, 46(1):12–19, 2004c.

A. Kremling, R. Heermann, F. Centler, K. Jung, and E. D. Gilles. Analysis of two-component signal transduction by mathematical modeling using the Kdpd/Kdpe system of *Escherichia coli*. *BioSystems*, 78(1-3):23–37, 2004d.

A. Kremling, J. Stelling, K. Bettenbrock, S. Fischer, and E.D. Gilles. Metabolic networks: Biology meets engineering sciences. In L. Alberghina and H.V. Westerhoff, editors, *Systems Biology Definitions and Perspectives*, volume 13 of *Topics in Current Genetics*, pages 215–234. 2005.

A. Kremling, K. Bettenbrock, and E. D. Gilles. Analysis of global control of *Escherichia coli* carbohydrate uptake. *BMC Systems Biology*, 1:42, 2007.

A. Kremling, K. Bettenbrock, and E. D. Gilles. A feed-forward loop guarantees robust behavior in *Escherichia coli* carbohydrate uptake. *Bioinformatics*, 24:704–710, 2008.

M. Mangold, O. Angeles-Palacios, M. Ginkel, **A. Kremling**, R. Waschler, A. Kienle, and E. D. Gilles. Computer-aided modeling of chemical and biological systems: Methods, tools, and applications. *Ind. Eng. Chem. Res.*, 44(8):2579–2591, 2005.

J. Saez-Rodriguez, **A. Kremling**, H. Conzelmann, K. Bettenbrock, and E. D. Gilles. Modular analysis of signal transduction networks. *IEEE CSM*, 24(4), 2004.

J. Saez-Rodriguez, **A. Kremling**, and E. D. Gilles. Dissecting the puzzle of life: Modularization of signal transduction networks. *Computers & Chemical Engineering*, 29(3):619–629, 2005.

J. W. Schmid, K. Mauch, M. Reuss, E.D. Gilles, and **A. Kremling**. Metabolic design based on a coupled gene expression-metabolic network model of tryptophan production in *Escherichia coli*. *Metab. Eng.*, 6(4):364–377, 2004.

Literaturverzeichnis

U. Alon, M. G. Surette, N. Barkai, and S. Leibler. Robustness in bacterical chemotaxis. *Nature*, 397:168–171, jan 1999. doi: 10.1038/16483. Letters to Editor.

S. P. Asprey and S. Macchietto. Statistical tools for optimal dynamic model building. *Comput. Chem. Eng.*, 24((2-7)):1261–1267, 2000.

N. Barkai and S. Leibler. Robustness in simple biochemical networks. *Nature*, 387, 1997.

D. Blangy, H. Buc, and J. Monod. Kinetics of the allosteric interactions of phosphofructokinase from *Escherichia coli*. *Journal of Molecular Biology*, 31:13–35, 1968.

B. H. Chen and S.P. Asprey. On the design of optimally informative dynamic experiments for model discrimination in multiresponse nonlinear situations. *Ind. Eng. Chem. Res.*, 42(7): 1379–1390, 2003.

T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3), 1996.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Chapman and Hall, 1993.

E.D. Gilles. Network theory for chemical processes. *Chem. Eng. and Technol.*, 21: 121–132, 1998

A. Kremling. *Strukturierung zellulärer Funktionseinheiten – ein signalorientierter Modellierungsansatz für zelluläre Systeme am Beispiel von* Escherichia coli. PhD thesis, Universität Stuttgart, 2002. Shaker Verlag Aachen.

A. Kremling and E.D. Gilles. The organization of metabolic reaction networks: II. Signal processing in hierarchical structured functional units. *Metab. Eng.*, 3(2):138–150, 2001.

J. W. Lengeler and K. Jahreis. Phosphotransferase systems or PTSs as carbohydrate transport and as signal transduction systems. In W.N. Konings, H.R. Kaback, and J.S. Lolkema, editors, *Handbook of Biological Physics*, chapter 25, pages 573–598. Elsevier Science B.V., 1996.

L. Ljung. *System Identification – Theory for the user.* Prentice Hall PTR, Upper Saddle River, New Jersey, second edition edition, 1999.

M. Mangold, A. Kienle, E.D. Gilles, and K.D. Mohl. Nonlinear computation in diva - methods and applications. *Chem. Eng. Sci.*, 55(2):441–454, 2000.

D. C. Montgomery, G. C. Runger, and N. F. Hubele. *Engineering Statistics.* John Wiley and Sons, Inc., 2001.

A. Munack. Some improvements in the identification of bioprocesses. In M. N. Karim and G. Stephanopoulos, editors, *Modeling and control of biotechnical processes 1992*, IFAC Symposia series, pages 89–94. IFAC, Pergamon Press, 1992.

F.C. Neidhardt, J.L. Ingraham, and M. Schaechter. *Physiology of the bacterial cell: A molecular approach.* Sinauer Associates, Sunderland, Massachusetts, 1990.

J. Plumbridge. Expression of *ptsG*, the gene for the major glucose PTS transporter in *Escherichia coli*, is repressed by Mlc and induced by growth on glucose. *Mol. Microbiol.*, 29(4): 1053–1063, 1998.

C. Posten and A. Munack. On-line application of parameter estimation accuracy to biotechnical processes. In *Proceedings of the American Control Conference*, volume 3, pages 2181–2186, 1990.

P.W. Postma, J.W. Lengeler, and G.R Jacobson. Phosphoenolpyruvate: Carbohydrate phosphotransferase systems. In F. C. Neidhardt (Editor in Chief), editor, Escherichia coli *and* Salmonella. ASM Press, Washington, D.C., 1996.

J. Stelling, A. Kremling, M. Ginkel, K. Bettenbrock, and E. D. Gilles. Towards a virtual biological laboratory. In H. Kitano, editor, *Foundations of Systems Biology*, chapter 9, pages 189–212. The MIT Press, 2001.

J. Stelling, U. Sauer, Z. Szallasi, F. J. Doyle, and J. Doyle. Robustness of cellular functions. *Cell*, 118, 2004.

J.J. Tyson, K.C. Chen, and B. Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr Opinion in Cell Biol*, 5(12):221–231, 2003.

E. B. Waygood and B. D. Sanwal. The control of pyruvate kinase of *Escherichia coli*. *J. Biol. Chem.*, 249(1):265–274, 1974.

Publikationen 2003-2008

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 281, NO. 5, pp. 2578–2584, February 3, 2006
© 2006 by The American Society for Biochemistry and Molecular Biology, Inc. Printed in the U.S.A.

A Quantitative Approach to Catabolite Repression in *Escherichia coli**[S]

Received for publication, July 25, 2005, and in revised form, October 17, 2005 Published, JBC Papers in Press, November 1, 2005, DOI 10.1074/jbc.M508090200

Katja Bettenbrock[‡1,2], **Sophia Fischer**[‡1], **Andreas Kremling**[‡], **Knut Jahreis**[§3], **Thomas Sauter**[‡], **and Ernst-Dieter Gilles**[‡]

From the [‡]*Systems Biology Group, Max-Planck-Institut für Dynamik komplexer technischer Systeme, 39106 Magdeburg, Germany and* [§]*AG Genetik, Universität Osnabrück, 49069 Osnabrück, Germany*

A dynamic mathematical model was developed to describe the uptake of various carbohydrates (glucose, lactose, glycerol, sucrose, and galactose) in *Escherichia coli*. For validation a number of isogenic strains with defined mutations were used. By considering metabolic reactions as well as signal transduction processes influencing the relevant pathways, we were able to describe quantitatively the phenomenon of catabolite repression in *E. coli*. We verified model predictions by measuring time courses of several extra- and intracellular components such as glycolytic intermediates, EIIA^{Crr} phosphorylation level, both LacZ and PtsG concentrations, and total cAMP concentrations under various growth conditions. The entire data base consists of 18 experiments performed with nine different strains. The model describes the expression of 17 key enzymes, 38 enzymatic reactions, and the dynamic behavior of more than 50 metabolites. The different phenomena affecting the phosphorylation level of EIIA^{Crr}, the key regulation molecule for inducer exclusion and catabolite repression in enteric bacteria, can now be explained quantitatively.

Catabolite repression in *Escherichia coli* designates the observation that if different carbohydrates are present in a medium under unlimited conditions, one of them is usually taken up preferentially. Although the fundamental biochemical principles of the regulatory network have been revealed, a quantitative description of this growth behavior is still missing. The center of the regulatory network is formed by the phosphoenolpyruvate (PEP)[4]:carbohydrate phosphotransferase systems (PTS). These systems are involved in both transport and phosphorylation of a large number of carbohydrates, in movement toward these carbon sources (chemotaxis), and in regulation of a number of metabolic pathways (1–3). The PTS in *E. coli* consist of two common cytoplasmatic proteins, EI (enzyme I) and HPr (histidine-containing protein), as well as an array of carbohydrate-specific EII (enzyme II) complexes. Because all components of the PTS, depending on their phosphorylation status, can interact with various key regulator proteins, the output of the PTS is represented by the degree of phosphorylation of the proteins involved in phosphoryl group transfer, *e.g.* unphosphorylated EIIA^{Crr} inhibits the uptake of other non-PTS carbohydrates by a process called inducer exclusion. Phosphorylated EIIA^{Crr} activates the adenylate cyclase (CyaA) and leads to an increase in the intracellular cAMP level.

Understanding the regulation of carbohydrate uptake requires a quantitative description of the PTS. In this context it is important that the degree of phosphorylation of EIIA^{Crr} is proportional to the PEP/pyruvate ratio, when no carbohydrates are transported (4) and the respective equilibrium constant is an upper boundary when the PTS is active (5). The PTS should therefore not be regarded as a measure for the transport of PTS substrates but more as a general measure for carbohydrate availability. One feature of our contribution in this study is the dynamic simulation of uptake and degradation of several carbohydrates using a single set of parameters and the validation of the model with measurements including the degree of phosphorylation of EIIA^{Crr}.

Previously, a detailed mathematical model describing diauxic growth behavior of *E. coli* on a mixture of glucose and lactose has been presented (6, 7). It describes the dynamics of a number of important metabolic components and enzymes, including gene expression. In the present study, using a number of isogenic strains with defined mutations we have extended the model to include the uptake of additional carbohydrates (glycerol, sucrose, and galactose). By combining these carbohydrate pathways with the signal transduction processes, it was possible to set up a comprehensive mathematical model of catabolite repression in *E. coli*. Fig. 1 sketches the metabolic pathways and regulatory interactions considered in the model.

The intention of this contribution is to model a well known biological system in a detailed, realistic, and quantitative manner, to demonstrate the use of appropriate tools and methods, and to show how the iterative process of experimental verification and model adaptations can lead to a deeper understanding of biological systems as well as the formulation of new biological problems. This is in contrast to other models of inducer exclusion, which describe the system simply as a switch between different steady states (8), exclude the dynamics of the regulatory network (9), or describe only small parts of the system (10, 11).

The overall strategy for experimental model verification comprises three ideas: (i) stimulating the system in different time frames by slow or fast changes of the environmental situation; (ii) providing conditions that allow growth under carbohydrate limited and unlimited conditions, respectively; and (iii) constructing a set of isogenic mutant strains with defined mutations in genes involved in the pathway of interest. Model predictions were verified by measuring several extra- and intracellular components, *i.e.* glycolytic intermediates, EIIA^{Crr} phosphorylation levels, protein concentrations of both LacZ and PtsG, and total cAMP concentrations under various growth conditions. The entire data base consists of 18 experiments performed with nine different strains.

---

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplements 1 and 2. Supplement 1 contains detailed model documentation, and supplement 2 provides additional information about the modeled units and the experiments used for model validation.

[1] Both authors contributed equally to this work.

[2] To whom correspondence should be addressed: Systembiologie, Max-Planck-Institut für Dynamik komplexer technischer Systeme, Sandtorstr. 1, 39106 Magdeburg, Germany. Tel.: 49-391-6110-249; E-mail: bettenbrock@mpi-magdeburg.mpg.de.

[3] Supported by the Deutsche Forschungsgemeinschaft through Sonderforschungsbreich 431.

[4] The abbreviations used are: PEP, phosphoenolpyruvate; PTS, phosphotransferase system; Hpr, histidine-containing protein, cAMP-CRP, receptor protein.

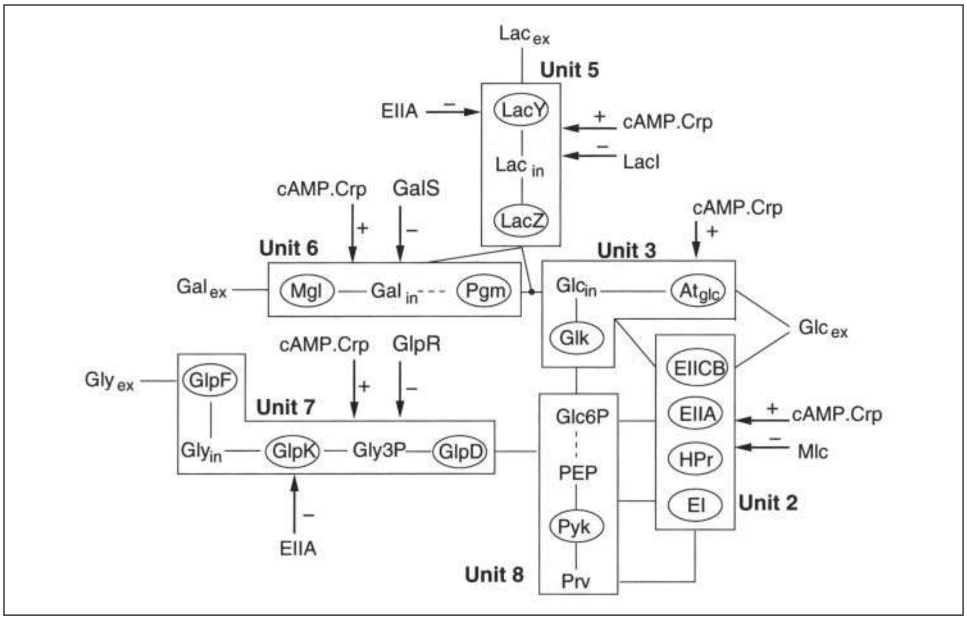


FIGURE 1. **Systems diagram of pathways and regulatory interactions considered in the model.** The graphic shows schematically the pathways and regulatory interactions included in the model. *Boxes* show functional units with similar regulation. The units correspond to the classification of equations in supplement 1. *Ellipses* symbolize enzymes, and *arrows* indicate regulatory signals (transcriptional regulation or regulation of enzyme activity). Some of these arrows correspond to unit 1 in supplement 1, as this unit describes the regulation by the cAMP·CRP transcriptional regulator. Not all enzymes and regulatory actions included in the model are shown in the graphic, *e.g.* sucrose uptake and metabolism are omitted.

MATERIALS AND METHODS

Strains, Media, and Growth Conditions—All strains used were isogenic derivatives of LJ110, constructed by standard P1*kc* transduction techniques (7). Strains and the relevant mutations are listed in Table 1. The *glk*::*cat* mutation of BL2 was taken from strain DM1000 (12). The *dgsA*::Tn*10kan* mutation was taken from strain KM563, kindly provided by W. Boos (University of Konstanz). The sucrose positive derivative of LJ110, named LJ210, was constructed essentially as follows. The 9.2-kb EcoRV/HindIII fragment from pJoe*637* (13) carrying the genes for the PTS-dependent transport and metabolism of sucrose was cloned between the inverted repeat regions of the transposon Tn*1721* yielding plasmid pKJL710. This plasmid with the artificial mini-Tn1721::*scr*$^+$, which lacks the *tnpA* gene for a transposase, was used to transform strain CSH28 F′lac (14). The *tnpA* gene was introduced into this strain by transformation of plasmid pPSO110 *tnpA*$^+$ (3). To select for transposition of the mini-Tn*1721*::*scr*$^+$ onto the F′lac, strain CSH28/F′lac/pPSO110/pKJL710 was crossed with the prototrophic Scr-negative strain PS5 (15). Standard minimal medium plates with 0.2% sucrose as sole carbon source were used for the selection of PS5/F′lac::mini-Tn*1721*::*scr*$^+$ transconjugants. A P1*kc* lysate was generated from one of the transconjugants and used to transfer the mini-Tn*1721*::*scr*$^+$ into the chromosome of LJ110. Using standard Hfr and P1 mapping techniques the insertion of the mini-Tn*1721*::*scr*$^+$ into the *E. coli* chromosome was located at 6 min (corresponding to position zag). All other strains and mutations and the construction of the plasmid F′8gal::ϕ(*ptsGop-lacZ*) have been described earlier (7).

Strains were grown in phosphate-buffered minimal medium as described previously (16). Carbohydrates were sterilized by filtration and added to the concentrations indicated in the figures. All experiments were performed at 37 °C. They were performed either in shaker flasks with volumes at least 5 times higher than the culture volume under vigorous shaking or in a Biostad B reactor (B. Braun, Biotech International). Cultures were stirred at 400 rpm and aerated with 1 liter of air/l culture volume/min. Antibiotics were added to the precultures but were omitted from the experimental culture to avoid side effects. Tetracycline was added to 10 mg/liter, chloramphenicol to 25 mg/liter, and kanamycin to 25 mg/liter.

Analytical Methods—The concentration of biomass was determined by measuring the absorbance (optical density) of the culture at either 420 or 560 nm in an Ultrospec 3000 spectrophotometer (Amersham Biosciences). Extracellular carbohydrates and acetate were determined enzymatically with the respective test kits of r-Biopharm GmbH (Germany). cAMP was quantified with the cAMP enzyme immunoassay kit from Sigma-Aldrich. All tests were performed as recommended by the manufacturer. Measurement of β-galactosidase activities was performed as described (17). The analysis of the EIIA^{Crr} phosphorylation state was carried out by Western blotting essentially as described by Takahashi *et al.* (18). Contrary to the protocol, proteins were precipitated at −80 °C at least overnight. Detection was performed with polyclonal EIIA^{Crr} antibodies from rabbit. As secondary antibodies, goat anti-rabbit antibodies conjugated with horseradish peroxidase were used, and detection was carried out by using the SuperSignal West Femto maximum sensitivity substrate (Pierce) and a cooled charge-coupled device camera system (Intas) or by exposure to films. The sum of the two EIIA^{Crr}-specific bands in each lane was set to 100%. Time courses of glycolytic metabolites (glucose 6-phosphate, fructose, fructose 6-phosphate, PEP, and pyruvate) were measured as described elsewhere (19–21).

Simulation Environment and Parameter Identification—To set up the equations and perform simulation studies, the ProMoT/Diva environment was used (22). Parameter estimation was also performed using the program Diva as described previously (7).

According to the reference model (7) the specific growth rate, μ, is assumed to be dependent on all incoming substrate fluxes r_{si}, weighted with yield coefficients Y_{si}, which leads to the formula.

$$\mu = Y_{s1}r_{s1} + Y_{s2}r_{s2} + \ldots + Y_{sn}r_{sn} \qquad \text{(Eq. 1)}$$

One problem, however, is that the yield coefficients are not constant for different growth conditions. Thus, for some experiments it was not possible to simulate the time course of biomass simultaneously with the time courses of sugar uptake. Because sugar uptake is proportional to biomass concentration and, additionally, the balances for the enzymes and proteins, rates of synthesis, degradation, and dilution term are in the same order of magnitude, incorrect predictions of the growth rate lead to large deviations in a number of state variables. To overcome this problem the following strategy has been used here. Based on the experimental data, the specific growth rate was determined as a piecewise constant function or as a spline function for each experiment and, finally, was used as an input into the system.

RESULTS

Model Formulation and Validation—To extend and validate the reference model (7), additional experiments with differing carbon sources, conditions, and isogenic mutants were performed. Additional intracellular and extracellular states (extracellular acetate, galactose, sucrose, and glycerol, degree of phosphorylation of EIIA$^{\text{Crr}}$, concentration of EIICB$^{\text{Glc}}$) were measured. Thus the PTS and its interactions with several uptake systems were observed under varying conditions. Metabolic pathways for sucrose, glycerol, and galactose were included in the model. Furthermore, a new biochemical scheme for induction of *ptsG* (23) was incorporated. The mathematical model is composed of ordinary differential and algebraic equations (DAE system). A detailed documentation of the model is included in supplement 1.

Although many kinetic parameters have been published in the past by other groups, the use of different strains and conditions in those studies made it necessary to repeat a number of experiments with a defined wild-type strain and isogenic mutant strains. The experiments provide an excellent basis for parameter estimation and model validation. Batch experiments with single growth substrates as well as with mixtures of two substrates were conducted. To analyze the influence of starting conditions, the same experiment was performed with varying preculture conditions. Pulse and fed batch experiments provided information about fast processes. Additionally, continuous cultures were used to study the behavior of the bacterial culture under limited carbohydrate conditions as well as the transition from saturating to limited carbohydrate supply. To summarize, the results from 18 different experiments were applied in the modeling process. Figures showing measurements and simulations of these experiments are shown in supplement 2. On the basis of these experiments and by application of the ProMoT/Diva environment (22) with sophisticated methods for sensitivity analysis and parameter analysis and estimation, 55 parameters could be estimated, which represents about 32% of all the parameters. The following sections of this article will demonstrate some of the results and discuss some alterations regarding the initial model.

Regulation of pts Operon and ptsG Expression by cAMP·CRP and Mlc—The glucose-specific PTS in *E. coli* consists of the cytoplasmic protein EIIA$^{\text{Crr}}$, encoded by the *crr* gene (part of the *ptsHIcrr* operon) and the membrane-bound protein EIICB$^{\text{Glc}}$ (gene *ptsG*), which transport and concomitantly phosphorylate glucose. The phosphoryl groups are transferred from PEP via successive phosphorelay reactions involving EI, HPr, EIIA$^{\text{Crr}}$, and EIICB$^{\text{Glc}}$ to the substrate. The regulation of the *ptsG* gene and of *pts* operon expression is very complex. Among others, two major regulators, the cAMP·CRP complex and the repressor Mlc (also called DgsA, gene *dgsA*) (3, 23–25), are involved. It was demonstrated that unphosphorylated EIICB$^{\text{Glc}}$ can relieve the repression of *ptsG* gene expression by sequestering Mlc from its binding sites through a direct protein-protein interaction in response to glucose (3, 26–29) (reviewed in Ref. 1). In contrast to Mlc, the cAMP·CRP complex activates *ptsG* gene expression. Because intracellular cAMP levels are low during growth on glucose, these two antagonistic regulatory mechanisms guarantee a precise adjustment of *ptsG* expression levels under various growth conditions.

Regulation of the *ptsG* gene by Mlc was incorporated into the model in order to analyze its effect on overall growth behavior. This is realized in the model as follows. For the glucose phosphorylation step of PTS, an irreversible bi-bi mechanism for the two substrates, extracellular and intracellular glucose, was applied (see Ref. 7). Either glucose or phosphorylated EIIA$^{\text{Crr}}$ bind to EIICB$^{\text{Glc}}$ at first. This means that the following conformations of the PTS transporter occur: free EIICB$^{\text{Glc}}$, the unphosphorylated complexes $\text{EIICB}^{\text{Glc}}\cdot Glc_{ex}$ and $\text{EIICB}^{\text{Glc}}\cdot Glc_{in}$ and the phosphorylated complexes $\text{EIICB}^{\text{Glc}}\cdot\text{EIIAP}$, $\text{EIICB}^{\text{Glc}}\cdot Glc_{ex}\cdot\text{EIIAP}$, and $\text{EIICB}^{\text{Glc}}\cdot Glc_{in}\cdot\text{EIIAP}$. All unphosphorylated conformations are now able to bind to Mlc but with different affinities. As a result of the parameter fitting, affinities of unphosphorylated complexes with glucose were much stronger than affinity of free EIICB$^{\text{Glc}}$. This corresponds qualitatively to experimental results.[5]

TABLE 1

Strains

Strain	Genotype	Origin
LJ110	W3110, F$^-$, Fnr$^+$	(3)
BKG47	LJ110 Δ(*ptsG*)∷*cat*	(7)
KB51	LJ110 Δ*cyaA854*∷*Tn10/6*	(7)
KB7	LJ110 *dgsA*∷*Tn10kan*	This study
LJT172	LJ110 F'8:: ϕ(*ptsG*$_{o,p}$−*lacZ*), Δ(*manXYZ*)∷*cat*	This study
LJT171	LJ110 F'8:: ϕ(*ptsG*$_{o,p}$−*lacZ*), *dgsA*∷*Tn10kan*, Δ(*manXYZ*)∷*cat*	This study
LZ110	LJ110 Δ*(argF-lacZ)169*	(3)
KB53	LJ110 *lacI3098*∷*Tn10kan*	(7)
BL2	LJ110 *glk*∷*cat*	This study
LJ210	LJ110 *zag*∷miniTn1721 *scr*$^+$	This study
PS5	S136 *recA56*	(13)
CSH28	F′lac	(14)

Using strains LZ110 and LJT172 (both *dgsA*$^+$) and the *dgsA*-negative strain LJT171 (see Table 1), *ptsG-lacZ* expression as an indicator for the amount of EIICB$^{\text{Glc}}$ was monitored by measuring the β-galactosidase activity. Bacteria were pre-grown on glycerol or glucose and inoculated on glucose, glycerol, or a mixture of glucose and glycerol, respectively. These experiments were used to estimate parameters for *ptsG* expression.

A basal activity of *ptsG* expression can be detected if cells are grown with glycerol alone (Fig. 2*F, dash-dotted line*). In cultures of strain LZ110 growing on glucose and glycerol, induction effects could be observed during the first (glucose) phase. EIICB$^{\text{Glc}}$ accumulated until the supply of glucose had been exhausted. During this growth phase, uptake of glycerol was prevented by inducer exclusion and low cAMP values. After depletion of glucose, Mlc became active and prevented further synthesis of EIICB$^{\text{Glc}}$. By dilution through growth, the amount of EIICB$^{\text{Glc}}$ was now diminished (Fig. 2*F, dashed line*). In the *dgsA* mutant strain LJT171, during growth with glucose EIICB$^{\text{Glc}}$ was synthesized in approximately the same amounts as in LZ110. Differences became obvious during the second (glycerol) growth phase. Although in LZ110 *ptsG* is repressed during growth with glycerol, in LJT171 it is induced even more strongly. This can be attributed to the lack of inhibition by Mlc and to higher intracellular cAMP levels during growth with glycerol. Extracellular cAMP accumulated in large amounts in the medium (Fig. 2*D*).

Transcription of one of the two major promoters of the *ptsHIcrr* operon (*ptsH* P0) is regulated in the same way as transcription of *ptsG*. However, the concentrations of the encoded enzymes increase only by a factor 2 or 3 (30–32) for EI and HPr, and EIIA$^{\text{Crr}}$ is almost constant, because *crr* is, in addition, transcribed by a constitutive promoter located within the *ptsI* gene (30, 32). Because of the weak effect of regulation by Mlc and cAMP·CRP on *pts* operon expression, it was neglected in the previous model (7). The concentrations of the PTS proteins were set to be constant. This model variant allowed the simulation of a number of experiments, but interestingly, one type of experiment, the so-called "disturbed" batch, could not be reproduced. In

---

[5] K Jahreis, unpublished data.

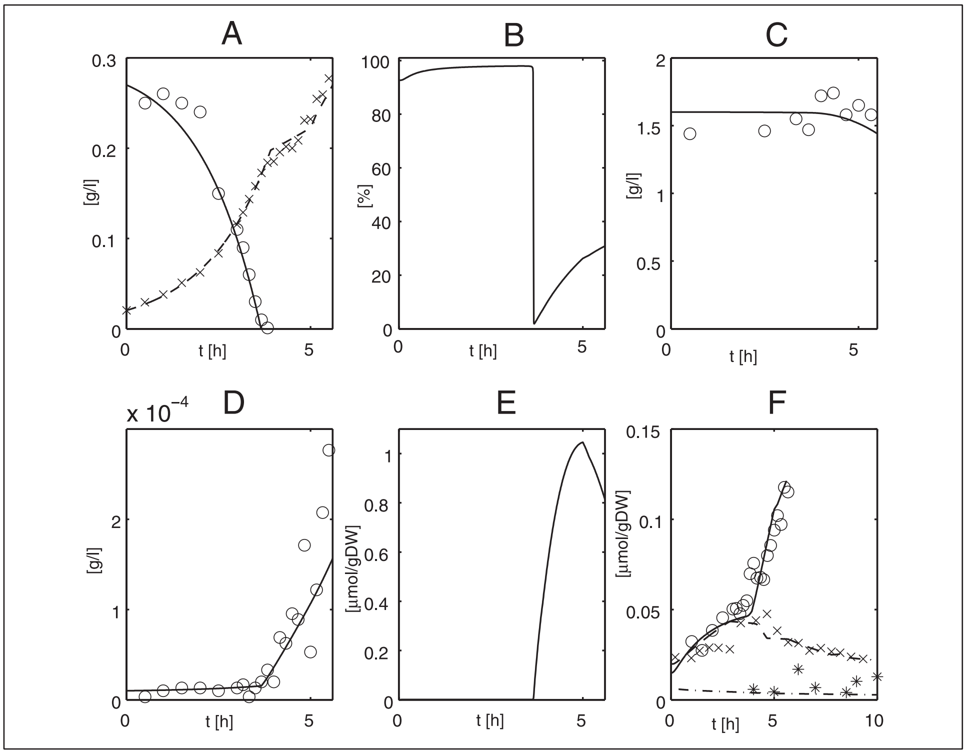


FIGURE 2. **Diauxic growth on glucose and glycerol (preculture glucose) of strain LJT171 (*dgsA*$^-$) (*plots A–F*), LZ110 (*plot F*), and LJT172 (*plot F*).** Time courses are as follows: *A*, concentrations of biomass (*dashed line, X*) and extracellular glucose (*solid line, open circles*); *B*, fraction of unphosphorylated EIIA^{Crr}; *C*, concentration of extracellular glycerol. *D*, concentration of extracellular cAMP; *E*, concentration of intracellular cAMP; *F*, concentration of EIICB^{Glc} during diauxic growth (glucose/glycerol, preculture glucose) with strain LJT171 (*solid line*), during diauxic experiment (glucose/glycerol, preculture glucose) with strain LZ110 (*dashed line*), and during growth of strain LJT172 on glycerol (preculture glycerol) (*dotted line*). *Symbols* denote measurements, and *lines* denote simulation results.

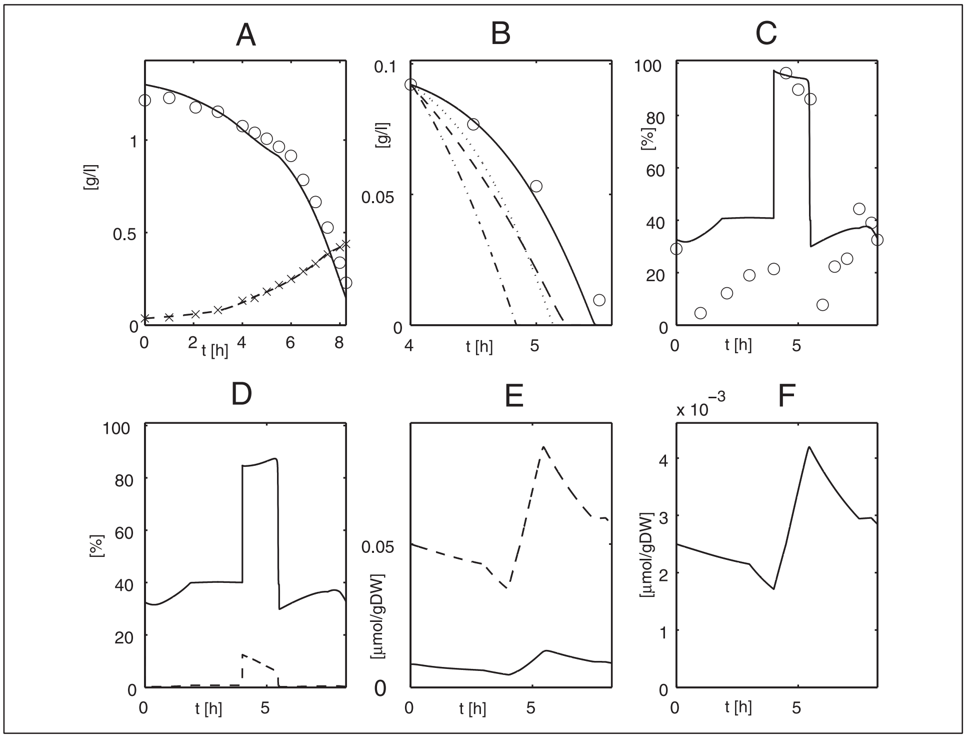


FIGURE 3. **Disturbed batch experiment with glucose pulse on culture of LJ110 (wild type) growing exponentially on glycerol (preculture glycerol).** Time courses are as follows. *A*, concentrations of biomass (*dashed line, x*) and extracellular glycerol (*solid line, open circles*). *B*, concentration of extracellular glucose. *Solid line*, model with detailed *pts* operon expression; *dashed line*, model with constant EIICB^{Glc} (0.04 μmol/gDw, grams dry weight); *dotted line*, model with constant concentrations of EI and HPr (0.006 and 0.12 μmol/gDW, grams dry weight); and *dashed-dotted line*, model with constant concentrations of EIICB^{Glc}, EI, and HPr (0.04, 0.006, and 0.12 μmol/gDW, grams dry weight). *C*, fraction of unphosphorylated EIIA^{Crr}. *D*, fractions of free EIIA^{Crr} (*solid line*) and EIIA^{Crr} bound to GlpK (*dashed line*). *E*, concentrations of EIICB^{Glc} (*solid line*) and HPr (*dashed line*). *F*, concentration of EI. *Symbols* denote measurements, and *lines* denote simulation results. *gDW*, grams dry weight.

these experiments the cells are grown on a single carbon source and glucose is pulsed in the midlog phase of growth. The feeding of glucose results in dramatic changes of the phosphorylation level of EIIA^{Crr}. Disturbed batch experiments with glycerol and lactose were performed (Fig. 3).The glucose uptake rate after the pulse was drastically reduced (to less than one-third of that during growth on glucose). Two phenomena might be responsible for this effect. First, it has been verified experimentally for *Salmonella typhimurium* (33) that PTS velocity and activity are reduced by EIIA^{Crr} binding to LacY and GlpK. These interactions are considered in the model. As a second possibility, the concentrations of PTS enzymes could be reduced down to rate-limiting conditions.

Using the set of parameters describing glucose uptake in the other experiments, simulated glucose uptake proceeded much too fast. It was not possible to fit the curve by modulating the GlpK interaction with EIIA^{Crr} or by using different fixed concentrations of the PTS proteins. Therefore regulation of the *pts* operon expression by Mlc and by

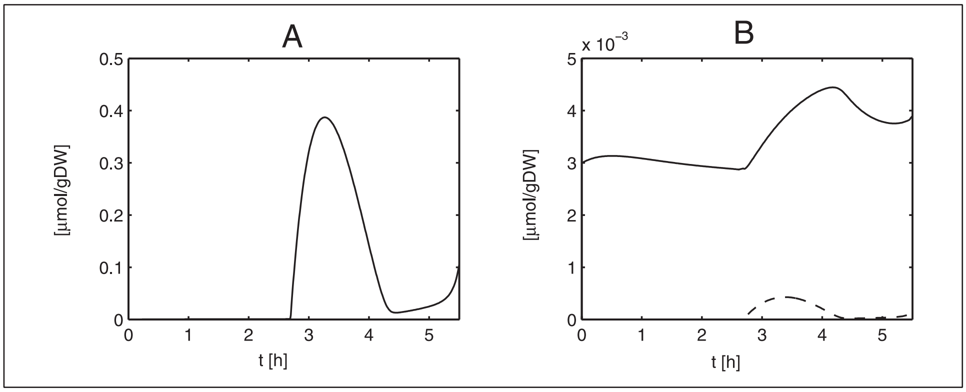


FIGURE 4. **LJ110 (wild type).** The diauxic growth experiment (glucose/lactose, preculture glucose) is shown. Time courses are as follows: *A*, concentration of cAMP; *B*, concentrations of CRP (*solid line*) and cAMP·CRP (*dashed line*). For simulation results and measurements of additional time courses see legend for Fig. 5. *gDW*, grams dry weight.

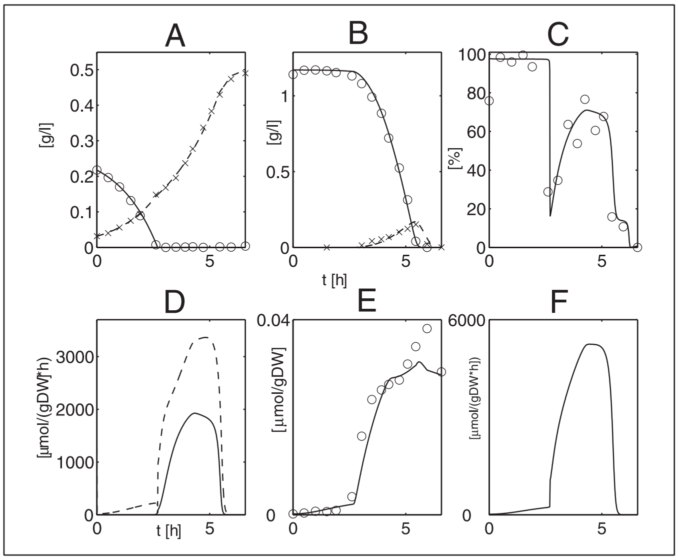


FIGURE 5. **Diauxic growth on glucose and lactose (preculture glucose) with LJ110 (wild type).** Time courses are as follows: *A*, concentrations of biomass (*dashed line, x*) and extracellular glucose (*solid line, open circles*); *B*, concentration of extracellular lactose (*solid line*) and extracellular galactose (*dashed line*); *C*, fraction of unphosphorylated EIIA^{Crr}; *D*, rates of phosphorylation of intracellular glucose by PTS (*solid line*) and glucokinase (*dashed line*); *E*, concentration of β-galactosidase; *F*, rate of lactose uptake. *Symbols* denote measurements, and *lines* denote simulation results. *gDW*, grams dry weight.

cAMP·CRP was introduced into the model. Mlc activity was modeled as described and validated for *ptsG* expression. The slow uptake of glucose during disturbed batch experiments could then be described. Fig. 3*B* presents the results of different model variants (with and without regulated expression of EIICB^{Glc}, HPr, and EI).

It can now be stated that although the expression of the *pts* operon varies only by a factor of 2–3, this variation is important for the activity of the PTS under certain growth conditions. It may be that this is most obvious when cells are shifted quickly from poor substrates like glycerol to good (PTS) substrates. Although such situations are seldom monitored in experiments, they might occur regularly in nature where the supply of nutrients can change very rapidly and dramatically. The degree of phosphorylation of EIIA^{Crr} depends on the rate of phosphorylation of glucose by the PTS, the PEP/Prv ratio (5, 4), the amount of complexation of EIIA^{Crr} with LacY, GlpK, and other enzymes, and as a result of this study, also considerably on the concentrations of the common PTS enzymes. Some of these factors might influence the activity of the PTS to a greater or lesser extent, but all contribute to the overall behavior. The model could now be used to dissect these different influences, which would be tedious to do with experiments only. The example shows that even in a very complex model with many parameters not every result can be obtained simply by fitting of parameters. If a model is carefully validated, mistakes or incorrect simplifications become obvious.

Analysis of Glucose-Lactose Diauxic Growth—The phosphorylated form of EIIA^{Crr} directly or indirectly activates the adenylate cyclase CyaA, which generates cAMP from ATP. Thus, high cAMP levels are the consequence of carbohydrate-limiting conditions. The alarmone cAMP binds to CRP, a global carbon catabolite regulator responsible for the induction of various genes. Because cAMP·CRP is a negative transcription factor for adenylate cyclase and a positive transcription factor for CRP and the PTS proteins, the system shows a number of feedback loops and is therefore highly complex.

In the case of glucose repression of the lactose uptake system, regulation due to inducer exclusion is generally believed to be the most important mechanism (34), and the cAMP·CRP complex is supposed to be mainly required for autoregulation of the lactose uptake system (4, 35) in order to prevent lactose uptake rates that are too high. The relevance of the cAMP·CRP complex during diauxic growth on glucose and lactose is seen in an acceleration of expression of the *lacZYA* operon after depletion of glucose and therefore a minimization of the lag phase between glucose and lactose utilization. However, the question has arisen as to how regulation of cAMP is realized in such a dynamic fashion (34).

Measurements for the validation of the model presented here allowed, in combination with simulations, a deep insight into the occurring processes. The model has been validated by measurements of β-galactosidase, extracellular cAMP, and the degree of phosphorylation of EIIA^{Crr} during diauxic growth of the wild type and a *lacI* mutant (strain KB53, see Table 1) on glucose and lactose. Fig. 4 shows simulation results of intracellular cAMP, CRP, and the complex of both during glucose/lactose diauxie of the wild type (Fig. 5). The results are in agreement with data presented previously (34). After depletion of glucose, cAMP rose sharply and subsequently decreased rapidly during exponential growth on lactose. Simulation results of the CRP concentration show a slight increase (factor 1.3) during the lactose uptake phase. The complex of both, which is crucial for regulation, shows qualitatively the same behavior as cAMP. This indicates that during diauxic growth the variation of cAMP concentration is much more important than the variation of CRP concentrations, which have almost no effect. This is in accordance with results from Ishizuka *et al.* (19) who reported that variation of cAMP concentration is responsible for transient repression, whereas the lowering of the CRP concentration becomes important for permanent repression. This indicates that the regulation by variation of cAMP concentrations is important for fast, dynamic processes, *e.g.* during the lag phase, whereas variation of CRP concentration becomes important for the adaptation to a certain carbon source.

In contrast to regulation of the lactose uptake system, cAMP-dependent catabolite repression seems to be the dominant mechanism of glucose repression of glycerol uptake (36). This corresponds to simulated higher cAMP and CRP levels during growth on glycerol (see supplement 2, Fig. 11).

In Fig. 5 , measured phenomena are described for growth of wild-type cells on glucose and lactose. EIIA$^{\text{Crr}}$ was unphosphorylated mainly during the uptake of glucose. Once the glucose had been consumed, EIIA$^{\text{Crr}}$ shifted very quickly to its phosphorylated form and afterward became increasingly unphosphorylated during induction of the *lac* operon and uptake of lactose. In Fig. 5, *plots C* and *F*, the correlation of unphosphorylated EIIA$^{\text{Crr}}$ and the rate of lactose uptake during the second growth phase can be seen. Uptake of glucose and lactose proceeded in the same manner for all studied strains, apart from BKG47 (supplement 2, Fig. 20). However, the measured degree of phosphorylation of EIIA$^{\text{Crr}}$ differed. Strain BL2 (supplement 2, Fig. 12) showed a very low degree of phosphorylation, whereas in KB7, a *dgsA* mutant (supplement 2, Fig. 18) EIIA$^{\text{Crr}}$ was much more phosphorylated. Strain BKG47 showed a reverse diauxic growth on glucose and lactose, and lactose was now taken up preferentially. The degree of phosphorylation of EIIA$^{\text{Crr}}$ during lactose uptake was nearly as low as for the wild type.

Another interesting feature of glucose-lactose diauxie is the production of intracellular glucose during growth on lactose. This intracellular glucose may be phosphorylated by a glucokinase (gene *glk*) or by the PTS (2, 37). Growth of two mutant strains, one of them lacking the *glk* gene (strain BL2, see Table 1) and the other one lacking the *ptsG* gene (strain BKG47, see Table 1), on lactose shows that phosphorylation via glucokinase and phosphorylation via PTS are possible *in vivo*. To solve this problem, measurements of the phosphorylation degree of EIIA$^{\text{Crr}}$ during growth of the wild type and of various mutants on lactose combined with model analysis were suitable to answer the question of the ratio of both of these fluxes.

According to the reference model (7), the predominant part of intracellular glucose had to be phosphorylated by the PTS in order to reproduce the measured degree of phosphorylation of EIIA$^{\text{Crr}}$ (not shown). However, this model was not able to reproduce measurements with BKG47 because it predicted 100% phosphorylation of EIIA$^{\text{Crr}}$. By a refinement of the pyruvate kinase kinetics, which was Michaelis-Menten-type kinetics in the reference model, to a Hill-type kinetics and subsequent fitting of the glycolysis parameters with different experiments, including a sucrose pulse experiment with measurements of glycolytic intermediates (21), the model was improved significantly and can now be used to explain the growth of all strains on lactose. Fig. 5*D* shows that with the new model about two-thirds of the intracellular glucose is phosphorylated by glucokinase and one-third by the PTS. The lower degree of EIIA$^{\text{Crr}}$ phosphorylation of strain BL2 can be explained by a higher flux via the PTS caused by the lack of glucokinase (supplement 2, Fig. 12). However, this flux was almost as high as that in strain KB7, being caused by higher concentrations of HPr and EI and resulting in a degree of phosphorylation which is even higher than in the wild type (supplement 2, Fig. 18). The most interesting result of this set of experiments was that the exchange of the pyruvate kinase kinetics from Michaelis-Menten-type to Hill-type kinetics was necessary to reproduce the measured data. The simplification of the model set up by using formal kinetic approaches leads to false results. Choosing the wrong kinetics might be without effect for many enzymes, but for enzymes that modulate key metabolites, choosing the right kinetics is crucial. This was obvious for this set of experiments because the PTS phosphorylation state is influenced mainly by the PEP to pyruvate ratio. By changing the pyruvate kinase kinetics this ratio is also influenced directly.

DISCUSSION

In this study we present a comprehensive model of *E. coli* metabolism that is able to describe uptake and degradation of several carbohydrates (glucose, lactose, glycerol, sucrose, and galactose) and to reproduce measurements of intracellular enzyme concentrations (LacZ and EIICB$^{\text{Glc}}$), glycolytic intermediates, and the phosphorylation state of EIIA$^{\text{Crr}}$. The different phenomena influencing the EIIA$^{\text{Crr}}$ phosphorylation level, which is the key regulator for catabolite repression and inducer exclusion, can now be analyzed by using this model.

In contrast to other modeling approaches, important parts of our model could be validated by a comprehensive set of experiments, *i.e.* based on the experimental data a number of uncertain or even unknown kinetic parameters could be estimated. The organism was stimulated by providing different mixtures of carbon sources using different preculture conditions, *i.e.* modifying the intracellular initial conditions and by altering the biochemical network by constructing a set of isogenic mutant strains. All experiments are described with a single set of parameters. Moreover, the dynamic behavior of the strains in a number of growth situations (diauxic growth, batch, continuous culture, disturbed batch, pulse response) could be reproduced.

The biological knowledge represented in the model has been collected from a number of publications: kinetic studies on enzymes and transcription factors as a starting point for parameter identification; genetic studies to set up a possible model structure and experimental data from array studies and proteomics to decide whether gene expression for proteins in the model has to be included.

The strategy presented in this study shows that it is possible to estimate a relatively high number of parameters even if only a limited set of measurements is available. This is based on the fact that the analysis of the system was performed under varying conditions and stimuli. Moreover, the experiments revealed that some modules (submodels) had to be refined by including more detailed knowledge of molecular biology; mainly the description of regulatory processes had to be improved. This demonstrates that mathematical models can help access regulatory processes if they are described very accurately and are validated with appropriate experiments. In addition, the data presented reveal that to set up a realistic model of good quality it is also important to carefully choose the correct kinetics. As presented here for the enzyme pyruvate kinase, the selection of the wrong kinetics (Michaelis-Menten) leads to problems in the model validation procedure. Although the use of formal kinetic approaches simplifies the model formulation, it is also a source of error. Testing different possible reaction kinetics would be necessary for the set-up of a good model.

The model presented confirms the current knowledge about catabolite repression and glucose-lactose diauxie in *E. coli*. Inducer exclusion is the most important regulatory mechanism in glucose-lactose diauxie, as described by Inada *et al*. (38), but in the case of glycerol the situation is different. This model will be used in the future to analyze these differences more thoroughly. The model hints at an important effect of cAMP concentrations during switches like that from glucose to lactose in diauxic growth as reported by Ishizuka *et al*. (19). CRP concentrations seem to be less important during such dynamic processes and might be more important for long term adaptation to different growth substrates. It has been reported that the *pts* operon is regulated by cAMP·CRP as well as by Mlc (30, 31). This regulation is weak (factor of 2–3) (32) and hence has often been neglected. The quantitative analysis, with the help of a dynamic mathematical model, was able to show the effects of this regulation. Obviously it is important if cells are shifted from feast to famine conditions. The analysis of intracellular glucose phosphorylation is another good example for the application of a quantitative mathematical model. Intracellular glucose phosphorylation by glucokinase and by PtsG has been reported in the past (33–41). The different publications propose a different share of both systems depending on the source of intracellular glucose. But such analyses are mostly qualitative. A quantitative mathematical model that is able to consider dynamic enzyme concentrations and fluxes can help to evaluate these differences. The analysis performed for the phosphorylation derived from splitting of lactose is shown in this study, but the model would allow the same analyses to be performed for growth with maltose or mellibiose as well. This might be another application of the model in the future.

As shown in this article the mathematical model confirms biological knowledge about catabolite repression and allows additional quantitative analyses. Dynamic models for cellular systems that are validated with a comprehensive set of experiments are seldom found in the literature. However, there is pressing need for "good" models, those able to describe phenomena relevant to biotechnology or medicine. The model at hand can help to develop strategies for model set-up and validation for these systems.

Acknowledgments—We thank J. W. Lengeler for many helpful and critical discussions and M. Fleischer, H. Tietgens, B. Laube, and A. Focke for technical assistance.

REFERENCES

1. Plumbridge, J. (2002) *Curr. Opin. Microbiol.* **5,** 187–193
2. Postma, P. W., Lengeler, J. W., and Jacobson, G. R. (1993) *Microbiol. Rev.* **57,** 543–594
3. Zeppenfeld, T., Larisch, C., Lengeler, J. W., and Jahreis, K. (2000) *J. Bacteriol.* **182,** 4443–4452
4. Hogema, B. M., Arents, J. C., Bader, R., Eijkemanns, K., Yoshida, H., Takahashi, H., Aiba, H., and Postma, P. W. (1998) *Mol. Microbiol.* **30,** 487–498
5. Kremling, A., Fischer, S., Sauter, T., Bettenbrock, K., and Gilles, E. D. (2004) *BioSystems* **73,** 57–71
6. Kremling, A., and Gilles, E. D. (2001) *Metab. Eng.* **3,** 138–150
7. Kremling, A., Bettenbrock, K., Laube, B., Jahreis, K., Lengeler, J. W., and Gilles, E. D. (2001) *Metab. Eng.* **3,** 362–379
8. Covert, M. W., and Palsson, B. O. (2002) *J. Biol. Chem.* **277,** 28058–28064
9. Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K., and Reuss, M. (2002) *Biotechnol. Bioeng.* **79,** 53–73
10. Setty, Y., Mayo, A. E., Surette, M. G., and Alon, U. (2003) *Proc. Natl. Acad. Sci.* **100,** 7702–7707
11. Wong, P., Gladney, S., and Keasling, J. D. (1997) *Biotechnol. Prog.* **13,** 132–143
12. Meyer, D., Schneider-Fresenius, C., Horlacher, R., Peist, R., and Boos, W. (1997) *J. Bacteriol.* **179,** 1298–1306
13. Schmid, K., Ebner, R., Altenbuchner, J., Schmitt, R., and Lengeler, J. W. (1988) *Mol. Mocrobiol.* **2,** 1–8
14. Miller, J. H. (1972) *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
15. Hochhut, B., Jahreis, K., Lengeler, J. W., and Schmid, K. (1997) *J. Bacteriol.* **179,** 2097–2102
16. Tanaka, S., Lerner, S. A., and Lin, E. C. C. (1967) *J. Bacteriol.* **93,** 642–648
17. Pardee, B., and Prestige, L. S. (1961) *Biochim. Biophys. Acta* **49,** 77–88
18. Takahashi, H., Inada, T., Postma, P., and Aiba, H. (1998) *Mol. Gen. Genet.* **259,** 317–326
19. Ishizuka, H., Hanamura, A., Kunimura, T., and Aiba, H. (1993) *Mol. Microbiol.* **10,** 341–350
20. Bergmeyer H. U. (1979) *Methoden der Enzymatischen Analyse*, Verlag Chemie, Weinheim (Bergstr.), Germany
21. Sauter, T., and Gilles, E. D. (2004) *J. Biotechnol.* **110,** 181–199
22. Ginkel, M., Kremling, A., Nutsch, T., Rehner, R., and Gilles, E. D. (2003) *Bioinformatics* **19,** 1169–1176
23. Plumbridge, J. (1998) *Mol. Microbiol.* **29,** 1053–1063
24. Roehl, R. A., and Vinopal, R. T. (1980) *J. Bacteriol.* **142,** 120–130
25. Kimata, K., Inada, T., and Aiba, H. (1998) *Mol. Microbiol.* **29,** 1509–1519
26. Lee, S.-J., Boos, W., Bouche, J. P., and Plumbridge, J. (2000) *EMBO J.* **19,** 5353–5361
27. Tanaka, Y., Kimata, K., and Aiba, H. (2000) *EMBO J.* **19,** 5344–5352
28. Nam, T. W., Cho, S. H., Shin, D., Kim, J. H., Jeong, J. Y., Lee, J. H., Roe, J. H., Peterkofsky, A., Kang, S. O., Ryu, S., and Seok, Y.J. (2001) *EMBO J.* **20,** 491–498
29. Seitz, S., Lee, S. J., Pennetier, C., Boos, W., and Plumbridge, J. (2003) *J. Biol. Chem.* **278,** 10744–10751
30. Tanaka, Y., Kimata, K., Inada, T., Tagami, H., and Aiba, H. (1999) *Genes Cells* **4,** 391–399
31. Kim, S.-Y., Nam, T.-W., Shin, D., Koo, B.-M., Seok, Y.-J., and Ryu, S. (1999) *J. Biol. Chem.* **274,** 25398–25402
32. DeReuse, H., and Danchin, A. 1988) *J. Bacteriol.* **170,** 3827–3837
33. Rohwer, J. M., Bader, R., Westerhoff, H. V., and Postma, P. W. (1998) *Mol. Microbiol.* **29,** 641–652
34. Inada, T., Kimata, K., and Aiba, H. (1996) *Genes Cells* **1,** 293–301
35. Brueckner, R., and Titgemeyer, F. (2002) *FEMS Microbiol. Lett.* **209,** 141–148
36. Holtman, C. K., Pawlik, A. C., Meadow, N. D., and Pettigrew, D. W. (2001) *J. Bacteriol.* **183,** 3336–3344
37. Nuoffer, C. B., Zanolari, B., and Erni B. (1988) *J. Biol. Chem.* **263,** 6647–6655
38. Inada, T., Takahashi, H., Mizuno, T., and Aiba, H. (1996) *Mol. Gen. Genet.* **253,** 198–204
39. Buhr, A., Daniels, G. A., and Erni, B. (1992) *J. Biol. Chem.* **267,** 3847–3851
40. Rephaeli, A. W., and Saier, M. H., Jr. (1980) *J. Bacteriol.* **141,** 658–663
41. Curtis, S. J., and Epstein, W. (1975) *J. Bacteriol.* **122,** 1189–1199

JOURNAL OF BACTERIOLOGY, Oct. 2007, p. 6891–6900 Vol. 189, No. 19
0021-9193/07/$08.00+0 doi:10.1128/JB.00819-07
Copyright © 2007, American Society for Microbiology. All Rights Reserved.

Correlation between Growth Rates, EIIA^{Crr} Phosphorylation, and Intracellular Cyclic AMP Levels in *Escherichia coli* K-12[▽]

Katja Bettenbrock,[1]* Thomas Sauter,[2] Knut Jahreis,[3] Andreas Kremling,[1]
Joseph W. Lengeler,[1] and Ernst-Dieter Gilles[1]

MPI für Dynamik Komplexer Technischer Systeme, Sandtorstr. 1, 39106 Magdeburg, Germany[1]; Institut für Systemdynamik, Universität Stuttgart, Pfaffenwaldring 9, 70569 Stuttgart, Germany[2]; and AG Genetik, Universität Osnabrück, 49069 Osnabrück, Germany[3]

Received 25 May 2007/Accepted 20 July 2007

In *Escherichia coli* K-12, components of the phosphoenolpyruvate-dependent phosphotransferase systems (PTSs) represent a signal transduction system involved in the global control of carbon catabolism through inducer exclusion mediated by phosphoenolpyruvate-dependent protein kinase enzyme IIA^{Crr} (EIIA^{Crr}) (= EIIA^{Glc}) and catabolite repression mediated by the global regulator cyclic AMP (cAMP)-cAMP receptor protein (CRP). We measured in a systematic way the relation between cellular growth rates and the key parameters of catabolite repression, i.e., the phosphorylated EIIA^{Crr} (EIIA^{Crr}~P) level and the cAMP level, using in vitro and in vivo assays. Different growth rates were obtained by using either various carbon sources or by growing the cells with limited concentrations of glucose, sucrose, and mannitol in continuous bioreactor experiments. The ratio of EIIA^{Crr} to EIIA^{Crr}~P and the intracellular cAMP concentrations, deduced from the activity of a cAMP-CRP-dependent promoter, correlated well with specific growth rates between 0.3 h^{-1} and 0.7 h^{-1}, corresponding to generation times of about 138 and 60 min, respectively. Below and above this range, these parameters were increasingly uncoupled from the growth rate, which perhaps indicates an increasing role executed by other global control systems, in particular the stringent-relaxed response system.

In *Escherichia coli*, the phosphoenolpyruvate (PEP)-dependent phosphotransferase systems (PTSs) represent important uptake systems for a number of carbohydrates which mediate transport and concomitant phosphorylation of their respective substrates (10, 44). In addition to their transport function, all components of the various PTSs of a cell form an important signal transduction system. The signal transduction properties of the PTS depend on the phosphorylation state of its proteins (26, 49). The PTSs usually consist of two general proteins, i.e., the PEP-dependent protein kinase enzyme I (EI), and the histidine-containing protein (HPr), and up to 20 different, substrate-specific enzymes II (EII). EII usually comprise two soluble domains EIIA and EIIB involved in phosphotransfer and the membrane-bound transporter domain EIIC (44). The major regulatory output signal of the PTS depends on the phosphorylation level of EIIA^{Crr} (according to its genetic nomenclature), also designated EIIA^{Glc} due to its function as the EIIA domain for the glucose-specific PTS (9, 23, 52). EIIA^{Crr} inhibits the activity of a number of non-PTS transporters and enzymes (8, 32, 33, 35, 36), a process referred to as inducer exclusion. Furthermore, the phosphorylated form of EIIA^{Crr} (EIIA^{Crr}~P) activates adenylate cyclase (1, 13, 41, 57), which in turn synthesizes cyclic AMP (cAMP) (59). The indicator molecule or alarmone cAMP is the coactivator of the important global transcription factor CRP (cAMP receptor protein). Together, they regulate in a process called cAMP-CRP-dependent catabolite repression efficient transcription of different genes involved in the synthesis of a large number of catabolic enzymes (4, 39, 43). The central role of EIIA^{Crr}~P in the activation of adenylate cyclase is largely based on mutant analysis (13, 23, 33).

The phosphorylation state of the PTS and hence the intracellular cAMP concentrations are postulated to depend largely on two major factors: (i) the uptake rate of any PTS substrate which determines the dephosphorylation rate of EI (this kinase autophosphorylates in a reversible process with PEP to generate pyruvate [49]) and (ii) the ratio of PEP to pyruvate, two central intermediate metabolites in glycolysis and gluconeogenesis which directly influence the EI autophosphorylation reaction. This ratio, however, is especially difficult to measure in vivo, and there is little corresponding data available for cells growing under different conditions (18, 27). In one thorough study, starved cells were used (16, 17). The results indicated a correlation between the EIIA^{Crr} phosphorylation level and the PEP-to-pyruvate ratio, but it is not clear how these results reflected the conditions in growing cells. Furthermore, recent in vitro reconstitution experiments indicated the putative existence of additional factors which might also modulate adenylate cyclase activity (38, 41, 42, 47). Therefore, the determination of the phosphorylation level of EIIA^{Crr} was considered an alternative test for the intracellular PEP-to-pyruvate ratio during steady-state conditions. Metabolic reactions are very fast, and hence, the PTS phosphorylation levels as well as the PEP, pyruvate, and also intracellular cAMP concentrations should quickly reach a quasi-steady-state level during growth with nonlimiting concentrations of carbohydrates.

In this paper we systematically tested the correlation between growth rates, EIIA^{Crr} phosphorylation levels (meaning in this case the ratio of EIIA^{Crr} to EIIA^{Crr}~P), extracellular

---

* Corresponding author. Mailing address: MPI für Dynamik Komplexer Technischer Systeme, Sandtorstr.1, 39106 Magdeburg, Germany. Phone: 49 391 6110249. Fax: 49 391 6110510. E-mail: bettenbrock@mpi-magdeburg.mpg.de.

▽ Published ahead of print on 3 August 2007.

Downloaded from jb.asm.org at MPI FUER DYNAMIK KOMPLEXER on May 7, 2009

cAMP concentrations, and the activity of a cAMP-CRP-dependent promoter in *E. coli* K-12 grown on different carbohydrates and with various carbohydrate concentrations but only for growth rates (μ) between 0.3 h^{-1} to 0.7 h^{-1}. Above and below these growth rates, the correlation was less clear, which supports the idea that additional regulatory elements become relevant under these conditions.

MATERIALS AND METHODS

Bacterial strains, media, and growth conditions. The two strains used in this study were LJ110, an Fnr^+ derivative of the *E. coli* K-12 mutant W3110 (22), and its genetically engineered derivative LJ210, which carries the *scr* genes for PTS-dependent transport and metabolism of sucrose integrated within its chromosome (2, 54).

Strains were grown in phosphate-buffered minimal medium (MM) as described by Tanaka et al. (58). For some bioreactor experiments, the ammonium concentration of the medium was increased to 90 mM to enable growth to higher cell densities. Carbohydrates were sterilized by filtration. If not indicated otherwise, they were added to 2 g/liter for experiments in shake flasks and to 5 g/liter for batch experiments in bioreactors. For the reporter gene assays, kanamycin was added to the cultures to 25 mg/liter. Biomass concentrations were determined by measuring the absorbance at 420 nm or at 560 nm in an Ultrospec3000 (Amersham Biosciences).

Strains were pregrown overnight in MM supplied with the same carbohydrate to be used in the experimental culture. The cultures were washed in fresh MM without the addition of carbohydrates. For experiments in shake flasks, the volumes in the flasks exceeding the culture volume at least five times, the washed cells were inoculated to 2.5×10^7 cells/ml. The cultures were incubated at 37°C under vigorous shaking (250 rpm) if not indicated otherwise. Growth was monitored by measurement of the absorbance at 420 nm. For batch experiments in bioreactors, preculture conditions were the same as for experiments in shake flasks. The cells were added to approximately 1×10^8 to 2×10^8 cells/ml. The cultures were continuously stirred under aerobic conditions (partial O_2 pressure of >20% of saturation). For experiments with various carbohydrate concentrations, the reactor was set up with 3 liters of MM supplied with 0.1 g/liter of the respective carbohydrate. Cells were added to approximately 4×10^8 cells/ml, and 1 liter/h of MM supplied with 0.8 g/liter of the respective carbohydrate was continuously fed into the bioreactor while the same volume was withdrawn. Carbohydrate and extracellular cAMP concentrations were monitored by either directly taking supernatant from the reactor with the help of a filtration module in the reactor or by taking culture samples and removing cells quickly by centrifugation at low temperatures (4°C).

Measurements of metabolite concentrations and enzyme activities. Measurements of extracellular carbohydrate concentrations were performed either enzymatically with the test kits from r-Biopharm GmbH (Germany) or on a Dionex DX-600 system (Dionex Corp.) equipped with an electrochemical detector and a Carbopac PA-100 column. Extracellular cAMP concentrations were measured with the cAMP enzyme immunoassay system (GE Healthcare) as recommended by the manufacturer. For these tests, the cells were precultured in the same medium and carbon source as used during the test. Cells washed free of cAMP as described above were inoculated to 5×10^7 cells per ml, the changes in cAMP concentrations were determined throughout a complete growth curve, and the cAMP concentration for a cell density of 5×10^8 cells per ml was obtained after interpolation. The β-galactosidase activities were determined essentially by the method of Pardee and Prestidge (37) and modified by Miller (31) and expressed in micromoles per milligram of protein and per minute. Determination of intracellular PEP and pyruvate concentrations were performed as described previously (53).

Analysis of the $EIIA^{Crr}$ phosphorylation state. The $EIIA^{Crr}$ phosphorylation state was analyzed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and Western blotting essentially as described previously (57). Deviating from the protocol, protein precipitation was carried out at −80°C overnight. Detection was performed with polyclonal $EIIA^{Crr}$ antiserum from a rabbit. As secondary antibodies, goat anti-rabbit antibodies conjugated to horseradish peroxidase were used, and detection was performed with the SuperSignal West Femto Maximum sensitivity substrate (Pierce) and a cooled charge-coupled-device camera (INTAS, Germany). The sum of the $EIIA^{Crr}$-specific bands was set at 100%.

Reporter gene studies. To analyze the activity of the cAMP-dependent *scrYp* and cAMP-independent *scrKp* promoters, fragments covering the promoter regions of both genes and approximately 200 bp upstream and downstream of the start codon including the known cAMP-CRP binding site of *scrYp* (55), were amplified by PCR, and each promoter fragment was cloned separately in front of the *luxCDABE* genes into pCS26 (3). The plasmids were transformed into *E. coli* LJ110, and the cells were grown in MM supplied with different carbohydrates. At various time points throughout a growth curve, samples were taken. The biomass concentrations of these samples were determined by measurement of the absorbance at 420 nm, and the relative luminescence units of a 100-μl sample was measured in a luminescence reader (Mithras; Berthold Technologies) (measurement time of 0.1 s). For analysis, the relative luminescence per 5×10^8 cells and the growth rates of the cultures were calculated. Average values for a specific experiment were taken from at least three measurements during the exponential phase of a culture.

RESULTS

Analysis of $EIIA^{Crr}$ phosphorylation levels and of extracellular cAMP concentrations during growth on different carbohydrates. To analyze the influence of growth rates on the $EIIA^{Crr}$ phosphorylation level and on extracellular cAMP concentrations, growth was tested first in shake flasks. Various hexoses, pentoses, and organic acids, which feed into different parts of central metabolism, were used as single carbon sources, among them PTS and non-PTS substrates. Strain LJ110 or its sucrose-positive relative LJ210 were grown in standard minimal medium supplied with saturating amounts (2 g/liter) of the carbohydrate. Additionally, the $EIIA^{Crr}$ phosphorylation levels were determined from experiments in bioreactors. As expected, the growth rates varied with different carbohydrates, but identically for both strains. Furthermore, no significant differences in growth rate and $EIIA^{Crr}$ phosphorylation could be observed between both types of experiments. Consequently, they were summarized and presented together.

If, as hypothesized initially (13, 33), adenylate cyclase was activated only by $EIIA^{Crr}{\sim}P$, then $EIIA^{Crr}$ phosphorylation levels and cAMP concentrations should correlate closely. According to the data in Table 1 and Fig. 1, high growth rates seemed to correspond to low cAMP and low $EIIA^{Crr}$ phosphorylation levels, and low growth rates seemed to correspond to high cAMP and high $EIIA^{Crr}$ phosphorylation levels.

Considering $EIIA^{Crr}$ phosphorylation levels, for relatively high growth rates, no clear distinction between PTS substrates and non-PTS substrates could be seen. Thus, about 20% of $EIIA^{Crr}$ remained phosphorylated during growth on the PTS substrates *N*-acetylglucosamine and mannitol and a similar percentage during fast growth on the non-PTS substrates lactose, L-arabinose, and gluconate.

Furthermore, a closer analysis revealed that the close correlation between growth rates, $EIIA^{Crr}$ phosphorylation levels, and cAMP concentrations was valid only for cells growing with specific growth rates (μ) between 0.3 h^{-1} and 0.7 h^{-1} (corresponding to about 140- to 60-min generation time). This was in contrast to cells growing very slowly ($\mu < 0.3$ h^{-1}), in particular those growing on acetate, D-mannose, and D-glucosamine. One major difference with these slow-growing cells was not only the $EIIA^{Crr}$ phosphorylation levels but also the extracellular cAMP concentrations deviated strongly from experiment to experiment, although both were sampled from the same cultures, and the deviations in growth rates were minor (see the error bars in Fig. 1 and Table 1). It is not clear whether this represents a systematic behavior of starved cells or whether this high variability was caused by experimental procedures, though these were highly standardized as described in Mate-

Downloaded from jb.asm.org at MPI FUER DYNAMIK KOMPLEXER on May 7, 2009

TABLE 1. EIIA^{Crr} phosphorylation levels and extracellular cAMP concentrations[a]

Carbon source	No. in figures[b]	μ (h^{-1})	EIIA^{Crr}~P level (%)	Extracellular cAMP concn (nM)	PEP-to-pyruvate ratio
D-Glucose-6-phosphate	1	0.74 ± 0.09	20 ± 4	55 ± 25	0.21 ± 0.29
D-Glucose	2	0.68 ± 0.03	5 ± 2	142 ± 24	0.15 ± 0.01
Sucrose	3	0.65 ± 0.02	10 ± 7	148 ± 18	0.45 ± 0.29
Lactose	4	0.60 ± 0.04	17 ± 7	137 ± 48	0.12 ± ND
N-Acetyl-D-glucosamine	5	0.57 ± 0.03	17 ± 10	114 ± 19	ND
D-Mannitol	6	0.58 ± 0.06	14 ± 6	137 ± 23	0.22 ± ND
L-Arabinose	7	0.51 ± 0.04	25 ± 2	111 ± 78	ND
D-Gluconate	8	0.50 ± 0.03	30 ± 9	99 ± 31	0.04 ± 0.06
Maltose	9	0.48 ± 0.07	41 ± 10	395 ± 140	0.27 ± ND
sn-Glycerol	10	0.43 ± 0.02	47 ± 4	374 ± 125	ND
D-Fructose	11	0.41 ± 0.05	24 ± 3	400 ± 450	0.40 ± 0.29
Succinate	12	0.35 ± 0.05	65 ± 16	271 ± 133	1.07 ± ND
D-Glucitol	13	0.32 ± 0.02	26 ± 10	ND	ND
D-Galactose	14	0.28 ± 0.06	59 ± 9	473 ± 96	ND
Acetate	15	0.17 ± 0.02	41 ± 24	1342 ± 1350	6.68 ± 1.95
D-Mannose	16	0.15 ± 0.03	48 ± 22	1447 ± 1142	1.4 ± 1.03
D-Glucosamine	17	0.12 ± 0.01	64 ± 9	ND	ND

[a] The growth and test conditions from batch experiments with various carbon sources were as described in Materials and Methods using strains LJ110 and LJ210. Growth rates and phosphorylated EIIA^{Crr}~P and extracellular cAMP concentrations measured at 5×10^8 cells per ml represent means ± standard deviations from at least two independent experiments. ND, not determined.

[b] Numbers in the symbols in Fig. 1 and 2.

rials and Methods. Poor correlation could perhaps indicate that under these growth conditions, extracellular cAMP concentrations do not correspond directly to intracellular cAMP concentrations, e.g., because additional factors modulate cAMP excretion from the cell and subsequent uptake into the cell, degradation of cAMP, or cAMP production, respectively (4, 12).

Deviations of another type could be seen during growth on D-fructose, D-glucitol, and D-glucose-6-phosphate. (i) On fructose and on glucitol, the phosphorylation levels of EIIA^{Crr}~P (24%) were too low, and the extracellular cAMP level during growth on fructose was too high compared to those of other carbohydrates in cells growing at similar growth rates. Growth on fructose, and in particular Δ*ptsH* mutants in which fructose-specific protein (FPr) replaces the missing histidine-containing protein (HPr), have already been reported to cause enhanced

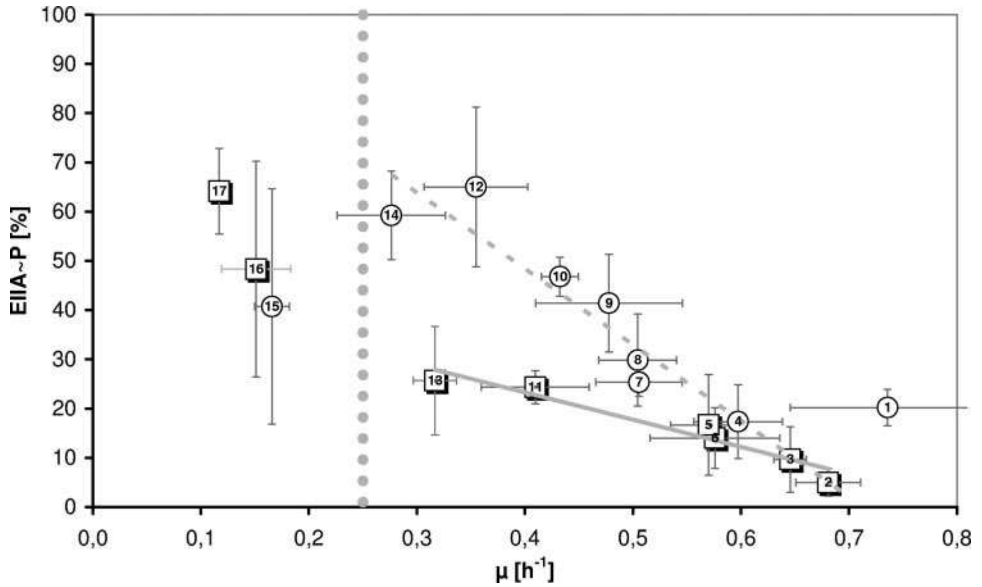


FIG. 1. Correlation of EIIA^{Crr} phosphorylation state and growth rate during growth of *E. coli* LJ110 and LJ210 with various carbon sources. EIIA^{Crr} phosphorylation levels were determined by Western blotting as described in Materials and Methods. The data represent mean values from at least two independent cultures and from at least four samples taken during exponential growth of one culture. The numbers in the symbols refer to the carbon sources as indicated in Table 1 with error bars indicating standard deviations. Circles correspond to non-PTS substrates, while squares represent PTS substrates. The gray line connecting the data points for PTS substrates represents a trend line considering all PTS data points. The dashed gray line represents a trend line considering all non-PTS substrates with exceptions glucose-6-phosphate and acetate. The trend lines show almost linear correlations between EIIA^{Crr} phosphorylation levels and growth rate. The obtained R^2s were 0.93 for the PTS and 0.86 for the non-PTS trend line. The gray dots at $\mu = 0.25\ h^{-1}$ are drawn to point out the two areas mentioned in the Discussion.

Downloaded from jb.asm.org at MPI FUER DYNAMIK KOMPLEXER on May 7, 2009

cAMP production (6, 26). Moreover, in contrast to other PTS substrates, e.g., *N*-acetylglucosamine and mannitol, which allow high growth rates, growth on fructose and on glucitol is slow and the influence of the transport reactions on the EIIA^{Crr} phosphorylation state might become visible. Different characteristic curves of the correlation of growth rate and the level of EIIA^{Crr}~P have also been predicted by a mathematical model that we have set up (2). This model predicts that the difference in both curves becomes more pronounced at lower growth rates (A. Kremling, unpublished results). These predictions are supported by the data presented in Fig. 1. (ii) Dephosphorylation was at its maximum (90 to 95%) during growth on sucrose ($\mu = 0.65\ h^{-1}$) and, in particular, on glucose ($\mu = 0.68\ h^{-1}$), but not on glucose-6-phosphate ($\mu = 0.74\ h^{-1}$; 80% dephosphorylation), the fastest growth substrate tested here. Glucose-6-phosphate, the only phosphorylated carbon source tested here, stimulates a EIICBA^{Glc}-dependent glucose/glucose-6-phosphate exchange (51). When present at high intracellular concentrations, it causes back phosphorylation of EIICB^{Glc} by glucose-6-phosphate, which would consequently result in an elevated level of EIIA^{Crr}~P.

Determination of the activity of a cAMP-CRP-dependent promoter compared to a cAMP-CRP-independent promoter. Accurate and fast measurements of intracellular cAMP concentrations are difficult and further complicated by the high extracellular cAMP concentrations which amount to 95% of the total cAMP (12, 40). Intracellular cAMP determinations always require extensive washing. Such methods are impossible to validate, as no standards for intracellular cAMP exist and the influences of washing on cAMP levels are poorly understood. To minimize the problems, "in vivo" measurements were performed by using the cAMP-dependent *scrYp* promoter and the cAMP-independent *scrKp* promoter of the *scr* regulon from pUR400 (54). Both promoters were fused independently and in the absence of the specific repressor gene *scrR* and independently, to the *luxCDABE* genes of the low-copy-number vector pCS26, as described in Materials and Methods. The usage of the *lux* reporter genes, either behind a cAMP-dependent or cAMP-independent promoter, should allow accurate measurements of the activity of the cAMP-CRP complex. This in turn should closely correlate with the active intracellular cAMP concentrations. Consequently, constitutive expression of the *scrKp* promoter represents the overall capacity of the transcriptional and translational machinery of the cells. On the other hand, transcription from *scrYp* is very low in the absence of cAMP, and transcription should increase strictly correlated to increasing intracellular cAMP levels (54). Also, because both promoter activities were measured with the same reporter genes, from the same vectors, and in the same host strain, changes in the activity of the cAMP-independent promoter *scrKp* can be used to correct for changes in *scrYp* activity due to altered growth rates and to changes in plasmid copy numbers.

Cells of strain LJ110 carrying either of both constructs were grown in parallel on minimal medium with different carbohydrates. The relative luminescence units were determined throughout batch experiments in shake flasks, and all measurements were carried out in parallel to limit further day-to-day variations. Analysis of the units measured during the exponential growth phase revealed activity variations with changing growth rates, but the changes differed in a characteristic way for the two promoters, being more pronounced for *scrYp* than for *scrKp* (Fig. 2). In addition, while the *scrKp* promoter showed considerable activity at all growth rates, the *scrYp* activity was only marginal at high growth rates.

In growing cells, proteins are diluted constantly because of the increase in cell volume followed by cell division. Therefore, to correct for higher dilution rates of proteins in faster growing cultures, the activities of both promoters were expressed in relative luminescence units multiplied with the corresponding growth rate. Analysis of these corrected units revealed a rather constant basal activity of about 4,000 corrected units for the cAMP-independent promoter *scrKp* on all carbon sources, except for acetate (1,570 units) with its exceedingly slow growth rate (Fig. 2C). This indicated that the overall capacity of transcription and translation correlated with growth rate, except for very slow growth rates. In contrast, the corrected activities of the cAMP-dependent promoter varied drastically (≥100-fold [Fig. 2D]). As before, two distinct ranges could be detected in the experiments. For growth rates higher than 0.6 h^{-1}, low intracellular cAMP concentrations were indicated by marginal activities of the cAMP-dependent *scrYp* promoter. This was expected in view of the low extracellular cAMP concentrations measured for these growth rates (Table 1). In contrast, decreasing growth rates correlated with higher *scrYp* activities, and intracellular cAMP concentrations peaked around 0.3 h^{-1} or 140-min generation time. At very slow growth rates ($\leq 0.3\ h^{-1}$), *scrYp* still showed significant activities, but here no clear correlation between growth rate and cAMP could be seen. The plot representing the corrected *scrYp* activity, i.e., intracellular cAMP (Fig. 2D), resembled the ratio of the EIIA^{Crr} phosphorylation level to growth rate (Fig. 1). Although the slope of the two curves differed for growth rates between $\geq 0.3\ h^{-1}$ to $\leq 0.7\ h^{-1}$, the plot showed the same ranges. This indicated that EIIA^{Crr} phosphorylation levels correlated, but not directly, with the intracellular cAMP concentrations.

Extracellular cAMP concentrations (Table 1) correlated less well, perhaps due to variable excretion or metabolism of cAMP. In addition, the reporter gene assays determined the activation of a promoter by the cAMP-CRP complex. This activation does not exclusively depend on the intracellular cAMP but also on the CRP concentrations (reference 10 and references therein). Data based solely on extracellular cAMP measurements thus do not necessarily mirror the true intracellular cAMP concentrations and must be considered with great caution.

Bioreactor experiments with various carbohydrate concentrations. Up to now, we investigated the correlation of growth rate, EIIA^{Crr} phosphorylation level, and cAMP concentration when the growth rate of the cells was limited by the quality of the carbon source. An alternative method to vary the growth rate is to grow the cells under different or limiting concentrations of a specific substrate. Using batch cultures as well as a chemostat-like construction by means of dialysis bags, Notley-McRobb et al. (34) reported drastic changes in intra- and extracellular cAMP concentrations at external glucose concentrations around 300 μM for *E. coli*. According to the K_m value of 3 to 10 μM for the Glc-PTS transport activity in whole cells (7, 23), a change in the phosphorylation state of EIIA^{Crr} at this

Downloaded from jb.asm.org at MPI FUER DYNAMIK KOMPLEXER on May 7, 2009

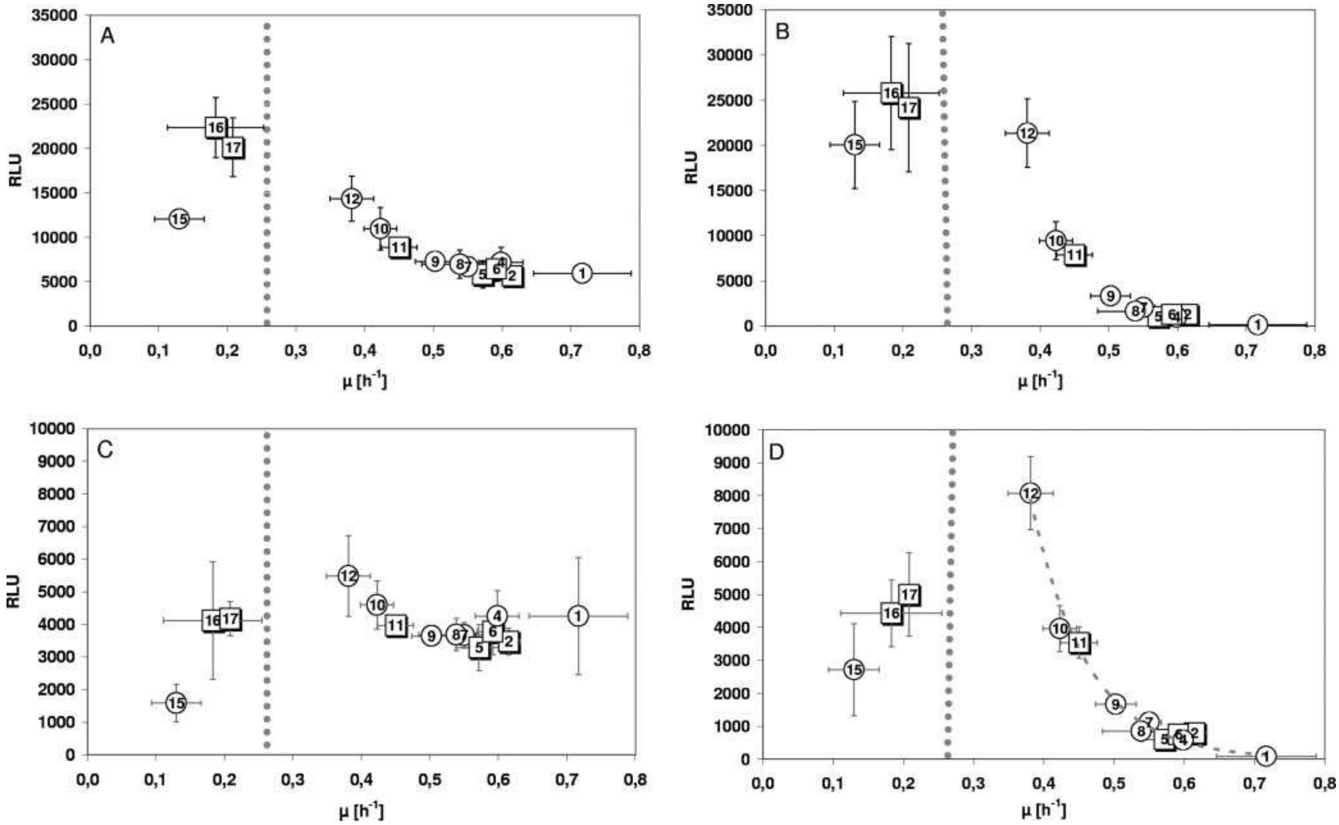


FIG. 2. Correlation of the activities of the cAMP-independent *scrKp* and cAMP-dependent *scrYp* promoters to growth rates resulting from growth on different substrates. (A and B) Relative luminescence activities (in relative light units [RLU]) of *E. coli* LJ110 carrying either the constitutive and cAMP-independent *scrKp* promoter (A) or the constitutive and cAMP-dependent *scrYp* promoter (B) fused to *luxCDABE* genes of pCS26 during batch cultures with various carbon sources. (C and D) *scrK* and *scrYp* activities, respectively, after multiplication of the RLUs with the corresponding growth rate. This standardization is done to account for differences in dilution rate of the proteins that vary with growth rate. (D)Activity of the *scrYp* promoter after multiplication of the RLUs with growth rate. By using an exponential fit, we were able to obtain a R^2 of 0.95, showing good correlation. The trend line was added to this plot as a dotted gray line. Values represent mean values from at least three independent experiments, and the error bars indicate standard deviations. The numbers in symbols correspond to carbon sources as in Table 1.

high glucose concentration seemed unlikely. Unfortunately, the EIIA^{Crr} phosphorylation state was not determined in these experiments.

During the starting phase of continuous bioreactor experiments, the carbohydrate concentration drops until it becomes limiting, i.e., growth rates are mostly determined by the decreasing external carbon source concentrations. This decrease is much slower than it is in batch experiments, allowing for a better resolution of data in the low carbohydrate concentration ranges, in particular those related to changing growth rates, cAMP concentrations, and EIIA^{Crr} phosphorylation levels. Therefore, such an experiment was performed with glucose as the carbon source, and a typical time course is shown in Fig. 3. During the first 3 hours and with high glucose concentrations, the EIIA^{Crr} phosphorylation level remained low. The slow rise in phosphorylation level probably reflects adaptation to stronger aeration within the bioreactor. At about 2.7 h, when the extracellular glucose concentration had dropped to about 40 μM, the phosphorylation level changed within minutes from about 10 to 70%, finally reaching more than 90%. At about the same time, the extracellular cAMP concentrations began to increase considerably faster than before, which we interpret as due to higher cAMP production rates. Thus, the changes in the EIIA^{Crr} phosphorylation level and in the extracellular cAMP concentration occurred simultaneously, again indicating a correlation between both parameters. Growth rates could not be calculated precisely from the data because, due to the high dilution rate, changes in biomass were very small within the relevant time window. However, growth rates as calculated from the dilution rate were estimated to vary between 0.6 h^{-1} for growth under nonlimiting glucose concentrations and about 0.33 h^{-1} for growth under limiting glucose concentrations.

A set of continuous bioreactor experiments similar to the one shown in Fig. 3 was performed with glucose, sucrose, and mannitol (Fig. 4). These three PTS carbohydrates have similar K_m values as determined in transport assays, i.e., 5 to 12 μM for glucose and the Glc-PTS (7, 23), 10 μM for sucrose and the Scr-PTS (55), and 2 to 11 μM for mannitol and the Mtl-PTS (15, 19, 25, 48), respectively. Therefore, they could be expected to give similar results. For each substrate, the phosphorylation levels of EIIA^{Crr} and the measured extracellular cAMP concentrations (data not shown) correspondingly and drastically began to increase when the external carbohydrate concentrations reached a level of 10 to 50 μM. As before, growth rates could be calculated only when based on dilution rates and

Downloaded from jb.asm.org at MPI FUER DYNAMIK KOMPLEXER on May 7, 2009

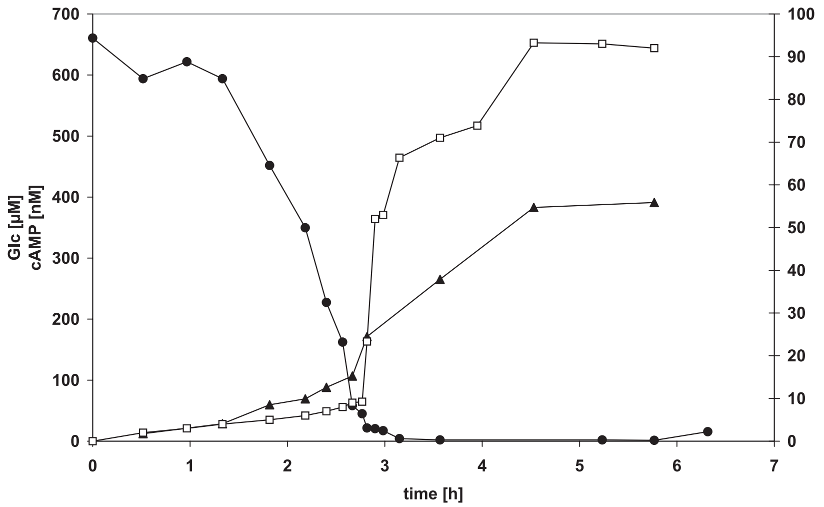


FIG. 3. Time course of an experiment with various glucose concentrations. Show are the measurements from a continuous bioreactor experiment with *E. coli* LJ110 with glucose as the carbon source. Bioreactor setup and measurements were as described in Materials and Methods. cAMP samples were taken by taken samples from the bioreactor and by rapid centrifugation of these samples at 4°C. Open circles indicating the glucose concentration (in micromolar) are given on the leftmost *y* axis as well as extracellular cAMP concentrations given in nanomolar and represented by open triangles. The EIIA^{Crr} phosphorylation level represented by filled squares is plotted on the rightmost *y* axis.

Downloaded from jb.asm.org at MPI FUER DYNAMIK KOMPLEXER on May 7, 2009

decreasing carbohydrate concentrations. They were estimated to change from about 0.55 h^{-1} to 0.33 h^{-1} within the relevant time interval.

In summary, these data show a correlation between EIIA^{Crr} phosphorylation levels, cAMP production rates, and the extracellular carbohydrate concentrations which determine the growth rates. Using a mathematical model for the glucose-PTS which was able to reproduce the experiments (2), we calculated that an apparent K_m value of 12 µM extracellular carbohydrate corresponds to a level of 50% phosphorylated EIIA^{Crr}. This was in general agreement with the known kinetics of the three PTSs. Although at a first glance the experimental setup of Notley-McRobb et al. (34) seems to allow growth experiments equivalent to our experiments in a continuous bioreactor, deviations in the experimental setup may account for the different results. Thus, at least for their dialysis cultures, it is not clear whether glucose and oxygen diffusion within the dialysis bag was sufficient. Furthermore, due to the known problems in the measurement of extracellular cAMP concentrations, it is also not clear how accurate cAMP concentrations could be determined under their experimental conditions. Comparing phosphorylation assays performed with cell extracts to growth and transport assays with whole cells, high deviations (10-fold) in the K_m values have been measured for the glucose-PTS (50)

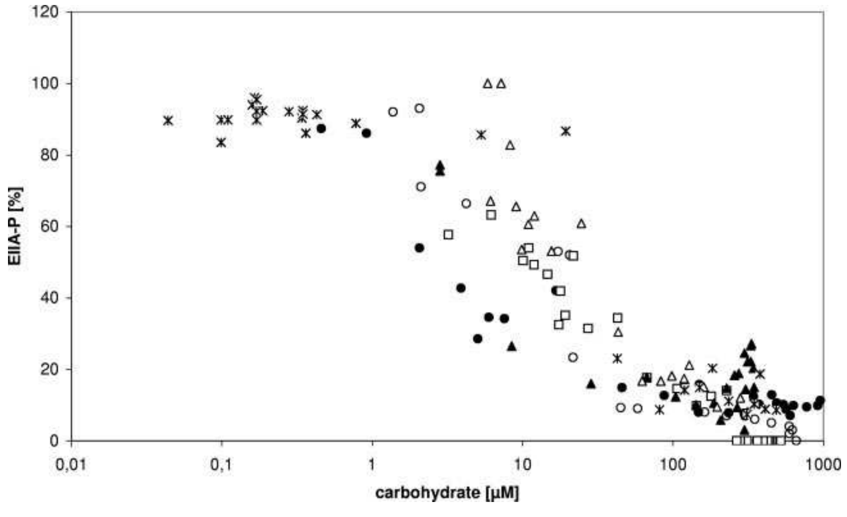


FIG. 4. Growth experiments with changing carbohydrate concentrations. The figure shows measurements from continuous bioreactor experiments with strain LJ110 or LJ210 and with glucose, sucrose, or mannitol as the carbon source. The experimental setup and measurements were as described in Materials and Methods. The EIIA^{Crr} phosphorylation level is plotted semilogarithmically against the carbohydrate concentrations. Symbols: ■, ○, and ●, experiments with glucose as the carbon source; △ and ▲, experiments with sucrose as the carbon source; *, experiment with mannitol as the carbon source.

and the three hexitol-PTSs (25). Yet it is difficult to envision, how these observations could explain the 30- to 90-fold deviation between the measured apparent K_m value for the glucose-PTS and the onset of extracellular cAMP increase as observed by Notley-McRobb et al. Besides this discrepancy, both sets of results indicated a similar correlation between growth rates, the ratio of EIIA^{Crr} to EIIA^{Crr}~P, and cAMP concentrations.

DISCUSSION

The data presented within this study demonstrate, to our knowledge for the first time, why the PTS is suited as a sensor system for the physiological state of the cell. The data show a strict correlation of growth rate, determined by the quality of the carbon source, and the EIIA^{Crr} phosphorylation state, at least for medium to high growth rates. We hypothesize that if growth is limited solely by the quality of the carbon source, i.e., by the cell's capacity to take up and metabolize the carbon source, then the PEP-to-pyruvate ratio in the cell is a direct measure of this growth rate. This PEP-to-pyruvate ratio is reflected by the phosphorylation state of EIIA^{Crr}, making this molecule an ideal candidate for sensing the physiological state of the cell. This information is subsequently transduced by the modulation of enzymatic activities, most importantly the activity of adenylate cyclase. In contrast to previous assumptions, we could show that the EIIA^{Crr} phosphorylation state not only represents a measure for the presence or absence of a PTS substrate but that the phosphotransferase system represents a universal sensor for the physiological state of the cell with respect to the carbohydrate metabolism. This is possible because the PTS phosphorylation state is directly linked to the PEP-to-pyruvate ratio and hence to the central metabolism.

In enteric bacteria, carbon and energy metabolism are controlled largely by two global regulatory mechanisms called cAMP-CRP-dependent catabolite repression and inducer exclusion. The combined phosphotransferase systems of a cell together constitute an expedient signal transduction system that senses intracellular changes in carbon catabolism and energy metabolism as changes in the phosphorylation levels of its components. All phosphoryl-transfer reactions within the PTS are reversible. Therefore, the major key parameters which determine the phosphorylation level of its components are the PEP-to-pyruvate ratio during growth on any carbon source, even a non-PTS carbon source, together with the uptake activity of the various PTSs during growth on PTS substrates. Furthermore, the EIIA^{Crr} phosphorylation level should reflect directly the PEP-to-pyruvate ratio in the cell regardless of the carbon source used for growth (49). Unfortunately, determinations of the true intracellular PEP and pyruvate concentrations are difficult to perform in growing cells. A network of reactions is coupled to the so-called pyruvate node, and the control of activity or of synthesis of the corresponding genes and enzymes has not yet been elucidated in detail. For an estimation of the true intracellular concentrations of PEP and pyruvate that may change in the millisecond range (28), a careful determination of all fluxes would be needed. At present, such an analysis can be performed only by using mutants (reference 20 and references therein) and mathematical models (18). We used measurement of the EIIA^{Crr} phosphorylation level as an alternative method. This measurement is based on the strict coupling of the PTS to the PEP-to-pyruvate ratio in the cell (49).

Correlation between growth rates and EIIA^{Crr} phosphorylation levels. Our data show a good correlation of growth rates, as determined by the quality or quantity of the carbon source, and the EIIA^{Crr} phosphorylation states, at least for medium to high growth rates. Different growth rates were obtained first by using various carbohydrates. In such cells, growth was limited by the affinity and capacity of the uptake systems, as well as by the capacity of the first metabolic reactions. High growth rates clearly correlated with lower EIIA^{Crr} phosphorylation levels, indicating a lower PEP-to-pyruvate ratio, and vice versa, but only for cells grown with generation times of about 140 to 60 min, i.e., specific growth rates between 0.3 h^{-1} and 0.7 h^{-1}.

Three apparent exceptions among the carbon sources were D-fructose, D-glucitol, and D-glucose-6-phosphate (substrates 11, 13, and 1, respectively, in Fig. 1). During growth on fructose, phosphorylation of EIIA^{Crr} was too low (~20% determined versus ~50% estimated) and extracellular cAMP levels were too high, compared to other carbon sources allowing similar growth rates. The low phosphorylation level of EIIA^{Crr} might be related to the involvement of the HPr-like protein FPr in the fructose-PTS. Thus, Δ*ptsH* mutants, in which FPr replaces the missing HPr, also show enhanced cAMP production (6, 26). Furthermore, the repressor protein FruR (alternatively Cra) of the *fru* operon is involved as an activator in the regulation of the *pps* gene, encoding PEP synthase (14; our unpublished results), and of some other glycolytic and gluconeogenic genes in *E. coli* and *Salmonella enterica* serovar Typhimurium (5, 45, 46). Another possibility is that the low phosphorylation level of EIIA^{Crr} during growth on fructose simply results from dephosphorylation of the PTS during fructose transport. This is corroborated by the fact that with glucitol, the other PTS substrate, which allows medium growth rates, the same deviation was observed. Hence we postulate a distinction between PTS substrates and non-PTS carbon sources. This distinction was very weak for substrates allowing fast growth but became more pronounced for substrates resulting in medium growth rates. This is corroborated by modeling studies which predict that at high growth rates, the PTS phosphorylation activity should have a low impact on the phosphorylation level of EIIA^{Crr}, while at low growth rates this impact increases considerably (22). Although the comparison to glucitol and the analysis with the help of the model provide an explanation for the low phosphorylation state of EIIA^{Crr} during growth on fructose, they cannot explain the high levels of extracellular cAMP during growth on fructose that we and others have observed.

The third substrate that displayed a deviating behavior was glucose-6-phosphate. During fast growth on glucose-6-phosphate, the phosphorylation level of EIIA^{Crr} was higher than expected. It had already been reported that glucose-6-phosphate did not elicit catabolite repression, although cAMP levels were very low during growth with glucose-6-phosphate (11). It is tempting to speculate that glucose-6-phosphate interferes with the PTS phosphorylation level. Glucose-6-phosphate is the product of PTS-mediated glucose uptake. High intracellular levels of glucose-6-phosphate have been reported to inhibit PTS-mediated glucose uptake (21, 24). In addition, EIICB^{Glc}/

Downloaded from jb.asm.org at MPI FUER DYNAMIK KOMPLEXER on May 7, 2009

$EIIA^{Crr}$-mediated cross phosphorylation of glucose by glucose-6-phosphate has been reported (51). During growth on glucose-6-phosphate, high intracellular glucose-6-phosphate levels apparently allow rephosphorylation of the PTS, explaining the elevated phosphorylation levels of $EIIA^{Crr}$.

Compared to the batch experiments, the continuous bioreactor experiments allow a better resolution of data within the transition phase from fast growth on high substrate concentrations to lower growth rates caused by decreasing external substrate concentrations. We used three PTS substrates, i.e., D-glucose, sucrose, and D-mannitol, with similar transport K_m values, of which the former two PTSs use $EIIA^{Crr}$ as their phosphate donor. The corresponding results did show the same general trend as the previous experiments, i.e., higher growth rates correlated closely with decreased phosphorylation of $EIIA^{Crr}$. In particular, they did not show a significant increase in the $EIIA^{Crr}$ phosphorylation level before the substrate concentrations decreased to concentrations in the range of the K_m values (Fig. 3 and 4). Obviously, cells coordinate $EIIA^{Crr}$ phosphorylation and cAMP levels with growth rates in a similar way, whether cell growth was limited because of the nature or amount of the carbon source used.

Correlation between growth rates and intracellular cAMP levels. A second key parameter besides the PEP-to-pyruvate ratio in the control of carbon catabolism is the alarmone cAMP. This coactivator of the global transcription factor CRP is essential in controlling the synthesis of several hundred genes and catabolic enzymes. In a Δ*cyaA* mutant of *E. coli*, increasing growth rates on glucose could be obtained by adding increasing amounts of cAMP to such cells. Furthermore, different cAMP concentrations corresponded to the growth rate on diverse carbon sources (12). Unfortunately, we and others have been unable until now to test intracellular cAMP concentrations in growing cells directly, rapidly, and in a reliable way (30, 40). As shown in Table 1, extracellular cAMP concentrations have no simple relation to the intracellular, i.e., biologically active, cAMP amounts. Attempts to correlate intracellular cAMP levels with β-galactosidase activities transcribed from a constitutively expressed *lacZp* promoter (12) have not sufficiently taken into consideration indicator protein dilution, thus presenting an incomplete picture of the cell's physiology under various growth conditions. In an extension of such in vivo studies, we compared the activities from a cAMP-dependent promoter and a cAMP-independent promoter of the *scr* regulon. These constructs allowed correction for changes in promoter activities due to altered growth rates and concomitant protein dilution rates and to plasmid copy number effects. In agreement with the $EIIA^{Crr}$ phosphorylation tests, these corrected data also indicated, first a major (central) phase, valid from medium to high growth rates (Fig. 2). Within this phase, the constitutively expressed and cAMP-independent promoter *scrKp* showed variations of less than twofold in its corrected promoter activities over the range of growth rates between 0.3 h^{-1} and 0.7 h^{-1}. Because the corrected *scrKp* activities represent basically the general transcriptional and translational capacity of the cell, this capacity seems to correlate strictly with growth rates within this central range. The marked deviation during growth on acetate probably indicates the increasing starvation stress. This contrasted with the cAMP-dependent promoter *scrYp*, whose equally corrected, activities varied more than 100-fold within this central range of growth rates. Such drastic changes must obviously be attributed largely to changes in the intracellular cAMP concentrations and more precisely in the amount of active cAMP-CRP complexes. Similar to the $EIIA^{Crr}$ phosphorylation level, the intracellular cAMP levels had a clear maximum around a μ of 0.3 h^{-1} (or 140-min generation time) (Fig. 2), and no difference between PTS and non-PTS substrates could be seen. Apparently, the physiologically relevant intracellular cAMP concentrations (Fig. 2) cannot be deduced easily from the extracellular cAMP concentrations (Table 1).

Conclusions. To the best of our knowledge, we show here for the first time that below and above a specific growth rate of 0.3 h^{-1} and 0.7 h^{-1}, the $EIIA^{Crr}$ phosphorylation state and the activity of the cAMP-CRP-dependent promoter as represented by the *scrYp* activity became increasingly uncoupled from the growth rate. These deviations were not simply a consequence of the large inaccuracies in the corresponding tests. Rather, the results seem to indicate additional factors also modulating the PEP-to-pyruvate ratio, $EIIA^{Crr}{\sim}P$ levels, and cAMP production or the activity of cAMP-CRP-dependent promoters. For low growth rates, both promoter activities and the $EIIA^{Crr}{\sim}P$ level decreased roughly in parallel, while for high growth rates, the three key parameters became constant. Apparently, cAMP-dependent gene activation becomes less and less relevant during either very fast ($\mu \geq 0.7\ h^{-1}$), or very slow growth ($\mu \leq 0.3\ h^{-1}$), perhaps indicative of more global cellular changes in the corresponding cells. Thus, between specific growth rates of 0.4 h^{-1} to 0.8 h^{-1}, the mean cell volume increases from 0.55 to 1.38 μm^3, the number of ribosomes per cell doubles, and the ppGpp concentration decreases from 150 to 25 μM (29, 56). Such extreme physiological conditions which trigger stress responses seem to be increasingly controlled by other global antistress regulatory networks, e.g., RpoS in slow-growing and prolonged starving cells, and the "stringent-relaxed" control system which in fast-growing cells mainly determines growth rates. Similarly, other global regulatory systems can be expected to become active in cells growing under, e.g., phosphate or nitrogen limitation, and hence might override the regulation by cAMP-CRP. Finally, highly different $EIIA^{Crr}$ phosphorylation levels during growth on poor growth substrates seem to indicate that under such extreme conditions PTS phosphorylation can be uncoupled from the PEP-to-pyruvate ratio, corroborating data pointing to additional factors which also control adenylate cyclase activity (34, 38).

ACKNOWLEDGMENTS

We thank Andrea Focke, Helga Tietgens, and Marion Fleischer for excellent technical assistance.

K.J. was supported by the Deutsche Forschungsgemeinschaft through Sonderforschungsbereich 431, TP14.

REFERENCES

1. **Amin, N., and A. Peterkofsky.** 1995. A dual mechanism for regulating cAMP levels in *Escherichia coli*. J. Biol. Chem. **270:**11803–11805.
2. **Bettenbrock, K., S. Fischer, A. Kremling, K. Jahreis, T. Sauter, and E. Gilles.** 2006. A quantitative approach to catabolite repression in *Escherichia coli*. J. Biol. Chem. **281:**2578–2584.
3. **Bjarnason, J., C. M. Southward, and M. G. Surette.** 2003. Genomic profiling of iron-responsive genes in *Salmonella enterica* serovar Typhimurium by high-throughput screening of a random promoter library. J. Bacteriol. **185:** 4973–4982.
4. **Botsford, J. L., and J. G. Harman.** 1992. Cyclic AMP in procaryotes. Microbiol. Rev. **56:**100–122.

Downloaded from jb.asm.org at MPI FUER DYNAMIK KOMPLEXER on May 7, 2009

5. **Chin, A. M., D. A. Feldheim, and M. H. Saier, Jr.** 1989. Altered transcriptional patterns affecting several metabolic pathways in strains of *Salmonella typhimurium* which overexpress the fructose regulon. J. Bacteriol. **171:**2424–2434.
6. **Crasnier-Mednansky, M., M. C. Park, W. K. Studley, and M. H. Saier, Jr.** 1997. Cra-mediated regulation of *Escherichia coli* adenylate cyclase. Microbiology **143:**785–792.
7. **Curtis, S. J., and W. Epstein.** 1975. Phosphorylation of D-glucose in *Escherichia coli* mutants defective in glucosephosphotransferase, mannosephosphotransferase, and glucokinase. J. Bacteriol. **122:**1189–1199.
8. **de Boer, M., C. P. Broekhuizen, and P. W. Postma.** 1986. Regulation of glycerol kinase by enzyme $\mathrm{II}^{\mathrm{Glc}}$ of the phosphoenolpyruvate:carbohydrate phosphotransferase system. J. Bacteriol. **167:**393–395.
9. **De Reuse, H., and A. Danchin.** 1988. The *ptsH*, *ptsI*, and *crr* genes of *Escherichia coli* phosphoenolpyruvate-dependent phosphotransferase system: a complex operon with several modes of transcription. J. Bacteriol. **170:**3827–3837.
10. **Deutscher, J., C. Francke, and P. W. Postma.** 2006. How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. Microbiol. Mol. Biol. Rev. **70:**939–1031.
11. **Dumay, V., A. Danchin, and M. Crasnier.** 1996. Regulation of *Escherichia coli* adenylate cyclase activity during hexose phosphate transport. Microbiology **142:**575–583.
12. **Epstein, W., L. B. Rothman-Denes, and J. Hesse.** 1975. Adenosine 3:5-cyclic monophosphate as mediator of catabolite repression in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **72:**2300–2304.
13. **Feucht, B. U., and M. H. Saier, Jr.** 1980. Fine control of adenylate cyclase by the phosphoenolpyruvate:sugar phosphotransferase systems in *Escherichia coli* and *Salmonella typhimurium*. J. Biol. Chem. **141:**603–610.
14. **Geerse, R. H., J. Vanderpluijm, and P. W. Postma.** 1989. The repressor of the PEP-fructose phosphotransferase system is required for the transcription of the *pts* gene of *Escherichia coli*. Mol. Gen. Genet. **218:**348–352.
15. **Grenier, F. C., E. B. Waygood, and M. H. Saier, Jr.** 1986. The bacterial phosphotransferase system: kinetic characterization of the glucose, mannitol, glucitol and N-acetylglucosamine systems. J. Cell. Biochem. **31:**97–105.
16. **Hogema, B. M., J. C. Arents, R. Bader, K. Eijkemans, H. Yoshida, H. Takahashi, H. Alba, and P. W. Postma.** 1998. Inducer exclusion in *Escherichia coli* by non-PTS substrates: the role of the PEP to pyruvate ratio in determining the phosphorylation state of enzyme $\mathrm{IIA}^{\mathrm{Glc}}$. Mol. Microbiol. **30:**487–498.
17. **Hogema, B. M., J. C. Arents, T. Inada, H. Aiba, K. vanDam, and P. W. Postma.** 1997. Catabolite repression by glucose 6-phosphate, gluconate and lactose in *Escherichia coli*. Mol. Microbiol. **24:**857–867.
18. **Holms, H.** 1996. Flux analysis and control of the central metabolic pathways in *Escherichia coli*. FEMS Microbiol. Rev. **19:**85–116.
19. **Jacobson, G. R., C. A. Lee, J. E. Leonard, and M. H. Saier, Jr.** 1983. Mannitol-specific enzyme II of the bacterial phosphotransferase system. I. Properties of the purified permease. J. Biol. Chem. **258:**10748–10756.
20. **Kao, K. C., L. M. Tran, and J. C. Liao.** 2005. A global regulatory role of gluconeogenic genes in *Escherichia coli* revealed by transcriptome network analysis. J. Biol. Chem. **280:**36079–36087.
21. **Kornberg, H. L.** 1973. Nature and regulation of sugar uptake by *Escherichia coli*. Proc. Aust. Biochem. Soc. **6:**4.
22. **Kremling, A., K. Bettenbrock, B. Laube, K. Jahreis, J. Lengeler, and E. Gilles.** 2001. The organization of metabolic reaction networks. III. Application for diauxic growth on glucose and lactose. Metab. Eng. 362–379.
23. **Lengeler, J., A. M. Auburger, R. Mayer, and A. Pecher.** 1981. The phosphoenolpyruvate-dependent carbohydrate-phosphotransferase system enzymes II as chemoreceptors in chemotaxis of *Escherichia coli* K-12. Mol. Gen. Genet. **183:**163–170.
24. **Lengeler, J., and H. Steinberger.** 1978. Analysis of regulatory mechanisms controlling activity of hexitol transport-systems in *Escherichia coli* K-12. Mol. Gen. Genet. **167:**75–82.
25. **Lengeler, J. W.** 1975. Nature and properties of the hexitol transport systems in *Escherichia coli*. J. Bacteriol. **124:**26–38.
26. **Lengeler, J. W., K. Bettenbrock, and R. Lux.** 1994. Signal transduction through phosphotransferase systems, p. 182–188. *In* A.-M. Torriani-Gorini, E. Yagil, and S. Silver (ed.), Phosphate in microorganisms. Cellular and molecular biology. ASM Press, Washington, DC.
27. **Lowry, O. H., J. Carter, J. B. Wood, and L. Glaser.** 1971. The effect of carbon and nitrogen sources on the level of metabolic intermediates in *Escherichia coli*. J. Biol. Chem. **246:**6511–6521.
28. **Lux, R., V. R. N. Munasinghe, F. Castellano, J. W. Lengeler, J. E. T. Corrie, and S. Khan.** 1999. Elucidation of a PTS-carbohydrate chemotactic signal pathway in *Escherichia coli* using a time-resolved behavioral assay. Mol. Biol. Cell **10:**1133–1146.
29. **Marr, A. G.** 1991. Growth rate of *Escherichia coli*. Microbiol. Rev. **55:**316–333.
30. **Matin, A., and M. K. Matin.** 1982. Cellular levels, excretion, and synthesis rates of cyclic AMP in *Escherichia coli* grown in continuous culture. J. Bacteriol. **149:**801–807.
31. **Miller, J. H.** 1972. Experiments in molecular genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
32. **Misko, T., W. Mitchell, N. Meadow, and S. Roseman.** 1987. Sugar transport by the bacterial phosphotransferase system. Reconstitution of inducer exclusion in *Salmonella typhimurium* membrane vesicles. J. Biol. Chem. **262:**16261–16266.
33. **Nelson, S., J. Wright, and P. Postma.** 1983. The mechanism of inducer exclusion. Direct interaction between purified $\mathrm{III}^{\mathrm{Glc}}$ of the phosphoenolpyruvate-sugar phosphotransferase system and the lactose carrier of *Escherichia coli*. EMBO J. **2:**715–720.
34. **Notley-McRobb, L., A. Death, and T. Ferenci.** 1997. The relationship between external glucose concentration and cAMP levels inside *Escherichia coli*: implications for models of phosphotransferase-mediated regulation of adenylate cyclase. Microbiology **143:**1909–1918.
35. **Novotny, M. J., W. L Frederickson, E. B. Waygood, and M. H. Saier, Jr.** 1985. Allosteric regulation of glycerol kinase by enzyme $\mathrm{III}^{\mathrm{Glc}}$ of the phosphotransferase system in *Escherichia coli* and *Salmonella typhimurium*. J. Bacteriol. **162:**810–816.
36. **Osumi, T., and M. H. Saier, Jr.** 1982. Regulation of the lactose permease activity by the phosphoenolpyruvate:sugar phosphotransferase system: evidence for direct binding of the glucose specific enzyme III to the lactose permease. Proc. Natl. Acad. Sci. USA **79:**1457–1461.
37. **Pardee, A. B., and L. S. Prestidge.** 1961. Initial kinetics of enzyme induction. Biochim. Biophys. Acta **49:**77–88.
38. **Park, Y. H., B. R. Lee, Y. J. Seok, and A. Peterkofsky.** 2006. In vitro reconstitution of catabolite repression in *Escherichia coli*. J. Biol. Chem. **281:**6448–6454.
39. **Pastan, I., and R. L. Perlman.** 1968. The role of the *lac* promoter locus in the regulation of beta-galactosidase synthesis by cyclic 3′,5′-adenosine monophosphate. Proc. Natl. Acad. Sci. USA **61:**1336–1342.
40. **Peterkofsky, A., and C. Gazdar.** 1971. Glucose and metabolism of adenosine 3′–5′-cyclic monophosphate in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **68:**2794–2798.
41. **Peterkofsky, A., Y. J. Seok, N. Amin, R. Thapar, S. Y. Lee, R. E. Klevit, E. B. Waygood, J. W. Anderson, J. Gruschus, H. Huq, and N. Gollop.** 1995. The *Escherichia coli* adenylyl-cyclase complex: requirement of PTS proteins for stimulation by nucleotides. Biochemistry **34:**8950–8959.
42. **Peterkofsky, A., I. Svenson, and N. Amin.** 1989. Regulation of *Escherichia coli* adenylate cyclase activity by the phosphoenolpyruvate:sugar phosphotranferase system. FEMS Microbiol. Rev. **63:**103–108.
43. **Postma, P. W., A. R. Broekhuizen, J. Schuitema, A. P. Vogler, and J. W. Lengeler.** 1988. Carbohydrate transport and metabolism in *Escherichia coli* and *Salmonella typhimurium*: regulation by the PEP:carbohydrate phosphotransferase system, p. 43–52. *In* F. Palmieri and E. Quagliariello (ed.), Molecular basis of biomembrane transport. Elsevier Science Publishing, Amsterdam, The Netherlands.
44. **Postma, P. W., J. W. Lengeler, and G. R. Jacobson.** 1996. Phosphoenolpyruvate:carbohydrate phosphotransferase systems, p. 1149–1174. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed., vol. 1. American Society for Microbiology, Washington, DC.
45. **Prost, J. F., D. Negre, C. Oudot, K. Murakami, A. Ishihama, A. J. Cozzone, and J. C. Cortay.** 1999. Cra-dependent transcriptional activation of the *icd* gene of Escherichia coli. J. Bacteriol. **181:**893–898.
46. **Ramseier, T. M., D. Negre, J. C. Cortay, M. Scarabel, A. J. Cozzone, and M. H. Saier, Jr.** 1993. In vitro binding of the pleiotropic transcriptional regulatory protein, FruR, to the *fru*, *pps*, *ace*, *pts* and *icd* operons of *Escherichia coli* and *Salmonella typhimurium*. J. Mol. Biol. **234:**28–44.
47. **Reddy, P., N. Meadow, S. Roseman, and A. Peterkofsky.** 1985. Reconstitution of regulatory properties of adenylate-cyclase in *Escherichia coli* extracts. Proc. Natl. Acad. Sci. USA **82:**8300–8304.
48. **Roossien, F. F., M. Blaauw, and G. T. Robillard.** 1984. Kinetics and subunit interaction of the mannitol-specific enzyme II of the *Escherichia coli* phosphoenolpyruvate-dependent phosphotransferase system. Biochemistry **23:**4934–4939.
49. **Roseman, S., and N. D. Meadow.** 1990. Signal transduction by the bacterial phosphotransferase system: diauxie and the *crr* gene (J. Monod revisited). J. Biol. Chem. **265:**2993–2996.
50. **Ruijter, G. J. G., G. Vanmeurs, M. A. Verwey, P. W. Postma, and K. Vandam.** 1992. Analysis of mutations that uncouple transport from phosphorylation in enzyme $\mathrm{II}^{\mathrm{Glc}}$ of the *Escherichia coli* phosphoenolpyruvate-dependent phosphotransferase system. J. Bacteriol. **174:**2843–2850.
51. **Saier, M. H., Jr., B. U. Feucht, and W. K. Mora.** 1977. Sugar phosphate:sugar transphosphorylation and exchange group translocation catalyzed by the Enzyme II complexes of the bacterial phosphoenolpyruvate:sugar phosphotransferase system. J. Biol. Chem. **252:**8899–8907.
52. **Saier, M. H., Jr., and S. Roseman.** 1976. Sugar transport. The *crr* mutation: its effect on repression of enzyme synthesis. J. Biol. Chem. **251:**6598–6605.
53. **Sauter, T., and E. D. Gilles.** 2004. Modeling and experimental validation of the signal transduction via the *Escherichia coli* sucrose phosphotransferase system. J. Biotechnol. **110:**181–199.

Downloaded from jb.asm.org at MPI FUER DYNAMIK KOMPLEXER on May 7, 2009

54. **Schmid, K., R. Ebner, J. Altenbuchner, R. Schmitt, and J. W. Lengeler.** 1988. Plasmid-mediated sucrose metabolism in *Escherichia coli* K12: mapping of the *scr* genes of pUR400. Mol. Microbiol. **2:**1–8.
55. **Schmid, K., R. Ebner, K. Jahreis, J. W. Lengeler, and F. Titgemeyer.** 1991. A sugar-specific porin, ScrY, is involved in sucrose uptake in enteric bacteria A. Mol. Microbiol. **5:**941–950.
56. **Shehata, T. E., and A. G. Marr.** 1971. Effect of nutrient concentration on the growth of *Escherichia coli*. J. Bacteriol. **107:**210–216.
57. **Takahashi, H., T. Inada, P. Postma, and H. Aiba.** 1998. CRP down-regulates adenylate-cyclase activity by reducing the level of phosphorylated IIA^{Glc}, the glucose-specific phosphotransferase protein, in *Escherichia coli*. Mol. Gen. Genet. **259:**317–326.
58. **Tanaka, S., S. A. Lerner, and E. C. C. Lin.** 1967. Replacement of a phosphoenolpyruvate-dependent phosphotransferase by a nicotinamide adenine dinucleotide-linked dehydrogenase for the utilization of mannitol. J. Bacteriol. **93:**642–648.
59. **Yang, J. K., and W. Epstein.** 1983. Purification and characterization of adenylate cyclase from *Escherichia coli* K12. J. Biol. Chem. **258:**3750–3758.

Downloaded from jb.asm.org at MPI FUER DYNAMIK KOMPLEXER on May 7, 2009

BIOINFORMATICS

Vol. 19 no. 9 2003, pages 1169–1176
DOI: 10.1093/bioinformatics/btg128

Modular modeling of cellular systems with ProMoT/Diva

M. Ginkel, A. Kremling*, T. Nutsch, R. Rehner and E. D. Gilles

Max-Planck-Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1,
D-39106 Magdeburg, Germany

Received on April 9, 2002; revised on August 22, 2002; November 18, 2002; accepted on November 29, 2002

ABSTRACT

Motivation: Need for software to setup and analyze complex mathematical models for cellular systems in a modular way, that also integrates the experimental environment of the cells.

Results: A computer framework is described which allows the building of modularly structured models using an abstract, modular and general modeling methodology. With this methodology, reusable modeling entities are introduced which lead to the development of a modeling library within the modeling tool ProMot. The simulation environment Diva is used for numerical analysis and parameter identification of the models. The simulation environment provides a number of tools and algorithms to simulate and analyze complex biochemical networks. The described tools are the first steps towards an integrated computer-based modeling, simulation and visualization environment.

Availability: Available on request to the authors. The software itself is free for scientific purposes but requires commercial libraries.

Contact: ginkel@mpi-magdeburg.mpg.de

Supplementary information: http://www.mpi-magdeburg.mpg.de/projects/promot

INTRODUCTION

The success in modern biology in analyzing the genetic structure of many organisms has allowed bioinformatics to become a very popular science. Consequentially, a number of database systems have been developed to organize the large amount of data occurring during research (e.g. Kanehisa and Goto, 2000; Salgado *et al.*, 2001). New measurement techniques like cDNA microarrays and 2D-gelelectrophoreses have been also established and are used to obtain insight into the overall cellular state. Moreover, new measurement techniques which allow samples to be taken within a time window of 2/100 seconds are possible now and are used to analyze the intracellular dynamics of small metabolites if the system is shifted from one steady state to another (Schaefer *et al.*, 1999). These facts—availability of knowledge of the genetic structure and new measurement techniques—smooth the transition of biology from a qualitative to a quantitative science. However, to analyze and possibly predict cellular behavior based on the the increasing quantity of knowledge and therefore more complex cellular system models the application of mathematical modeling is necessary.

For dynamical systems, we previously introduced a suitable modeling framework (Kremling *et al.*, 2000) based on the definition of submodels called *modeling objects*. These modeling objects cover a broad range from single enzymatic reaction steps to rather complex structures, which are called *operons* and *modulons* in bacterial genetics (Neidhardt *et al.*, 1990). This paper deals with two computational aspects in modeling cellular systems: (i) the modular assembly of dynamic model equations; and (ii) model validation based on parameter identification from available measurements. Although different other modeling and simulation tools like GEPASI (Mendes, 1997), Jarnac (Sauro, 2000), VCell (Schaff *et al.*, 1997), DBSolve (Goryanin *et al.*, 1999), E-Cell (Tomita *et al.*, 1999) and others also solve systems of differential equations, they don't provide a modular approach for model setup.

A problem not discussed here in detail is the exchange of models between different modeling and simulation tools. We take part in an international initative of simulation tool developers to define a practical standard for mathematical models of cells that is called systems biology markup language (SBML Hucka *et al.*, 2000). It is planned to import and export SBML in ProMot.

After introducing the modeling concept, a software environment, combining two tools, namely ProMoT and Diva will be presented. The Process Modeling Tool ProMoT (Tränkle *et al.*, 2000) was originally designed for the computer-aided modeling of chemical processes as well as for the implementation of libraries that contain reusable modeling entities. The differential-algebraic

---

*To whom correspondence should be addressed.

© Oxford University Press 2003; all rights reserved.

models created with ProMoT are added to the model library of the simulation environment Diva (Mohl *et al.*, 1997). Differential Algebraic Equations—DAE, sometimes also called ODE–NAE models are a combination of ODE that are simultaneously solved with algebraic constraints. The numerical methods provided by Diva are applied to the numerical analysis, dynamic and steady state simulation and identification of model parameters.

SYSTEM AND METHODS

Systems biology seeks to combine experimental and theoretical work for a better understanding of the overall behavior of cellular systems. This implies that not only the cellular interior has to be modeled but also the environment, e.g. the fluxes into and out of a bioreactor which allow exposure of the organism to defined and reproducible conditions. When analyzing complex systems with a high number of elements and several interconnected levels, e.g. a fermentation plant with a bioreactor containing liquid and biophase, where the biophase again is decomposed in metabolic units, a common base is required which is applicable to all levels. Therefore, network theory was proposed for analysis and synthesis problems in chemical and biochemical engineering.

Network theory

Network theory (Gilles, 1998) gives a fundamental way to decompose various processes into hierarchical units in a systematic manner. The hierarchical structure of the process is represented in several levels (see Fig. 1). All levels consist of two basic types of elements, namely *components*, representing the holdup of different physical quantities (drawn as circles in the figure), and *coupling elements* describing the interactions and transports between the different components (rectangles in the figure). The top level can be, for example a device level consisting of components like reactors and other devices and coupling elements like valves and pumps etc. The devices again consist of phases that are coupled by phase-boundaries or membranes and finally the phases consist of *storages* that are coupled via reactions or diffusive and convective relations. Network theory integrates all these levels into the same theoretical concept in a modular way with well defined interfaces.

How are these models computed? There is a division of tasks between the basic elements. Components provide information about their *potentials*, i.e. their concentration and require information about the fluxes coming in and leading out of them. Essentially they balance the potentials with regard to the fluxes. Coupling elements calculate the *fluxes*, i.e. the reaction rates, depending on the potentials and provide the flux information to the components. This two-directional information exchange of potentials and fluxes forms a *potential-flux vector*. If a potential-flux vector is passed across phase or device borders it must consist of extensive quantities only, like mass flux, molar flux or volume flux, otherwise it is not possible to achieve modularity.

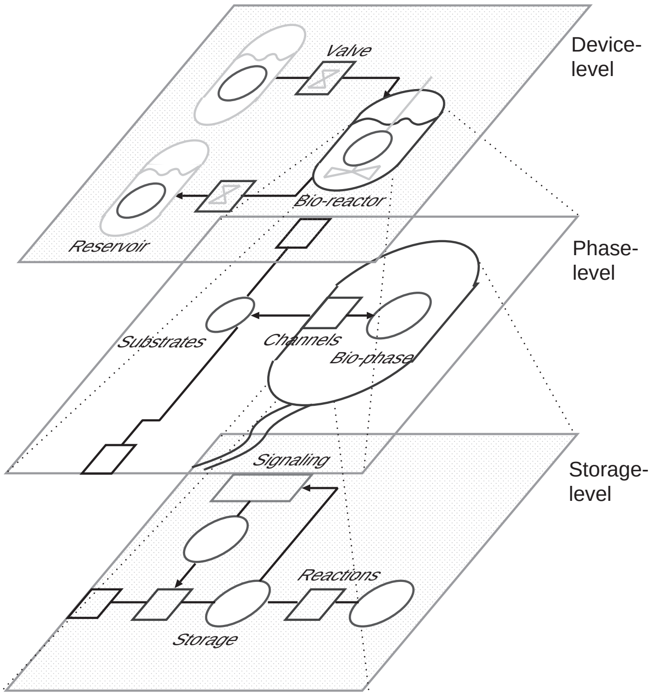


Fig. 1. Representation of different detail levels of a bioreactor model by the means of network theory. The different layers are all represented, based on components and coupling elements. These elements provide a modular structure and interfaces for the more detailed levels.

Modeling concept for biochemical plants

A continuously stirred tank reactor with several substrate feeds and an outflow is considered as an example for a biochemical plant and is depicted in Figure 1. The plant is composed of process devices namely the reservoirs and the bioreactor and their coupling elements that are valves. The bioreactor is modeled with two phases: the liquid-phase and the biophase. The liquid-phase model comprises its volume as an extensive reference quantity V_l (unit $[l]$) and storages for the substrates (concentrations $\underline{c_l}$ (unit $[g/l]$)). The biophase contains storages for intracellular metabolites, and the biochemical reactions as coupling elements. As an extensive reference quantity, e.g. biomass (m_{bio}) or total volume of the cells should be used.

For an exact formulation of the cell growth, all existing exchange-fluxes through transporters between the two phases have to be summed up. But due to the fact that biological models can probably never comprise every single transport pathway connecting the cell to the liquid phase, a slightly different approach has to be chosen. We suggest balancing the exchange-fluxes by using yield coefficients $\underline{Y}$ (unit $[g_{DW}/g]$) for the substrates taken into account in the model. The balance of biomass then results in:

$$\dot{m_{bio}} = \underbrace{\underline{Y}^{\mathrm{T}}\, M\, \underline{j_{ex}}[c_l]}_{\mu}\, m_{bio} - m_{bio}\, \frac{J_{out}[V]}{V_l}\ , \qquad (1)$$

where M is a diagonal matrix of the molar weights of the substrates, $\underline{j_{ex}}[c_l]$ the substrate exchange between biomass and liquid phase, $J_{out}[V]$ the liquid flow out of the reactor and μ the growth rate of the cells. This allows for modeling the cell growth realistically while representing only the main substrate transport pathway (called channel) in the model.

Modeling framework for cellular systems

In microbiology, the thinking in functional units (describing a subset of the cellular processes) has become popular and has resulted in the definition of subnetworks that are under control of a common regulator protein (Neidhardt *et al.*, 1990). The combination of these ideas with network theory leads to a modeling framework which was previously introduced (Kremling *et al.*, 2000). At the highest level of resolution, elementary submodels (modeling objects) are defined. Important elementary modeling objects are substance storages and substance transformers for the metabolic network and signal transformers for the regulatory network.

Two or more storages can be connected by a substance transformer that represents a biochemical reaction. Transformers are treated as two complementary aspects: (i) the representation of the stoichiometric structure of the reaction with interfaces for substrates and products; and (ii) the reaction kinetics together with the participating and controlling ligands (activators and inhibitors).

Since the understanding of signal transduction and processing is the key for describing the overall behavior of cellular systems, these processes are described in a separate class named signal transformers. Elementary modeling objects can now be aggregated to describe more complex processes like gene expression or signal transdcution cascades (Kremling and Gilles, 2001).

Modular model representation

ProMoT enables the use of object-oriented modeling techniques including encapsulation, aggregation, and inheritance. In ProMoT, dynamic models are built by aggregating *structural* and *behavioral modeling entities*. Structural modeling subdivides a model into *modules*. Examples for modules in systems biology are process units (e.g. fermentation reactors), balanced volumes (e.g. phases), functional units of the metabolism (e.g. glycolysis) and elementary entities (e.g. reactions). In general they represent components and coupling elements of network theory on different hierachy levels. Modules are encapsulated and therefore separated from their environment whereas their interfaces are defined by *terminals*. The behavior of a module is characterized by aggregated variables and equations in a module-local DAE.

To establish connections in a modularized model, groups of variables are assigned to terminals. When terminals are linked, additional linking equations connect the different behavioral subsystems. Terminals are not required to have a specified direction (e.g. input or output). In case of substance flows in biological reaction networks they represent a bidirectional information exchange of a concentration and a flow rate in the sense of potential flow vectors. Another important form of terminals in biological systems are cellular signals which represent only a concentration. Modules, terminals and links are structural modeling entities, whereas variables and equations are behavioral modeling entities.

The emphasis on modularity has several advantages in modeling complex biological systems:

- the user works with comprehensible networks of modules rather than with reaction networks with hundreds or thousands of parts. With this feature it is also easier to divide tasks between different modelers working on parts of the same system, which is desirable for large scale biological models;

- the interface of a module can be specified separately from its implementation. This leads to a simplified exchangeability of different module implementations with the same interface. This can be used e.g. for implementations of a module which differ in the detail-level;

- an important and often neglected task in model development is model debugging. With the depicted structure it is easier to debug an individual module with its input-output behavior in a well defined test frame first before the modules are combined to a larger system. Since the couplings are explicit in the modular system, removal of feedback can be easily carried out as simplification to isolate errors in the model.

The modeling entities in ProMoT are organized as an object-oriented class hierarchy with multiple inheritance. This concept from computer science was adopted to allow a better organization of complex modeling libraries and flexible implementation of large scale models. Every entity in this hierarchy inherits all parts and attributes from their respective superclasses. With this method abstraction is possible and more general and reusable entities can be formed.

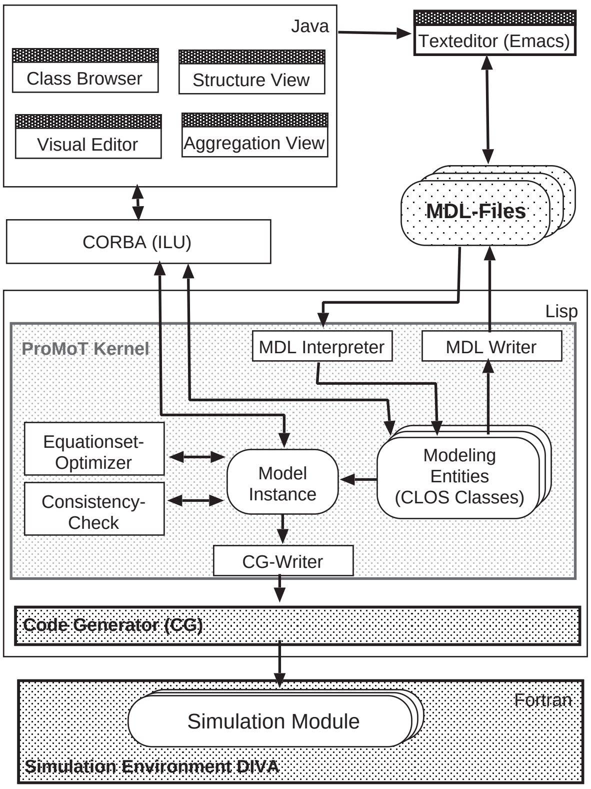


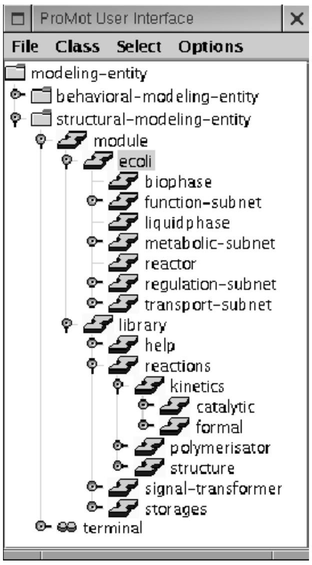
Fig. 2. Software architecture of ProMoT. The kernel provides all model handling including reading and writing modeling language, instantiation, consistency-check and writing of Diva models. The GUI and the kernel interact through a Corba-middleware called ILU.

IMPLEMENTATION OF ProMoT

ProMoT provides a special modeling language as well as a graphical user interface (GUI) for interactive modeling. The modeling tool, as well as the simulation environment, are developed under different Unix-derived operating systems, however the main platform is Linux. As shown in Figure 2 the kernel of the system is implemented as a modeling server in object-oriented Common Lisp (using the Common Lisp Object System CLOS). Further information about availability and requirements of the software can be found on the web page. Although Lisp is currently not a very popular language, it has certain qualities that are adjuvant for an easy and flexible implementation. ProMoT's modeling entities are classes and use multiple inheritance. Therefore they are internally represented by specialized classes in CLOS, which handles inheritance and creation of instances. The classes represent aggregation and composition of aggregated parts explicitly, which allows construction of complex containment hierarchies and their analysis in the final model. Since Lisp classes themselves can be programmed, conforming to the CLOS Meta Object Protocol (Kiczales *et al.*, 1991), this foundation of ProMoT is implemented as an extension of standard Lisp classes. The classes are dynamic meta-objects in the Lisp runtime environment; that is why it is also possible to edit them at runtime using either the graphical editor or through changing the source code.The possibility to do this is rarely found in programming languages: most languages provide class meta objects only for reflection (i.e. read-only introspection in Java), if they provide any at all. For example, common Lisp and Smalltalk also allow one to change classes (write access), which is one of the main reasons to build the modeling environment in Lisp. The representation of the mathematical model is done in a symbolic way. This makes it possible to manipulate the formulae, e.g. for normalization of the differential equations or during optimization of the final simulation model, which can be easily implemented in Lisp.

The modeling language MDL (Model Description Language) of ProMoT is a declarative, object-oriented language that allows a symbolic implementation of variables and equations rather than the programming of imperative code. ProMoT interprets MDL to create the class representation and can serialize the classes to MDL. Thus the modeling language is used as the storage format for the modeling libraries. Because every aspect of a model can be described within MDL, the modeling language is the most powerful way to model in ProMoT. The GUI is a client that is implemented in Java using the Java Foundation Classes (Swing). It interacts with the kernel in a Model View Controller (MVC) fashion and has the role of a view and controller for the models in the kernel. With the GUI, users can explore and manipulate the modeling entities by their graphical representation. Therefore views of the inheritance hierarchy and the topology of submodules and their connections can be presented. The visual aspect is very important especially for the communication in interdisciplinary teams, to have a common notion of the considered modeling entities. Besides that also graphical editing of the topological structure of modeling entities can be done interactively with flow-chart diagrams. In this way new higher structured modules can be created easily. For changes on behavioral modeling entities the GUI launches a text editor in order to change the MDL source code of a single modeling entity. Thus modeling language and graphical editor can be used alternately to change modeling entities from the user interface.

The internal processes in a typical modeling scenario of ProMoT are as follows: The user loads necessary libraries (the details are introduced in the next subsection) with basic module definitions from MDL files using the class browser of the GUI (Fig. 3). Then he creates or extends a module that should be simulated in Diva (the main model). Therefore he builds the module structure out of predefined



ProMot User Interface

File Class Select Options

- modeling-entity
  - behavioral-modeling-entity
  - structural-modeling-entity
    - module
      - ecoli
        - biophase
        - function-subnet
        - liquidphase
        - metabolic-subnet
        - reactor
        - regulation-subnet
        - transport-subnet
      - library
        - help
        - reactions
          - kinetics
            - catalytic
            - formal
          - polymerisator
          - structure
        - signal-transformer
        - storages
    - terminal

Fig. 3. ProMoT class browser showing the class library discussed in the text.

modules from the library and adds special parameter and initial values for the variables. If the user needs a special module with a behavior not available in the library, he can create this using the modeling language. He should use standardized abstract superclasses for the new module to stay compatible with the rest of the library. Finally the main model is written to Diva. In this process the Lisp class representing the main model is instantiated with all contained submodules. Then a consistency check is performed, that unveils logical errors in the model structure and also structural errors in the underlying equation system.

If these checks can be passed successfully, a compact DAE is generated from the structured representation within ProMoT by aggregating all equations together with coupling relations. The modular structure is only used during modeling: since the Diva simulator does not allow changes to the structure at runtime, it is not needed during simulation. The structured way of modeling in ProMoT and the use of modeling libraries often introduce unnecessary algebraic equations in the resulting model for couplings in links and calculations of variables to achieve flexible modules. The resulting DAE can be divided into a differential part (2) and a purely algebraic part (3):

$$\boldsymbol{B_1}(\boldsymbol{x},\,\boldsymbol{p},t)\dot{\boldsymbol{x}} = \boldsymbol{f_1}(\boldsymbol{x},\,\boldsymbol{p},t) \tag{2}$$

$$0 = \boldsymbol{f_2}(\boldsymbol{x},\,\boldsymbol{p},t), \tag{3}$$

where $\boldsymbol{x}$ is the vector of states, $\boldsymbol{p}$ the vector of parameters, $\boldsymbol{u}$ the input vector and $\boldsymbol{B_1}$ is the descriptor matrix. The modeling system analyzes the algebraic part $\boldsymbol{f_2}$ and identifies implicit algebraic equations ($\boldsymbol{f_2}'$), that have to be calculated simultanously with the differential equations.

$$\boldsymbol{h} := \boldsymbol{g}(\boldsymbol{x},\boldsymbol{h},\,\boldsymbol{p},t) \tag{4}$$

$$\boldsymbol{B_1}(\boldsymbol{x},\boldsymbol{h},\,\boldsymbol{p},t)\dot{\boldsymbol{x}} = \boldsymbol{f_1}(\boldsymbol{x},\boldsymbol{h},\,\boldsymbol{p},t) \tag{5}$$

$$0 = \boldsymbol{f_2}'(\boldsymbol{x},\boldsymbol{h},\,\boldsymbol{p},t). \tag{6}$$

Explicit equations in the algebraic part are sorted according to their dependencies and are directly calculated as assignments to intermediate variables $\boldsymbol{h}$ by the functions $\boldsymbol{g}$. Additionally constant expressions and unnecessary variables are identified and eliminated through symbolic transformations. This produces a more compact and performant implementation of the model that also avoids numerical problems with inconsistent initial conditions of the DAE.

Finally ProMoT generates Fortran source code that can be used within the simulation environment Diva. Therefore the Code Generator (Köhler *et al.*, 1997) is invoked which translates the symbolic representation of ProMoT to Fortran subroutines and prepares the initialization of the sparse matrix numerics of Diva.

Library for metabolic models

All modeling entities are held in a knowledge base that comprises elementary modeling objects like terminals, storages, transformers and channels as well as predefined higher structured modules, e.g. for gene-expression. The user-defined models can be based on the predefined modules and are added also to the knowledge base. In this way the setup of new models is simplified and sped up considerably. It is less error-prone and in addition the models become standardized, what enables exchange and reusability of models. The library for modeling biological systems in ProMoT contains several categories which are presented to the user in the tree-structure of the class browser (see Fig. 3).

For the representation of basic modules like storages and substance transformers basic terminals are defined, e.g. `term-reaction-flux` and `term-storage-flux` for connections of intracellular reactions with intracellular storages. These terminals define a potential variable `c` for a concentration and a flux variable `r` for the reaction rate. Subclasses of these basic terminal types add additional attributes, e.g. in the terminal `term-liq-storage-flux` a variable for the molar weight is aggregated which is necessary to convert cell-external concentrations at the border of the biophase.

The predefined elementary modules are represented beneath `module/library` with the `storages` as an important subgroup. There are storages defined for the liquid-phase and for the biophase with different kinds of terminals. For example `storage-intra_x` is a storage with a `term-storage-flux` terminal. It contains a differential equation for a substance storage that automatically takes

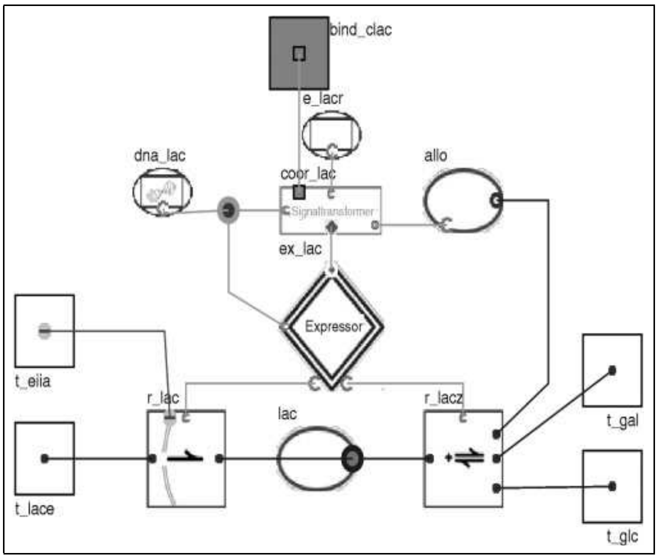


Fig. 4. Setup of a model for lactose transport as it appears in the visual editor of ProMoT. The whole drawing is the interior of a module. The white and gray boxes at the outer edges are external terminals (in this case referring to terminals of the submodules). The elements inside the drawing are aggregated submodules, lines represent links.

the dilution by the cell-growth μ into account. In Figure 4 the modules `lac` and `allo` are instances of this class. The module `storage-const-enz_c` like `e_lacr` in the figure is used for defining the constant enzyme-concentration as an input into a substance or signal transformer.

Another important subgroup of the library is `reactions`, where transformers and channels are defined. Their attributes are subdivided into kinetic and structural properties. Beneath the subclasses `kinetics` there are various predefined modules, e.g. `mm`: Michaelis–Menten, `mmea`: Michaelis–Menten with essential activation, `pp`: ping-pong mechanism, `sr2`: sequential random 2 substrate mechanism and of course simple formal first- and second-order reactions. In these modules the appropriate variables, equations and terminals for the calculation of the reaction kinetics are defined. Kinetics define terminals as well for connecting storages that actually do not take part in the reaction as substrate or product but affect the rate as e.g. enzyme, activator or inhibitor. For the structural (stoichiometric) part of the reactions there is a separate group of modules below `structure`. These classes contain terminals and stoichiometric parameters for subtrates and products and most of the the geometry-information (position of substrate and product terminals, iconic representation). Finally the complete transformer or channel inherits the relevant kinetic and structural properties from one element of both groups respectively that form the aforementioned complementary aspects. This enables a selection of the required transformer either trough the branch `kinetics` or `structure` from the browser tree. For example, the module `trans2a-mm_l` is a transformer that connects two storages with a Michaelis–Menten-reaction and a graphical representation to the left, `channel2a-mmui` is a channel with a Michaelis–Menten reaction and uncompetitive inhibition, where the first connected storage has to be extracellular and the product is intracellular. This channel class is instantiated as submodule `r_lac` in Figure 4.

The user-defined modules are also held in the knowledge base. There is no formal difference between user-defined and library modules. Library modules are just designed having more generality in mind. As an example a model of the carbohydrate uptake of *Escherichia coli* (Kremling and Gilles, 2001; Kremling *et al.*, 2001) has been implemented in ProMoT. The top module of the *E. coli* model is the class `reactor` that consists of the biophase and the liquid-phase (as shown in Fig. 1). The class `biophase` itself is highly structured and comprises some central parts of the catabolism and different transport pathways with their respective interacting regulation networks. One of these pathways is the uptake of lactose as shown in Figure 4.

Lactose is taken up, coming in with the `t_lace` terminal on the left, through the `r_lac` channel on the left and cell-internal lactose `lac`. It is further degraded up by the transformer `r_lacz` into glucose and galactose leaving the module through the terminals on the right hand side. A by-product of this reaction is allolactose (`allo`), which is important for the control of transcription of the enzymes LacY and LacZ catalyzing `r_lac` and `r_lacz` in `coor_lac`. This provides, that the enzymes only get expressed, if lactose is present in the medium. The signal-transformer `coor_lac` models the interactions at the promoter binding site of the DNA-sequence for the enzymes and also integrates the signal of a global activator that is included through the terminal `bind_clac`. The diamond-shaped expressor contains a model for the translation and the degradation of the two enzymes. The reaction `r_lac` interacts with other transport pathways for glucose (namely the phosphotransferase system) via the inhibitor EIIA, which enters the model as a concentration signal through the terminal `t_eiia`. As long as the glucose transport is active the resulting high concentration of EIIA inhibits the uptake of lactose.

NUMERICAL MODEL ANALYSIS WITH DIVA

The numerical analysis of the models is done with the simulation environment Diva (Mohl *et al.*, 1997). Within Diva many different numerical computations are possible, based on facilities to calculate the steady state and dynamic behavior of the model using non-linear equation solvers and integrators. For metabolic models two methods are of special interest: (i) parameter analysis with respect to experimental data; and (ii) identification of parameters and model accuracy.

Parameter analysis and sensitivities

Sensitivity analysis of cellular models is often associated with the calculation of flux and concentration control coefficients from ‘Metabolic Control Analysis’ (Heinrich and Schuster, 1996). Diva uses parameter sensitivities $w_{ij} = \partial x_i / \partial p_j$ for another purpose.

The aim of the parameter analysis is to find a parameter vector $\boldsymbol{p}$, that is a subset of all parameters contained in the model. It should be possible to estimate $\boldsymbol{p}$ with a given variance γ. The choice of γ depends on the accuracy of the measurement data and the demands on the model. To identify the elements of $\boldsymbol{p}$, given some initial vector $\boldsymbol{p}_0$ the user is interested in, the following approach is used: The model and the measured data are analyzed with the Fisher information matrix (Ljung, 1999). The Fisher information matrix is defined by:

$$\boldsymbol{F} = \sum_{k=1}^{N} [\boldsymbol{W}(t_k)^{\mathrm{T}} \boldsymbol{C}(t_k)^{-1} \boldsymbol{W}(t_k)], \qquad (7)$$

with the matrix of the sensitivities $\boldsymbol{W}(t_k) = \partial \boldsymbol{x} / \partial \boldsymbol{p}$ and the covariance matrix $\boldsymbol{C}(t_k)$. The covariance matrix $\boldsymbol{C}$ is assumed as a diagonal matrix with the variance of the states σ_i as elements. It is assumed that σ_i do not depend from time point t_k while it is taken as a constant.

Applying a method introduced by Posten and Munack (1990), $\boldsymbol{p}$ can be determined by analyzing the eigenvalues λ and eigenvectors of $\boldsymbol{F}$: The parameter p_i out of $\boldsymbol{p}$, most contributing to the eigenvector that corresponds to the smallest eigenvalue λ_{min}, is removed successively from p until $\gamma \geqslant \sqrt{\frac{1}{\lambda_{min}}}$. Although this provides only a local estimate of the lower bound for the variance of parameter estimation, the method was applied successfully in optimal experimental design for a biotechnological process (Baltes *et al.*, 1994).

Parameter estimation

Identification with Diva is restricted to the estimation of parameters in a fixed model structure. Measurement data (z_{ik}) is available for a subset of the states at time point t_k. The aim of the estimation method is to minimize the objective function:

$$\Phi(\boldsymbol{p}) = \sum_{k=1}^{N} \sum_{i=1}^{n} w_{ik}{}^2 \left(\frac{x_i(\boldsymbol{x}_o, \boldsymbol{u}, \boldsymbol{p}, t_k) - z_{ik}}{z_i^{\max}} \right)^2 \qquad (8)$$

where w_{ik}, $z_i^{\max}$ are scaling factors of the individual measurement data and for each experiment respectively. For the optimization a SQP (**S**equential **Q**uadratic **P**rogramming) method from the NAG library (Moré and Wright, 1993) is used.

Numerical analysis of the example

A mathematical model for catabolite repression was introduced previously (Kremling and Gilles, 2001; Kremling *et al.*, 2001). The model describes glucose and lactose uptake as well as the control of gene expression of the respective enzymes. The model comprises 22 ODE’s and seven algebraic equations. According to the underlying modeling concept, the equations are assigned to modeling objects. One of these objects, Lac transport has been introduced above.

Based on the available measurements in first run—biomass, extracellular glucose, lactose and intracellular LacZ—and the experiments performed it is expected that parameters which could be estimated are closely related to glucose and lactose transport kinetics and to LacZ synthesis. The result of the analysis shows that 16 parameters could be estimated In a second run three additional measurements—intracellular concentration of protein EIIA which is involved in glucose uptake and intracellular and extracellular cAMP concentration—are included in the analysis (the time course of these experiments are not yet published). With these measurements available, 20 parameters could be estimated.

DISCUSSION

An overview was given for a workbench of software tools that supports modeling and numerical analysis of cellular systems. The modeling tool ProMoT provides an approved methodology and the possibility to use ready made modeling entities out of knowledge-bases. Efficient modeling is supported by the use of a graphical user interface and a modeling language. Sophisticated methods for the numerical analysis of the resulting models are provided by the simulation environment Diva. Besides dynamic simulation the identification of parameters and the analysis of sensitivities are possible. The workbench is different from other tools like Gepasi and Jarnac because it deals with modular models and can handle DAE like VCell and DBSolve. It is well suited for larger simulation and parameter estimation problems with up to 10 000 differential equations, because of the advantages of modular model development and the efficient numerical routines of the simulator. The main advantage of our software is the modular modeling approach which is to our knowledge not provided by any other aforementioned tool. Since the development cycle for the models includes a compilation step, the advantage of the efficient computation in Diva is bought with a longer preparation time, which does not scale well for rapid prototyping of small text-book models.

For the modeling of cellular signaling the current approach is to simplify the interactions of proteins, DNA binding sites and other substances to form signal transformers. This simplification is often not easily possible, due to a deficiency of quantitative knowledge. Therefore the only correct way is to model all possible interactions with explicit reactions and to explore the behavior of the system interactively in simulation and experiment. Since in complex regulatory networks the number of reactions increases exponentially due to combinatoric effects of binding sites in complex macromolecules, the models become very complicated. For a solution of this problem current research aims to describe the basic interactions of binding sites and the compound structure of the molecules and to let the modeling system generate the complete reaction network automatically. It is planned to implement this approach as a specialized class of modules in ProMoT.

The goal of the described software tools is a virtual laboratory containing facilities for modeling, simulating and visualizing parts of intracellular metabolisms. Therefore other computer aided methods have to be integrated. For rising amounts of knowledge the use of databases is necessary which allow the sharing of discoveries between workgroups at different locations connected by the internet. For effective development of complex metabolic calculation models the use of standardized modeling entities is only one aspect. Others are facilities which allow the visualization of different aspects of the resulting special models and support especially the debugging of models under construction.

REFERENCES

Baltes,M., Schneider,R., Sturm,C. and Reuss,M. (1994) Optimal experimental design for parameter estimation in unstructured growth models. *Biotechn. Progr.*, **10**, 480–488.

Gilles,E.D. (1998) Network theory for chemical processes. *Chemical Engineering and Technology*, **21**, 121–132.

Goryanin,I., Hodgman,T. and Selkov,E. (1999) Mathematical simulation and analysis of cellular metabolism and regulation. *Bioinformatics*, **15**, 749–758.

Heinrich,R. and Schuster,S. (1996) *The Regulation of Cellular Processes*. Chapman and Hall, London.

Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H. and the rest of the SBML forum, (2003) The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kiczales,G., des Rivières,J. and Bobrow,.D.G. (1991) *The Art of the Metaobject Protocol*. MIT Press, Cambridge, MA.

Köhler,R., Räumschüssel,S. and Zeitz,M. (1997) Code generator for implementing differential algebraic models used in the process simulation tool DIVA. In *Proceedings of 15th IMACS World Congress*. Berlin, pp. 621–626.

Kremling,A. and Gilles,E.D. (2001) The organization of metabolic reaction networks: II. Signal processing in hierarchical structured functional units. *Metab. Eng.*, **3**, 138–150.

Kremling,A., Jahreis,K., Lengeler,J.W. and Gilles,E.D. (2000) The organization of metabolic reaction networks: a signal-oriented approach to cellular models. *Metab. Eng.*, **2**, 190–200.

Kremling,A., Bettenbrock,K., Laube,B., Jahreis,K., Lengeler,J.W. and Gilles,E.D. (2001) The organization of metabolic reaction networks: III. Application for diauxic growth on glucose and lactose. *Metab. Eng.*, **3**, 362–379.

Ljung,L. (1999) *System Identification—Theory for the User*, Second edn, Prentice Hall PTR, Upper Saddle River, NJ.

Mendes,P. (September 1997) Biochemistry by numbers: simulation of biochemical pathways with GEPASI 3. *Trends Biochem. Sci.*, **22**, 361–363.

Mohl,K.D., Spieker,A., Köhler,R., Gilles,E.D. and Zeitz,M. (1997) DIVA—A simulation environment for chemical engineering applications. *ICCS Collect. Vol. Sci. Pap*. Donetsk State Techn. University, Ukraine, pp. 8–15.

Moré,J.J. and Wright,S.J. (1993) *Optimization Software Guide*. SIAM, Philadelphia, PA.

Neidhardt,F.C., Ingraham,J.L. and Schaechter,M. (1990) *Physiology of the Bacterial Cell: a Molecular Approach*. Sinauer, Sunderland, MA.

Posten,C. and Munack,A. (1990) On-line application of parameter estimation accuracy to biotechnical processes. In *Proceedings of ACC*. **3**, pp. 2181–2186.

Salgado,H., Santos,A., Gama-Castro,S., Millan-Zarate,D., Diaz-Peredo,E., Sanchez-Solano,F., Perez-Rueda,E., Bonavides-Martinez,C. and Collado-Vides,J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* k-12. *Nucleic Acids Res.*, 72–74.

Sauro,H.M. (2000) Jarnac: a system for interactive metabolic analysis. In Hofmeyr,J.-H.S., Rohwer,J.M. and Snoep,J.L. (eds), *Animating the Cellular Map*. Stellenbosch University Press, South Africa, pp. 221–228.

Schaefer,U., Boos,W., Takors,R. and Weuster-Botz,D. (1999) Automated sampling device for monitoring ntracellular metabolite ynamics. *Analyt. Biochem.*, **270**, 88–96.

Schaff,J., Fink,C.C., Slepchenko,B., Carson,J.H. and Loew,L.M. (1997) A general computational framework for modeling cellular structure and function. *Biophys. J.*.

Tomita,M., Hashimoto,K., Takahashi,K., Shimizu,T.S., Matsuzaki,Y., Miyoshi,F., Saito,K., Tanida,S., Yugi,K., Venter,J.G. and Hutchison III,C.A. (1999) E-CELL: software environment for whole-cell simulation. *Bioinformatis*, **15**, 72–84.

Tränkle,F., M.,Zeitz, Ginkel,M. and Gilles,E.D. (2000) Promot: a modeling tool for chemical processes. *MCMDS*, **6**, 283–307.

BIOINFORMATICS

Vol. 19 no. 4 2003, pages 524–531
DOI: 10.1093/bioinformatics/btg015

The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models

M. Hucka[1, 2,*], *A. Finney*[1, 2], *H. M. Sauro*[1, 2], *H. Bolouri*[1, 2, 3], *J. C. Doyle*[1], *H. Kitano*[1, 2, 4, 16, 18], and the rest of the SBML Forum*: A. P. Arkin*[5], *B. J. Bornstein*[6], *D. Bray*[7], *A. Cornish-Bowden*[8], *A. A. Cuellar*[9], *S. Dronov*[10], *E. D. Gilles*[11], *M. Ginkel*[11], *V. Gor*[6], *I. I. Goryanin*[10], *W. J. Hedley*[9], *T. C. Hodgman*[10], *J.-H. Hofmeyr*[12], *P. J. Hunter*[9], *N. S. Juty*[10], *J. L. Kasberger*[5], *A. Kremling*[11], *U. Kummer*[13], *N. Le Novère*[7], *L. M. Loew*[14], *D. Lucio*[14], *P. Mendes*[15], *E. Minch*[19], *E. D. Mjolsness*[20], *Y. Nakayama*[16], *M. R. Nelson*[17], *P. F. Nielsen*[9], *T. Sakurada*[16], *J. C. Schaff*[14], *B. E. Shapiro*[6], *T. S. Shimizu*[7], *H. D. Spence*[10], *J. Stelling*[11], *K. Takahashi*[16], *M. Tomita*[16], *J. Wagner*[14] *and J. Wang*[17]

[1]Control and Dynamical Systems, MC 107-81, California Institute of Technology, Pasadena, CA 91125, USA, [2]ERATO Kitano Symbiotic Systems Project, Tokyo, Japan, [3]University of Hertfordshire, Hertfordshire, UK, [4]The Systems Biology Institute, Tokyo, Japan, [5]University of California, Berkeley, CA, USA, [6]NASA JPL, Pasadena, CA, USA, [7]University of Cambridge, Cambridge, UK, [8]CNRS-BIP, Marseille, France, [9]University of Auckland, Auckland, New Zealand, [10]GlaxoSmithKline, Stevenage, UK, [11]Max-Planck-Institute for Complex Technical Systems, Magdeburg, Germany, [12]University of Stellenbosch, Stellenbosch, South Africa, [13]EML, Heidelberg, Germany, [14]University of Connecticut Health Center, Farmington, CT, USA, [15]Virginia Bioinformatics Institute, Blacksburg, VA, USA, [16]Keio University, Tokyo, Japan, [17]Physiome Sciences Inc., Princeton, NJ, USA, [18]Sony Computer Science Laboratories, Inc., Tokyo, Japan, [19]LION bioscience AG, Heidelberg, Germany and [20]School of Information and Computer Science, University of California, Irvine, CA, USA

Received on May 17, 2002; revised on October 15, 2002; accepted on October 25, 2002

ABSTRACT

Motivation: Molecular biotechnology now makes it possible to build elaborate systems models, but the systems biology community needs information standards if models are to be shared, evaluated and developed cooperatively.

Results: We summarize the Systems Biology Markup Language (SBML) Level 1, a free, open, XML-based format for representing biochemical reaction networks. SBML is a software-independent language for describing models common to research in many areas of computational biology, including cell signaling pathways, metabolic pathways, gene regulation, and others.

Availability: The specification of SBML Level 1 is freely available from http://www.sbml.org/.

Contact: sysbio-team@caltech.edu.

1 INTRODUCTION

Systems biology is characterized by synergistic integration of theory, computational modeling, and experiment (Kitano, 2002). Many contemporary research initiatives demonstrate the growing popularity of this kind of multidisciplinary work (e.g. Abbott, 1999). There now exists a variety of computational tools for the budding systems biologist (see below); however, the diversity of software has been accompanied by a variety of incompatibilities, and this has lead to numerous problems. For example:

---

[*] To whom correspondence should be addressed.

Bioinformatics 19(4) © Oxford University Press 2003; all rights reserved.

- Users often need to work with complementary resources from multiple simulation/analysis tools in the course of a project. Currently this involves manually re-encoding the model in each tool, a time-consuming and error-prone process.
- When simulators are no longer supported, models developed in the old systems can become stranded and unusable. This has already happened on a number of occasions, with the resulting loss of usable models to the community. Continued innovation and development of new software tools will only aggravate this problem unless the issue is addressed.
- Models published in peer-reviewed journals are often accompanied by instructions for obtaining the model definitions. However, because each author may use a different modeling environment (and model representation language), such model definitions are often not straightforward to examine, test and reuse.

1.1 Approach

The current inability to exchange models between different simulation and analysis tools has its roots in the lack of a common format for describing models. To address this, we formed a *Software Platforms for Systems Biology* forum under the auspices of the ERATO Kitano Systems Biology Project (funded by the Japan Science and Technology Corporation and hosted in part at the California Institute of Technology). The forum initially included representatives from the teams developing the software packages *BioSpice* (Arkin, 2001), *Cellerator* (Shapiro and Mjolsness, 2001), *DBsolve* (Goryanin *et al.*, 1999), *E-CELL* (Tomita *et al.*, 2001), *Gepasi* (Mendes, 1997), *Jarnac* (Sauro, 2000), *StochSim* (Morton-Firth and Bray, 1998), and *Virtual Cell* (Schaff *et al.*, 2001), and later grew to include the developers of *ProMoT/DIVA* (Ginkel *et al.*, 2000) and the CellML language at the University of Auckland and Physiome Sciences (Hedley *et al.*, 2001).

The forum decided at the first meeting in April 2000 to develop a simple, XML-based language for representing and exchanging models between simulation/analysis tools: the *Systems Biology Markup Language* (SBML). We chose XML, the eXtensible Markup Language (Bray *et al.*, 1998), because of its portability and increasingly widespread acceptance as a standard data language for bioinformatics (Achard *et al.*, 2001). SBML is formally defined using UML, the Unified Modeling Language (Object Management Group, 2002), and this in turn is used to define a representation in XML. The base definition, *SBML Level 1*, is the result of analyzing common features in representation languages used by several ODE-, DAE- and stochastic-based simulators, and encompasses the minimal information required to support non-spatial biochemical models. Subsequent releases of SBML (termed *levels*) will add additional structures and facilities to Level 1 based on features requested and prioritized by the SBML community. By freezing sets of features in SBML definitions at incremental levels, we hope to provide software authors with stable standards and allow the simulation community to gain experience with the language definitions before introducing new elements.

1.2 Benefits to Biologists

Widespread use of SBML in software packages would benefit users as well as developers, by helping to address the problems of interoperability listed earlier in this introduction. With greater interaction between tools, and a common format for publications and databases, users would be better able to spend more time on actual research rather than on struggling with data format issues. (Note that biologists and other software users are *not* intended to write their models in SBML by hand—it is the software tools that read and write the format.)

2 OVERVIEW OF SBML LEVEL 1

A chemical reaction can be broken down into a number of conceptual elements: reactant species, product species, reactions, stoichiometries, rate laws, and parameters in the rate laws. To analyze or simulate a network of reactions, additional components must be made explicit, including compartments for the species, and units on the various quantities. A definition of a model in SBML simply consists of lists of one or more of these various components:

Compartment: A container of finite volume for well-stirred substances where reactions take place.

Species: A chemical substance or entity that takes part in a reaction. Some example species are ions such as calcium ions and molecules such as ATP.

Reaction: A statement describing some transformation, transport or binding process that can change one or more species. Reactions have associated rate laws describing the manner in which they take place.

Parameter: A quantity that has a symbolic name. SBML provides the ability to define parameters that are global to a model, as well as parameters that are local to a single reaction.

Unit definition: A name for a unit used in the expression of quantities in a model. This is a facility for both setting default units and for allowing combinations of units to be given abbreviated names.

Rule: A mathematical expression that is added to the model equations constructed from the set of reactions. Rules can be used to set parameter values, establish constraints between quantities, etc.

```
 1  <?xml version="1.0" encoding="UTF-8"?>
 2  <sbml xmlns="http://www.sbml.org/sbml/level1"
 3        level="1" version="2">
 4    <model name="gene_network_model">
 5      <listOfUnitDefinitions>
 6          ...
 7      </listOfUnitDefinitions>
 8      <listOfCompartments>
 9          ...
10      </listOfCompartments>
11      <listOfSpecies>
12          ...
13      </listOfSpecies>
14      <listOfParameters>
15          ...
16      </listOfParameters>
17      <listOfRules>
18          ...
19      </listOfRules>
20      <listOfReactions>
21          ...
22      </listOfReactions>
23    </model>
24  </sbml>
```

Fig. 1. The skeleton of a model definition expressed in SBML, showing all possible top-level elements.

A software package can read in a model expressed in SBML and translate it into its own internal format for model analysis. For instance, a package might provide the ability to simulate a model by constructing a set of differential equations representing the network and then performing numerical integration on the equations to explore the model's dynamic behavior.

Figure 1 shows the skeleton of an SBML model description. It exhibits the standard characteristics of an XML data stream (Bray *et al.*, 1998): it is plain text, each element consists of a matched pair of start/end tags enclosed by '`<`' and '`>`' characters, some elements can contain attributes of the form *`attribute='value'`*, and the first line contains a particular sequence of characters (beginning with '`<?xml`') declaring the rest of the data stream as conforming to the XML encoding standard.

The element `sbml`, beginning on line 2 of Figure 1, encapsulates an SBML model definition. The first attribute, `xmlns`, is required for tools that read XML to be able to verify the syntax of a given definition against the XML Schema for SBML. (This is an aspect of XML parsing that is beyond the scope of this article; interested readers may find more information in books such as that by Skonnard and Gudgin 2001.) The `level` attribute on element `sbml` identifies the SBML *level* in use; currently the only level defined is Level 1, but Level 2 is already under development. The attribute `version` is provided to enable updated versions of a given SBML level to be distinguished.

Inside `sbml`, there must be exactly one subelement: `model`, which itself can have a single optional attribute whose value specifies the name of the model (as shown on line 4). The `model` element can contain several different subelements; each acts as a container for a different kind of component in a model definition. The contents of these `listOf`_____ containers are the topic of Section 4.

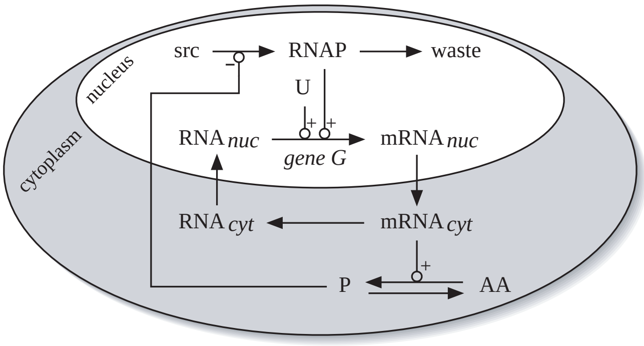


Fig. 2. Schematic diagram of the example model.

3 AN EXAMPLE MODEL

In the following sections, we describe the various components of SBML with the help of a concrete example. It illustrates one application of SBML, but it is by no means the only type of model that can be represented.

Our example is a two-compartment model of a hypothetical single-gene oscillatory circuit in a eukaryotic cell. The model is shown diagrammatically in Figure 2 and the reaction equations for the model are given in Table 1. In this highly simplified model, the nucleus of the cell is represented as one compartment and the surrounding cell cytoplasm as another compartment. Let us suppose that there is a gene G which encodes its own repressor and is transcriptionally activated at a constant rate, V_i, by a ubiquitous transcription factor U. Transcriptional activation involves several enzymatic reactions summarized here as the production of active *RNAP* (from source material, *src*) and its degradation (to *waste*). The transcribed *mRNA* is then transported out of the nucleus and into the cytoplasm, where it is translated into the product (*P*) of the gene G from constituent amino acids (*AA*) and where it is also subject to degradation. *P* travels from the cytoplasm back into the nucleus to repress further transcription of G, but is itself also subject to degradation. Eventually, the concentration of *P* becomes so low that G can be reactivated by U, and the cycle repeats itself.

Table 1. Reactions in the example model. $mRNA_{nuc}$: mRNA in nucleus. $mRNA_{cyt}$: mRNA in cytoplasm. RNA_{cyt}, RNA_{nuc}: RNA constituents. The terms beginning with the letters '*K*' and '*V*' are parameters given values in Section 4.4.

Reaction	Rate
$src \rightarrow RNAP$	$V_i/(1+P/K_i)$
$RNAP \rightarrow waste$	$V_{kd} \cdot RNAP$
$RNA_{nuc} \rightarrow mRNA_{nuc}$	$\dfrac{V_{m1} \cdot RNAP \cdot RNA_{nuc}}{K_{m1} + RNA_{nuc}}$
$mRNA_{nuc} \rightarrow mRNA_{cyt}$	$k_1 \cdot mRNA_{nuc}$
$mRNA_{cyt} \rightarrow RNA_{cyt}$	$\dfrac{V_{m2} \cdot mRNA_{cyt}}{mRNA_{cyt} + K_{m2}}$
$RNA_{cyt} \rightarrow RNA_{nuc}$	$k_2 \cdot RNA_{cyt}$
$AA \rightarrow P$	$\dfrac{V_{m3} \cdot mRNA_{cyt} \cdot AA}{AA + K_{m3}}$
$P \rightarrow AA$	$(V_{m4} \cdot P)/(P + K_{m4})$

4 THE COMPONENTS OF SBML

Our goal in this section is to describe SBML in enough detail that readers can gain a general sense for its capabilities. This description summarizes SBML's major elements but omits many details; a detailed definition is presented in the SBML specification (Hucka *et al.*, 2003).

At the outset, we need to elaborate on two data type issues. The first concerns the definitions of basic data types such as `double`, `integer`, etc. Whenever these are used in SBML, they simply refer to the definitions of these data types in XML Schema (Biron and Malhotra, 2000; Thompson *et al.*, 2000). The second issue concerns the allowable syntax of names in name attributes. Names are used throughout SBML to allow different components of a model to have meaningful labels. When an SBML model definition is converted by a simulation/analysis software tool into the tool's native internal form, these names are typically turned into symbols in the software's representation of the model. However, some simulation and analysis tools place restrictions on the characters allowed in symbolic names. To support these packages, names in SBML Level 1 are restricted to character strings having the following syntax: a name is case-sensitive and must begin with either a letter or an underscore ('_') character, followed by any number of letters, digits or underscore characters in any combination. The minimum length for a name is one letter, or one underscore followed by one letter if the first character of the name is an underscore. A 'letter' can be either upper or lower case. Also, though XML permits the use of Unicode characters (Unicode Consortium, 1996), SBML limits the set of characters allowed in names to plain ASCII text characters for compatibility with existing simulation software.

4.1 Compartments

A *compartment* in SBML represents a bounded volume in which species are located. Compartments do not necessarily have to correspond to actual structures inside or outside of a cell, although models are often designed that way. The following fragment of SBML defines the compartments for our example model:

```
<listOfCompartments>
  <compartment name="Cyt" volume="1.5" />
  <compartment name="Nuc" outside="Cyt" />
</listOfCompartments>
```

There is one required attribute for a `compartment` element, `name`, to give it a unique name by which other parts of an SBML model definition can refer to it. A `compartment` can also have an optional `volume` attribute giving the total volume of the compartment. This enables concentrations of species to be calculated in the absence of spatial geometry information. The `volume` attribute defaults to a value of '`1`' (one). The units of volume may be explicitly set using the optional attribute `units`. The value of this attribute must be one of the following: a predefined unit name from Table 2, the term '`volume`' (which, if used, signifies that the default units of volume should be used—see Section 4.5), or the name of a unit defined by a unit definition in the enclosing `model`. If absent, as in the example above, the units default to the value set by the built-in '`volume`'.

The optional attribute `outside` can be used to express containment relationships between compartments. If present, the value of `outside` for a given compartment must be the name of another compartment enclosing it, or in other words, the compartment that is 'outside' of it. This enables the representation of simple topological relationships between compartments, for those simulation systems that can make use of the information (e.g. for drawing simple diagrams of compartments). Although containment relationships are partly taken into account by the compartmental localization of reactants and products, it is not always possible to determine purely from the reaction equations whether one compartment is meant to be located within another. In the absence of a value for `outside`, compartment definitions in SBML Level 1 do not have any implied spatial relationships between each other. (We hope to introduce support for additional spatial characteristics in a future level of SBML.)

As with the other top-level components, compartments are optional in an SBML model definition. If no compartment is defined, the model is assumed to be located within a single compartment of unit volume.

4.2 Species

The `species` element in SBML is used to represent entities such as ions and molecules that participate in reactions. The following is the list of species for our example:

```
<listOfSpecies>
  <species name="mRNA_nuc" compartment="Nuc"
           initialAmount="0.0032834" />
  <species name="RNA_nuc" compartment="Nuc"
           initialAmount="96.117" />
  <species name="RNAP" compartment="Nuc"
           initialAmount="0.66349" />
  <species name="mRNA_cyt" compartment="Cyt"
           initialAmount="3.8742"/>
  <species name="P" compartment="Cyt"
           initialAmount="22.035" />
  <species name="RNA_cyt" compartment="Cyt"
           initialAmount="0.0054068" />
  <species name="AA" compartment="Cyt"
           initialAmount="90.465" />
  <species name="src" compartment="Nuc"
           initialAmount="1"
           boundaryCondition="true" />
  <species name="waste" compartment="Nuc"
           initialAmount="1"
           boundaryCondition="true" />
</listOfSpecies>
```

The `species` element has two required attributes: `name` and `initialAmount`. The attribute `name` is required to give each species a unique name in a model. The attribute `initialAmount`, of type `double`, is used to define the initial quantity (as a total molar amount, not concentration) of the species in the compartment where it is located. The units of this quantity may be set explicitly using the optional attribute `units`. The value of `units` must be one of the following: a predefined unit name from Table 2, the term '`substance`' (which, if present, signifies that the default units of quantity should be used—see Section 4.5), or a new unit name defined by a unit definition in the enclosing `model`. If absent, the units default to the value set by the built-in '`substance`'.

The attribute `compartment` is a string that names the compartment within which the species is located. The attribute can be omitted only if the model does not define any compartments (and thus assumes the default; see Section 4.1); otherwise, each species must have a value for `compartment`.

The optional attribute `boundaryCondition` takes on a boolean value to indicate whether the amount of the species is fixed or variable over the course of a simulation. The value of `boundaryCondition` defaults to a value of '`false`', indicating that by default, the amount is not fixed. If the amount of a species is defined as being fixed, it implies that some external mechanism maintains a constant quantity in the compartment throughout the course of a reaction. (The term *boundary condition* alludes to the role of this constraint in a simulation.)

A final optional attribute of `species` is `charge`, an integer indicating a charge value (in terms of electrons, not the SI unit Coulombs). This may be useful when the species is a charged ion such as calcium (Ca^{2+}).

4.3 Reactions

A *reaction* represents some transformation, transport or binding process, typically a chemical reaction, that can change one or more chemical species. In SBML, reactions are defined using lists of reactant species and products, their stoichiometric coefficients, and kinetic rate laws. Space limitations permit us to give only one SBML reaction definition as an example:

```
<listOfReactions>
 <reaction name="R1" reversible="false">
   <listOfReactants>
     <species Reference species="src" />
   </listOfReactants>
   <listOfProducts>
     <species Reference species="RNAP"/>
   </listOfProducts>
   <kineticLaw formula="Vi/(1+P/Ki)" />
 </reaction>
 ...

</listOfReactions>
```

The required `name` attribute gives the reaction a unique name to identify it in the model. The optional attribute `reversible` takes a boolean value indicating whether the reaction is reversible. If unspecified, the default value is '`true`'. An explicit flag is necessary because the kinetic law expression for a reaction is optional. Information about reversibility is useful in certain kinds of analyses such as elementary mode analysis (Schuster *et al.*, 2000).

The optional attribute `fast` is another boolean attribute in the `reaction` element; a value of '`true`' signifies that the given reaction is a 'fast' one. This may be relevant when computing equilibrium concentrations of rapidly equilibrating reactions. Simulation/analysis packages may choose to use this information to reduce the number of ODEs required and thereby optimize such computations. The default value of `fast` is '`false`'.

The reactants and products of a reaction are identified by references to species using `speciesRef` elements inside `listOfReactants` and `listOfProducts` containers. A `speciesRef` has one required attribute, `species`, whose value must be the name of a species defined in the `model`'s `listOfSpecies`. Stoichiometric numbers for the products and reactants can be specified using two optional attributes on the `speciesRef` element: `stoichiometry` and `denominator`. Both attributes take positive integers as values, and both have default values of ‘`1`’ (one). The absolute value of the stoichiometric number is the value of `stoichiometry` divided by `denominator`, and the sign is implicit from the role of the species (i.e. positive for reactants and negative for products). The use of separate numerator and denominator terms allows a simulator to employ rational arithmetic if it is capable of it, potentially reducing round-off errors and other problems during computations. In our example model above, we only needed to use the default values.

Finally, the optional `kineticLaw` element is used to provide a mathematical formula describing the rate at which the reactants combine to form the products. (In general there is no useful default value that can be substituted in place of a missing kinetic law, but the element is optional because certain kinds of network analysis are still possible in the absence of information on reaction kinetics.) The `kineticLaw` element has one required attribute, `formula`, of type `string`, that expresses the rate of the reaction in $substance/time$ units. The allowable syntax of formula strings is described in the SBML Level 1 specification; it consists of basic operators such as multiplication, addition, exponentiation, etc., as well as a number of predefined functions for common kinetic rate laws.

A `kineticLaw` element can optionally have attributes `substanceUnits` and `timeUnits` to specify the units of substance and time. If these attributes are not used in a given reaction, the units are taken from the defaults defined by the built-in terms ‘`substance`’ and ‘`time`’ of Table 3 in Section 4.5. Although not used in our two-compartment example model, a `kineticLaw` element can also contain zero or more optional `parameter` elements that define new terms used only in the `formula` string.

Readers may wonder why formulas in SBML are not expressed using MathML (W3C, 2000). Although using MathML would be more in the spirit of XML, it would introduce new complexity for software tools. Most contemporary simulation software tools for systems biology represent mathematical formulas simply using text strings. To keep SBML Level 1 simple and maximally compatible with known software, we chose to represent formulas as strings as well. This does not preclude a later level of SBML from introducing the ability to use MathML.

4.4 Parameters

The `parameter` element in SBML is used to associate a name with a floating-point value, so that the name can be used in formulas in place of the value. Here are the parameter definitions for our example:

```
<listOfParameters>
  <parameter name="Vi"  value="10" />
  <parameter name="Ki"  value="0.6"/>
  <parameter name="Vkd" value="1" />
  <parameter name="Vm1" value="50" />
  <parameter name="Km1" value="1" />
  <parameter name="k1"  value="10000" />
  <parameter name="Vm2" value="50" />
  <parameter name="Km2" value="1" />
  <parameter name="k2"  value="10000" />
  <parameter name="Vm3" value="50" />
  <parameter name="Km3" value="80" />
  <parameter name="Vm4" value="50" />
  <parameter name="Km4" value="1" />
</listOfParameters>
```

The `parameter` element has one required attribute, `name`, representing the parameter’s name in the model. The optional attribute `value` is of type `double` and determines the numerical value assigned to the parameter. The units on the `value` may be specified by the optional attribute `units`. The string used for `units` must be chosen from one of the following: a predefined unit name from Table 2; one of the three terms ‘`substance`’, ‘`time`’, or ‘`volume`’ (see Section 4.5); or the name of a new unit defined in the list of unit definitions in the enclosing `model`.

Parameters can be defined in two places in SBML: in lists of parameters defined at the top level in a `model`-type structure (in the `listOfParameters` described in Section 2), and within individual reaction definitions (as described in Section 4.3). Parameters defined at the top level are *global* to the whole model; parameters that are defined within a reaction are local to the particular reaction and (within that reaction) *override* any global parameters having the same names.

4.5 Unit Definitions

Although we did not need to define any special units in our example model, SBML does provide a way to define new units and redefine default units.

A unit definition consists of a `name` attribute and an optional `listOfUnits` subelement that in turn contains one or more `unit` elements. For example, the following definition illustrates how an abbreviation named ‘`mmls`’ can be defined for the units mmol l^{-1} s^{-1}:

```
<listOfUnitDefinitions>
  <unitDefinition name="mmls">
    <listOfUnits>
      <unit kind="mole"   scale="-3"/>
      <unit kind="liter"  exponent="-1"/>
      <unit kind="second" exponent="-1"/>
    </listOfUnits>
  </unitDefinition>
  ...
</listOfUnitDefinitions>
```

Table 2. The possible values of `kind` in a `unit` element. All are names of base or derived SI units, except for '`dimensionless`' and '`item`', which are SBML additions. '`Dimensionless`' is needed for cases where a quantity does not have units, and '`item`' is needed to express such things as 'N items' (e.g. '100 molecules'). Although '`Celsius`' is capitalized, for simplicity, SBML requires that these names be treated in a case-insensitive manner. Also, note that the gram and liter/litre are not strictly part of International System of Units (BIPM, 2000); however, they are so commonly used in SBML's areas of application that they are included as predefined unit names.

`ampere`
`becquerel`
`candela`
`Celsius`
`coulomb`
`dimensionless`
`farad`
`gram`
`gray`
`henry`
`hertz`
`item`
`joule`
`katal`
`kelvin`
`kilogram`
`liter`
`litre`
`lumen`
`lux`
`meter`
`metre`
`mole`
`newton`
`ohm`
`pascal`
`radian`
`second`
`siemens`
`sievert`
`steradian`
`tesla`
`volt`
`watt`
`weber`

As this illustrates, SBML uses a compositional approach to defining units. The definition of mmol l^{-1} s^{-1} is constructed by combining a `unit` element representing millimoles with a `unit` element representing liter^{-1} and another `unit` element representing second^{-1}.

The `unit` element has one required attribute, `kind`, whose value must be a name taken from the list of units in Table 2. The optional `exponent` attribute has a default value of '`1`' (one). A unit such as liter^{-1} is obtained by using attributes `kind="liter"` and `exponent="-1"`. Finally, a `unit` element also accepts an optional `scale` field; its value must be an integer used to set the scale of the unit. For example, a unit that has a `kind` value of '`gram`' and a `scale` value of '`-3`' signifies $10^{-3} *$ gram, or milligrams. The default value of `scale` is zero.

There are three special unit names in SBML, listed in Table 3, corresponding to the three types of quantities that play roles in biochemical reactions: amount of substance, volume and time. SBML defines default units for these quantities, all with a default `scale` value of `0`. The various components of a model, such as parameters, can use only the predefined units from Table 2, new units defined in unit definitions, or the three predefined names '`substance`', '`time`', and '`volume`' from Table 3. The latter usage signifies that the units to be used should be the designated defaults. A model may change the default scales by reassigning the keywords '`substance`', '`time`', and '`volume`' in a unit definition.

Table 3. SBML's built-in quantities.

Name	Allowable Units	Default Units
`substance`	moles *or* no. of molecules	moles
`volume`	liters	liters
`time`	seconds	seconds

4.6 Rules

Rules in SBML provide a way to create constraints on variables and parameters for cases in which the constraints cannot be expressed using the reaction components described in Section 4.3. There are two orthogonal dimensions by which rules can be described. First, there are three different possible functional forms, corresponding to the following three general cases (where x is a variable, f is some arbitrary function, and W is a vector of parameters and variables that may include x):

1. left-hand side is zero: $0 = f(W)$
2. left-hand side is a scalar: $x = f(W)$
3. left-hand side is a rate-of-change: $\mathrm{d}x/\mathrm{d}t = f(W)$

The second dimension concerns the role of variable x in the equations above: x can be the name of a compartment (to set its volume), the name of a species (to set its concentration), or the name of a parameter (to set its value).

The approach taken to covering these cases in SBML is to define separate kinds of elements for each of the cases, and to allow these within a single `listOfRules` container within a `model` definition (see Table 1). Each contains a `name` attribute that specifies the quantity being referenced, and a `formula` attribute that holds the right-hand side expression of the rule. For the actual details, we refer readers to the SBML Level 1 specification.

5 STATUS AND FUTURE PLANS

As mentioned above, SBML Level 1 is intended to provide only a basic representation of biochemical reaction networks. Space constraints prevent us from giving a detailed description of SBML here; the full definition is available in a separate document (Hucka *et al.*, 2003). A number of simulation and analysis packages already support SBML Level 1 or are in the process of being extended to support it. At the time of this writing, the tools include: *Cellerator* (Shapiro and Mjolsness, 2001), *DBsolve* (Goryanin *et al.*, 1999), *E-CELL* (Tomita *et al.*, 2001), *Gepasi* (Mendes, 1997), *Jarnac* (Sauro, 2000), *NetBuilder* (Brown *et al.*, 2002), *ProMoT/DIVA* (Ginkel *et al.*, 2000), *StochSim* (Morton-Firth and Bray, 1998), and *Virtual Cell* (Schaff *et al.*, 2001).

Future levels of SBML will add more features requested by the modeling community. The process for feature selection involves a request for proposals from the *Software Platforms for Systems Biology* forum, followed by discussions and votes during subsequent meetings, and finally the drafting of a specification by selected members. Some of the features under discussion for SBML Level 2 are the introduction of MathML and metadata support. The latter will add a systematic mechanism for recording such information as author and publication references; it will also provide a way to annotate a model with information such as cross-references to biological data sources.

Finally, the project is moving from a primarily Caltech/ERATO-led effort toward a community-led and -maintained model for the future of SBML. We invite all interested parties to join us.

ACKNOWLEDGEMENTS

This work has been supported by: the Japan Science and Technology Corporation's ERATO Kitano Symbiotic Systems Project; an International Joint Research Grant from the NEDO/Japanese Ministry of Economy, Trade and Industry; and the Rice Genome and Simulation Project from the Japanese Ministry of Agriculture.

REFERENCES

Abbott,A. (1999) Alliance of US labs plans to build map of cell signaling pathways. *Nature*, **402**, 219–220.

Achard,F., Vaysseix,G. and Barillot,E. (2001) XML, bioinformatics and data integration. *Bioinformatics*, **17**, 115–125.

Arkin,A.P. (2001) *Simulac & Deduce*. Available via the World Wide Web at http://gobi.lbl.gov/∼aparkin.

Biron,P.V. and Malhotra,A. (2000) XML Schema part 2: Datatypes, Available via the World Wide Web at http://www.w3.org/TR/xmlschema-2/.

Bray,T., Paoli,J. and Sperberg-McQueen,C.M. (1998) Extensible markup language (XML) 1.0, Available via the World Wide Web at http://www.w3.org/TR/1998/REC-xml-19980210.

Brown,C.T., Rust,A.G., Clarke,P.J.C., Pan,Z., Schilstra,M.J., De Buysshcher,T., Griffin,G., Wold,B.J., Cameron,R.A., Davidson,E.H. and Bolouri,H. (2002) New computational approaches for analysis of cis-regulatory networks. *Developmental Biology*, **246**, 86–102.

Bureau International des Poids et Mesures (2000) *The International System of Units (SI) supplement 2000*. Available via the World Wide Web at http://www.bipm.fr/pdf/si-supplement2000.pdf.

Ginkel,M., Kremling,A., Tränkle,F., Gilles,E.D. and Zeitz,M. (2000) Application of the process modeling tool ProMoT to the modeling of metabolic networks. In Troch,I. and Breitenecker,F. (eds), *Proc. of the 3rd MATHMOD*.

Goryanin,I., Hodgman,T.C. and Selkov,E. (1999) Mathematical simulation and analysis of cellular metabolism and regulation. *Bioinformatics*, **15**, 749–758.

Hedley,W.J., Nelson,M.R., Bullivant,D.P. and Nielson,P.F. (2001) A short introduction to CellML. *Phil. Trans. Roy. Soc. London A*, **359**, 1073–1089.

Hucka,M., Finney,A., Sauro,H.M. and Bolouri,H. (2003) Systems Biology Markup Language (SBML) Level 1: Structures and facilities for basic model definitions, Available via the World Wide Web at http://www.sbml.org/.

Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.

Mendes,P. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.*, **22**, 361–363.

Morton-Firth,C.J. and Bray,D. (1998) Predicting temporal fluctuations in an intracellular signalling pathway. *J. Theor. Biol.*, **192**, 117–128.

Object Management Group (2002) *UML Specification documents available via the World Wide Web at http://www.omg.org/uml/*.

Sauro,H.M. (2000) Jarnac: a system for interactive metabolic analysis. In Hofmeyr,J.-H., Rohwer,J. and Snoep,J. (eds), *Animating the Cellular Map*. Stellenbosch Univ. Press.

Schaff,J., Slepchenko,B., Morgan,F., Wagner,J., Resasco,D., Shin,D., Choi,Y.S., Loew,L., Carson,J., Cowan,A. *et al.* (2001) *Virtual Cell*. Available via the World Wide Web at http://www.nrcam.uchc.edu.

Schuster,S., Fell,D.A. and Dandekar,T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, **18**, 326–332.

Shapiro,B.E. and Mjolsness,E.D. (2001) Developmental simulations with Cellerator. In Yi,T.-M., Hucka,M., Morohashi,M. and Kitano,H. (eds), *Proceedings of the Second International Conference on Systems Biology (ICSB2001)*. Omnipress.

Skonnard,A. and Gudgin,M. (2001) Essential XML Quick Reference, Addison-Wesley.

Thompson,H.S., Beech,D., Maloney,M. and Mendelsohn,N. (2000) *XML Schema part 1: Structures*. Available via the World Wide Web at http://www.w3.org/TR/xmlschema-1/.

Tomita,M., Nakayama,Y., Naito,Y., Shimizu,T., Hashimoto,K., Takahashi,K., Matsuzaki,Y., Yugi,K., Miyoshi,F., Saito,Y. *et al.* (2001) *E-Cell*. Available via the World Wide Web at http://www.e-cell.org/.

Unicode Consortium, (1996) *The Unicode Standard, Version 2.0*. Addison-Wesley Developers Press.

W3C, (2000) *W3C's math home page*. Available via the World Wide Web at http://www.w3.org/Math/.

Chemical Engineering Science 61 (2006) 7805–7818

Chemical Engineering Science

www.elsevier.com/locate/ces

Model based statistical analysis of adsorption equilibrium data

M. Joshi[b], A. Kremling[a], A. Seidel-Morgenstern[a,b,*]

[a] *Max-Planck-Institut für Dynamik komplexer technischer Systeme, Magdeburg, Germany*
[b] *Institut für Verfahrenstechnik, Otto-von-Guericke-Universität, Magdeburg, Germany*

Received 5 July 2006; received in revised form 18 August 2006; accepted 18 August 2006
Available online 3 September 2006

Abstract

A large group of separation problems can be solved using selective adsorption on suitable solids. A mathematical description of adsorption isotherms, which relate the equilibrium concentrations in the fluid phase to the loadings of the solid, could be used to design, observe and control such processes in an efficient way. However, the determination of the isotherms typically requires the identification of unknown parameters in postulated models from experimental data. While for the estimation of the parameters a number of tools and methods are available, a comprehensive analysis of the quality of the parameters is seldom performed. To estimate and characterize parameters obtained from adsorption measurements in this work a non-linear regression analysis was explored in combination with an extended statistical analysis. Hereby, the non-linearity method ("intrinsic" and "parameter-effect" non-linearity) proposed by Bates and Watts [1980. Relative curvature measures of non-linearity. Journal of the Royal Statistical Society: Series B (Methodological) 42, 1–25] was used to check the quality of parameters and the suitability of model/data combinations. The variances of the parameters are determined with the bootstrap method originally proposed by Efron and Tibshirani [1993. An Introduction to the Bootstrap. Chapman and Hall, CRC Press, London, Boca Raton.]. The later approach clearly overcomes some limitation of classical Fisher-information-matrix (FIM) method. By applying these statistical methods to different adsorption models and data sets, it was found that non-linearity method is a good tool to check the quality of the model/data combination. Furthermore, it was found that the confidence intervals of the parameters determined based on the bootstrap are larger than predicted by traditional methods.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Adsorption isotherms; Parameter estimation; Non-linearity; Bias; Correlation coefficient; Confidence intervals; Bootstrap

1. Introduction

In many separation and purification processes like preparative chromatography, potable water purification and waste water treatment, adsorption phenomena play an important role. Knowledge about the underlying adsorption equilibrium is the important pre-requisite in order to design and optimize adsorption processes (Ruthven, 1984). To solve a specific separation problem, typically, adsorption equilibrium functions have to be determined experimentally. Several static and dynamic methods are available to acquire this information (Seidel-Morgenstern, 2004). A frequently applied method for determining single solute adsorption isotherms is the conventional batch method based on mixing known amounts of adsorbent with solutions of various initial concentrations and measuring the equilibrium concentrations. Solving the mass balance, corresponding equilibrium loadings can be simply calculated. As a result of such standard experiments, a certain number of pairs "concentration in the liquid phase vs loading" are available which are analyzed with various isotherm models. The models are subsequently used to analyze the system behavior under consideration and allow design, control and optimization of the process.

The aim of the paper is to analyze the quality of the parameters of isotherm models after their estimation using standard optimization methods. In general, besides the estimation of the parameters, the determination of parameter accuracy is a further step before model analysis and application (Isermann, 1992). Uncertainties of the model parameters may lead to wrong conclusions and make it difficult to design a certain adsorption process. Therefore, a proper quantification of the uncertainties is

---

[*] Corresponding author. Lehrstuhl für Chemische Verfahrenstechnik, Otto-von-Guericke Universität Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany. Tel.: +49 391 671 8643; fax: +49 391 671 2028.

E-mail address: anseidel@vst.uni-magdeburg.de (A. Seidel-Morgenstern).

0009-2509/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ces.2006.08.052

required. A standard method frequently used in parameter estimation procedures is a simple transformation of the adsorption isotherms that are non-linear in the parameters into a form that is linear in the parameters. This linear form allows an easy calculation of the parameters and parameter variances. However, given a model f that relates a concentration c to the loading a by

$$a = f(c, \underline{\theta}) + \varepsilon \tag{1}$$

(where $\underline{\theta}$ is a vector of parameters, and ε the overall error) a transformation

$$g(a) = g(f(c, \underline{\theta}) + \varepsilon) \tag{2}$$

also transforms the overall error term of the model. Since the method of linear regression normally assumes an independently and identically distributed (i.i.d) error, the application of the standard linear regression method is not longer reasonable. Consequently, methods of nonlinear regression have to be applied to estimate the parameters and to determine parameter uncertainties.

In this contribution a non-linearity analysis according to work of Bates and Watts (1980) is used to analyze the non-linear properties of the function f, which can be quantified in terms of "intrinsic" and "parameter-effect" non-linearity. These measures describe the extent of the non-linear behavior of the model and can be used to characterize the quality of the "model/data combination". Theory used so far for the analysis of uncertainties in parameters is mainly based on the Fisher-information-matrix (FIM) (e.g., Ljung, 1999), which uses a linear approximation of the function f. Therefore, the calculated variance of the free parameters represent only a lower bound. A method to overcome this limitation, the so called bootstrap approach (Efron and Tibshirani, 1993), is used in this study to get a better approximation of confidence intervals.

The combination of the non-linear regression analysis with the bootstrap approach is applied to equilibrium data for the adsorption of indol from aqueous solution on activated carbon measured under static conditions. In a previous study parameters of different isotherm models were already estimated by a least-square approach (Seidel et al., 1985). However, the quality of the parameters obtained was not further analyzed. As an extension to the previous mentioned study, the non-linear characteristic properties of the parameters were analyzed in this work in detail, confidence intervals were calculated and a comparison of the different models with respect to the available experimental data was performed. The non-linearity in combination with Box's bias estimate was used to check the quality of model/data combinations. Further, confidence interval for non-linear models were determined based on bootstrap method.

2. Experimental data

Experimental data were taken from Seidel (1987). Adsorption equilibrium data of indol dissolved in water have been measured in a concentration range of 10^{-3}–2 mmol/l on four different activated carbons at 20 °C. The following activated carbon samples were analyzed: Hydraffin 71 (carbon 1), TVAX 1 (carbon 2), Filtrasorb 400 (carbon 3) and AG3 (carbon 4). Details regarding the experimental procedure and the characteristics of the activated carbon can be found in Seidel (1987). The data sets for all carbons analyzed and the parameter values determined earlier by Seidel et al. (1985) are given in Tables A.1 and A.2 respectively.

3. Methods

3.1. Model formulation

Commonly used isotherm models applied to describe adsorption from aqueous solution on activated carbon are the Langmuir, Freundlich and Redlich-Peterson isotherm equations (Ruthven, 1984). The well-known Langmuir isotherm reads

$$a_L = a_s \frac{k_L \cdot c}{1 + k_L \cdot c}. \tag{3}$$

The Freundlich isotherm reads

$$a_F = k_F \cdot c^n. \tag{4}$$

The Redlich-Peterson isotherm reads

$$a_{\mathrm{RP}} = \frac{H_{\mathrm{RP}} \cdot c}{1 + k_{\mathrm{RP}} \cdot c^p}. \tag{5}$$

The disadvantage of the Freundlich isotherm is that it does not follow Henry's law at concentrations approaching zero. This condition is fulfilled by the Redlich-Peterson isotherm. For $p=1$, Eq. (5) converts to the Langmuir isotherm, for $1 \geqslant k_{\mathrm{RP}} \cdot c^p$ it simplifies to Henry's law and for $1 \ll k_{\mathrm{RP}} \cdot c^p$ it becomes identical to the Freundlich isotherm. For $p \neq 1$, it is not possible to transform the three parameter Redlich-Peterson model into a linear form.

A more comprehensive theory for comparing the three adsorption models given above (and other models) is provided by the theory of adsorption on heterogeneous surfaces (Jaroniec and Madey, 1988; Cerofolini and Rudzinski, 1997).

3.2. Parameter estimation and model accuracy

The free parameters in Eqs. (3)–(5) were estimated using a least-square (LS) approach. These parameters should minimize the quadratic error between the experimental data $Y^{\exp}$ and the model output Y^{mod} for all sample points. Typically, the squared error is further normalized by the experimental data $Y^{\exp}$ (i.e. $a^{\exp}$) to bring all measurements into the same scale. As in Seidel et al. (1985) in this study the following objective function (Θ) was used

$$\Theta = \sum_{i=1}^{m} \left(\frac{Y_i^{\exp} - Y_i^{\mathrm{mod}}}{Y_i^{\exp}} \right)^2, \tag{6}$$

where m is the number of sample points.

Model accuracy was tested (i) with the relative standard deviation

$$\% \, \sigma_r = \sqrt{\frac{\Theta}{\mathrm{df}}} \cdot 100, \tag{7}$$

with degree of freedom df $= m - k$, where k is the number of model parameters and with the coefficient of regression R_c^2 (Montgomery et al., 2001)

$$R_c^2 = R^2 - \frac{k \cdot (1 - R^2)}{\text{df} - 1} \tag{8}$$

with

$$R^2 = 1 - \frac{\Theta}{C_T} \quad \text{and} \quad C_T = \sum_{i=1}^{m} \left(\frac{Y_i^{\exp} - 1/m \sum_{i=1}^{m} Y_i^{\exp}}{Y_i^{\exp}} \right)^2 . \tag{9}$$

The relative standard deviation ($\% \sigma_r$) was used to compare the results with Seidel et al. (1985). In the analysis R_c^2 is used instead of R^2 since it takes into account the number of free parameters, which is necessary to compare various models with different degrees of freedom.

The methods described so far were applied already in Seidel et al. (1985) for analyzing the data and the models. In this study we performed additional statistical tests to check the quality of parameters and the suitability of model/data combination. In the following section, we will describe these methods shortly.

3.3. Non-linearity analysis

The non-linearity analysis gives a better understanding regarding the extent to which non-linear models differ from linear models. The concept was first introduced by Bates and Watts (1980). The approach is illustrated below with an example model η that is non-linear in a parameter θ. Using only two measuring points it is possible to draw the solution locus for a simple model given by

$$\begin{bmatrix} \eta_1(\theta) \\ \eta_2(\theta) \end{bmatrix} = \begin{bmatrix} 3 \cdot x_1^{\theta} \\ 3 \cdot x_2^{\theta} \end{bmatrix}, \quad \theta \geqslant 0. \tag{10}$$

in a two-dimensional phase plot. Parameter θ is now varied while the other values are fixed: e.g. $x_1 = 0.015$, $x_2 = 0.7$. The solution locus shown in Fig. 1 describes the dependency of given values for x on parameter θ. Since the model is non-linear in θ, the expectation surface/solution locus is curved and equal spaced values of θ map to unequally spaced values on $\eta(\theta)$. The shape of the solution locus and the spacing of the values of constant $\Delta\theta$ on the solution locus in the vicinity of a fixed value of θ, are used as a measure of the degree to which the non-linear model differs from the linear model. In-general the non-linearity of a model can be separated into two components (Bates and Watts, 1980):

1. "Intrinsic non-linearity" (IN) is associated with the curvature of the solution locus in the sample space. It represents the inverse of the radius of a circle which best approximates the solution locus in the direction of the tangent plane for fixed parameter values. For a linear model, IN is zero since the solution locus is e.g. a straight line in the two-dimensional case.
2. "Parameter-effect non-linearity" (PE) is associated with the projections of the parameter lines on the tangent plane to the solution locus and is a measure of the lack of parallelism and the inequality of the parameter lines spacing on the solution locus at the least square solution. For a linear model the parameters are equally spaced on the solution locus.

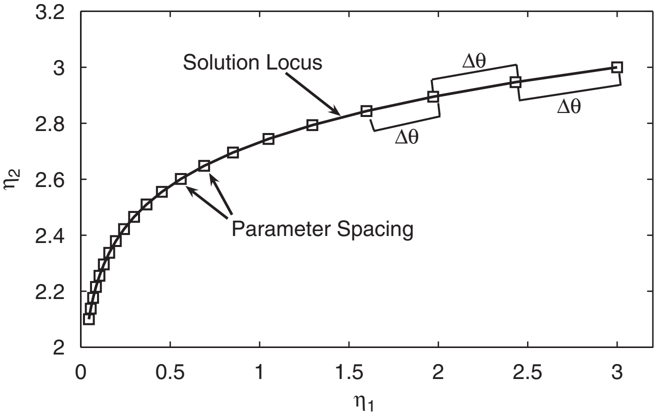


Fig. 1. Plot of the expectation surface (solid line) in the response space (Eq. (10)). The symbols correspond to equal spaced values for θ $(0, 0.05, 0.1, \ldots, 1)$.

IN and PE can be calculated as suggested by Bates and Watts (1988). Briefly, one calculates the tangent plane (first derivative with respect to the parameters) for function $\eta(\theta)$ at a fixed value for θ. Then the acceleration (second derivative with respect to the parameters) can be decomposed in the direction parallel to the tangent plane and normal to the tangent plane. To determine the curvature measures for each parameter, both components of the acceleration are divided by the squared length of the tangent vector in direction of the selected parameter. The curvature measure that is in direction of the tangent vector is designed as PE while the curvature measure normal to the tangent vector is designed as IN. Since IN and PE are normalized, they are dimensionless. The significance of IN and PE, i.e., the close-to-linear behavior, can be assessed by comparing the values with a corresponding significance level from the F-distribution (critical values IN_c and PE_c)

$$\text{IN} \leqslant \frac{1}{\sqrt{F(\alpha/2, \text{df}_1, \text{df}_2)}} = \text{IN}_c,$$

$$\text{PE} \leqslant \frac{1}{\sqrt{F(\alpha/2, \text{df}_1, \text{df}_2)}} = \text{PE}_c, \tag{11}$$

where $\text{df}_1 = k$ and $\text{df}_2 = m - k$, and e.g. $\alpha = 0.05$, representing a 95% confidence interval. Values smaller than the statistical limits IN_c, PE_c indicate a close-to-linear behavior. The theory will be used below to access the non-linearity in adsorption models with real data sets. For a number of examples, Bates and Watts show that IN is typically small and that the major contribution to the non-linearity is due to the parameterization, i.e., PE. Moreover, they found that PE is closely associated with the measure of bias introduced by Box (1971).

3.4. Box bias calculation

For single response systems, the bias in the least square estimate of parameters in non-linear regression models can be calculated as follows according to the method proposed by Box (1971)

$$\mathrm{Bias}(\hat{\theta}) = \frac{-\sigma_a^2}{2}\left[\sum_{i=1}^{m} W_i W_i^T\right]^{-1} \times \sum_{i=1}^{m} W_i \,\mathrm{tr}\left[\left\{\sum_{i=1}^{m} W_i W_i^T\right\}^{-1} H_i\right], \tag{12}$$

where the standard deviation σ_a is defined as

$$\sigma_a = \sqrt{\frac{\sum_{i=1}^{m} (Y_i^{\mathrm{exp}} - Y_i^{\mathrm{mod}})^2}{\mathrm{df}}}, \tag{13}$$

and W_i and H_i are the first and second derivative of the model function with respect to the parameters, respectively. The bias given in Eq. (12) is a vector of dimension of parameter (k), representing the discrepancy between the estimates of the parameters and the true parameter values. The bias expressed as a percentage of a LS estimate is a good measure for the non-linearity in parameters. The percentage bias can be expressed as:

$$\%B_j = 100 \cdot \frac{\mathrm{Bias}(\hat{\theta}_j)}{\hat{\theta}_j}, \quad j = 1, \ldots, k. \tag{14}$$

Ratkowsky (1983) quantifies a model as linear if $\%B$ is below 1.

3.5. Fisher-information-matrix

The classical way to estimate the confidence intervals for parameters is based on parameter sensitivities ω_j, which describe an infinitesimal change of a state variable x according to a change of parameter θ_j

$$\omega_j = \frac{\partial x}{\partial \theta_j}. \tag{15}$$

FIM is calculated by the sum over all sample points with the sensitivity vector $\underline{\omega}$ and the inverse of variance of the measurements σ_a^2 (Ljung, 1999)

$$\underline{\omega} = [\omega_1 \;\; \omega_2 \;\; \cdots \;\; \omega_k], \tag{16}$$

$$\mathrm{FIM} = \frac{1}{\sigma_a^2} \sum_m \underline{\omega}^T \cdot \underline{\omega}, \tag{17}$$

This expression for FIM appears if one calculates the variance $\sigma_{\hat{\theta}}^2$ of estimated parameters $\hat{\theta}$

$$\sigma_{\hat{\theta}} = E[(\hat{\theta} - E[\hat{\theta}])^2], \tag{18}$$

with $E[\bullet]$ expectation. The following equation holds true for the variance of single parameter $\sigma_{\hat{\theta}_j}^2$ based on the Cramer–Rao inequality

$$\sigma_{\hat{\theta}_j}^2 \geqslant (\mathrm{FIM}^{-1})_{jj}. \tag{19}$$

Thus the confidence interval for the parameters is given by

$$\hat{\theta}_j - \sigma_{\hat{\theta}_j} \cdot t_{\alpha/2}^{\mathrm{df}} < \theta_j < \hat{\theta}_j + \sigma_{\hat{\theta}_j} \cdot t_{\alpha/2}^{\mathrm{df}}. \tag{20}$$

The variance–covariance matrix (FIM^{-1}) can be used to calculate correlations between parameters

$$\Omega_{hj} = \begin{cases} 1 & \text{if } h = j, \\ \dfrac{(\mathrm{FIM}^{-1})_{hj}}{\sqrt{(\mathrm{FIM}^{-1})_{hh}} \cdot \sqrt{(\mathrm{FIM}^{-1})_{jj}}} & \text{if } h \neq j, \end{cases} \tag{21}$$

where Ω_{hj} represent the correlation coefficients between parameters h and j.

3.6. Re-parameterization

Re-parameterization is highly recommended, if a model is found to be far from linear (i.e., if $\mathrm{PE}/\mathrm{PE}_c > 1$, Eq. (11)) (Ratkowsky, 1990). For re-parameterization, the parameters of the model may be expressed as a function of the parameters of a second model. Below re-parametrization for the Redlich-Peterson isotherm is shown as an example. New parameters ϕ_1, ϕ_2 as a function of the actual parameters H_{RP} and k_{RP} can be defined as follows:

$$\phi_1 = \frac{1}{H_{\mathrm{RP}}}; \quad \phi_2 = \frac{k_{\mathrm{RP}}}{H_{\mathrm{RP}}}. \tag{22}$$

Then Redlich-Peterson isotherm can be written as

$$a_{\mathrm{RP}} = \frac{c}{\phi_1 + \phi_2 \cdot c^p}. \tag{23}$$

Model (5) and (23) are "re-parameterizations" of each other. As the shape of the solution locus is independent of the parameterization, the process of re-parameterization does not alter IN. Different re-parameterizations of the basic model produce the same goodness of fit and the same fitted values, but parameterization has beneficial effect by making confidence regions narrower for the parameters ϕ_1 and ϕ_2 and convergence faster (Table 1).

It should be noted that transformation of the parameters is different from the transformation of the response variable into a linear form. Transformation of a response variable distort the response space and creates a new expectation surface, thereby affects the disturbance term and the validity of the assumptions on it. On the other hand, transformation of the parameters does not affect the assumption of the deterministic part and the error term.

3.7. Bootstrap approach

The bootstrap approach introduced by Efron and Tibshirani (1993) is a data-based simulation method for statistical

Table 1
Comparison of convergence using normal Redlich-Peterson (Eq. (5)) and re-parameterized Redlich-Peterson (Eq. (23))

D	Whole data set			
	Normal Redlich-Peterson Eq. (5)		Re-parameterized Redlich-Peterson Eq. (23)	
	No. of iterations	Function count	No. of iterations	Function count
$\mathbf{C}_2$	500	2000	30	124
C_1	118	476	68	276
C_3	500	2000	64	260
C_4	250	2000	49	200

Termination criterion given to the optimizer was, maximum number of iterations 1000 and maximum number of function evaluations 2000.

inference. A main application of the method is the estimation of confidence regions for a non-parametric distribution. To perform the analysis, the set of experimental data $\mathbf{Y}^{\text{exp}}$ is used as a data base. Due to measurement errors a repetition of the experiment would lead to a slightly different set of data $\mathbf{Y}_1^*$ and therefore to a different set of estimated parameters. The bootstrap approach now uses a large set of B times replicated experimental data $(\mathbf{Y}_1^*, \mathbf{Y}_2^*, \mathbf{Y}_3^* \ldots, \mathbf{Y}_B^*)$ to calculate statistical properties of the resulting distribution of the (re)-estimated sets of parameters. Since it is not practical to repeat an experiment very often in reality, a Monte–Carlo simulation is used to generate the data. The approach is described below in brief; more details can be found in Joshi et al. (2006).

3.8. Reconstruction of the experimental data

For every run of the Monte-Carlo simulation, the set of data has to be replaced by a new one. As the considered experiments were performed only once (Seidel et al., 1985), standard deviations for all sample points were not available. The following procedure was used to generate additional quasi-experimental data sets: a constant average standard deviation (σ) was assumed for the whole data set, which was calculated by Eq. (13). The bootstrap data were generated with an additive absolute error $(\sigma_a \cdot r)$ for the model f with optimal parameters $\hat{\theta}$:

$$\mathbf{a}^* = f(c, \hat{\theta}) + \sigma_a \cdot r, \tag{24}$$

where r is a random number generated from normal distribution with mean 0 and standard deviation 1.

3.9. Simulation, parameter estimation and outlier analysis

To estimate the parameters, Matlab environment with “lsqnonlin” solver, which applies the “Levenberg–Marquardt” method for optimization was used to minimize Θ in Eq. (6). Simulations and parameter (re-)estimation was performed B times to generate a sufficient data sets using Eq. (24).

The re-estimated parameters are first analyzed with respect to outliers. Outliers are extreme cases of one variable, or a combination of variables, which have a strong influence on the calculation of statistics. Sometimes the data sets include one or more values that appear unusually large or small and are out of place when compared with the other data values. These values are known as outliers and are often erroneously included in the analysis of data sets. A single outlier is capable in changing considerably the confidence interval of parameters. We computed outliers as described in (Montgomery et al., 2001). Quartiles Q_i divide the sorted data set into four equal parts where 25% of the data can be found between Q_1 and Q_2 (representing the median) and 25% of the data can be found between Q_2 and Q_3. The spread sp is defined as $\text{sp} = Q_3 - Q_1$. Outliers are defined as such values that are beyond the borders given by $Q_1 - 1.5 \cdot \text{sp}$ and $Q_3 + 1.5 \cdot \text{sp}$. In case of parametric histograms the outliers with respect to parameter can be found because of extreme data sets generated by random error $(\sigma_a \cdot r)$.

3.10. Confidence intervals in the bootstrap framework

Confidence interval analysis in the field of parameter estimation is one of the important statistical tests to evaluate parameter reliability. The goal of bootstrap confidence interval theory is to calculate confidence limits for a parameter θ_j from the bootstrap distribution which is represented in form of a parametric histogram (DiCiccio and Efron, 1996).

The set of replicated experimental data $\mathbf{a}_1^*, \mathbf{a}_2^*, \mathbf{a}_3^*, \ldots, \mathbf{a}_B^*$ is used to calculate statistical properties of the resulting distribution of the (re)-estimated set of parameters $\hat{\theta}_1^*, \hat{\theta}_2^*, \hat{\theta}_3^*$ and $\hat{\theta}_B^*$ where B is the number of bootstrap replications. Let $\hat{\theta}^{*(\alpha)}$ indicate the $100 \cdot (1-\alpha)$th percentile of B bootstrap replication. Then the percentile interval $(\overline{\theta}_{\text{lo}}, \overline{\theta}_{\text{up}})$ of intended coverage $1 - 2\alpha$, is obtained by

$$(\overline{\theta}_{\text{lo}}, \overline{\theta}_{\text{up}}) = (\hat{\theta}^{*(\alpha/2)}, \hat{\theta}^{*(1-\alpha/2)}). \tag{25}$$

Once the intervals are calculated by the above given procedure it is necessary to describe the length L, the shape sh of the confidence interval $(\overline{\theta}_{\text{lo}}, \overline{\theta}_{\text{up}})$ and the shape of the histogram sh_H which can be calculated as follows:

$$L = \overline{\theta}_{\text{up}} - \overline{\theta}_{\text{lo}}, \tag{26}$$

$$\text{sh} = \frac{\overline{\theta}_{\text{up}} - \overline{\theta}}{\overline{\theta} - \overline{\theta}_{\text{lo}}}, \tag{27}$$

$$\text{sh}_H = \frac{\overline{\theta}_{\text{lg}} - \overline{\theta}}{\overline{\theta} - \overline{\theta}_{\text{sm}}}. \tag{28}$$

Shape sh measures asymmetry of the confidence interval about the point estimate. Shape $\mathrm{sh} > 1.0$ indicates greater distance from $\overline{\theta}_{\mathrm{up}}$ to $\overline{\theta}$ than from $\overline{\theta}$ to $\overline{\theta}_{\mathrm{lo}}$. The corresponding shape of the histogram, sh_H, is based on the deviation of mean from the smallest value ($\overline{\theta}_{\mathrm{sm}}$) and the largest value ($\overline{\theta}_{\mathrm{lg}}$). If length is based on the current value of $\overline{\theta}$, $\%L$ can be used for normalization:

$$\%L = \frac{L}{\overline{\theta}} \cdot 100. \tag{29}$$

From the data obtained, correlation coefficients Ω_b are calculated with the standard method (Constantinides and Mostoufi, 1999)

$$\Omega_b = \frac{\mathrm{Cov}(\underline{\theta}^h, \underline{\theta}^j)}{\hat{\sigma}_{\underline{\theta}^h}\hat{\sigma}_{\underline{\theta}^j}}. \tag{30}$$

3.11. Effect of sample size

The quality of the estimated parameters determined strongly depends on the number of data points and the range of concentration where the measurements are performed. To analyze the influence of the number of data points, the available data set was reduced, by taking every second point. In this way the covered range of concentrations remain same. All methods introduced so far will be also applied for data sets modified in this manner. IN and PE (Eq. (11)) values are useful quantities to study the effect of sample size.

4. Results and discussions

Four data sets presented in Table A.1 were analyzed with the three adsorption isotherm models described above. Below at first, Carbon 2 data set with 13 data points is analyzed in detail. Subsequently the results for the remaining data sets (Carbon 1, 3, and 4) are summarized.

4.1. Analysis of Carbon 2 data

4.1.1. Langmuir isotherm

In the field of adsorption thermodynamics, frequently, nonlinear isotherm models are transformed to models that are linear in parameters. It is obvious that the Langmuir model (Eq. 3, $a_L(c)$) can be transformed to a linear form in three different ways Laszlo, 2005, (Chairata et al., 2005), and (Senthilkumaar et al., 2006), e.g.

$$\frac{1}{a_L} = \frac{1}{a_s \cdot k_L} \cdot \frac{1}{c} + \frac{1}{a_s}, \tag{31}$$

$$\frac{c}{a_L} = \frac{1}{a_s \cdot k_L} + \frac{1}{a_s} \cdot c, \tag{32}$$

$$a_L = a_s - \frac{1}{k_L} \cdot \frac{a_L}{c}. \tag{33}$$

Table 2
Comparison between the parameters of transformed linear Langmuir models vs non-linear Langmuir model and transformed Freundlich model vs non-linear Freundlich model for Carbon 2 data set

Equation No.	$\underline{\theta}_j$	Parameter $\hat{\theta}_j$	Standard deviation $\%\sigma_r$ (Eq. (7))
Langmuir isotherm			
Eq. (3)	a_s	2.3	19.32
	k_L	321.6	
Eq. (31)	a_s	2.0	22.33
	k_L	492.4	
Eq. (32)	a_s	3.0	44.85
	k_L	59.5	
Eq. (33)	a_s	2.4	20.62
	k_L	344.8	
Freundlich isotherm			
Eq. (4)	k_F	3.6	19.09
	n	0.26	
Eq. (34)	k_F	3.6	20.07
	n	0.25	

Parameters can be then estimated via linear regression. It is important to note again that, if an additive error term is assumed as an appropriate representation of the overall error, the transformation of the function involves transformation of the error term, too (see Eq. (2)). All three transformable Langmuir models deliver different sets of parameters (see Table 2). Results in Fig. 2a, show that three different isotherms $a_L(c)$ are obtained based on the Eqs. (31)–(33). Although the agreement for small concentrations is good, for large values of the concentration significant deviations from the measurements can be detected. This is also true if Eq. (3) is applied. In general it was found that Langmuir parameters have the highest bias of all parameters (e.g. Bias = 9.84% for k_L) compared to Freundlich and Redlich-Peterson isotherm parameters. Furthermore, R_c^2 is 0.78 (Eq. (8)) and $\%\sigma_r$ is high (Table 2), showing the worst agreement of this isotherm with measurements. The correlation coefficient ($\Omega_{a_s k_L} = 0.52$, (Eq. (21))) indicates that these two parameters are not strongly correlated.

4.1.2. Freundlich isotherm

The Freundlich model (Eq. (4)) can also be transformed to a linear form (e.g. Laszlo, 2005):

$$\ln a_F = n \cdot \ln c + \ln k_F. \tag{34}$$

In case of Freundlich model, only small deviations between transformed and non-linear model (Eq. (4)) were observed (Fig. 2b, Table 2).

A detailed statistical analysis e.g. IN and PE etc. was performed with Freundlich isotherm (Eq. (4)). Tables 3 and 4 summarizes the results. IN and PE both are below the critical limit. Furthermore, agreement between the models and data set is good with high regression coefficient $R_c^2 = 0.98$. The percentage bias for both parameters is below 1%. The confidence intervals and correlation between parameters ($\Omega_{k_F n} = 0.66$) are small revealing a relative high precision of the estimated parameters.

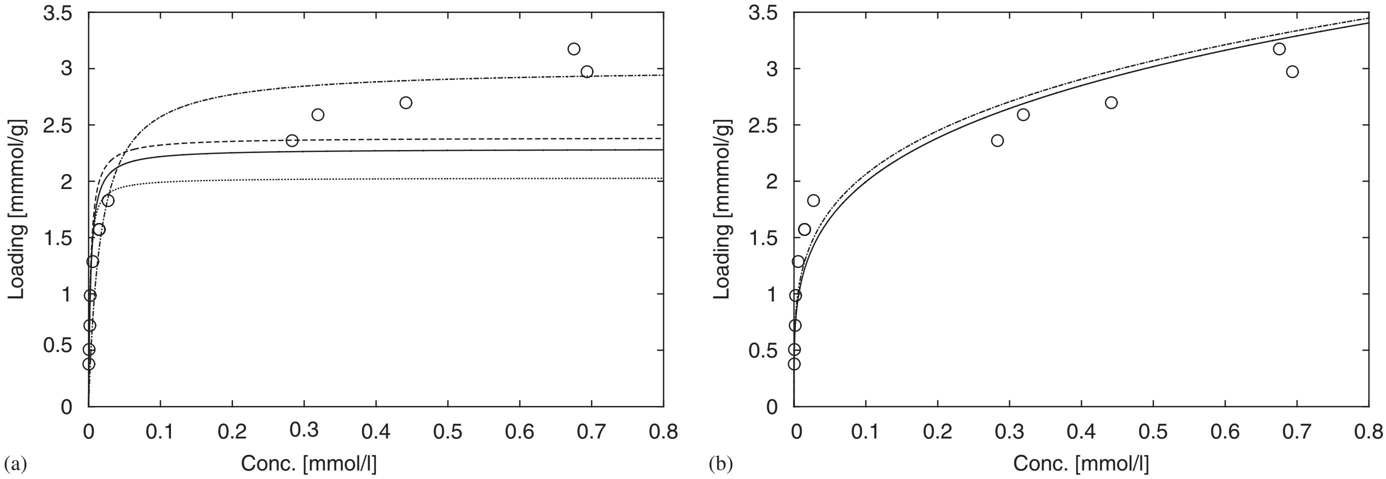


Fig. 2. (a) transformed Langmuir models as described in Eqs. (31)–(33). Points are experimental data (Carbon 2). ···: Eq. (31), —· : Eq. (32), and — — : Eq. (33). (b) —· : Eq. (34). Solid lines represents non-linear models in original form (Eqs. (3)–(4)).

Table 3
Nonlinearity estimates for the normal, re-parameterized model with whole and reduced data sets for Freundlich and Redlich-Peterson isotherm

D	m	Property	Limit IN_c & PE_c	Freundlich	Limit IN_c & PE_c	Redlich-Peterson normal	Redlich-Peterson re-parameterized
Whole data set							
$\mathbf{C}_2$	13	IN	0.4362	0.0617	0.4552	0.1968	0.1967
		PE		0.1450		1.7443	0.3254
C_1	20	IN	0.4683	0.0417	0.4993	0.2291	0.2285
		PE		0.0730		3.1992	0.4229
C_3	21	IN	0.4710	0.0530	0.5029	0.2348	0.2284
		PE		0.0799		11.2271	0.1282
C_4	22	IN	0.4734	0.0219	0.5061	0.1579	0.1462
		PE		0.0326		7.4871	0.3824
Reduced data set							
$\mathbf{C}_2$	7	IN	0.3443	0.0923	0.3166	0.3220	0.3210
		PE		0.2335		2.7429	0.4930
C_1	10	IN	0.4062	0.0715	0.4120	0.4063	0.4067
		PE		0.1237		4.6467	0.6837
C_3	11	IN	0.4183	0.0851	0.4297	0.4333	0.4325
		PE		0.1335		16.4186	0.2193
C_4	11	IN	0.4183	0.0370	0.4297	0.2246	0.2173
		PE		0.0521		22.9301	0.5912

$$\mathrm{Limit} = \frac{1}{\sqrt{F(\alpha/2, \mathrm{df}_1, \mathrm{df}_2)}}.$$

4.1.3. Redlich-Peterson isotherm

Fig. 3 and Table 3 summarizes the results for the normal Redlich-Peterson model (Eq. (5)). IN is below the critical limit, but PE is above the limit. The agreement between the model and the data set is also confirmed by the high regression coefficient $R_c^2 = 0.99$. However, the quality of the parameters cannot be assured because of the high PE value. To know which parameter behaves non-linearly in the model, the bias was calculated. Percentage bias values exceed 1% for both parameters H_{RP} and k_{RP}, whereas for p is less than 1%. Furthermore, broad confidence intervals were obtained (Table 4). Basically this is because of a high correlation between the two parameters H_{RP} and k_{RP} ($\Omega_{12} = 0.99$).

A re-parameterization was performed for the Redlich-Peterson model (Eq. (23)) in order to decrease PE, bias, correlation between parameters and size of confidence intervals. The results obtained after re-parameterization are also given in Table 3 (IN and PE). From the results summarize in Fig. 3a, it can be seen that IN remains the same as explained before. However, drastic decrease in PE value is observed (Fig. 3b). However still, for parameter ϕ_1 (Eq. (22)) the percentage bias value still exceeds 1% (Table 5). The correlation between the parameters has improved ($\Omega_{H_{\mathrm{RP}}k_{\mathrm{RP}}} = 0.35$) and the confidence intervals became smaller. Parameters that characterize the goodness of fit have comparable values.

Table 4
Comparison between Freundlich and Redlich-Peterson confidence intervals calculated with FIM method (Eq. (17))

D	$\underline{\theta}_j$	Freundlich						$\underline{\theta}_j$	Redlich-Peterson					
		R_c^2	σ_a	$\%\sigma_r$	$\hat{\theta}_j$	%L	%B		R_c^2	σ_a	$\%\sigma_r$	$\hat{\theta}_j$	%L	%B
Whole data set														
$\mathbf{C}_2$	k_F	0.98	0.27	19.09	3.61 ± 0.40	22.02	0.12	H_{RP}	0.99	0.12	6.49	1297.1 ± 832.03	128.3	6.2
	n				0.26 ± 0.06	42.92	0.40	k_{RP}				418.1 ± 260.96	124.8	6.0
								p				0.86 ± 0.04	8.6	0.09
C_1	k_F	0.98	0.16	10.57	2.14 ± 0.12	11.59	0.0	H_{RP}	0.99	0.13	8.64	2639.4 ± 3727.10	282.4	38.2
	n				0.18 ± 0.03	29.22	0.11	k_{RP}				1264.1 ± 1780.40	281.7	38.1
								p				0.85 ± 0.03	6.87	0.05
C_3	k_F	0.98	0.21	14.14	3.22 ± 0.21	13.08	0.03	H_{RP}	0.98	0.20	14.48	9594.7 ± 39293.44	819.1	359.1
	n				0.23 ± 0.03	26.52	0.13	k_{RP}				3028.9 ± 12345.45	815.2	357.3
								p				0.79 ± 0.04	10.46	0.08
C_4	k_F	0.99	0.08	7.48	2.51 ± 0.07	5.90	0.003	H_{RP}	0.99	0.08	7.82	9632.5 ± 31044.46	644.6	222.9
	n				0.24 ± 0.01	11.21	0.02	k_{RP}				3898.7 ± 12544.64	643.5	222.5
								p				0.77 ± 0.02	4.6	0.04
Reduced data set														
$\mathbf{C}_2$	k_F	0.98	0.28	20.84	3.49 ± 0.63	36.28	0.28	H_{RP}	0.99	0.11	7.60	1350.4 ± 1606.24	237.9	13.3
	n				0.26 ± 0.10	74.98	1.03	k_{RP}				443.5 ± 513.23	231.4	12.9
								p				0.85 ± 0.07	15.6	0.22
C_1	k_F	0.97	0.21	12.84	2.18 ± 0.25	22.78	0.002	H_{RP}	0.98	0.16	10.46	2289.9 ± 5202.94	454.4	74.2
	n				0.18 ± 0.05	55.50	0.31	k_{RP}				1083.0 ± 2451.06	452.6	73.9
								p				0.86 ± 0.06	13.2	0.22
C_3	k_F	0.98	0.25	14.06	3.12 ± 0.35	22.22	0.05	H_{RP}	0.98	0.24	15.06	7358.2 ± 50141.01	1362.9	955.3
	n				0.23 ± 0.05	45.02	0.33	k_{RP}				2398.1 ± 16272.09	1357.1	951.4
								p				0.78 ± 0.08	19.6	0.22
C_4	k_F	0.99	0.09	7.70	2.46 ± 0.13	10.83	0.009	H_{RP}	0.99	0.10	8.18	19945.9 ± 208667.71	2092.3.7	1995.1
	n				0.23 ± 0.02	20.28	0.05	k_{RP}				8185.7 ± 85529.38	2089.7	1992.6
								p				0.77 ± 0.03	8.73	0.08

Values for R_c^2 (Eq. (8)) and standard deviation (Eqs. (13) and (7)) are also given. All results are shown for whole and reduce data sets.

4.1.4. Bootstrap analysis

4.1.4.1. Bootstrap confidence intervals The standard FIM method and the bootstrap method are compared with respect to parameter accuracy. With Freundlich and Redlich-Peterson models $B = 2000$ bootstrap replication are performed (MATLAB 7 is used; each run lasts approx. 4–6 h on a workstation with dual processor 2.2 GHz each, AMD Opteron 248, and with 4 GB RAM). For higher values of B, no changes in characteristics of the histograms could be observed. For the calculation of the confidence intervals $\alpha = 0.05$ is used. The σ_a values used in Eq. (24) to generate data are given in Table 4.

Table 6 and Fig. 4, summarizes the results of the bootstrap analysis. The comparison between the two approaches reveals, that with the standard FIM method the confidence intervals of the parameter are underestimated. The size of the confidence intervals, calculated by the bootstrap method and with the FIM differ up to a factor of ≈ 2 (Fig. 4a, Freundlich model, parameter k_F). Further, maximum probability values ($\hat{\theta}_j^{\max}$) for the parameters are also given in Table 6.

The non-linearity of the model can be observed by high values of the shape factor ($\mathrm{sh}_H > 1$). It can be seen from Fig. 4b that Redlich-Peterson parameters are highly biased (asymmetric histograms) showing non-linear dependency of the parameters, which agrees quite well with Box's bias criterion ($\%B > 1$). The correlation coefficient between parameters (H_{RP} and k_{RP}) from bootstrap parametric data is $\Omega_b = 0.99$.

4.1.4.2. Comparison of the confidence intervals by a simulation study To illustrate the usefulness of bootstrap method, simulations were performed with the Freundlich model in order to determine confidence intervals for loadings obtained from both methods i.e., bootstrap and FIM. Extreme profiles for loadings are calculated using upper bounds and lower bounds of the parameters. Fig. 5a shows the intervals calculated for loading a for Carbon 2. The confidence region for a calculated using bootstrap approach contains more simulated data points (≈ 95%) in comparison to the profiles calculated with FIM method which contains only ≈ 76% of data points. This reveals that intervals calculated with the bootstrap method are more appropriate.

4.1.5. Effect of sample size

The effect of sample size on the non-linear behavior of model/data combination was examined by reducing the

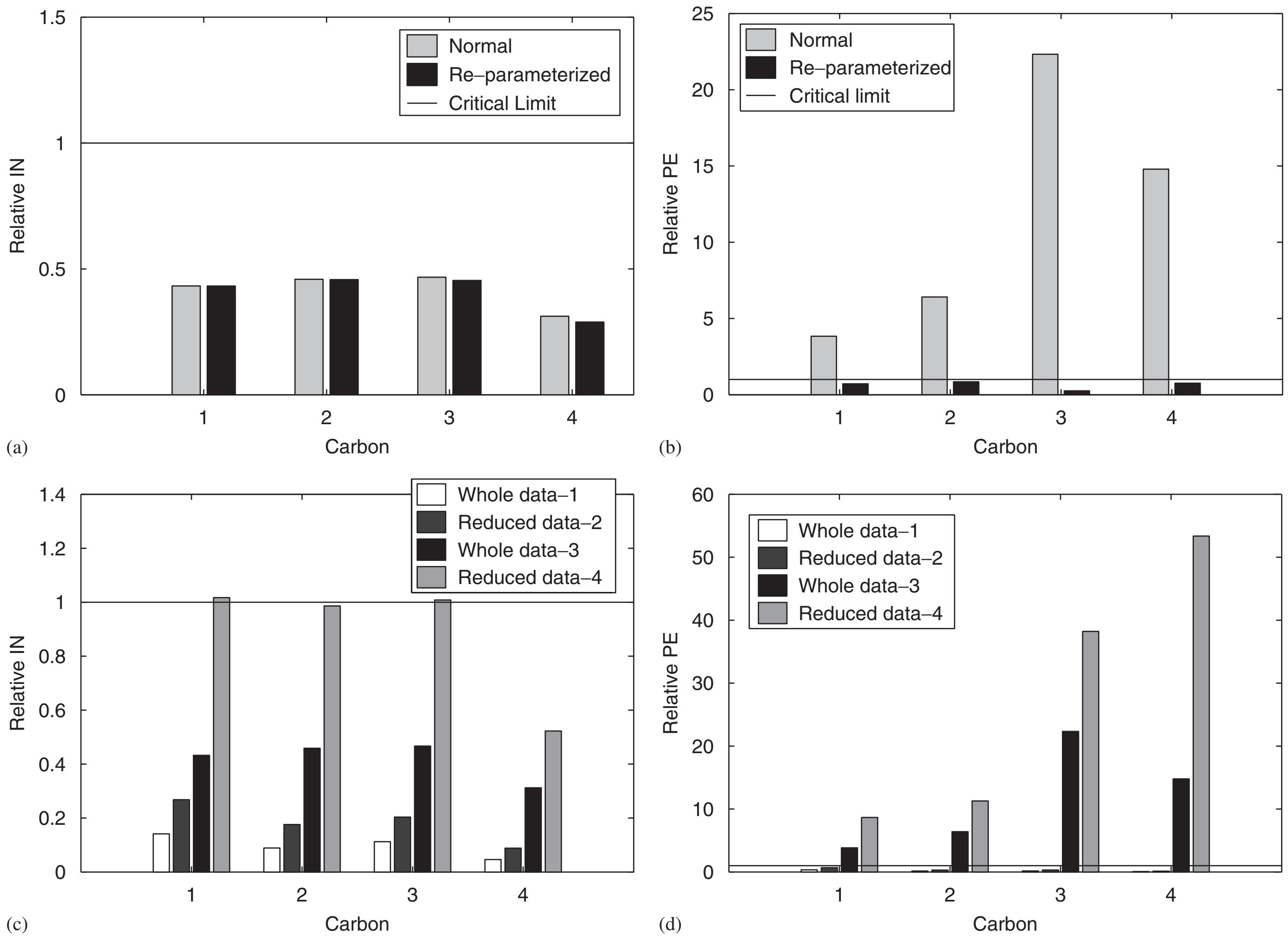


Fig. 3. Comparison between all data set (Carbon 1–4): (a) relative IN values (IN/IN_c) before and after re-parameterization for Redlich-Peterson isotherm; (b) relative PE values (PE/PE_c) before and after re-parameterization for Redlich-Peterson isotherm; (c) relative IN values for whole and reduced data set; (d) relative PE values for whole and reduce data set. (1–2) Freundlich isotherm and (3-4) Redlich-Peterson isotherm.

Table 5
FIM confidence intervals for Redlich-Peterson after re-parameterization

D	N	$\underline{\theta}_j$	R_c^2	σ_a	Point estimation				
					$\hat{\theta}_{\mathrm{lo}}$	$\hat{\theta}_j$	$\hat{\theta}_{\mathrm{up}}$	%L	%B
Whole data set									
$\mathbf{C_2}$	13	ϕ_1	0.99	0.12	0.0003	0.0008	0.0013	128.3	2.13
		ϕ_2			0.304	0.322	0.340	11.2	0.098
		p			0.82	0.86	0.89	8.63	0.09
C_1	20	ϕ_1	0.99	0.13	−0.0002	0.0004	0.0009	286.8	6.76
		ϕ_2			0.457	0.479	0.500	9.2	0.0596
		p			0.82	0.85	0.88	6.9	0.08
C_3	21	ϕ_1	0.98	0.21	−0.0004	0.0001	0.0005	1433.4	30.07
		ϕ_2			0.292	0.314	0.334	13.5	0.0137
		p			0.74	0.78	0.83	10.9	0.09
C_4	22	ϕ_1	0.99	0.08	−0.0003	0.00002	0.0003	2575.9	14.7
		ϕ_2			0.388	0.399	0.412	6.12	0.028
		p			0.742	0.767	0.778	4.6	0.03

Table 6
Comparison between confidence intervals of parameters of Freundlich and Redlich-Peterson models calculated with Bootstrap method ($B = 2000$)

D	$\underline{\theta}_j$	Freundlich								$\underline{\theta}_j$	Redlich-Peterson							
		$\tilde{\theta}_j$	$\bar{\theta}_{lo}$	$\bar{\theta}_j$	$\bar{\theta}_{up}$	%L	sh$_H$	θ_j^{max}	O		$\tilde{\theta}_j$	$\bar{\theta}_{lo}$	$\bar{\theta}_j$	$\bar{\theta}_{up}$	%L	sh$_H$	θ_j^{max}	O
Whole data set																		
C_2	k_F	3.68	3.07	3.70	4.47	37.82	1.11	3.55	75	H_{RP}	1171.88	536.59	1235.59	2312.35	143.72	1.59	1150.00	109
	n	0.29	0.21	0.30	0.45	77.28	1.46	0.27	157	k_{RP}	377.94	181.24	398.05	726.19	136.90	1.53	345.00	108
										p	0.86	0.81	0.86	0.92	12.46	1.03	0.86	44
C_1	k_F	2.12	1.98	2.13	2.28	14.22	1.07	2.13	26	H_{RP}	2058.08	643.53	2456.59	6476.27	237.43	2.57	1350.00	234
	n	0.19	0.16	0.19	0.24	42.96	1.15	0.19	80	k_{RP}	997.56	315.87	1185.11	3121.19	236.71	2.52	700.00	237
										p	0.85	0.82	0.85	0.89	8.90	1.01	0.85	39
C_3	k_F	3.30	2.98	3.31	3.72	22.41	1.19	3.23	80	H_{RP}	4976.54	529.48	8386.70	24842.47	289.90	2.72	1000.00	4
	n	0.25	0.20	0.26	0.33	50.07	1.23	0.25	127	k_{RP}	1570.91	178.50	2663.52	7888.80	289.48	2.80	250.00	6
										p	0.79	0.72	0.79	0.88	19.29	0.99	0.78	78
C_4	k_F	2.53	2.39	2.53	2.68	11.43	1.02	2.51	34	H_{RP}	7715.34	1268.24	8295.70	18302.99	205.34	2.03	2500.00	0
	n	0.25	0.23	0.25	0.28	20.72	1.04	0.25	38	k_{RP}	3129.44	525.09	3387.30	7530.58	206.82	2.06	1250.00	1
										p	0.77	0.75	0.77	0.79	5.62	1.01	0.77	40

O stands for number of outliers. Further given are median ($\tilde{\theta}_j$), mean ($\bar{\theta}_j$), shape of parametric histogram (sh$_H$), percent length of confidence interval (%L), maximum probability values (θ_j^{max}).

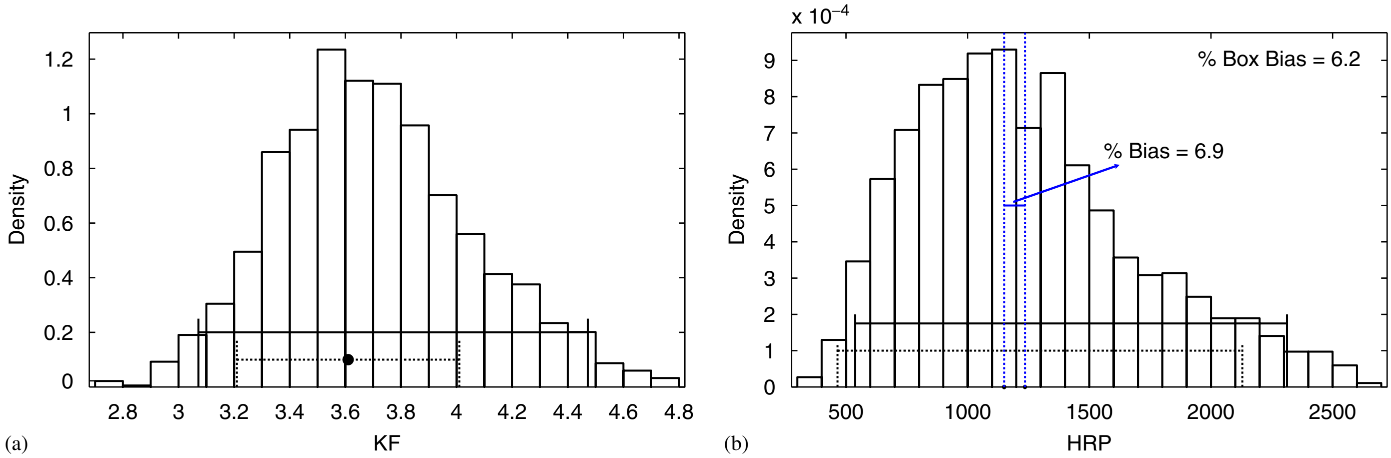


Fig. 4. Comparison between bootstrap confidence intervals (solid line) with classical FIM confidence intervals (dotted line) for Carbon 2: (a) parameter k_F of Freundlich model; (b) parameter H_{RP} of Redlich-Peterson model.

sample size for the Carbon 2 data set from 13 to 7 (using every second point). This study was intended to explore whether the sample size was enough for estimating parameters with satisfying statistical properties. The analysis was performed with the Freundlich and the Redlich-Peterson isotherm models as described before. Table 3 in addition with Table 4 show the effect of sample size. It can be concluded, that the reduction in data set increases IN (Fig. 3c) and PE (Fig. 3d), and therefore also the bias. In case of Freundlich the curvature measure of non-linearity is still below the critical value for the reduced data set. Obviously, the length of confidence intervals (Table 4) increases for all the parameters.

A general statement on required sample sizes cannot be given because it not only depends on model structure but also on the region where the experiments are performed.

4.1.6. *Analysis of Carbon 1, 3, and 4*

Analysis of the data sets for Carbon 1, 3, and 4 (Table A.1) reveals in case of the Freundlich model, that the IN and PE values are always below the critical values, indicating a close-to-linear behavior.

In case of the Redlich-Peterson model, a small discrepancy between IN values before and after re-parameterization is found (Table 3 and Fig. 3a). The reason behind this is, that parameters H_{RP} and k_{RP} are strongly correlated before re-parameterization and a global minimum is difficult to obtain. Negative limits of the confidence intervals are partly observed for Carbon 1, 3 and 4. This is again due to a high correlation between the parameter H_{RP} and k_{RP}. PE values in all cases are higher than the critical values (Fig. 3b). Re-parameterization is performed to improve PE, correlation and confidence regions of the

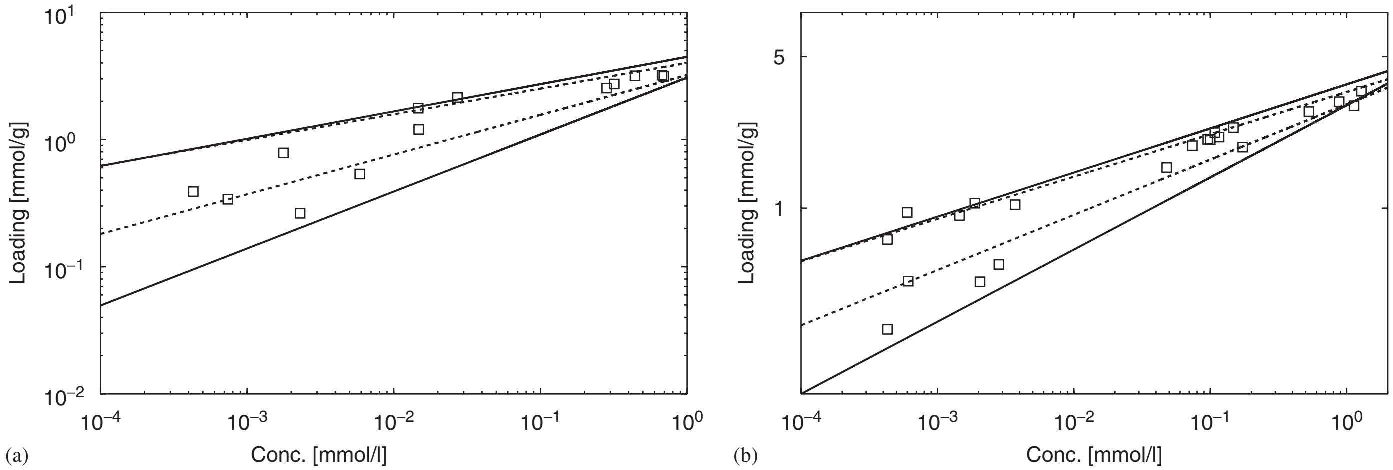


Fig. 5. Application of bootstrap confidence intervals with Freundlich model for Carbon 2 and 3. Confidence region for the loading based on the bootstrap approach (solid line) and based on FIM (dashed): (a) Carbon 2; (b) Carbon 3.

parameters. As for Carbon 2, here, again a drastic reduction in PE values and correlation is observed. The confidence intervals are improved for ϕ_2 as shown in Table 5. Now, only for parameter ϕ_1 the confidence interval includes negative values. It can be due to two reason. Firstly, the parameter ϕ_1 is highly biased. Secondly, more data points are needed in high concentration region, to reduce the size of confidence intervals. As for Carbon 2, the reduction in data sets causes an increase in the curvature measures of the non-linearity for Carbon 1, 3, and 4 Fig. 3(c–d).

Application of bootstrap confidence interval for Carbon 3 (Fig. 5) also shows that ≈ 95% of data points lie inside profiles of loadings calculated from bootstrap approach in comparison to trajectories calculated from FIM method which contain only ≈ 71% of the data points.

5. Conclusions

In this study a methodology for the evaluation of the quality of free parameters in adsorption isotherm models was applied. It comprises a non-linearity analysis and the analysis of the parameter confidence intervals by two methods, the classical approach with the Fisher-Information-Matrix and the bootstrap approach. The analysis was performed on single component adsorption equilibrium data (indol from aqueous solution on activated carbon) measured in a previous study (Seidel et al., 1985).

In all cases considered slightly better minimal values of the objective function (σ_r) used in the parameter estimation is obtained in comparison to the old study. Two approaches, i.e., transformable linear and non-linear least square, show that the Langmuir model does not give a satisfying description of the considered experimental data. In spite of the fact, that the Freundlich isotherm does not follow Henry's law, it has the smallest values of IN and PE as well as the smallest confidence intervals in comparison to the other models. This indicates that for the covered range of the liquid phase concentrations, this simple model provides an adequate description. The three parameter Redlich-Peterson model fits the data best but the non-linearity of the parameters (PE values) is high, giving high bias, and contains partly negative values in the confidence intervals. Although the model fits the data excellently from statistical point of view (R_c^2 close to 1), the model is not acceptable in the current form in terms of quality of parameters. The concept of re-parameterization was used to reduce the PE values for the Redlich-Peterson isotherm. Significant reduction in PE and bias values are found leading to smaller correlation coefficient values and better confidence intervals for parameter ϕ_2. However confidence interval for ϕ_1 remains negative. Thus, more data points would be needed in high concentration range to obtained better results.

To overcome the limits of the Fisher-information-matrix (FIM), which gives only a lower bound of the parameter confidence intervals, the bootstrap method was applied. In general, the non-linearities detected with the IN and PE values can also be found in the bootstrap parametric histograms. Non-linear parameters show a bias and have shape values different from 1. The confidence intervals determined with the bootstrap method are broader then the intervals obtained with FIM (in some cases up to a factor of 2). Using the FIM approach, low sensitivities of the parameters lead to negative values in the confidence intervals. Benefits from the bootstrap approach becomes obvious if the limit-values of the parameter confidence intervals are used in a simulation study predicting limits of equilibrium loadings. Using the bootstrap interval values, ≈ 95% of the measured data points lie in between the predicted limits. In contrast, using the values from the FIM approach, only ≈ 71–76% data are covered. The application of the bootstrap method is also a good possibility to verify the bias estimates calculated with Box formula. As can be seen from the tables given, high values of the shape are in accordance with high bias values.

A study of the sample size for the Freundlich model shows, that half of the size of data set used is enough to get values of

IN and PE below the critical limit. Clearly with the reduction of the data sets, a trend to higher IN and PE and broader confidence intervals is observed. Model/data combination analysis for Carbon 1, 3, and 4 with respect to the Freundlich model reveals a close-to-linear behavior as IN and PE values were found below the critical limits. For the Redlich-Peterson model, the non-linearity due to the parameterization was found again to be high, concluding that re-parameterization of the model is necessary to improve the quality of parameters.

Non-linearity analysis together with a bias estimate was found to be a useful method to evaluate the accuracy of parameter estimates. Together with the bootstrap method the quality of the parameters can be characterized in a much better way then with traditional methods typically used so far. This concept can be applied not only to analyze such simple adsorption isotherms models but also for any model which is non-linear in parameters.

Notation

Symbols

a	loading,mmol/g
a_s	saturation loading in Eq. (3),mmol/g
c	concentration in liquid phase,mmol/l
C_T	variability of observed values
D	data set
df	degree of freedom
E	expectation
f	non-linear function
F	F-distribution
g	transformed function
H	Hessian matrix of the model function
H_{RP}	constant in Eq. (5),l/g
k	number of parameters
k_F	constant in Eq. (4),$\mathrm{mmol}^{1-\mathrm{n}} \cdot \mathrm{l}^{\mathrm{n}/\mathrm{g}}$
k_L	constant in Eq. (3)
k_{RP}	constant in Eq. (5),$(\mathrm{l/mmol})^{p}$
L	length of confidence interval
m	number of data points
n	exponent in Eq. (4)
p	exponent in Eq. (5)
r	generates random number with normal distribution with mean zero and standard deviation one
R^2	coefficient of regression
R_c^2	corrected coefficient of regression
sh	shape of confidence interval
sh_H	shape of histogram
t	student *t*-distribution
tr	trace
W	Jacobian matrix of the model function
x	state variable
Y	data points

Greek letters

α	significance level (e.g. 0.05)
ε	overall error
η	non-linear function
σ_a	absolute standard deviation in Eq. (13)
$\%\sigma_r$	relative standard deviation in Eq. (7)
θ	real parameter
$\hat{\theta}$	estimated parameters
$\theta^{\max}$	maximum probability value from the parametric histogram
$\bar{\theta}$	mean value from the parametric histogram
$\tilde{\theta}$	median value from the parametric histogram
Θ	objective function
ϕ	function of old parameters in Redlich-Peterson case
φ	function of old parameters in Langmuir case
Ω	correlation coefficient
ω	parameter sensitivity in Eq. (15)

Subscripts

a	absolute
b	bootstrap
B	number of bootstrap
F	Freundlich
h	indices for parameter
i	running number of experimental data
j	indices for number of parameters
L	Langmuir
lo	lower bound
lg	largest value in parametric histogram
r	relative
RP	Redlich-Peterson
sm	smallest value in parametric histogram
up	upper bound

Superscripts

exp	experimental values
mod	model values
$*$	new generated bootstrap data

Acknowledgement

Financial support provided by German Research Foundation (DFG FOR-447) and Fonds der Chemische Industrie is gratefully acknowledge. We would like to thank Prof. D. Flockerzi for fruitful discussions.

Appendix A

The data sets for all carbons analyzed and the parameter values determined earlier by Seidel et al. (1985) are given in Tables A.1 and A.2 respectively.

Table A.1
Equilibrium data for the adsorption of indol from aqueous solution at 20 °C for Carbons C_1–C_4 from Seidel (1987)

N	Carbon 1		Carbon 2		Carbon 3		Carbon 4	
	c mmol/l	*a* mmol/g	*c* mmol/l	*a* mmol/g	*c* mmol/l	*a* mmol/g	*c* mmol/l	*a* mmol/g
1	0.00061	0.518	0.00043	0.378	0.00043	0.514	0.00032	0.368
2	0.00064	0.509	0.00074	0.507	0.00043	0.596	0.00036	0.347
3	0.00265	0.69	0.00177	0.72	0.00060	0.541	0.00070	0.473
4	0.00299	0.689	0.00230	0.986	0.00061	0.772	0.00144	0.453
5	0.0032	0.83	0.00587	1.287	0.00145	0.863	0.00384	0.649
6	0.00384	0.769	0.0147	1.572	0.00188	0.974	0.00405	0.678
7	0.00702	1.002	0.0148	1.572	0.00205	0.649	0.00512	0.812
8	0.00917	0.974	0.0272	1.828	0.00282	0.911	0.00811	0.823
9	0.0433	1.335	0.2832	2.360	0.00371	0.735	0.0118	0.781
10	0.0523	1.438	0.3191	2.590	0.0479	1.891	0.0125	0.793
11	0.0523	1.732	0.4416	2.697	0.0739	2.055	0.0314	1.200
12	0.1408	1.62	0.6754	3.174	0.0957	2.149	0.0320	1.074
13	0.2858	1.793	0.6936	2.972	0.0999	1.960	0.0493	1.330
14	0.3221	1.723			0.1080	2.223	0.0503	1.360
15	0.4318	1.948			0.1160	2.241	0.1599	1.518
16	0.5468	1.833			0.1478	2.293	0.1877	1.726
17	0.7364	1.957			0.1727	2.200	0.2666	1.707
18	1.058	2.092			0.5284	2.775	0.2704	1.880
19	2.289	2.226			0.8812	2.609	0.2941	1.893
20	2.292	2.232			1.132	3.177	0.5674	2.095
21					1.282	3.160	1.497	2.795
22							1.540	2.760

Table A.2
Parameter estimates for the Freundlich and Redlich-Peterson isotherms models using the data from Table A.1

D	$\underline{\theta}_j$	Freundlich			$\underline{\theta}_j$	Redlich-Peterson		
		$\hat{\theta}_j$	$\%\sigma_r$	σ_a		$\hat{\theta}_j$	$\%\sigma_r$	σ_a
Whole data set								
$\mathbf{C}_2$	k_F	2.17	19.63	0.27	H_{RP}	2378	6.7	0.12
	n	0.18			k_{RP}	1128		
					p	0.86		
C_1	k_F	3.64	10.7	0.17	H_{RP}	1299	9.41	0.12
	n	0.25			k_{RP}	416		
					p	0.86		
C_3	k_F	3.27	14.7	0.21	H_{RP}	7139	15.12	0.19
	n	0.22			k_{RP}	2247		
					p	0.8		
C_4	k_F	2.52	7.6	0.08	H_{RP}	18248	8.6	0.08
	n	0.24			k_{RP}	7289		
					p	0.77		

Small differences between the parameters and standard deviations given below and parameters in Seidel et al. (1985) are due to truncating primary data differently and different optimization algorithm used.

References

Bates, D.M., Watts, D.G., 1980. Relative curvature measures of non-linearity. Journal of the Royal Statistical Society: Series B (Methodological) 42, 1–25.

Bates, D.M., Watts, D.G., 1988. Nonlinear Regression Analysis and its Application. Wiley, New York.

Box, M.J., 1971. Bias in nonlinear estimation. Journal of the Royal Statistical Society: Series B (Methodological) 33, 171–201.

Cerofolini, G.F., Rudzinski, W., 1997. Equilibria and dynamics of gas adsorption on heterogeneous solid surfaces. In: Rudzinski, W., Steele, W.A., Zgrablich, G. (Eds.), Theoretical Principles of Single- and Mixed-gas Adsorption Equilibria on Heterogeneous Solid Surfaces. Elsevier, Amsterdam.

Chairata, M., Rattanaphania, S., Bremnerb, J.B., Rattanaphani, V., 2005. An adsorption and kinetic study of lac dyeing on silk. Dyes and Pigments 64, 231–241.

Constantinides, A., Mostoufi, N., 1999. Numerical Methods for Chemical Engineers with Matlab Applications. Prentice-Hall, Englewood Cliffs.

DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence interval. Statistical Science 11, 189–212.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, CRC Press, London, Boca Raton.

Isermann, R., 1992. Identifikation dynamischer systeme, Vol. Band 1. Springer, Berlin.

Jaroniec, M., Madey, R., 1988. Physical Adsorption on Heterogeneous Solids. Elsevier Science, Amsterdam.

Joshi, M., Seidel-Morgenstern, A., Kremling, A., 2006. Exploiting bootstrap method for quantifying parameter confidence intervals in dynamical systems. Metabolic Engineering 8, 447–455.

Laszlo, K., 2005. Adsorption from aqueous phenol and aniline solutions on activated carbons with different surface chemistry. Colloids and Surfaces A: Physicochemical Engineering Aspects 265, 32–39.

Ljung, L., 1999. Systems Identification: Theory for the Users. Prentice Hall, New York.

Montgomery, D.G., Runger, G.C., Humbele, N.F., 2001. Engineering Statistics. Wiley, New York.

Ratkowsky, D.A., 1983. Nonlinear Regression Modelling: a Unified Practical Approach. Marcel Dekker, New York.

Ratkowsky, D.A., 1990. Handbook of Non-linear Regression Models. Marcel Dekker, New York.

Ruthven, D.M., 1984. Principles of Adsorption and Adsorption Processes. Wiley, New York.

Seidel, A., 1987. Ph.D. Thesis: Adsorptionsgleichgewichte von in Wasser gelösten organischen Stoffen an Aktivkohlen. Dissertation, Zentralinstitut für physikalische Chemie, Akademie der Wissenschaften der DDR, Berlin.

Seidel, A., Tzscheutschler, E., Radeke, K.-H., Gelbin, D., 1985. Adsorption equilibrium of aqueous phenol and indol solutions on activated carbons. Chemical Engineering Science 40, 215–222.

Seidel-Morgenstern, A., 2004. Experimental determination of single solute and Competitive adsorption isotherms. Journal of Chromatography A 1037, 255–272.

Senthilkumaar, S., Kalaamani, P., Subburaam, C., 2006. Liquid phase adsorption of crystal violet onto activated carbons derived from male flowers of coconut tree. Journal of Hazardous Materials 136, 800–808.

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Metabolic Engineering 8 (2006) 447–455

METABOLIC
ENGINEERING

www.elsevier.com/locate/ymben

Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems

M. Joshi[a], A. Seidel-Morgenstern[a,b], A. Kremling[b,*]

[a] *Institut für Verfahrenstechnik, Otto-von-Guericke UniversitätMagdeburg, Universitätsplatz 2, Magdeburg, Germany*
[b] *Max-Planck-Institut für Dynamik komplexer technischer Systeme, Sandtorstr. 1, 39106 Magdeburg, Germany*

Received 22 November 2005; received in revised form 6 February 2006; accepted 3 April 2006
Available online 6 May 2006

Abstract

A quantitative description of dynamical systems requires the estimation of uncertain kinetic parameters and an analysis of their precision. A method frequently used to describe the confidence intervals of estimated parameters is based on the Fisher-Information-Matrix. The application of this traditional method has two important shortcomings: (i) it gives only lower bounds for the variance of a parameter if the solution of the underlying model equations is non-linear in parameters. (ii) The resulting confidence interval is symmetric with respect to the estimated parameter. Here, we show that by applying the bootstrap method a better approximation of (possibly) asymmetric confidence intervals for parameters could be obtained. In contrast to previous applications devoted to non-parametric problems, a dynamical model describing a bio-chemical network is used to evaluate the method.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Parameter confidence interval; Fisher-Information-Matrix; Bootstrap approach

1. Introduction

Measurements of small intermediates of primary metabolism as well as mRNA or proteome data are available for a large number of organisms (e.g., Schaefer et al., 1999; Richmond et al., 1999). To analyze the data and to detect new principles of cellular organization, often detailed mathematical models are set up to describe intracellular processes of interest. Besides the steady-state analysis to calculate the distribution of fluxes in a large cellular network, also dynamical models became very popular (e.g., Kremling et al., 2001). The models are mainly based on balance equations for intracellular components and describe the temporal changes of the intracellular concentrations by o.d.e.'s. The o.d.e.'s sum up all kinetic rates that synthesize or degrade the metabolites. Therefore, kinetic parameters are required that normally cannot be determined in vivo but have to be estimated from experimental data. Corresponding methods are discussed in the literature (e.g., Moles et al., 2003); the proposed algorithms have the ability to cover also large and complex systems. Besides the actual values determined for the different parameters it appears highly desirable to have some information on the confidence intervals of the estimated parameters. However, methods to analyze the accuracy of estimated parameters are scarce. Often, the following approach is used: the variance $\sigma^2_{p_j}$ of parameter p_j is given by

$$\sigma^2_{p_j} = (\mathbf{F}^{-1})_{jj}, \tag{1}$$

where $\mathbf{F}$ is Fisher-Information-Matrix (FIM). The confidence interval of parameter p_j is then given based on the estimate $\hat{p}_j$ of p_j by (Press et al., 2002):

$$\hat{p}_j - \sigma_{p_j} \cdot t^{\nu}_{\alpha/2} < p_j < \hat{p}_j + \sigma_{p_j} \cdot t^{\nu}_{\alpha/2}, \tag{2}$$

where $t^{\nu}_{\alpha/2}$ is given by Student's t-distribution, ν is the degree of freedom and α is the $(1-\alpha)$ 100% confidence interval selected by the user. However, the approach described has two major drawbacks. The given value in Eq. (1) is only a lower bound for the variance and the confidence interval is centered and symmetric to the

---

[*] Corresponding author. Systems Biology Group, Max-Planck-Institut für Dynamik komplexer technischer Systeme, Sandtorstr. 1, 39106 Magdeburg, Germany.

E-mail address: kremling@mpi-magdeburg.mpg.de (A. Kremling).

1096-7176/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymben.2006.04.003

estimated parameter. This is due to a linear approximation of the state variables with respect to the parameters. Nonlinear systems behave in a different way: they show non-normal distributions, often concomitant with a bias.

In the present study, we overcome the mentioned problems by exploiting the so-called bootstrap method (Efron and Tibshirani, 1993; DiCiccio and Efron, 1996). The method surmounts the theoretical limitations by assessing the uncertainties in statistics with data from finite samples. Like a Monte-Carlo method, the bootstrap uses stochastic elements and repeated simulations to analyze the properties of the system under consideration. The application of the method to dynamical systems is very infrequent. Therefore, to demonstrate the potentials, a small bio-chemical network is analyzed that describes the temporal changes of the concentration of a number of components with a set of o.d.e.'s.

Especially in a systems biology approach, the estimation of parameters becomes more and more important to set up "good" models (that is, models that are validated with a number of experiments). For biological systems it was shown that the sensitivity of state variables with respect to kinetic parameters is important for a better understanding of cellular dynamics. Here, we can show that the confidence intervals of the parameters are larger than predicted by traditional methods. Moreover, some of the intervals show asymmetric shapes. These findings allow to analyze the sensitivity of the parameters in a more precise way. Furthermore, also the design of new experiments is often based on directed modifications of the FIM to stimulate the system in such a way that the sensitivities of the parameters are improved. Applying an experimental design, it is shown that the confidence interval calculated with the bootstrap method reduces considerably.

2. Methods

2.1. Parameter estimation and model accuracy

Using a least-square approach, the kinetic parameters of the model should minimize the quadratic error between the simulation of the state variables x_i and the measured data x_i^M for all state variables. As the latter is only available at K discrete time points $\mathcal{T} = \{t_1, t_2, \ldots, t_K\}$, the errors at each measurement time point are summed. The squared error is furthermore normalized by the standard deviation of the corresponding measurement noise σ_{x_i} for every state variable x_i. Thus, less noisy signals are more weighted and all measurements are brought to the same scale. This results in the following objective function:

$$\Theta = \sum_{i=1}^{n} \Theta_i = \sum_{i=1}^{n} \sum_{t_k} \left(\frac{x_i(t_k) - x_i^M(t_k)}{\sigma_{x_i}} \right)^2, \tag{3}$$

where n is the number of state variables and t_k are time points where a sample was taken. To estimate the parameters according to Eq. (3) solver "lsqnonlin", which uses "Levenberg–Marquardt" method for optimization, from the MATLAB environment was chosen.

After estimation of the parameters, a model accuracy evaluation analysis should be performed to ensure that the model describes the experimental data sufficiently. This is done scarcely in a systems biology approach and is based mainly on large and often unknown measurement errors for intracellular components. If available, the standard deviation of the measured data is often based only on a low number of measurements (degree of freedom df_2). Therefore, an F-test for every state variable, considering the ratio of two χ^2 distributed variates, Θ_i with degree of freedom $df_{1i} = K - l_i$ with l_i is the number of parameters influencing state i, and σ_{xi} with degree of freedom df_2, is more appropriate. With

$$F = \frac{\Theta_i / df_{1i}}{\sigma_{xi}} \tag{4}$$

a model is accurate if for all state variables

$$F_{\alpha/2, df_{1i}, df_2} < \Theta_i < F_{1-\alpha/2, df_{1i}, df_2}. \tag{5}$$

2.2. Bootstrap approach

The bootstrap method is a data-based simulation method for statistical inference (Efron and Tibshirani, 1993). A main application of the method is the calculation of confidence intervals for a non-parametric distribution. The method is frequently applied to analyze data in medicine (DiCiccio and Efron, 1996). Here, we are interested in the confidence intervals of the kinetic parameters for dynamical models. To perform the analysis, an initial set of experimental data $\mathbf{S}$ is used as a database. Performing parameter estimation results in a first set of parameters. Due to measurement errors the repetition of the experiment leads to a slightly different set of data $\mathbf{S}_1^*$ and therefore to a different set of estimated parameters. The bootstrap approach now uses a large set of B-times replicated experimental data $\mathbf{S}_1^*, \mathbf{S}_2^*, \mathbf{S}_3^*, \ldots, \mathbf{S}_B^*$ to calculate statistical properties of the resulting distribution of the (re)-estimated set of parameters. Since in reality it is not possible to repeat the experiment a hundred times or more, a Monte-Carlo method is used to simulate the data.

2.2.1. Outliers

Outliers are extreme cases in which one variable, or a combination of variables, has a very strong influence on the calculation of statistics. We identified outliers as described in Montgomery et al. (2001). Quartiles Q_i divide the sorted data set into four equal parts where 25% of the data can be found between Q_1 and Q_2 (representing the median) and 25% of the data can be found between Q_2 and Q_3. The spread sp is defined as $sp = Q_3 - Q_1$. Outliers are defined as such values that are beyond the borders given by $Q_1 - 1.5 \cdot sp$ and $Q_3 + 1.5 \cdot sp$. The outlier procedure described is used to analyze the data of the estimated parameters resulting from the bootstap approach.

2.2.2. *Confidence intervals in the bootstrap framework*

Confidence interval analysis is one of the important statistical test to validate parameter reliability. The goal of bootstrap confidence interval theory is to calculate confidence limits for the parameters p_j from their distribution (number density function). Normally, the distribution is represented by a histogram. The number of bins in the histogram is calculated by the Freedman–Diaconis rule (Freedman and Diaconis, 1981).

The set of replicated experimental data $\mathbf{S}_1^*, \mathbf{S}_2^*, \mathbf{S}_3^*, \ldots, \mathbf{S}_B^*$ is used to generate a set of (re)-estimated parameters $\hat{p}_1^*, \hat{p}_2^*, \hat{p}_3^*, \ldots, \hat{p}_B^*$. The confidence intervals for the parameters are then calculated by the percentile method: let $\hat{p}^{*(\alpha)}$ indicate the $100 \cdot (1-\alpha)$ percentile of B bootstrap replications; then the percentile interval $(\bar{p}_{\mathrm{lo}}, \bar{p}_{\mathrm{up}})$ of intended coverage is obtained by

$$(\bar{p}_{\mathrm{lo}}, \bar{p}_{\mathrm{up}}) = (\hat{p}^{*(\alpha/2)}, \hat{p}^{*(1-\alpha/2)}). \tag{6}$$

The length L and shape sh of the confidence interval $(\bar{p}_{\mathrm{lo}}, \bar{p}_{\mathrm{up}})$ of the distribution are calculated based on the mean value $\bar{p}$ as follows:

$$L = \bar{p}_{\mathrm{up}} - \bar{p}_{\mathrm{lo}}, \tag{7}$$

$$sh = \frac{\bar{p}_{\mathrm{up}} - \bar{p}}{\bar{p} - \bar{p}_{\mathrm{lo}}}. \tag{8}$$

A shape $sh > 1.0$ indicates a greater distance from $\bar{p}_{\mathrm{up}}$ to $\bar{p}$ than from $\bar{p}$ to $\bar{p}_{\mathrm{lo}}$. Note, that, considering a confidence interval in a normal distribution, it is symmetric about $\bar{p}$, and has a shape $sh = 1.0$. If length L is based on the current value of parameter $\bar{p}$, symbol $\%L$ is used:

$$\%L = 100 \cdot \frac{L}{\bar{p}}. \tag{9}$$

2.3. *Fisher-Information-Matrix*

Traditionally, the solution of the non-linear o.d.e. system is linearized for small parameter perturbations to calculate the variances of the parameters with the help of the FIM (Faller et al., 2003). The definition of the FIM is based on parameter sensitivities w_{ij} which describe an infinitesimal change of the state variable x_i according to a change of parameter p_j:

$$w_{ij} = \frac{\partial x_i}{\partial p_j}. \tag{10}$$

Since the state variables are time dependent, the sensitivities w_{ij} are also. The o.d.e.'s for the w_{ij}'s for a model with l parameters of the form

$$\underline{\dot{x}} = \underline{f}\left(\underline{x}, \underline{p}\right) \tag{11}$$

are given by

$$\dot{W} = \frac{\partial \underline{f}}{\partial \underline{x}} \cdot W + \frac{\partial \underline{f}}{\partial \underline{p}} \tag{12}$$

with the matrix W:

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1l} \\ w_{21} & w_{22} & \cdots & \vdots \\ \vdots & & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nl} \end{bmatrix}. \tag{13}$$

The FIM is given now according to the following sum over all times t_k:

$$\mathbf{F} = \sum_{t_k} W^T \cdot C^{-1} \cdot W, \tag{14}$$

with the variance–covariance matrix of the measurements C. The expression for $\mathbf{F}$ appears if one calculates the variance $\sigma_{\hat{p}}^2$ of estimated parameters $\hat{p}$:

$$\sigma_{\hat{p}}^2 = E[(\hat{p} - E[\hat{p}])^2], \tag{15}$$

with $E[\bullet]$ is expectation. The following equation holds true for the variance of a single parameter $\sigma_{\hat{p}_j}^2$ based on the Cramer–Rao inequality (Ljung, 1999):

$$\sigma_{\hat{p}_j}^2 \geqslant (\mathbf{F}^{-1})_{jj}. \tag{16}$$

Note that by using Eq. (16) only a lower bound of the parameter variances can be calculated. This fact is neglected in many publications that use FIM for further parameter analysis and process improvement by experimental design (e.g., Baltes et al., 1994).

2.4. *Case study: generation of "experimental" data*

Simulated data were generated using a dynamical model describing a bio-reactor system. The model comprises the uptake of a carbohydrate S by biomass X, and the conversion of the substrate into intracellular components (Fig. 1). The uptake reaction is catalyzed by enzyme E_1 with product M_1. M_2 is the product of the second reaction, catalyzed by E_2. Enzyme E_3 catalyzes the reaction from M_2 to M_3, and enzyme E_4 catalyzes the degradation of M_3. The corresponding o.d.e.'s for this system read:

$$\begin{aligned} \dot{X} &= (\mu - D) \cdot X, \\ \dot{S} &= D \cdot (S_0 - S) - r_1 \cdot mw \cdot X, \\ \dot{M}_1 &= r_1 - r_2 - \mu M_1, \\ \dot{M}_2 &= r_2 - r_3 - \mu M_2, \\ \dot{M}_3 &= r_3 - r_4 - \mu M_3, \end{aligned} \tag{17}$$

with $D = q_{\mathrm{in}}/V$ is the dilution rate of the reactor, S_0 is the feed concentration, mw is the molecular weight of the substrate to convert from molar to g/L, and μ is the growth rate of the biomass. Equations for the rate laws r_j and the experimental conditions are summarized in Table S1 in the supplement.

The process described is controlled by input feed-rate q_{in} and feed concentration S_0. A first experiment was performed by starting with a high substrate concentration S_0 and given input/ output feed-rate q_{in} to drive the system

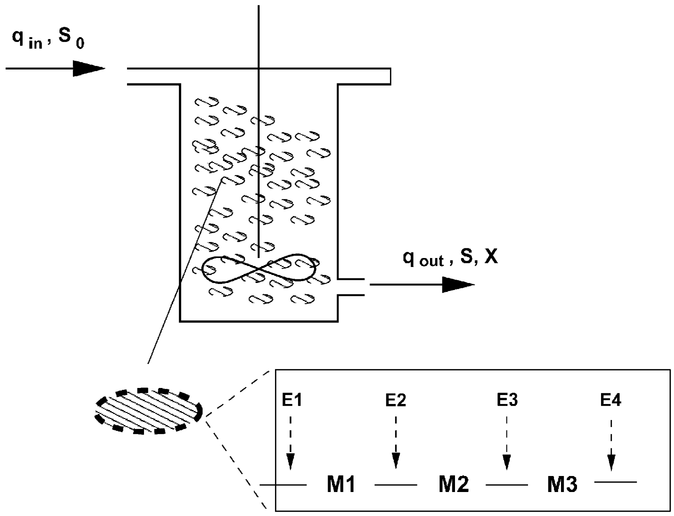


Fig. 1. Structure of the “real” model that is used to generate experimental data. The growth of the biomass depends on the substrate uptake reaction via enzyme E_1.

into a steady state. Samples for substrate, biomass and intracellular components M_1, M_2, and M_3 are taken 10 times per hour. The numerical values of the simulated data are added with random noise (using MATLAB random number generator with a t-distribution, with degree of freedom $df = 10$) within in the bounds given by relative errors (data in Table S1). The experiment was repeated five times and for all state variables the mean value $\bar{m}$ and standard deviation σ for each of the $K = 20$ time points were determined (Table S2 in the supplement).

2.4.1. *Reconstruction of the measured data distribution for the bootstrap approach*

To generate the data set $\mathbf{S}_1^*, \mathbf{S}_2^*, \mathbf{S}_3^*, \ldots, \mathbf{S}_B^*$ the variance of the measured data has to be known. The following approach is used: as described above a mean value $\overline{m_{ik}}$ and standard deviation σ_{ik} are available for every state variable. Especially for measurements of intracellular components in biological system, the type of error is seldom known. Therefore, based on the data given, two sets of bootstrap data are generated: one with a mean absolute error σ_i^{a} for every state variable (mean over all time points) and one with a mean relative error σ_i^{r}.

3. Results

3.1. *Model selection*

In contrast to the model description given above (Eq. (17), kinetic rate laws for r_j and parameter values in Table S1), we mimic a realistic approach by presuming that the knowledge on the system is incomplete: it is only known that the metabolites are converted in a linear chain from M_1 to M_3 and model Eq. (17) is valid. To show the influence of the choice of the kinetic rate laws on the bootstrap results, three different sets of kinetics were analyzed:

$$r_1 = r_{\max 1} \cdot \frac{S}{K_s + S}, \tag{18}$$

$$r_2 = r_{\max 2} \cdot g_1(M_1), \tag{19}$$

$$r_3 = r_{\max 3} \cdot g_2(M_2), \tag{20}$$

$$r_4 = r_{\max 4} \cdot M_3, \tag{21}$$

with

$$\text{Model 1}: \quad \begin{aligned} g_1 &= M_1, \\ g_2 &= \frac{M_2^n}{K_{m2}^n + M_2^n}, \end{aligned} \tag{22}$$

$$\text{Model 2}: \quad \begin{aligned} g_1 &= M_1, \\ g_2 &= M_2^n, \end{aligned} \tag{23}$$

$$\text{Model 3}: \quad \begin{aligned} g_1 &= M_1^m, \\ g_2 &= M_2^n. \end{aligned} \tag{24}$$

Model accuracy tests are performed with measurements as described above; results are summarized for every model variant in Table S3 in the supplement. Model 3 describes the data at best, and Model 1 at worst.

3.2. *Characteristics of parametric histograms*

With the three models, $B = 2000$ bootstrap replications were performed (MATLAB7 was used; each run lasted approx. three days on a workstation with dual processor, 2.2 GHz each (AMD Opteron 248) and with 4 GB RAM). For higher values of B, no changes of the characteristics of the histograms could be observed (see Fig. S1 in the supplement). For the calculation $\alpha = 0.05$ is used, that is, the confidence interval should contain 95% of the data. First, data for the parameters are analyzed with respect to outliers. For further analysis only those bootstrap data sets which does not include any outlier are considered (see Tables S4–S7 in the supplement).

Table 1 summarizes the results of the bootstrap data and compares it with the characteristics obtained with the point estimation by using FIM (see also Tables S4–S7 in the supplement). In general, the confidence intervals obtained from the bootstrap data are larger than those calculated based on FIM. In cases where the sensitivity of a model parameter is very low with respect to the experimental data, the lower bound of the confidence intervals obtained with FIM is negative. This can be seen exemplarily in Fig. 2A with parameter $r_{\max 3}$ in Model 1. Close inspection of the data revealed that only a small part of the kinetics $r_3 = r_3(M_2)$ is covered by the experimental data which result in large confidence intervals. High values of the shape sh indicate non-linear behavior of a system. This can be seen in Fig. 2B with parameter $r_{\max 3}$ in Model 2. The

Table 1
Summary of bootstrap results in comparison with a point estimation for the different models and for the type of errors used for the analysis

Rates/parameters		Model 1				Model 2				Robustness Model 2				Model 3			
		FIM		Bootstrap		FIM		Bootstrap		FIM		Bootstrap		FIM		Bootstrap	
Absolute error																	
r_i	p	$\hat{p}_j$	% L	$\hat{p}_j^{\max}$	% L	$\hat{p}_j$	% L	$p_j^{\max}$	% L	$\hat{p}_j$	% L	$p_j^{\max}$	% L	$\hat{p}_j$	% L	$p_j^{\max}$	% L
r_1	K_s	0.04	120.20	0.04	124.08	0.04	119.53	0.04	128.03	–	–	–	–	0.04	116.61	0.04	126.40
	$r_{\max 1}(10^{-5})$	0.18	8.25	0.18	18.19	0.18	8.17	0.18	18.31	–	–	–	–	0.17	8.04	0.18	18.48
r_2	$r_{\max 2}(10^{-5})$	0.43	17.97	0.42	21.66	0.43	17.97	0.42	22.60	–	–	–	–	0.36	65.73	0.38	68.65
	m	–	–	–	–	–	–	–	–	–	–	–	–	0.83	76.92	0.86	71.53
r_3	$r_{\max 3}(10^{-5})$	3.03	10620.82	0.50	264.56	0.49	66.45	0.45	68.01	–	–	–	–	0.49	66.35	0.45	67.71
	K_{m2}	1.80	4068.37	1.55	133.18	–	–	–	–	–	–	–	–	–	–	–	–
	n	2.96	284.10	2.95	52.34	2.88	42.65	2.83	41.69	–	–	–	–	2.87	42.68	2.83	41.07
r_4	$r_{\max 4}(10^{-5})$	0.49	12.11	0.48	17.24	0.49	12.11	0.48	17.52	–	–	–	–	0.49	12.11	0.48	17.42
Relative error																	
r_1	K_s	0.04	16.24	0.04	24.89	0.04	16.20	0.04	26.31	0.04	17.15	0.04	26.13	0.04	16.17	0.04	25.22
	$r_{\max 1}(10^{-5})$	0.17	2.11	0.18	9.45	0.17	2.09	0.17	9.62	0.18	2.12	0.17	9.18	0.17	2.08	0.17	9.45
r_2	$r_{\max 2}(10^{-5})$	0.51	20.19	0.53	46.82	0.51	20.19	0.54	45.34	0.51	20.60	0.54	46.09	0.37	35.67	0.39	72.48
	m	–	–	–	–	–	–	–	–	–	–	–	–	0.85	20.48	0.83*	37.71
r_3	$r_{\max 3}(10^{-5})$	1.26	3848.12	0.65	205.04	0.50	65.44	0.53	79.28	0.50	61.87	0.51	74.61	0.50	65.42	0.53	83.92
	K_{m2}	1.26	1598.30	0.93	96.27	–	–	–	–	–	–	–	–	–	–	–	–
	n	3.06	256.18	3.15	42.56	2.87	32.59	2.83	33.64	2.80	30.92	2.83	35.56	2.87	32.59	2.83	35.00
r_4	$r_{\max 4}(10^{-5})$	0.50	13.48	0.52	20.74	0.50	13.48	0.51	20.46	0.50	13.56	0.51	18.92	0.50	13.48	0.51	21.09

Values for the confidence region for the point estimation are calculated according Eq. (2); here $2.0 \cdot \sigma_{pj}$ is the 95% confidence interval. The confidence regions for the bootstrap data for the parameters are calculated according to the method described in the text. $p_j^{\max}$ is the value with highest probability.

*There are two maximum probability values, another is 0.89.

maximal difference between the point estimation and the value with highest probability that we have found is nearly 83% based on the value of the point estimation (see Table S4, parameter $r_{\max 3}$, absolute error). In 50% of all cases analyzed the shape *sh* differs significantly from $sh = 1$ indicating the non-linearity of the system with respect to parameters.

A bias "per se" is not sufficient to conclude for non-linearity. Fig. 2C shows the histogram of parameter $r_{\max 1}$ having a symmetric distribution. Although the mean value of the distribution is in close agreement with the point estimation, the standard deviations between both approaches differ. As it can be seen in Fig. 2D, parameter n in Model 2 has similar values for the mean and the standard deviation calculated by FIM.

The parameters in rate law r_1 can be used to check the accuracy of the optimization procedure: kinetics r_1 used to generate the data (see section "Case study") and in the model variants introduced in section "Results/Model selection" are the same. The point estimation as well as the bootstrap method show good agreement; both parameters $r_{\max 1}$ and K_s are estimated with high precision for all runs (Table 1, Table S1).

A comparison of the type of measurement error used revealed that relative errors for the state variables give smaller confidence intervals. For the parameters $r_{\max 1}$ and K_s the differences are very prominent (e.g., Table 1, values $\%L$ for K_s for Model 1 are 124.08 and 24.89), while for the other parameters only minor differences could be detected.

3.2.1. Correlation analysis

After generating the data, the bootstrap approach allows a fast detection of correlations between parameters. Often, using Michaelis–Menten kinetics, experimental data are insufficient to estimate both $r_{\max}$ and K_s with good quality (Baltes et al., 1994). Plotting the results of the bootstrap data of two parameters against each other is a good hint, if a correlation exist. Fig. 3 shows such a plot and compares it with a contour plot of the function $\Delta\Theta = \Delta p^T \mathbf{F} \Delta p$. Function $\Delta\Theta$ represents the change of the system when it is perturbed with small deviations Δp from the original set of parameters and represents at the same time the confidence region for the parameters under investigation. It can be seen that the parameters are only weakly correlated (correlation coefficient $r = 0.644$). Therefore, a simplification of the rate law r_1 is not necessary.

3.3. Robustness analysis

Chemical and bio-chemical processes are subject to different disturbances, e.g., for the dynamical system considered fluctuations in the valves or in the feed preparation may be present. To check the quality of the bootstrap method with respect to disturbances of the inputs, the feed-rate q_{in} was perturbed by addition of noise

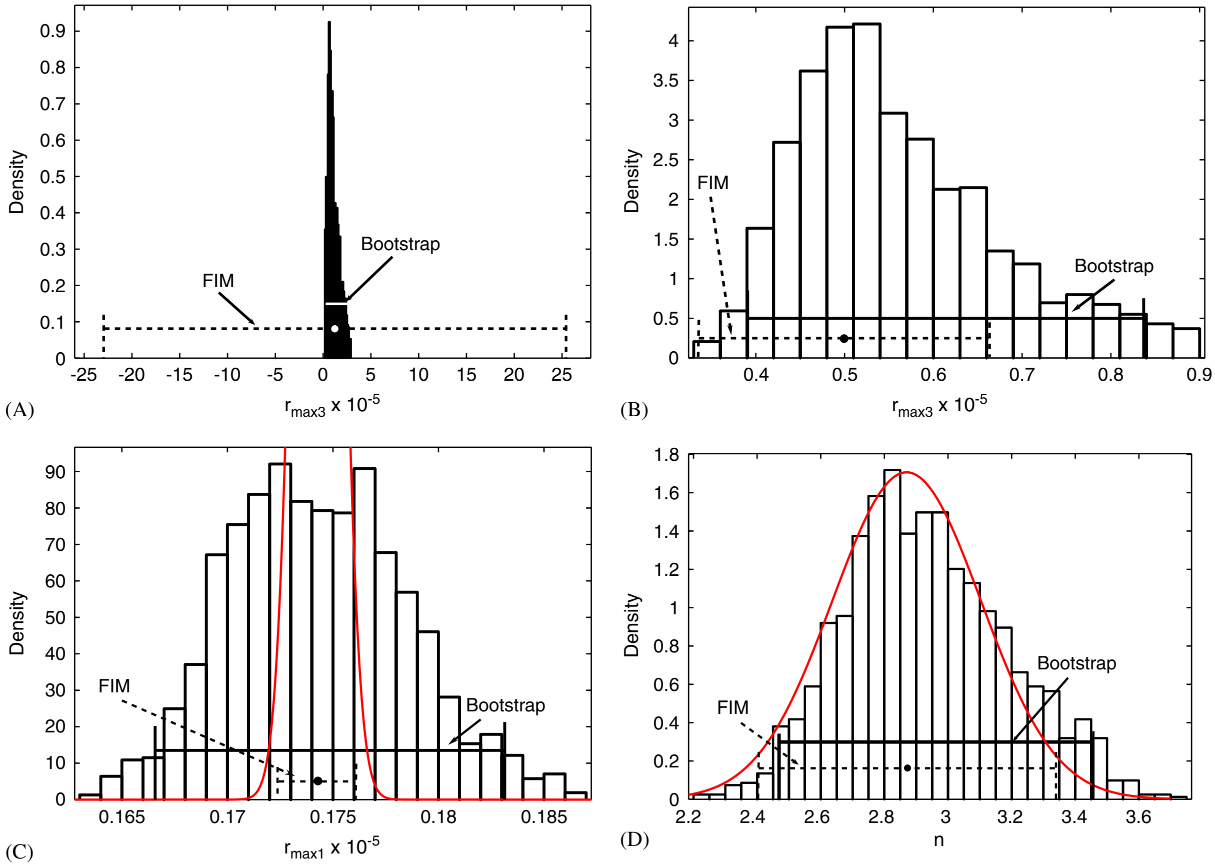


Fig. 2. Comparison of 2000 parametric bootstrap replications for parameter $r_{\max 3}$ (upper left: Model 1, upper right: Model 2), parameter $r_{\max 1}$ (lower left: Model 3), and n (lower right: Model 2). The normal distribution resulting from the point estimations as sketched in the lower plots. Confidence regions for the point estimation and bootstrap method were calculated according to Eqs. (2) and (6), respectively. The normal distribution corresponding to the values of the point estimation is sketched in the lower plots. Confidence regions for the point estimation and bootstrap method are calculated according to Eqs. (2) and (6), respectively.

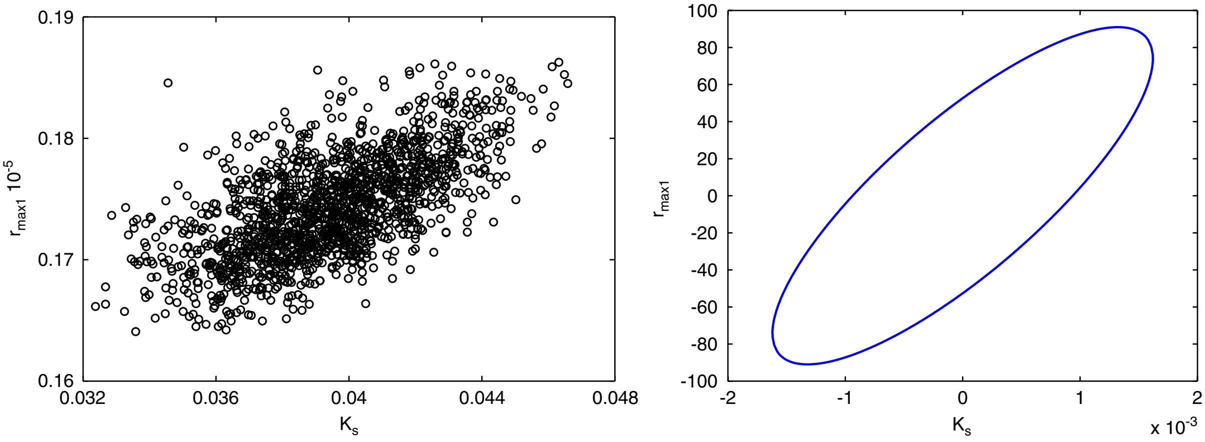


Fig. 3. Left: bootstrap results of parameter $r_{\max 1}$ plotted against the bootstrap results of parameter K_s in Model 1. Right: contour plot of the function $1 = \Delta p^T \mathbf{F} \Delta p$.

(relative error 10%). The bootstrap method was performed using Model 2. The parameter values do not change significantly and the main characteristics of the parameter histograms remain as before (Fig. 4) although slight changes can be observed (Table 1).

3.4. Evaluation of bootstrap method: simulation studies

For systems that are linear in the parameters, confidence regions of the state variables can be calculated based on the confidence intervals of the parameters while in the

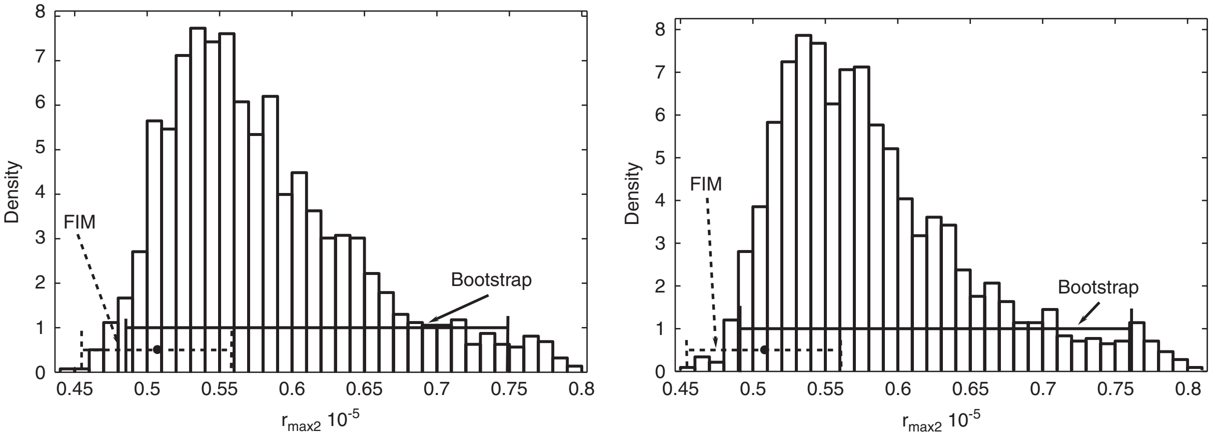


Fig. 4. Comparison of 2000 parametric bootstrap replications for parameter $r_{\max 2}$. No influence of disturbances of the feed rate on the distribution can be detected. Left: distribution with undisturbed feed. Right: distribution with disturbed feed.

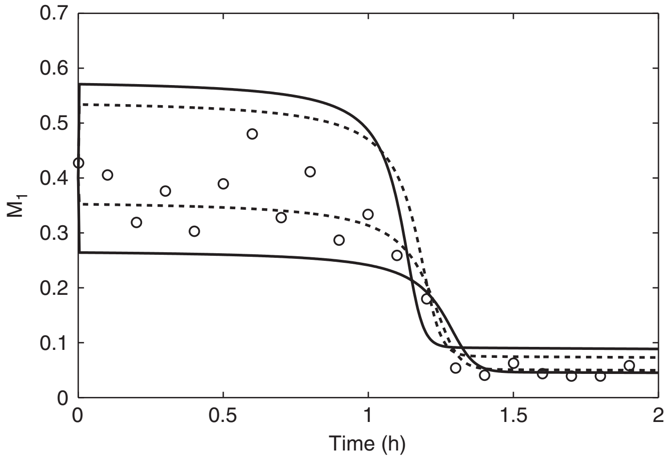


Fig. 5. Simulation of the system with two sets of parameters representing high and low fluxes through the network. With the bootstrap method (solid lines) more data lie in between the confidence region in comparison with FIM (dashed lines).

non-linear case, only simulations can be performed. This is done with upper bounds and lower bounds of the kinetic parameters in reaction rates r_1 and r_2 for Model 3 obtained from FIM and the bootstrap method, respectively. This guarantees low and high fluxes in the system and also large differences in the dynamical behavior of the state variables. Fig. 5 shows the time course of metabolite M_1 for the two conditions. It can be observed that more data points lie in between the interval given by the bootstrap approach than obtained from FIM.

3.5. Experimental design

A further evaluation of the bootstrap method is the comparison of the confidence intervals of the parameters if a standard experiment and an optimized experiment are available. The design of new experiments should allow to stimulate the system in such a way that the estimation of selected parameters becomes easier, that is, the sensitivity of these parameters becomes higher in the new experiment. Here, a new experiment was designed using a conventional approach. The designed input profile was used to generate new "in silico" experimental data. Again, these data were analyzed with the bootstrap method. As an example, rate law r_2 was used to estimate the parameters. With the previous models it turned out that the experimental data available are not sufficient to estimate the maximal rate of the enzyme, $r_{\max}$ (therefore no Michaelis–Menten kinetics are used in the models given so far). However, the knowledge on $r_{\max}$ may offer new insights in the overall function of the network. To incorporate this, the kinetics for r_2 was assumed

$$r_2 = r_{\max 2} \cdot \frac{M_1}{K_{m1} + M_1} \tag{25}$$

in Model 1 in contrast to Eq. (19). The new experiment was designed by optimizing the determinant of the FIM (D-optimality). As system theoretical input, the concentration of the first enzyme, represented by parameter $r_{\max 1}$, was used. A step change every hour for an overall time period of 6 h was allowed. The choice of parameter $r_{\max 1}$ is based on the fact that it is not possible to redirect the system in reasonable steady states alone with standard inputs q_{in} and S_0 (data not shown).

The optimization problem is formulated as

$$\min \det(\mathbf{F}^*), \tag{26}$$

where $\mathbf{F}^*$ is the 2×2 FIM which is obtained when only parameters $r_{\max 2}$ and K_{m1} are considered. Fig. 6 shows the optimized input profile for $r_{\max 1}$ and the improvement of the objective function by considering the function $\Delta\Theta = \Delta p^T \cdot \mathbf{F}^* \cdot \Delta p$. Fig. 7 shows a comparison of the bootstrap results of the two experiments. Since the intervals differ considerably, it is not possible to show both histograms in one plot.

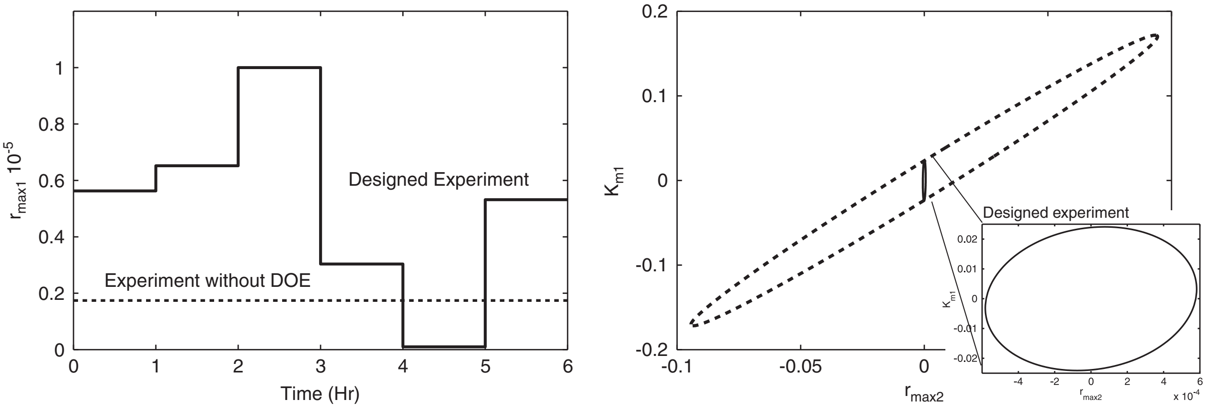


Fig. 6. Left: initial and optimized input profiles for $r_{\max 1}$. Right: contour plots of $1 = \Delta p^T \mathbf{F} \Delta p$ for the old and the newly designed experiment.

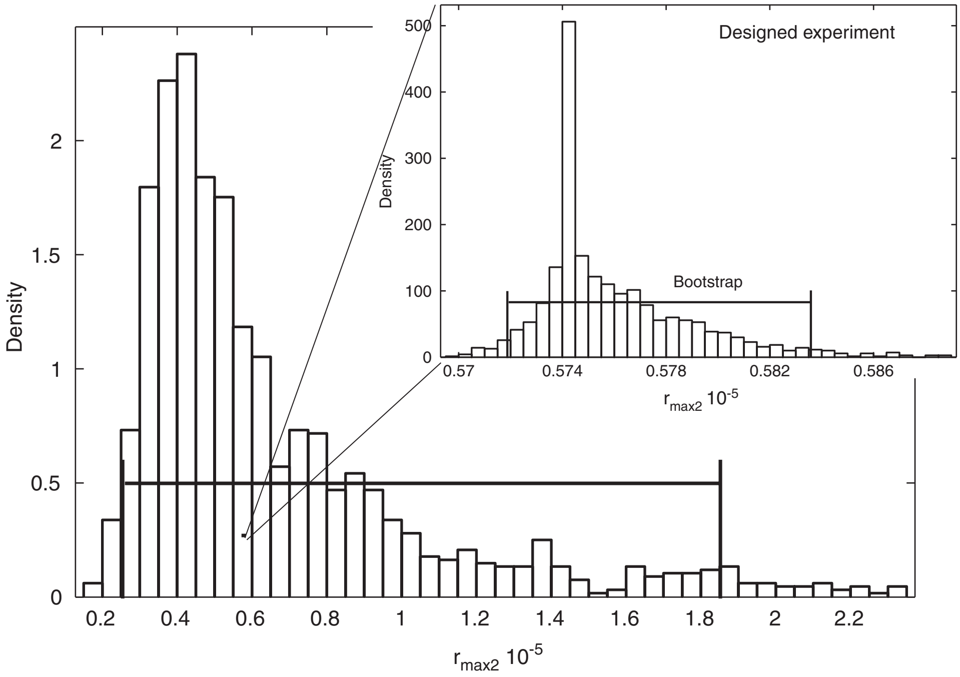


Fig. 7. Comparison of the old and the newly designed experiment to improve the sensitivity for $r_{\max 2}$. The histograms are performed with $B = 2000$ runs.

4. Discussion

In the present study the standard FIM method and the bootstrap method are compared to calculate parameter accuracy. The FIM is based on the calculation of parameter sensitivities and is used frequently in the literature to calculate confidence intervals of parameters in dynamical systems (Baltes et al., 1994; Asprey and Macchietto, 2000; Chen and Asprey, 2003; Faller et al., 2003; Zak et al., 2003). Hereby, the solution of the differential equation system is linearized with respect to the parameters. In contrast, the bootstrap method uses statistical methods based on data generated by Monte-Carlo simulations. For the case study presented, $B = 2000$ bootstrap replications were used. For problems with a higher number of equations and parameters this number might not be sufficient. The final number of runs needed to estimate the characteristics of the distribution properly may also depend on the quality of the parameters. In general, we suggest to choose the number of bins proportional to the estimated standard deviation from the point estimation. Using a simple bio-chemical network with a number of parameters, the non-linearity of the system is reflected by a non-normal distribution of the re-estimated parameters. Besides the choice of the kinetics, the type of error and the specific stimulation of the system show an influence on the size of the confidence interval and on the shape of the parameter distribution.

The comparison of the two approaches reveals that with the standard FIM method the confidence intervals of the parameters are underestimated. The size of the confidence interval, calculated by the bootstrap method and with the FIM method, differs up to a factor of 4 in this study. This is due to the non-linearity of the model with respect to the parameters. Although the data were generated by using mainly Michaelis–Menten type kinetics, the quality of the fit differs for each specific model. The best fit is reached with Model 3, describing the data only with approximated kinetics. However, comparing the quality of the parameters, the confidence intervals for the parameters in rate law r_2 are broader in Model 3 than in Model 2.

Comparison of the type of error reveals that using absolute errors for the state variables, large differences are detected for parameters $r_{\max 1}$ and K_s in rate law r_1. This might be due to the fact that for estimation of the parameters for r_1 the substrate concentration S is used while for the other kinetics the intracellular metabolites M_1, M_2, and M_3 are involved. In the experiment, S starts with a relative high value and ends at a low steady-state value. In our approach, the variances of the state variables are calculated by taking a mean over all data points. In case of absolute error this results in a large deviations for the substrate concentration, while for the intracellular metabolites the deviations are smaller (see Table 1).

In Fig. 5 the application of the results of the parameter confidence intervals is shown for a selected model variant. A confidence region for the state variable M_1 is predicted by simulating extreme cases. This is done by combining low and high fluxes through the network. With the bootstrap method approximately 90% of the data points are inside the confidence region while with the FIM method only 50% are inside. For a linear system, 95% of the data points are expected to be inside the region. With the bootstrap method this limit is approached very well. However, due to the complexity of large bio-chemical networks a general statement regarding the confidence regions for the state variables cannot be given.

Experimental design allows to redirect a system in such a way that hitherto insensitive parameters become sensitive. Figs. 6 and 7 show an example using a Michaelis–Menten type kinetics for r_2. With the initial experiment, it was not possible to estimate the parameters of r_2 in a reliable manner. Since the direct application of the bootstrap method to design new experiments is not reasonable due to computational burden (one optimization run need approx. 200 function evaluations, i.e., 200×3 days) a conventional design based on the minimization of the determinant of FIM was performed. This leads to a new input profile to efficiently stimulate the system. From the design with FIM, a reduction of the standard deviations for the parameters with factors approx. 120 and approx. 7, for parameters $r_{\max 2}$ and K_{M1}, respectively, is expected. Using the designed input profile, with the bootstrap method, as shown in Fig. 7, a factor of approx. 100 for parameter $r_{\max 2}$ is reached.

From our studies we conclude that the proposed bootstrap method is a valuable tool to determine confidence intervals for systems that are non-linear with respect to the parameters. It can be expected that a calculation of the confidence regions of the state variables becomes more precise if the parameter confidence intervals are determined more accurately. This will be an important step towards meaningful mathematical models, that is, models with high predictive power.

Acknowledgment

The financial support by the DFG (FOR 447) is gratefully acknowledged.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at 10.1016/j.ymben.2006.04.003.

References

Asprey, S.P., Macchietto, S., 2000. Statistical tools for optimal dynamic model building. Comput. Chem. Eng. 24 (2–7), 1261–1267.

Baltes, M., Schneider, R., Sturm, C., Reuss, M., 1994. Optimal experimental design for parameter estimation in unstructured growth models. Biotechnol. Prog. 10, 480–488.

Chen, B.H., Asprey, S.P., 2003. On the design of optimally informative dynamic experiments for model discrimination in multiresponse nonlinear situations. Ind. Eng. Chem. Res. 42 (7), 1379–1390.

DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals. Stat. Sci. 11(3).

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, London.

Faller, D., Klingmüller, U., Timmer, J., 2003. Simulation methods for optimal experimental design in systems biology. Simulation 79 (12), 717–725.

Freedman, D., Diaconis, P., 1981. On the histogram as a density estimator. Z. Wahrscheinlichkeitstheorie verw. Geb. 57, 453–476.

Kremling, A., Bettenbrock, K., Laube, B., Jahreis, K., Lengeler, J.W., Gilles, E.D., 2001. The organization of metabolic reaction networks: III. Application for diauxic growth on glucose and lactose. Metab. Eng. 3 (4), 362–379.

Ljung, L., 1999. System Identification—Theory for the User. Second ed., Prentice-Hall PTR, Upper Saddle River, NJ.

Moles, C.G., Mendes, P., Banga, J.R., 2003. Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Res. 13 (11), 2467–2474.

Montgomery, D.C., Runger, G.C., Hubele, N.F., 2001. Engineering Statistics. Wiley, New York.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 2002. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge.

Richmond, C., Glasner, J.D., Mau, R., Jin, H., Blattner, F.R., 1999. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acid Res. 27, 3821–3835.

Schaefer, U., Boos, W., Takors, R., Weuster-Botz, D., 1999. Automated sampling device for monitoring intracellular metabolite dynamics. Anal. Biochem. 270, 88–96.

Zak, D.E., Gonye, G.E., Schwaber, J.S., Doyle, F.J., 2003. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network. Genome Res. 13 (11), 2396–2405.

BIOTECHNOLOGY and BIOENGINEERING

Comment on Mathematical Models Which Describe Transcription and Calculate the Relationship Between mRNA and Protein Expression Ratio

A. Kremling
Max-Planck-Institute for Dynamics of Complex Technical Systems,
Sandtorstr. 1, 39106 Magdeburg, Federal Republic of Germany;
telephone: +49-0391-6110-466; fax: +49-0391-6110-526;
e-mail: kremling@mpi-magdeburg.mpg.de

Received 19 August 2005; accepted 1 June 2006

Published online 20 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bit.21065

ABSTRACT: Mathematical models to describe transcription (Arnold et al. (2001); Biotech Bioeng 72:548–561) and translation (Mehra et al. (2003); Biotech Bioeng 84:822–841) in bacteria are modified in order to improve reaction kinetics and to include the number of polymerase molecules that are active on the DNA, as well as to include the number of ribosomes that are active on the nascent and on the completed mRNA, respectively.

Biotechnol. Bioeng. 2007;96: 815–819.

© 2006 Wiley Periodicals, Inc.

KEYWORDS: transcription; translation; mathematical modeling

Introduction

Using detailed mathematical models to describe cellular processes has become very popular in recent years. This is based mainly on two reasons: (i) the availability of measurements of intracellular components and (ii) computational efforts to set up detailed and comprehensive mathematical models with the aid of software tools. However, the formulation of models is often based on a number of assumptions to reduce the complexity of the system at hand. As will be shown in the following with two examples describing the process of gene expression in procaryotic cells this often leads to over-simplification and basic conservation equations are thus violated.

In Arnold et al. (2001) a scheme to describe mRNA synthesis is presented and a reaction kinetic is derived. Figure A1 therein shows that genetic information, represented by D, is not subdivided into a control sequence and a structural genetic information. Therefore, component D is available only after termination of transcription. This leads to the fact, that in Equation (7) in Arnold et al. (2001), c_D appears as a substrate in the Michaelis–Menten kinetics, that is, a high concentration of promoter leads to a constant transcription rate, while the concentration of RNA polymerase appears in $v_{\max}$, that is, transcription rate is proportional to the number of RNA polymerase molecules. In contrast, it is shown below that the reaction kinetics can be interpreted as Michaelis–Menten kinetics, where the transcription rate is proportional to the number of promoters, and where RNA polymerase represents the substrate that is converted from a free form to an active form.

In Mehra et al. (2003) a mathematical model was introduced to describe bacterial gene expression. The model is used to elucidate the relationship between mRNA and protein expression ratio under different conditions. The basic reaction equations of the approach can be summarized as follows:

$$\begin{aligned} \mathrm{M}_i + \mathrm{R} &\underset{k_{-1,i}}{\overset{k_{1,i}}{\Leftrightarrow}} \mathrm{RM}_i \\ \mathrm{RM}_i &\xrightarrow{k_{2,i}} \mathrm{AR}_i \\ \mathrm{AR}_i &\xrightarrow{k_{3,i}} \mathrm{P}_i \end{aligned} \tag{1}$$

where M_i is the free ribosome binding site of mRNA i, R is the freely available ribosomes, RM_i is the ribosome

WILEY InterScience® DISCOVER SOMETHING GREAT

Correspondence to: A. Kremling

© 2006 Wiley Periodicals, Inc.

binding site that is occupied with a ribosome, AR_i is the concentration of the ribosome in the polysome, and P_i is the protein under consideration. For analysis, the overall conservation equation for the ribosomes is built by adding up RM_i and AR_i over all genes and adding the number of free ribosomes:

$$R_T = R + \sum^{n} RM_i + \sum^{n} AR_i \tag{2}$$

Analyzing the network given above, it can be seen, that for one gene, the number of active ribosomes are pooled in state AR_i. The concentration of AR_i is then calculated for the steady-state condition in dependence of parameters k_{2i} and k_{3i}. Since the stoichiometry of the given reaction scheme does not provide information on the movement of a RNA polymerase molecule on the mRNA, the parameters have to be adjusted in such a way, that the number of active ribosomes is estimated correctly. In Drew (2001) a Markov model was developed to describe protein synthesis for bacterial systems. As in the two contributions previously discussed, here, the DNA is also not subdivided into a control sequence and a structural sequence. Furthermore, ribosomes are considered to bind only to completed mRNA molecules. The model of Drew (2001) is extended in Heyd and Drew (2003) which describes the process of elongation in more detail.

In the following, a modification of the models is given to describe the distribution of the RNA polymerase on the DNA, and to include the distribution of the ribosomes on both the nascent and completed mRNA.

Part of these considerations are described in a PhD thesis (Kremling, 2002).

Results

A starting point for the model description is a situation provided in Figure 1 illustrating transcription and translation of a single gene. Here, RNA polymerase has started to move along the DNA strand. Depending on the size of the nascent mRNA, different numbers of ribosomes are bound to the individual chains. This situation represents a steady-state situation since the release of the polymerase at the end of the chain is accompanied with the binding of a new polymerase at the promoter binding site. Note, that a number of different reaction mechanisms are conceivable describing this situation. Although the velocities of transcription and translation processes are correlated considering an average gene, the set-up of a detailed step-by-step mechanism for a single gene seems to be difficult. However, a detailed model for translation is provided in a recent article by Mehra and Hatzimanikatis (2006). In the following, the scheme in Figure 1 is considered, since it represents the situation when RNA polymerase moves along the DNA template, independent from previous events. Reaction steps are defined that are necessary to describe the situation shown in the figure, comprising the distribution of the RNA polymerase on the DNA and comprising the distribution of ribosomes on the mRNA.

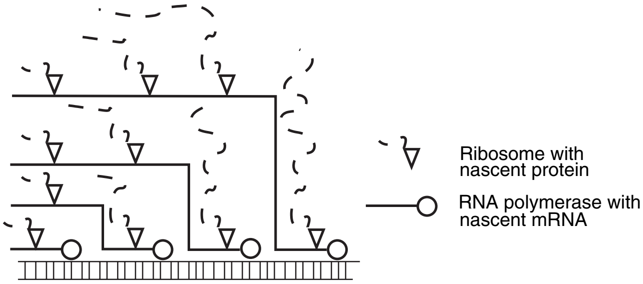

Ribosome with nascent protein
RNA polymerase with nascent mRNA

Figure 1. Scheme of the considered steady-state-situation of the transcription/translation process. The nascent mRNA is fixed by the RNA polymerase moving along the DNA strand. Depending on the length of the nascent mRNA, ribosomes can bind and move along the template.

Transcription

A single gene with length l is considered. RNA polymerase P with σ factor binds to the specific DNA binding site D. After binding, the polymerase clears the promoter (parameter k_{ctr}) and moves along the DNA (parameter k_{tr}; Eqs. (3) and (4)). Complexes Y and Y^i describe the moving polymerase. Binding of nucleotides enlarges the chain. The velocity of reaction Equations (3) and (4) are assumed to not depend on the concentration of the nucleotides. The completed RNA molecule is subject to degradation (parameter k_z):

$$P + D \overset{K_{Tr}}{\Leftrightarrow} PD \tag{3}$$

$$PD \xrightarrow{k_{ctr}} Y + D + \sigma \tag{4}$$

$$Y + Nu \xrightarrow{k_{tr}} Y^1 \tag{5}$$

$$Y^1 + Nu \xrightarrow{k_{tr}} Y^2 \tag{6}$$

$$\vdots$$

$$Y^{l-1} + Nu \xrightarrow{k_{tr}} P + RNA \xrightarrow{k_z} \text{degradation} \tag{7}$$

Using rapid equilibrium assumption for the reaction Equation (3), and steady-state assumptions for all complexes

DOI 10.1002/bit

Y^i one gets:

$$c_{PD} = \frac{c_P}{K_{\mathrm{Tr}} + c_P} c_{D0} \tag{8}$$

$$c_Y = k_{\mathrm{ctr}} \frac{c_{PD}}{k_{\mathrm{tr}}} \tag{9}$$

$$c_{Y^{l-1}} = c_{Y^{l-2}} = \cdots = c_Y \tag{10}$$

The rate of transcription r_{tr} is

$$r_{\mathrm{tr}} = k_{\mathrm{tr}} c_{Y^{l-1}} = k_{\mathrm{tr}} c_Y = k_{\mathrm{ctr}} \frac{c_P}{K_{\mathrm{Tr}} + c_P} c_{D0} \tag{11}$$

and the o.d.e. for mRNA is:

$$\dot{c}_{\mathrm{RNA}} = k_{\mathrm{ctr}} \frac{c_P}{K_{\mathrm{Tr}} + c_P} c_{D0} - k_z c_{\mathrm{RNA}} \tag{12}$$

RNA polymerase is distributed over all Y and Y^i complexes and the number of active RNA polymerase molecules can be calculated by:

$$c_{\mathrm{P\,activ}} = l \quad c_Y = l \frac{k_{\mathrm{ctr}}}{k_{\mathrm{tr}}} c_{PD} \tag{13}$$

Note, that the model does not account for the size of the polymerase and the number of nucleotides that are occupied with one polymerase molecule (in the reaction equation given, the polymerase occupies only one single nucleotide). However, since an undisturbed process is considered, the choice of the discretization of the DNA, here, one nucleotide, has no influence on the number of active polymerases. Considering, for example, a smaller number of stages on the gene, the velocity k_{tr} will decrease with the same ratio, that is, the ratio l/k_{tr} will be constant. The ratio l/k_{tr} represents the time the polymerase needs from the start to the end of the gene.

Translation

To set-up the scheme given in Figure 1, every Y^i complex (the moving RNA polymerase molecule) represents a starting point for translation. Free ribosome R binds to the (free) ribosome binding site $\mathrm{Y}^{i'}$. For the overall binding sites Y^i, the following conservation equation is valid:

$$\mathrm{Y}^i = \mathrm{Y}^{i'} + \mathrm{RY}^i \tag{14}$$

The chain is considered to grow maximal to length s with $s = i/3$ (in the following, the ratio $m/1$ instead of 1/3 is used). For elongation, loaded tRNA* is needed. Load of tRNA with amino acids is not included. However, the model can easily be extended in this direction. Components X^i and X^i_j describe the moving ribosome on the available nascent mRNA:

$$\mathrm{R} + \mathrm{Y}^{i'} \overset{K_{\mathrm{Tl}}}{\Leftrightarrow} \mathrm{RY}^i \tag{15}$$

$$\mathrm{RY}^i \xrightarrow{k_{\mathrm{ctl}}} \mathrm{X}^i + \mathrm{Y}^{i'} \tag{16}$$

$$\mathrm{X}^i + \mathrm{tRNA}^* \xrightarrow{k_{\mathrm{tl}}} \mathrm{tRNA} + \mathrm{X}^i_1 \tag{17}$$

$$\mathrm{X}^i_1 + \mathrm{tRNA}^* \xrightarrow{k_{\mathrm{tl}}} \mathrm{tRNA} + \mathrm{X}^i_2 \tag{18}$$

$$\vdots$$

$$\mathrm{X}^i_{s-1} + \mathrm{tRNA}^* \xrightarrow{k_{\mathrm{tl}}} \mathrm{tRNA} + \mathrm{X}^i_s \tag{19}$$

Parameter k_{ctl} describes the clearance of the ribosome-binding site and parameter k_{tl} describes the velocity of the moving ribosome (the reaction velocity does not depend on the concentration of the loaded tRNA).

RNA represents the completed mRNA molecule, free from DNA template and RNA polymerase. It can also be translated, but is subject to degradation. Degradation of the nascent mRNA is not considered. If one assumes that the nascent mRNA decays with the same time constant k_z as the completed mRNA, this has only minor effect on the steady-state concentration since the velocity of the RNA polymerase (parameter k_{tr}) is much greater than k_z. The process of ribosome binding is similar to the process described above. RNA′ represents a molecule with a free ribosome binding site. X and X_{j} describe the moving ribosome on the completed RNA. The protein is first synthesized, when the mRNA is complete (Eq. (24)):

$$\mathrm{R} + \mathrm{RNA}' \overset{K_{\mathrm{Tl}}}{\Leftrightarrow} \mathrm{RRNA} \tag{20}$$

$$\mathrm{RRNA} \xrightarrow{k_{\mathrm{ctl}}} \mathrm{X} + \mathrm{RNA}' \tag{21}$$

$$\mathrm{X} + t\mathrm{RNA}^* \xrightarrow{k_{tl}} \mathrm{tRNA} + \mathrm{X}_1 \tag{22}$$

$$\mathrm{X}_1 + \mathrm{tRNA}^* \xrightarrow{k_{tl}} \mathrm{tRNA} + \mathrm{X}_2 \tag{23}$$

$$\vdots$$

$$\mathrm{X}_{m-1} + \mathrm{tRNA}^* \xrightarrow{k_{\mathrm{tl}}} \mathrm{R} + \mathrm{protein} \xrightarrow{k_d} \mathrm{degradation} \tag{24}$$

As above, using the rapid equilibrium assumption for the binding of the ribosome and steady-state assumptions for the complexes X, the following equations will hold true:

$$c_{\mathrm{RY}^i} = \frac{c_\mathrm{R}}{K_\mathrm{Tl} + c_\mathrm{R}} c_{\mathrm{Y}^i} \tag{25}$$

$$c_\mathrm{RRNA} = \frac{c_\mathrm{R}}{K_\mathrm{Tl} + c_\mathrm{R}} c_\mathrm{RNA} \tag{26}$$

$$c_{\mathrm{X}^i} = c_{\mathrm{X}^i_j} = k_\mathrm{ctl} \frac{c_{\mathrm{RY}^i}}{k_\mathrm{tl}} \tag{27}$$

$$c_\mathrm{X} = c_{\mathrm{X}_j} = k_\mathrm{ctl} \frac{c_\mathrm{RRNA}}{k_\mathrm{tl}} \tag{28}$$

The rate of translation r_tl is:

$$r_\mathrm{tl} = k_\mathrm{tl} \quad c_\mathrm{X} = k_\mathrm{ctl} \frac{c_\mathrm{R}}{K_\mathrm{Tl} + c_\mathrm{R}} c_\mathrm{RNA} \tag{29}$$

and the o.d.e. for the protein is:

$$\dot{c}_P = k_\mathrm{ctl} \frac{c_\mathrm{R}}{K_\mathrm{Tl} + c_\mathrm{R}} c_\mathrm{RNA} - k_d \, c_P \tag{30}$$

To calculate the number of active ribosomes, two parts have to be considered. The first part considers the molecules on the single Y_i complex while the second part considers the molecules on the finished mRNA.

$$c_{R^i_\mathrm{active}} = s \quad c_{\mathrm{X}^i} = i \frac{m}{l} \frac{k_\mathrm{ctl} c_{\mathrm{RY}^i}}{k_\mathrm{tl}} \tag{31}$$

For all Y_i one gets:

$$\begin{aligned} \sum_i c_{R^i_\mathrm{active}} &= \frac{m}{l} \frac{k_\mathrm{ctl}}{k_\mathrm{tl}} \sum^i i c_{\mathrm{RY}^i} \\ &= m \frac{(l-1)}{2} \frac{k_\mathrm{ctl}}{k_\mathrm{tl}} \frac{c_\mathrm{R}}{K_\mathrm{Tl} + c_\mathrm{R}} c_\mathrm{Y} \end{aligned} \tag{32}$$

And the second part gives:

$$c_{R^\mathrm{RNA}_\mathrm{active}} = m c_\mathrm{X} = m \frac{k_\mathrm{ctl}}{k_\mathrm{tl}} c_\mathrm{RRNA} \tag{33}$$

For one single gene we need the following number of ribosomes:

$$\begin{aligned} c_{R_\mathrm{active}} &= m \tfrac{k_\mathrm{ctl}}{k_\mathrm{tl}} \tfrac{c_\mathrm{R}}{K_\mathrm{Tl} + c_\mathrm{R}} \left(\tfrac{l-1}{2} c_\mathrm{Y} + c_\mathrm{RNA} \right) \\ &= m \tfrac{k_\mathrm{ctl}}{k_\mathrm{tl}} \tfrac{c_\mathrm{R}}{K_\mathrm{Tl} + c_\mathrm{R}} \left(\tfrac{l-1}{2} \tfrac{k_\mathrm{ctr}}{k_\mathrm{tr}} \tfrac{c_P}{K_\mathrm{Tr} + c_\mathrm{P}} c_{D_0} + c_\mathrm{RNA} \right) \end{aligned} \tag{34}$$

If one considers mRNA and protein to be in steady state, the following relationship will hold true for the ratio of protein in a perturbed state to the reference state (index o):

$$f_P = \frac{c_P}{c_{P^0}} = \frac{k_\mathrm{ctl}}{k_{\mathrm{ctl}^0}} \frac{c_\mathrm{R} + K^0_\mathrm{Tl}}{c_\mathrm{R} + K_\mathrm{Tl}} \frac{k^0_d + \mu}{k_d + \mu} f_R \tag{35}$$

where μ is specific growth rate and f_R is the ratio of mRNA concentration for perturbed and reference state.

The concentration of free ribosomes c_R has to fulfill the following algebraic equation, if we consider n genes and a total concentration of ribosomes c_R0:

$$\begin{aligned} c_\mathrm{R0} = c_\mathrm{R} + \sum_{k=1}^{n} m_k \frac{k_{\mathrm{ctl}_k}}{k_{\mathrm{tl}_k}} \\ \cdot \frac{c_\mathrm{R}}{K_{\mathrm{Tl}_k} + c_\mathrm{R}} \left(\frac{l_k - 1}{2} \frac{k_{\mathrm{ctr}_k}}{k_{\mathrm{tr}_k}} \frac{c_P}{K_{\mathrm{Tr}_k} + c_P} c_{D0_k} + c_{\mathrm{RNA}_k} \right) \end{aligned} \tag{36}$$

Conclusion

In contrast to previous models for transcription and for translation (Arnold et al., 2001; Drew, 2001; Mehra et al., 2003), here a model taking into account the number of active RNA polymerase molecules, and the number of ribosomes on the nascent and completed mRNA is introduced. All models introduced so far are based on the same set of assumptions, given in Mehra et al. (2003). The model introduced here, takes into consideration, that for cellular polymerization processes the number of active catalysts like RNA polymerases and ribosomes has to be calculated very carefully. Although steric hindrance of neither a RNA polymerase molecule nor a ribosome molecule is considered in the proposed model, the volume of the molecules can be taken into account by dividing, for example, the DNA in l^* segments, where l^* represents the volume of the RNA polymerase. The proposed model will allow one to estimate the parameters for the processes in a better manner since the kinetic parameters have a clear interpretation.

Equation (34) which calculates the number of active ribosomes needed to produce a protein can be reformulated for steady-state using Equations (12) and (30). The active number of ribosomes depend on the number n_P of proteins, the number m of amino acids aggregated into the protein and an overall constant k_d describing protein degradation and dilution by growth. The equation reads:

$$n_{R_\mathrm{activ}} = \frac{m}{k_\mathrm{tl}} n_P k_d \left(\frac{3m}{2} \frac{k_z}{k_\mathrm{tr}} + 1 \right) \tag{37}$$

Further parameters k_tr, k_tl, k_z do not depend much on growth conditions and can be taken as constant values. The calculation of the number of ribosomes along with this rule of thumb is simulated in Figure 2. The Figure shows the number of ribosomes that are on the completed mRNA molecules and on nascent mRNA molecules. As can be seen, the number of molecules on nascent mRNA is approximately 20% of the overall number of active ribosomes. As shown on the right part of the Figure, with an increasing size

DOI 10.1002/bit

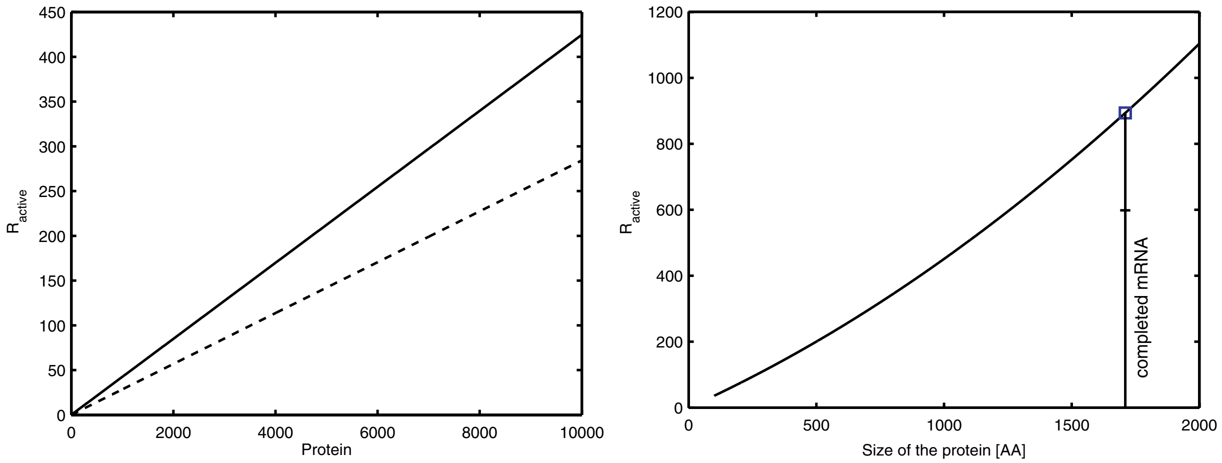


Figure 2. Simulation of the number of ribosomes active during protein synthesis. Left: Dependence from the number of proteins. The solid lines represent the overall number of active ribosomes while the dashed line represents the active ribosomes on the completed mRNA molecules. For the simulation no active degradation is taken into account, therefore k_d is represented by the growth rate $\mu = 1.944 \times 10^{-4}$ (s^{-1}) = 0.7 (h^{-1}). The length $m = 1{,}000$ (AA) is chosen for an average protein. Right: Dependence from the number of amino acids (AA) that build the protein. The symbol represents the LacZ protein. Considering 3,063 bp on the DNA and a factor of 1.66 for the number of DNA templates which are present for a growth rate $\mu = 1.944 \times 10^{-4}\ s^{-1} = 0.7\ h^{-1}$ the number of AA is 1,710. The calculation results in ca. 2,400 monomers, representing 6,000 tetramers in the cell. Values of the parameters are as follows: $k_{tr} = 40$ (Nu) s^{-1}, $k_{tl} = 11.7$(AA) s^{-1}, $k_z = 7.7 \times 10^{-3}\ s^{-1} = 27.72\ h^{-1}$ (Bremer and Dennis, 1987; Kennell and Riezman, 1977).

of the protein and therefore an increasing number of amino acids (AA) that have to be incorporated, the fraction of active ribosomes on the nascent mRNA increases. The symbol in the Figure represents the LacZ protein. Here, about one third of the active ribosomes are on the nascent mRNA.

The set-up of mathematical models is crucial if polymerization processes in cellular systems are being considered. They differ from models used in chemical engineering because the catalyst is bound to a component with limited size, that is, the length of the gene and/or the transcript has strong influences on the kinetics.

References

Arnold S, Siemann M, Scharnweber K, Werner M, Baumann S, Reuss M. 2001. Kinetic modeling and simulation of in vitro transcription by phage t7 RNA polymerase. Biotech Bioeng 72:548–561.

Bremer H, Dennis PP. 1987. Modulation of chemical composition and other parameters of the cell by growth rate. In: Neidhardt FC, (Editor in Chief), editor. Escherichia coli and Salmonella typhimurium. Washington, DC: ASM press. p 1527–1542.

Drew DA. 2001. A mathematical model for prokaryotic protein synthesis. Bull Math Biol 63:329–351.

Heyd A, Drew DA. 2003. A mathematical model for elongation of a peptide chain. Bull Math Biol 65:1095–1109.

Kennell D, Riezman H. 1977. Transcription and translation initiation frequencies of the Escherichia coli lac operon. J Mol Biol 114:1–21.

Kremling A. 2002. Strukturierung zellulaerer Funktionseinheiten—ein signalorientierter Modellierungsansatz fuer zellulaere Systeme am Beispiel von Escherichia coli. PhD thesis, Universitaet Stuttgart.

Mehra A, Hatzimanikatis V. 2006. An algorithmic framework for genome-wide modeling and analysis of translation networks. Biophys J 90:1136–1146.

Mehra A, Lee KH, Hatzimanikatis V. 2003. Insights into the relation between mRNA and protein expression patterns: I. Theoretical considerations. Biotech Bioeng 84:822–841.

Available online at www.sciencedirect.com

ScienceDirect

Journal of Biotechnology 129 (2007) 329–351

Journal of BIOTECHNOLOGY

www.elsevier.com/locate/jbiotec

Systems biology—An engineering perspective

A. Kremling*, J. Saez-Rodriguez

Systems Biology Group, Max-Planck-Institute for Dynamics of Complex Technical Systems, Germany

Received 31 August 2006; received in revised form 23 January 2007; accepted 19 February 2007

Abstract

The interdisciplinary field of systems biology has evolved rapidly over the last years. Different disciplines have aided the development of both its experimental and theoretical branches.

One field, which has played a significant role is engineering science and, in particular chemical engineering.

Here, we review and illustrate some of these contributions, ranging from modeling approaches to model analysis with a special focus on technique which have not yet been substantially exploited but can be potentially useful in the analysis of biochemical systems.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Modeling framework; Model reduction; Model verification; Control engineering

1. Introduction

Interdisciplinary research can be found in many areas to improve scientific work by combining tools and methods from different fields. Particularly, this has been observed in molecular biology, where researchers are confronted with large data sets (e.g. DNA sequences) requiring to cooperate with information scientists, leading to the establishment of bioinformatics. Additionally, biologists face a huge number of cellular components that have to be characterized by their functionality and the spatial/temporal behavior. Here, the emergent field of systems biology comes into play, developing a battery of experimental and theoretical approaches to solve problems in biotechnology and medicine (Kitano, 2002a,b).

Although, the origin of systems biology research as it is understood today is subject of controversy (Wolkenhauer, 2001), it is well accepted that two pillars can be defined: (i) a systematic collection of a large amount of experimental data for every type of component that can be found in a cell (Ideker et al., 2001) and (ii) a theoretical approach based on the view of a cell as a system, that can be characterized by state variables, inputs and outputs. Many theoretical fields like (bio-)physics, (bio-)mathematics and engineering science have contributed to systems biology with regard to the latter backbone, and it seems not pos-

---

* Corresponding author at: Systems Biology Group, Max-Planck-Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany. Tel.: +49 391 6110 466.

E-mail address: kremling@mpi-magdeburg.mpg.de (A. Kremling).

0168-1656/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2007.02.009

sible to assign a specific method to one of the fields. However, all fields have their specialties, and here we will focus on the activities initiated by the engineering community. In a series of papers, the groups of E.D. Sontag and J.C. Doyle outlined possibilities and perspectives in systems biology from a control engineers point of view (Sontag, 2004, 2005; Csete and Doyle, 2002). Sontag summarizes these activities as follows: (i) improving the design of instrumentation for high-precision measurements and manipulations; (ii) analysis of biological systems with respect to feedback characteristics, sensitivities, gain quantification and structure identification; (iii) sensor and actor design for technical systems based on structures found in biochemical systems; (iv) the formulation of entirely new theoretical control and systems theory problems. From our point of view, topic (ii) is the most important one, since it allows to improve the knowlegde on cellular systems, offering new approaches to reach this goal.

In systems biology, two main approaches have become accepted during the last years: bottom-up and top-down. The bottom-up approach is the most appropriate when all or most biochemical reactions of a process of interest are known and sufficient experimental data is available. The resulting mathematical description can be qualitative or quantitative, deterministic or stochastic depending on the aim. Quantitative models are characterized by a large number of kinetic parameters that have to be estimated from the experimental data. Optimization algorithms are available for these tasks, which can also be used for designing new experiments to improve the model structure and the parameter quality. A rationale followed in the bottom-up approach is a modular approach (Hartwell et al., 1999; Saez-Rodriguez et al., 2005a,b), where networks are decomposed into subunits (modules), which are thoroughly examined and afterwards aggregated into a larger model.

Collecting the huge amount of experimental data required for conducting a clean bottom-up approach is not feasible for many cases. Then, as an alternative to the bottom-up, the top-down approach, is useful, particularly when not much knowledge about the interactions of the elements involved is available. Here, these interactions between network components are examined first only based on experimental data. Based on correlation analysis or further cluster techniques, components that are tightly related can be identified, leading to the reconstruction of the networks. Bottom-up and top-down approaches complement each other, and probably the most efficient way to progress would be via a combination of both, following a so-called middle-out approach (Noble, 2002).

There are many general reviews on systems biology (Kitano, 2002a,b), from the point of view of a control-theorist (Sontag, 2005), a biologist (Sorger, 2005), drug discovery (Butcher et al., 2004), focusing on modeling approaches (Jong, 2002; Janes and Lauffenburger, 2006), etc. In this review, we focus specifically on contributions from (mainly chemical) engineering sciences, which comprises, among others, the development of modeling frameworks, characterization of closed loops, model reduction, model verification and experimental design. Many of these methods are well-established in the engineering discipline and applied to similar, but not-biologically inspired, problems. The goal of this contribution is (based on several examples of our own work and others), to illustrate how the engineering sciences can help the development of systems biology.

2. Systematic modelling—a framework for cellular systems

The set up of mathematical models to analyze complex systems is very common in many research fields. In chemical engineering, complex models comprising a large number of equations arise, for example, in the analysis of large plants or complex apparatus like distillation columns. Therefore, already in the eighties, modeling concepts and simulation tools with efficient numerical methods were developed. Additionally, systematic procedures for the model set-up of different types of processes and different types of chemical devices were established (Stephanopoulos et al., 1990; Marquardt, 1996). The approaches are mainly based on an object-oriented representation of the processes under consideration (see e.g. Mattsson et al., 1998; Ginkel et al., 2003). A particularly convenient modeling framework is based on network theory (Gilles, 1998). Network theory allows the definition of different levels, e.g. in chemical engineering, the plant level, or the level of a single apparatus. The concept was also applied to define different levels for cellular systems resulting in the idea of a modular view of the processes

(Kremling et al., 2000) that can be used as a basis for a computer tool. In the next section, we summarize some basic ideas based on this approach.

2.1. Network theory

Briefly, network theory considers all processes as a connection of components and coupling elements. Components possess a hold-up for physical quantities like energy, mass or momentum. Coupling elements describe the fluxes between components. Components and coupling elements can be defined on different hierarchical modeling levels. Consequently, systems of components and coupling elements can be aggregated to a single component on a higher level, or vice versa, units can be decomposed into more detailed subunits.

To apply network theory to cellular systems, the cellular processes have to be structured and characterized to set up a model library. Such a library can be seen as a construction kit that enables to build models from predefined submodels (Kremling et al., 2000). At the highest level of resolution, elementary submodels (modeling objects) are defined. Important elementary modeling objects are substance storages (metabolites, proteins, DNA and RNA), substance transformers (reactions) and signal transformers.

Elementary modeling objects can now be aggregated to describe more complex processes like gene expression or signal transduction cascades. Differential equations are typically used for the mathematical description of storages. However, the concept is very general and allows also to describe a component with the number of molecules or in a very simple manner only with “present”/“not present”. The latter case allows to set up more qualitative models as described in Saez-Rodriguez et al. (2006). For the substance transformers, normally, algebraic equations are used that relate the concentration of the substrates and effectors to the reaction rates.

A completely different effort aims to develop standardized mathematical models, allowing an exchange of models among different simulation tools. This standards will be described in Section 2.3.

2.2. Modularity

One of the ideas of the network theory is the simplification of the modeling procedure by providing subunits describing processes that have to be considered frequently in a model. For example, gene expression comprises the synthesis of proteins and is therefore a candidate for a submodel; in the signal transduction processes in eukaryotes, the MAP kinase cascade is a good example for a submodel that often appears, changing from instance to instance only the kinetic parameters.

This decomposition into modules not only speeds up the model set-up, but also facilitates the analysis of complex networks: since the smaller subunits are simpler to examine, one can expect to obtain new information from the submodules that provide insights into the properties of the whole network.

Hereby, the problem arises, in which way such submodels or functional units should be defined. Although the modularity of biological processes is generally accepted, a clear, unique definition of module is still lacking. Different proposals, such as evolutionary conservation, robustness and genetic co-expression have been suggested (Wolf and Arkin, 2003). There are also several efforts to rationalize from a mathematical point of view the definition of modules (Papin et al., 2004). We have proposed two different approaches: one is oriented on the biological knowledge and the functionality of the submodels that are described (Kremling et al., 2000), and the other on a conceptually more rigorous criterion, namely the absence of retroactivity (Saez-Rodriguez et al., 2005a,b). Both approaches are briefly introduced.

2.2.1. Biologically motivated criteria

The criteria defined are based on studies with prokaryotes but can be extended to describe processes for higher organisms. These criteria are based on denning functional units as those set of elements having in common three properties:

(i) *Physiological task*: This is the case when a number of elements work together towards to the same physiological task, for example, the different enzymes involved in the specific catabolic pathways for individual carbohydrates (lactose, galactose, etc.).
(ii) *Genetic unit*: In bacteria, genes encoding the enzymes of a functional unit are organized in genetical units. Furthermore, a hierarchical structure is commonly present: at the lowest level

in the hierarchy one can find operons: a group of genes expressed from the same promoter(s) and regulated individually by a common regulator and a specific stimulus. The paradigm is the *lac* operon of *Escherichia coli*. This well-known operon encodes the enzymes involved in lactose degradation, and is controlled by the repressor LacI and the inducer allolactose. There are also genetic units at a higher hierarchical level. For example, modulons are groups of operons and regulons controlled by global regulators that respond to more general stimuli, such as stress situations.

(iii) *Signal transduction network*: All elements of a functional unit are interconnected within a common signal transduction system. The signal flow over the unit border (“crosstalk” or “cross-regulation”) is small compared to the information exchange within the unit. Therefore, the coordinated response to a common stimulus (“stimulon”) helps to identify the members of a unit.

Fig. 1 illustrates the criteria. All molecular events involved in lactose uptake for *E. coli* are shown: the natural inducer of the system is intracellular allolactose, a by-product of the β-galactosidase reaction. Allolactose inactivates the lactose repressor LacI which leads to the production of more enzyme LacY (permease) and LacZ (lactose cleavage) and therefore also to the production of more allolactose. LacY and LacZ are metabolic enzymes responsible for the feeding of the carbohydrate into the central pathways. On the genetic levels, both genes (*lacy*, *lacZ*) are organized in an operon. Since allolactose promotes enzyme synthesis the elements are coupled by a positive feedback loop. Together with different carbohydrate uptake systems, the lactose pathway is under control of the transcription factor Crp. Crp is the last element of this global signal transduction unit that starts with the phosphotransferase system (PTS) as a sensory element.

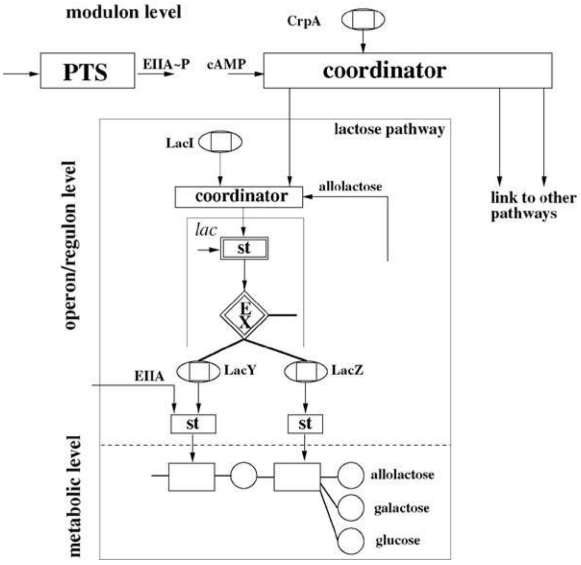


Fig. 1. The lactose pathway of *E. coli* represents a functional unit according to the criteria described.

The proposed criteria allow a rather rough subdivision of an overall cellular network. For smaller networks, a framework to demarcate modules from a system-theoretical point of view has also been developed that is introduced in the next section.

2.2.2. Absence of retroactivity as a criterion

The concept of absence of retroactivity considers unidirectional connections between the modules as interesting positions to separate signaling networks (Saez-Rodriguez et al., 2005a,b). Fig. 2 illustrates the concept with an example: the connection between two units *B* and *C* is free of retroactivity if there is an influence from a submodule B_i from B to a submodule C_j from C, but the submodule C_j does not influence B_i directly (Saez-Rodriguez et al., 2005a,b, 2004; Conzelmann et al., 2004). It can be shown that many signaling networks can be decomposed into units connected without (or with a weak) retroactivity. Particularly, interesting is the fact that the modules

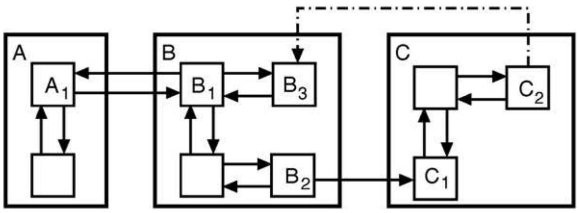


Fig. 2. The concept of absence of retroactivity: the connection between two units A and B is retroactive since there is a direct influence from a submodule A_1 from A to a submodule B_1 from B and vice versa. However, the connection between B and C is free of retroactivity since B_2 influences C_1 but not vice versa. A feedback to another subunit, e.g. from C_2 to B_3 (dashed line) is allowed (Saez-Rodriguez et al., 2005a,b, 2004).

obtained by applying the domain-oriented approach (see Section 5.1) are connected free of retroactivity. The properties of such units are independent of downstream elements and can be analyzed relatively straightforward by means of system theoretical tools (Saez-Rodriguez et al., 2004).

It results that there is a small number of modules, which via aggregation can describe almost any signal transduction network (Saez-Rodriguez et al., submitted). Once such a kit is defined, it would be reasonable to analyze its elements systematically form different points of view, such as stability, monotony, signal transfer properties and dynamics. As an illustration of this approach, a modular analysis of a model for the EGF-induced MAPK cascade (Schoeberl et al., 2002) allowed us to obtain insights into the signaling network, e.g. that the low sensitivity to ligand concentration can be traced back to the saturation of the ERK module (Saez-Rodriguez et al., 2004). Additionally, finding a less complex model for a certain module which retains its essential input/output behavior (Conzelmann et al., 2004; Saez-Rodriguez et al., 2004), it is possible to reduce the complexity of the model, as discussed in Section 5.2.

2.3. *Tools and formats in systems biology*

Nowadays, many computational tools tailored for research in systems biology are available. While some tools are available to analyze biochemical systems from a qualitative perspective (de Jong et al., 2003; Gonzalez et al., 2006; Klamt et al., 2006), most of them are devoted to kinetic modeling.

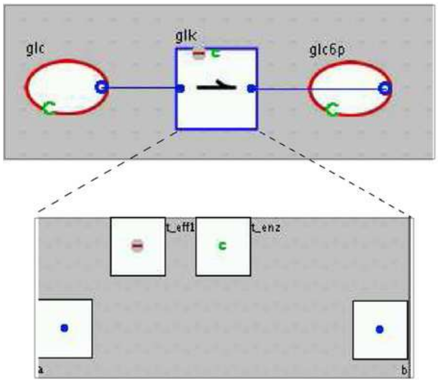
For reviews and a comparison of different approaches and tools for the latter, we refer to the recent reviews (Vacheva and Eils, 2006; Alves et al., 2006). Basic to many tools is the automatic generation of the balance equations that enables a dynamical simulation of the state variables. To set up the equations, the stoichiometric information for each reaction is used and the rate laws have to be defined. Many tools also provide a graphical user interface that allows to choose a modeling object, to parameterize it and to connect it to other modeling objects. A broad spectrum of methods for model analysis are provided: the analysis of the non-linear behavior to detect, e.g. bifurcations or oscillations, the calculation of sensitivities and optimization procedures for parameter estimation or experimental design (Vacheva and Eils, 2006).

A special interface that allows to use features from different tools is the systems biology workbench (SBW): SBW-enabled programs provide services to other client applications in such a way that programs can work out different tasks with the same model, e.g. simulation, graphical representation, model analysis, etc. Interestingly, there are also two systems biology toolboxes (SimBiology and SBToolbox; Schmidt et al., 2006) for MATLAB, a well known tool in control and engineering sciences.

Since models in systems biology are characterized by many components and interactions, visualization aspects are of great importance (Saraiya et al., 2005). One tool that provides a rich visualization support is the modeling environment ProMoT, originally set up in the field of chemical engineering, and subsequently extended to model kinetic (Ginkel et al., 2003), and recently also logical models of biochemical networks.

To allow an exchange between different tools, the XML-based formats SBML (Hucka et al., 2003) and CellML (Cuellar et al., 2003) are widely used. In SBML, the definition of a model consists of a list of elements like compartment, species, reaction, parameter, unit description and rule. The tool that uses SBML has to create the relevant mathematical equations, differential equations and algebraic equations from the information given in this description. Fig. 3 shows three elements, two substance storages and a substance transformer that connect both components. The left part shows a graphical representation of the individual elements. A closer look at the transformer emphasizes structural properties of the modeling object: besides the two terminals for the components, two further terminals are provided that allow to connect an enzyme and an effector. The right part of the plot gives the representation of the submodel in SBML.

It is important to note that exchange languages are very useful in terms of exchangeability but, at the same time, limit the form in which models can be set up (which is not the case, e.g. applying more general modeling approaches, such as the network theory). For example, a model describing the dynamics in a bioreactor cannot be represented in an adequate manner in SBML, where, e.g. valves and tubes can not be described.



```
<listOfSpecies>
<species name = "glc" compartment = "cyt"
initialAmount =  "0.1"/>
<species name = "glc 6-phosphate" compartment = "cyt"
initialAmount =  "0.2"/>
</listOfSpecies>
<listOfReactions>
<reaction name = "glk">
<listOfReactants>
<species Reactant species = "glc"/>
</listOfReactants>
<listOfProducts>
<species Reactantspecies= "glc 6-phosphate"/>
</listOfProducts>
<kineticlaw formula= "K*glc"/>
</reaction>
</listOfReactions>
```

Fig. 3. Example illustrating the representation in the modeling framework (Kremling et al., 2000) and in SBML (right) by means of a simple system where two components are connected by a reaction. Note that modifiers (e.g. enzymes) are not depicted in neither the left figure (where their potential connections to the reactions are shown) nor the right text.

3. Model verification and experimental design

New developments in metabolome, transcriptome and proteome measurement techniques lead to a huge amount of data that is used to set up detailed kinetic mathematical models. Often, information on the stoichiometry of the biochemical network describing the material flow or signaling flow is available while information on kinetic binding constants or turnover numbers are uncertain or not given. Therefore, one tries to find the best kinetic parameter values that describe the experimental data, leading to an optimization problem where the difference between simulation data and experimental data is minimized. Optimization problems are very frequently present in chemical engineering to design a plant or an apparatus in such a way that the quality and the quantity of a product is optimal, and thus systems biology can profit from the developments in the field of process engineering. Besides parameter estimation, optimization plays a key role in the design of new experiments, either to verify new hypothesis or to improve the structure and the kinetic parameters of a model. The determination of the structure of a (biochemical) network from the experimental data is referred as reverse engineering and is out of the scope of this contribution.

For all subsequent sections, a general single input non-linear model

$$\dot{\underline{x}} = \underline{f}(\underline{x}, u, \underline{p}); \quad \underline{y} = \underline{h}(\underline{x}), \tag{1}$$

is considered with state varibles $\underline{x}$ parameter vector $\underline{p}$, input u and measured output $\underline{y}$. The linearized model at an operating point $\underline{x}_{ss}$ u_{ss} is given by:

$$\dot{\underline{x}}' = A\,\underline{x}' + \underline{b}\,u'; \qquad \underline{y}' = C\,\underline{x}'. \tag{2}$$

3.1. Parameter estimation

Before the procedure of parameter estimation can start, the problem of parameter identifiability should be solved: parameter identifiability addresses the question whether a parameter of the model can be estimated given a set of experimental data (a priori identifiability), or to determine the accuracy which can be expected for each parameter (practical identifiability) (Gadkar et al., 2005). Determining a priori identifiability for general non-linear systems is solved only for special types of systems and depends on the inputs and the measured state variables. The concept of identifiability can be illustrated with a simple example (see Fig. 4).

The small network consists of two interconnected components. Assuming mass action kinetics, the linear ordinary differential equations (o.d.e.′s) are given by

$$\dot{x}_1 = -(k_1 + k_2)x_1 + k_3 x_2 + u \tag{3}$$

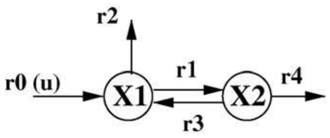


Fig. 4. Metabolic network with two components and five reactions.

$$\dot{x}_2 = k_2 x_1 - (k_3 + k_4) x_2, \tag{4}$$

where u is the constant input rate (r_0). If only the component x_1 can be measured, the system can be rewritten as one o.d.e. for x_1 of order 2[1] and the expected time course can be calculated by solving:

$$\ddot{x}_1 + p_1 \dot{x}_1 + p_2 x_1 = \dot{u} + p_3 u, \tag{5}$$

where parameters $\underline{p}$ depend on the original parameters: $p_1 = k_1 + k_2 + k_3 + k_4$, $p_2 = k_1(k_3 + k_4) + k_2 k_4$, $p_3 = k_3 + k_4$. Using a system identification approach for linear dynamical systems, the three parameters of vector $\underline{p}$ can be determined, but it is not possible to recalculate all the parameters k_i from the original system; that is, the parameters k_i are not identifiable.

Studies on practical identifiability often use the Fisher-Information-Matrix (FIM) to determine the accuracy that can be expected for a certain system. The FIM F is calculated as a sum over all time points t_k with the matrix of the parameter sensitivities W with $w_{ij} = \partial x_i / \partial p_j$ and the variance–covariance matrix C of the measured states by the following relationship:

$$F = \sum_{t_k} W^{\mathrm{T}} C^{-1} W. \tag{6}$$

In many cases, matrix C is a diagonal matrix with the variances of the measured state variables. For dynamical systems, the sensitivities are also time dependent and therefore the following additional o.d.e.′s have to be solved for a general system given in Eq. (1):

$$\dot{W} = \frac{\partial \underline{f}}{\partial \underline{x}} W + \frac{\partial \underline{f}}{\partial \underline{p}}, \tag{7}$$

where $\underline{f}$ are the right-hand side entries of the o.d.e.′s. The expected standard deviation σ_{pi} for each parameter p_i can be determined by the following relationship:

$$\sigma_{\hat{p}i} = \sqrt{(F^{-1})_{ii}}. \tag{8}$$

However, since the time course of the state variables are non-linear, the given relationship is due to the Cramer-Rao inequality only a lower bound (Ljung, 1999). Advanced methods therefore use statistical methods to improve the estimation of the parameter standard deviation.

A method recently introduced uses the so-called bootstrap method: like a Monte-Carlo method, the bootstrap uses stochastic elements and repeated simulations to analyze the properties of the system under consideration. Depending on the conditions, remarkable differences between the results obtained with the bootstrap and with the FIM can be detected (Joshi et al., 2006), arguing thus for the use of bootstrap methods for a rigorous analysis of parameter identifiability in biochemical systems.

Analysis of the FIM can also be used to detect correlations between parameters and thus obtain hints on which parameters can be estimated together. Several methods are introduced to group together parameters with a certain accuracy with respect to parameter estimation. The method introduced by Reichert and co-workers (Brun et al., 2001) uses the collinearity index γ_K of a modified FIM which is set up for K parameters: first, the sensitivities w_{ij} are scaled with appropriate values for parameter p_i and state x_j : $w^*_{ij} = \partial x_i / \partial p_j \, p^{\mathrm{s}}_j / x^{\mathrm{s}}_i$; afterwards the column of the matrix W^* with the scaled sensitivities are divided with the norm of the respective columns (Brun et al., 2001):

$$\underline{w}^{**}_j = \frac{\underline{w}_j}{\left\|\underline{w}_j\right\|}; \tag{9}$$

Finally, the minimal eigenvalue of $W^{**\mathrm{T}}_K W^{**}_K$ for a group of K parameters is the inverse of collinearity index γ_K. If the parameters are highly correlated, matrix W^{**}_K tends to be singular. Therefore, the minimal eigenvalue is very small and the collinearity index γ_K is high if some of the K parameters are correlated.

Parameter estimation is a classical optimization problem that is formulated in general with the objective function Θ as follows:

$$\min_{\underline{p}} \Theta = \sum_m \sum_n \sum_k \frac{(y^{\mathrm{ex}}_{nk} - y^{\mathrm{s}}_{nk}(\underline{p}))^2}{\sigma^2_{nk}}, \tag{10}$$

with y^{ex}_{nk} is the measured value for state n at time point k in the mth experiment, y^{s}_{nk} the corresponding simulated value that depend on the parameter vector $\underline{p}$ and σ_{nk} is the standard deviation of the measured values y^{ex}_{nk}. In many applications, the standard deviation is not known for every single sample point, and σ_n is considered a constant absolute error or proportional to the measured value.

---

[1] Calculation of the time derivative of both sides of Eq. (3) and inserting Eq. (4) leads to Eq. (5).

The engineering community has provided many scientific inputs to the development of optimization algorithms. Particularly, Banga and co-workers compared different strategies and concluded that a global stochastic optimization method, "Evolution Strategies" (ES), was the best one to solve a benchmark problem (Moles et al., 2003). In recent years, algorithms that combine local and global search strategies have become popular, to circumvent their problems (mainly, not finding the global optimum and slow conversion, respectively) (Katare et al., 2004; Rodriguez-Fernandez et al., 2006). Hybrid algorithms start the search with a global optimization and then switch to a local one. Using this approach, the computing time could be reduced significantly. Still problematic is to find the heuristics to switch from the global to the local search (Rodriguez-Fernandez et al., 2006).

3.2. Experimental design

The goal of experimental design is to define the most informative experiment, e.g. to distinguish between two or more model variants, to improve the validity of the model by reducing the parameter variances and to clarify the structure of the model. For many cases, again, an optimization problem can be formulated to solve one or in an iterative approach more of the problems, see, e.g. Asprey and Macchietto (2000). The first two issues are discussed here.

3.2.1. Discrimination between competing models

A typical case is the following: two or more model variants are available that describe one single experiment very well. Here, an experiment has to be performed that (ideally) allows a clear discrimination between the models. For a number of years, approaches have been developed to discern between model candidates, e.g. Boox and Hill (1967), Munack (1992), Cooney and McDonald (1995). The key idea is to find an input profile that maximizes the difference of the outputs of the competing models. In a series of papers, Asprey and co-workers describe and review methods for this purpose (Chen and Asprey, 2003). One of the approaches uses an extended weighting matrix including the variances of the measured state variables and the sensitivities of the parameters. In this case, the task can be formulated as the maximization of an objective function

$$\max_u = \int_{t_0}^{t_{\mathrm{end}}} [\Delta \mathbf{x}^{\mathrm{T}}(t) \mathbf{Q} \Delta \mathbf{x}(t)] \mathrm{d}t, \tag{11}$$

with $\mathbf{Q}$ being a general weighting matrix and $\Delta\mathbf{x}$ being the difference between the responses of the two competing models. Many different approaches for the choice of the weighting matrix can be found in the literature. Weighting should be done if the interesting state variables are within different orders of magnitude. In this case, it is useful to use a diagonal weighting matrix with elements:

$$Q_{ii} = \frac{1}{((x_{i1} + x_{i2})/2)^2}, \tag{12}$$

That is, to weight by the average of the two models. It is, however, also possible to include information about the measurement variances, the variances of the parameters of the model and the sensitivity of these parameters with respect to the interesting state variables. In Kremling et al. (2004a,b), the following approach was used:

$$Q = (C + \mathrm{VC}_1 + \mathrm{VC}_2)^{-1}, \tag{13}$$

where C is the variance of the measurements, and VC is the variance–covariance matrix for model predictions:

$$\mathrm{VC} = W F^{-1} W^{\mathrm{T}} \tag{14}$$

with sensitivity matrix W and FIM F. In this approach, only the diagonal elements of C and VC are used. This means that the squared model difference for a single state variable is weighted by a sum given by its measurement variance, and the square of the sensitivity of each fitted parameter with respect to the state variable, multiplied by the variance of the parameter. In other words, the difference of a state variable contributes less to the objective function if: (i) the measurement error of that state variable is large and (ii) the state variable in the designed experiment is very sensitive to parameters that could be estimated only with large errors using the experiment(s) performed so far. In Kremling et al. (2004a,b), the approach was successfully applied to a small biochemical network.

3.2.2. Improving parameter accuracy

The quality of the parameters, that is, the parameter variances can be determined with the FIM as described above. To improve the quality of the estimated parameters different criteria of optimality are defined:

- A-optimal design minimizes the trace of F^{-1}.
- D-optimal design maximizes the determinate of F.
- E-optimal design maximized the minimal eigenvalue of F.
- Modified E-optimal design minimizes the ratio between the maximal and the minimal eigenvalues of F.

The A- and D-optimal design are related to the arithmetic and the geometric mean of the expected errors. In contrast, the E-design tries to minimize the largest error. For the formulation of the optimization problem, some constraints have to be taken into account. Most important seems to be the input variable that is available to control the process. Using bio-reactors, often the feed rate and the feed concentration are used. Besides the range of possible input values, the time points where the input can be altered have to be taken into account (Kutalik et al., 2004).

4. Control and observation

In control engineering, the closed loop behavior is designed in such a way that a number of constraints are fulfilled. First, the closed loop is analyzed with respect to stability, which is the most important characteristic. Besides stability, and in order to design and to follow a process, observability (*how much can be seen*) and controllability (*how much can be changed*) are two additional concepts useful for analyzing the system. Some applications of engineering approaches to the stability analysis will be discussed in Section 6.1; here, we shall introduce the concepts of controllability and observability and their applicability to systems biology.

4.1. Controllability

Controllability is concerned with the question whether it is possible to calculate an input function u that allows to drive the system in finite time to certain desired final values of the state variables x or the output variables y. The problem is solved in two steps. First, controllability is checked that guarantees (at least for all linear systems) that an input function can be found, and second, the input is calculated in dependence of the desired final values. A fist attempt to analyze controllability in the S-systems framework was performed by Ervadi-Radhakrishna and Voigt (2005). They used an exact feedback linearization to transform their model into a controllable linear form. The procedure was applied to a small metabolic pathway with three state variables and with two and three input variables. Inputs in the system are the enzyme activities. A number of test scenarios were simulated and the authors conclude that the S-systems representation is a good basis to analyze controllability. Controllability could also be useful in experimental design to check if, for desired domains of the state variables, an input function can be found.

4.2. Observability

Observability can be seen as a measure to infer the state variables x of the system from its measured output y. As in the case of the controllability, the problem is solved in two steps. First, observability clarifies if it is possible to reconstruct state variables and second, an observer (or filter) has to be designed that allows to estimate the time course of the state variables from the measured outputs. This method has a high potential for systems biology, where the development of measurement devices is still time-consuming and expensive. Moreover, such a tool will allow to follow cellular events even if some components cannot be measured directly. Importantly, with information on internal state variables, the system can be controlled and redirected from the outside to reach a desired behavior as far as allowed by its controllability. Another field of application of this concept is model reduction: variables, which are not observable are in many cases not of interest, and can thus be removed from the model (see Section 5.1).

Let us illustrate this approach with a simple linear pathway, shown in Fig. 5A. The reaction kinetics are assumed to be irreversible. The system, of dimension n (i.e., with n state variables), reads

$$\dot{x}_1 = r_0 - r_1, \qquad \dot{x}_2 = r_1 - r_2, \qquad \dot{x}_3 = r_2 - r_3 \tag{15}$$

which can be expressed in a compact manner in the linear form if mass action is applied (Eq. (2)). The

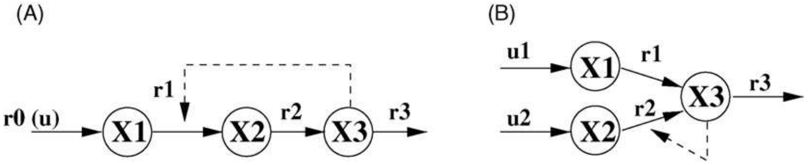


Fig. 5. (A) Linear pathway with feedback. The feedback allows to observe the system by measuring only X_1. (B) Pathway of a small network to check structural observability.

observability can be checked by calculating the rank of the matrix

$$P = [C^{\mathrm{T}} A^{\mathrm{T}} C^{\mathrm{T}} \cdots (A^{\mathrm{T}})^{n-1} C^{\mathrm{T}}], \tag{16}$$

where A, B and C are defined as in Eq. (2). If the rank of P is less than n, then not all states can be reconstructed. For the system without feedback ($r_1 = f(X_1)$), and measurements of X_1 or X_2, rank of P is equal to 2, and therefore does not allow to reconstruct all state variables. With the measurement of X_3 (i.e., $y = X_3$), it can be shown that P has full rank and all states can be reconstructed. Linear pathways are often regulated by feedback inhibition of one of the first enzymes in the pathway by the end-product ($r_1 = f(X_1, X_3)$). If this feedback is taken into account in the calculation, then rank of P is equal to 3 and the system is observable, independent from the choice of the output. This is due to the influence of X_3 on state X_1, or in other words, X_1 contains information from all components from the closed loop and therefore, the measurement of X_1 is enough to reconstruct all other state variables.

The concept of observability has been extended to make it independent from the choice of the numerical parameter values. For a comprehensive description of such a structural analysis, see Wend (1993) and Unger et al. (1995). Hereby, the analysis of structural observability analyzes the structural matrices S_A and S_C, where S_A and S_C have the same dimension as A and C in Eq. (16), respectively, but contain entries * instead of numerical values.

The system is structurally observable when: (i) representing the system as a graph, all nodes are linked directly or indirectly to the measured output and (ii) the structural rank of the matrix

$$S_{\mathrm{p}} = \begin{pmatrix} S_A \\ S_C \end{pmatrix} \tag{17}$$

is n. To determine the structural rank of matrix S_{p} one has to find columns with at least one entry *. For biochemical networks, this condition is almost always fullfilled since the components have an influence on their own degradation. To represent the dynamical system as a graph, the information of the Jacobian matrix can be used. The entries $J_{ij} = \partial f_i / \partial x_j$ indicate if element j has any influence on element i. If this is the case, the respective nodes in the graph are connected. A simple example is shown in Fig. 5 plot B. Two components representing two pathways are connected by a third component X_3. In the case that only component X_3 can be measured, matrix S_{p} reads

$$S_{\mathrm{p}} = \begin{pmatrix} * & 0 & 0 \\ 0 & * & * \\ * & * & * \\ \cdots & \cdots & \cdots \\ 0 & 0 & * \end{pmatrix}. \tag{18}$$

For this example, the conditions are fulfilled, since X_3 can be reached by X_1 and X_2 ($S_{\mathrm{p}}(3,1) = S_{\mathrm{p}}(3,2) = *$). Moreover, the structural rank is 3, since the diagonal elements of the upper parts of S_{p} have an entry. If it is assumed that reaction r_2 in Fig. 5 reads:

$$r_2 = k_2 \frac{X_2}{K_2 + X_2} X_3, \tag{19}$$

and if $X_2 \gg K_2$, r_2 is simplified:

$$r_2 = k_2 X_3. \tag{20}$$

In this case, the S_{p} is as follows:

$$S_{\mathrm{p}} = \begin{pmatrix} * & 0 & 0 \\ 0 & 0 & * \\ * & 0 & * \\ \cdots & \cdots & \cdots \\ 0 & 0 & * \end{pmatrix}. \tag{21}$$

and the structural rank is only 2. The system is not structural observable.

The structural considerations assure that a system is observable for *almost* all parameter sets, but cannot guarantee that the system is observable for all parameter values. If for this example, both metabolites X_1 and X_2 run on the same time scale, i.e. the entries in the Jacobian of the system J_{11} and J_{22} are similar, the systems matrix A reads for example

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & -1 \\ 1 & 1 & -3 \end{bmatrix}, \tag{22}$$

and the matrix P is then:

$$P = \begin{bmatrix} 0 & 1 & -4 \\ 0 & 1 & -4 \\ 1 & -3 & 8 \end{bmatrix}, \tag{23}$$

Matrix P has rank = 2, that is, the system is not fully observable.

Let us illustrate these ideas on a more realisitc example. The set of equations for a bacterium growing during a batch experiment in a bio-reactor has a special structure that facilitates the reconstruction of state variables. The equations for biomass B, substrate S and internal metabolites M_i read

$$\dot{B} = \mu B = Y r_{\mathrm{up}} B \tag{24}$$

$$\dot{S} = -w r_{\mathrm{up}} B \tag{25}$$

$$\dot{M}_i = \sum_j \gamma_{ij} r_{ij} - \mu M_i = \sum_j \gamma_{ij} r_{ij} - Y r_{\mathrm{up}} M_i, \tag{26}$$

with substrate uptake rate r_{up}, stoichiometric coefficients γ_{ij}, metabolic rates r_{ij}, molecular weight of the substrate w and yield coefficient Y. In general, the substrate uptake rate r_{up} depends on the substrate concentration. If the growth rate depends on the uptake rate $\mu = Y r_{\mathrm{up}}$, a graph based on the Jacobian can be drawn as shown in Fig. 6 indicating that all internal metabolites can be used to reconstruct at least biomass and substrate, since both state variables can be connected to every metabolite in the network. Furthermore, if the uptake rate depends on an internal metabolite, e.g. on ATP or on a member of the phosphotransferase system (e.g. the PEP phosphotransferase system in *E. coli* transfers a phosphoryl group from central metabolism to the incoming substrate), this metabolite influences the biomass concentration and hence, all metabolites that have a direct or indirect link to this metabolites "transfer" the information to the biomass (Fig. 6). Therefore, the system is observable by only measuring the biomass or the substrate concentration.

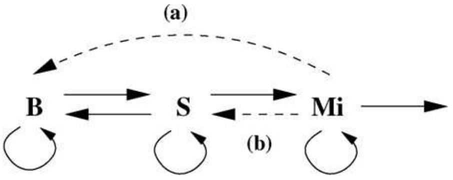


Fig. 6. General structure of a cellular model that includes the transport of the substrate. If one or more metabolites M_i are involved in the transport process, that is, $\gamma_{\mathrm{up}} = f(S, M_i)$ the graph has to extended by the dashed lines (a) and (b).

4.2.1. Observer/Kalman-filter

If the property observability is checked, an observer or filter can be constructed to estimate the state variables. A general scheme is shown in Fig. 7. The idea is to have the "real world" and the "model world" in parallel with the same input signals. The input will result in some response $\underline{y}$ of the real world system that can be measured by the measurement device; in the model world, the output is $\hat{\underline{y}}$. Both outputs are compared and the difference $\underline{y} - \hat{\underline{y}}$ is used in the observer/filter to redirect the model towards the measured output $\hat{\underline{y}}$. If the difference $\underline{y} - \hat{\underline{y}}$ is negligible, it can be expected that the simulated states $\hat{\underline{x}}$ corresponding to the outputs $\hat{\underline{y}}$ are close to the states $\underline{x}$ in the real world. If the equations for the model are given as in Eq. (1), the equations for the estimated state variables $\hat{\underline{x}}$ are determined by

$$\dot{\hat{\underline{x}}} = \underline{f}(\hat{\underline{x}}, u, \hat{p}) + K(\hat{\underline{x}})(\underline{y} - \hat{\underline{y}}). \tag{27}$$

where K is the gain of the observer.

Classical filter algorithms are based on the pure model, e.g. the Luenberger observer, or take into account that the model as well as the measurements are subject to uncertainties, as in the case of the Kalman filter. The latter and its extensions can therefore consider that the model is not a perfect description of the real world and that the measurements may be distorted with noise. Since the measured data is not available at any time point but at discrete sampling points, modern algorithms use two steps to first predict the course of the state variable in the next time step and second

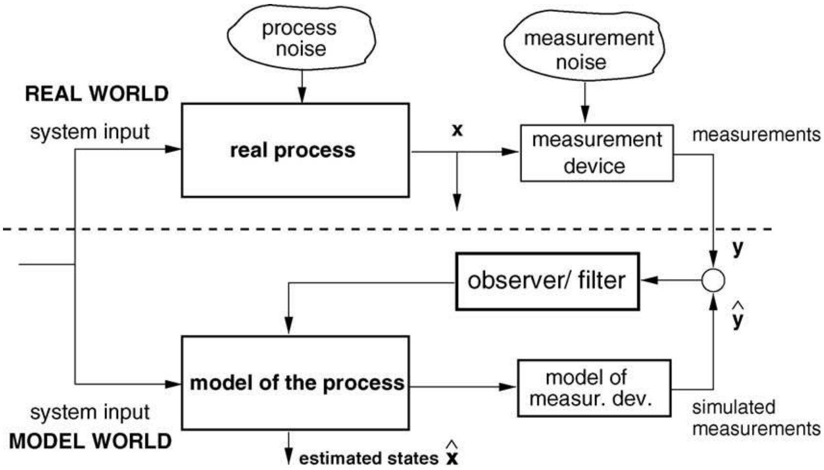


Fig. 7. Set up of a model based measurement system.

to correct the state variables with the current measurement. Although, sophisticated methods are available to estimate state variables, open points are, e.g. to take into account that parameters of the model are uncertain. This problem is raised by Dochain (2003). In his tutorial, problems of state and parameter estimation are discussed and recent findings are summarized.

An observer was designed for the experimental data from Bettenbrock et al. (2006). Measured time course of biomass, glucose, lactose, proteins LacZ and EIIA (PTS protein) are available during diauxic growth of *E. coli* on glucose and lactose. The system is observable by using only the measured biomass. Results with an extended Kalman filter (EKF) are shown in Fig. 8. Remarkable results are obtained and all state variables are reconstructed very well when only biomass is measured.

Note that the main difference to a standard parameter estimation approach is that with the EKF the model is adjusted dynamically to the experimental data as soon as new data is available. Therefore, it is suitable for on-line control of biochemical processes; for example, it could be applied to a replacement organ (e.g. a liver) to promptly monitor dangerous conditions.

5. Model reduction

Modeling strategies as those described in Chapter 2, together with the increasing amount of experimental data available and its exploitation with the parameter estimation and model discrimination methods introduced in Chapter 3, let us set up models of increasing size and detail. However, such complex models, which should actually be a tool to understand biochemical processes, become themselves difficult to understand. For many tasks, it is neither necessary nor desirable to work with such large and unmanageable models. For example, in many studies a specific question is to be addressed, and only smaller parts of an existing model are needed. In this case, one tries to simplify the structure of the rest of the model keeping the main characteristics of the original one. Therefore, the reduction of mathematical models of biochemical networks into manageable ones, without losing their essential properties, would be very useful (Saez-Rodriguez et al., 2005a,b).

One can distinguish between two types of model reduction. In the first type, only the behavior of the model, that is, the dynamics or the steady state characteristics, are needed and a phenomenological description is used. On the second type, however, model reduction is based on a simplified description of the system by grouping together components or by neglecting some of the molecular interactions; here, a model description is required. In this section, we shall introduce some approaches to the reduction of models describing biochemical networks, mainly inspired by engineering technics.

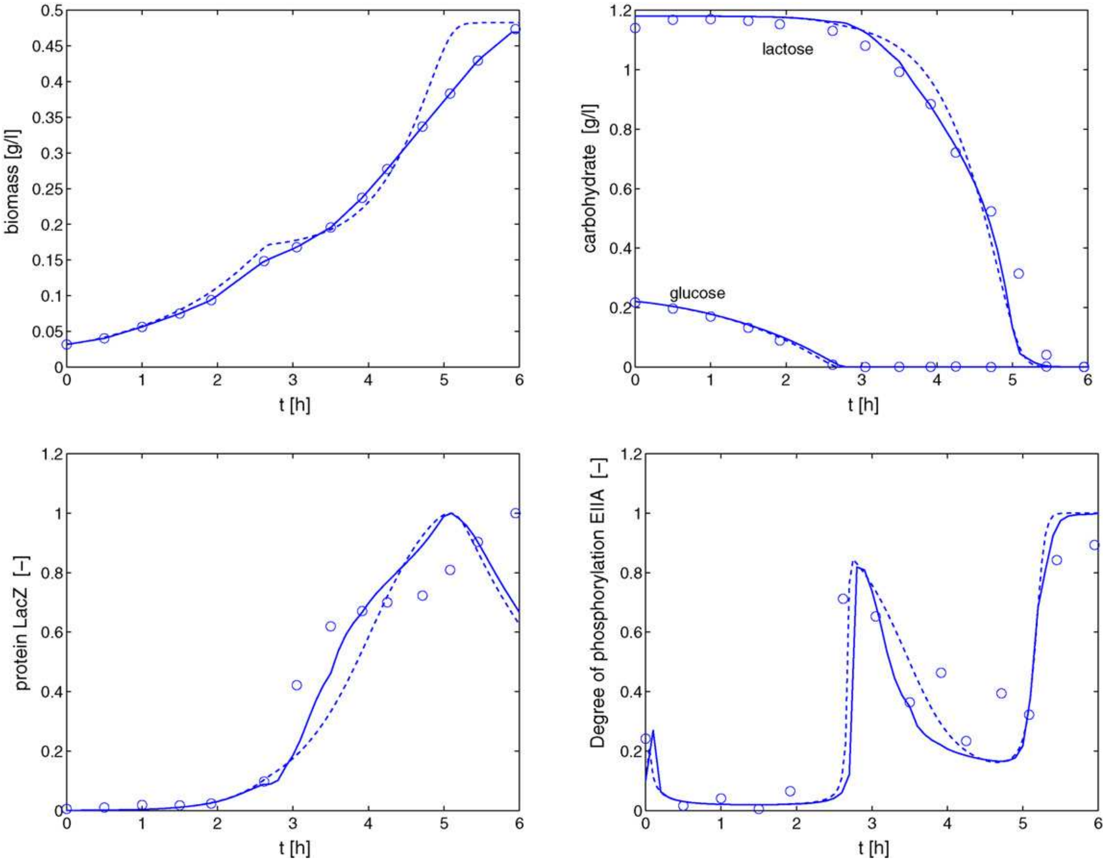


Fig. 8. Results of an extended Kalman filter (EKF) applied to experimental data of diauxic growth of *Escherichia coli* on glucose and lactose (Bettenbrock et al., 2006). Symbols represent measured data and solid lines is the output of the EKF. Dashed lines portray a pure simulation with a simplified model based on the model proposed in Bettenbrock et al. (2006). The filter was designed with MATLAB.

5.1. Reduction of combinatorial complexity in signal transduction networks

In signaling networks, particularly of mammalian cells, the receptors and adaptor proteins are characterized by their ability to bind different molecules via different domains (Pawson and Nash, 2003) (see Fig. 9(a)). If one considers all possible combinations of proteins, the number of feasible states (micro-states) increases exponentially (Blinov et al., 2004; Borisov et al., 2005; Conzelmann et al., 2004). This combinatorial complexity has typically been circumvented because it is very difficult to say a priori which micro-states are the important ones (Faeder et al., 2005). Therefore, a rigorous description has to include all possible micro-states (Fig. 9(a)). However, the number of equations is very high even for relatively simple models.

Recently, a new approach based on the macro-states (the states of the different domains) instead of the micro-states (the possible molecular combinations) has been proposed in Borisov et al. (2005) (Fig. 9(c)), and extended and formalized in Conzelmann et al. (2006). In this approach, a state space transformation allows a reversible move between the macroscopic and microscopic description.

The method operates as follows: starting with a system of the form of Eq. (1) describing the micro-states $\underline{x}$ (being $\underline{y}$ the macro-states of interest), first one adjusts the kinetic parameters according to domain interactions. For example, if a scaffold molecule A can bind to B and C, and both binding sites are independent, the kinetic parameters for the binding of A to B will be the same for all micro-states (i.e. the parameters for free A binding to B are equal to the parameters of the complex

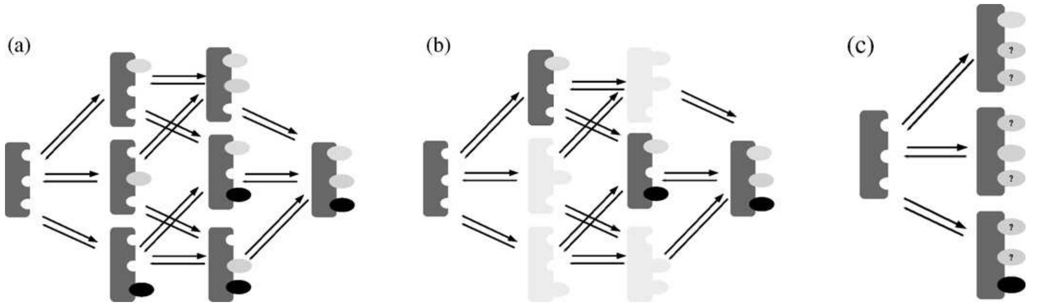


Fig. 9. Different approaches to the combinatorial complexity of signaling networks (Saez-Rodriguez et al., 2005a,b), illustrated by a scaffold protein with three binding sites for three different proteins. While a rigorous description has to include all possible combinatorial combinations (a), usually only some of them are modeled (b). According to Borisov et al. (2005) and Conzelmann et al. (2006), the system can be modeled using new states describing the different domains (c).

AC binding to B). Importantly, this process is independent of the parameter values; one only needs to know whether the binding sites influence each other.

Second, one performs a linear transformation $\underline{z} = T\underline{x}$, where T is a square, non-singular matrix. The resulting states z_i can be classified into levels or tiers, representing each tier a level of detail: 0th tier corresponds to the total concentration of the protein, the first tier the macro-states (the state of the individual domains), the 2nd tier the state of all pairs of domains (i.e. the concentration of proteins with concurrently occupied domains 1 and 2, 1 and 3, 2 and 3, etc.), 3rd tier of triples of domains, and so on. Importantly, this transformation is general, i.e. independent of domain interactions and kinetic parameters (Conzelmann et al., 2006). Furthermore, the new states of tiers 0 and 1 represent the number of free or occupied binding domains, a quantity biologists are used to work with, and that can be measured more easily than the concentration of particular species.

It results that in many relevant cases the transformed model equations for $\underline{z}$ can be decomposed into two sets $\underline{z}_1$ and $\underline{z}_2$ so that $\underline{z}_2$ is non-observable (see Section 4.2). Therefore, a reduced model only has to account for the o.d.e.′s describing $\underline{z}_1$ ($\dot{\underline{z}}_1 = \underline{g}_1(\underline{z}_1, \underline{u})$) (Conzelmann et al., 2006).

5.2. Modular model reduction

As mentioned in Section 2.2.2, modules defined according to the concept of retroactivity have the advantage that their signal transfer properties are independent of what they are connected to. Therefore, if a (simpler) substitute for a certain module can be found, it can replace it without altering the properties of the network as a whole, leading to a reduction of the complete system. In the case of (weak) retroactivity, a certain difference may appear.

If the module under study is simple enough, it can be analyzed analytically. For example, it can be shown that a module describing the double-phosphorylation of a MAPK can be reduced to a system with first order lag for low input values, and to an integrator for high values (Saez-Rodriguez et al., 2005a,b). For more complex modules, however, a more heuristical approach, based on simulation studies, has to be applied. Applied to the model for the EGF-induced MAPK cascade mentioned above (Schoeberl et al., 2002), simulation studies showed that the input/output behavior of several complex modules was remarkably similar. Thus, the more complex modules could be replaced by the structure of the most simple one, obtaining a very good approximation to the original model (Conzelmann et al., 2004).

5.3. Conservation relationships

A well-established form of reducing the number of equations of a dynamic model is based on conservation analysis. This method is based on the analysis of stoichiometric networks, a field started by the chemical engineering community. A conserved moiety is a molecular subgroup, which is conserved during the evolution of a network (Sauro and Ingalls, 2004). Since the total amount of a conservation moiety is constant, instead of describing all the states by a differential equation, one of them can be computed by an algebraic equation. Importantly, this reduction is exact and relies only on the network structure.

These structural conservations are particularly important in the case of models of signal transduction, which typically include many cycles of activation/deactivation of proteins. Consider the simple example of a protein which can exist in two states, active and non-active

$$P \rightleftharpoons P_{\mathrm{A}}. \tag{28}$$

In principle, one ODE would be written for the balance of both P and P_{A}. However, such a cycle represents a conservation moiety and it holds

$$[P] + [P_{\mathrm{A}}] = P_{\mathrm{T}}, \tag{29}$$

where P_{T} is a constant. Therefore, only one differential equation is needed (for either P or P_{A}), and the other one can be computed according to (29).

Algorithms for the identification of conservation moieties, based on an analysis of the stoichiometric matrix, are well established. Furthermore, several modeling tools incorporate them to automatically detect the moieties and reduce the dynamical model accordingly (Sauro and Ingalls, 2004).

5.4. Time scale hierarchies

Commonly, the time constants of biochemical systems span a large range, that is, the system is stiff. To analyze the system in a specific time window, the dynamics faster than those of interest can be approximated by the quasi steady-state assumption, and those slower can be neglected. There is a large body of literature on this topic. The mathematical foundations, e.g. singular perturbation methods and analysis of manifolds, are well-defined in a number of papers and textbooks (Fenichel, 1979; Segel and Slemrod, 1989; Wiggins, 1994) and application to chemical system can be found, see, e.g. Zagaris et al. (2004) and Powers et al. (2002).

The simplest approach analyzes the eigenvalues λ_i of a linear system of the form of Eq. (2) obtained from a non-linear system by linearization around an operating point (normally a steady-state $x_{\mathrm{ss}}, u_{\mathrm{ss}}$). The eigenvalues determine the time constants λ_i of the system:

$$\tau_i = \frac{1}{|\mathrm{Re}(\lambda_i)|}. \tag{30}$$

Thereby, a fast and a slow system can be defined by setting a threshold in a time constant τ^{S}. Fast modes have a time constant $\tau_i < \tau^{\mathrm{S}}$, while slow modes have $\tau_i > \tau^{\mathrm{S}}$. Unfortunately, fast and slow modes of the system do not correspond to state variables of the system in a one-to-one mapping; the modes are characterized by a linear combination of the state variables where the linear combinations are given by the matrix of the eigenvectors of the system.

For some applications, the respective modes can be assigned to a group of state variables that are involved in the same functionality. For example, in Kremling et al. (2004a,b), a model describing carbohydrate uptake, central metabolic reactions and gene expression for the uptake systems was analyzed with respect to the time scales. The authors concluded that, depending on the choice of the time window and the stimulus of the system, one or two of the subnetworks comprising the sensoring and metabolisms modules could be considered to be in steady state.

To circumvent the problems mentioned before, Hu and co-workers (Gerdtzen et al., 2004), first separate the system given with the stoichiometric matrix S and the vector of reaction rates $\underline{r}$

$$\dot{\underline{x}} = S\underline{r} - b\underline{x} \tag{31}$$

by defining fast and slow reactions $\underline{r}_{\mathrm{fast}}$ and $\underline{r}_{\mathrm{slow}}$. By applying singular perturbation arguments and scaling they rewrite the system in such a way that the fast reactions do not appear anymore and the slow (original) state variables can be grouped so that only the slow time scales are represented.

Recently, a method for automatically decomposing models of biochemical networks into a slow and fast part has been proposed (Zobeley et al., 2005). The method is based on: (i) a linearization of the system and (ii) conversion of the jacobian into a block-diagonal form, which allows the decomposition of the system into a fast and a slow block. This process is repeated along the simulated trajectory, providing hence a time-dependent analysis. The user can define a tolerance for the error of the approximation, and the method decides automatically what can be considered ‘slow’ accordingly to the user-defined error. As the authors point out, such a decomposition not only decreases the computational costs, but also allows the decomposition of the system in modules which can be analyzed independently. The method was shown to be able to reduce an oscillating, highly dynamic network for a peroxidase-oxidase reaction system from a dimension 11 to a dimension ranging from 2 to 6, depending on the regime of the system. It would be interesting to test the method on large signaling networks.

5.5. *Optimization approaches and set up of 'minimal' models*

There have been several efforts to reduce models of chemical networks models using genetic algorithms, e.g. Edwards et al. (1998), which can be useful for biochemical networks (Maurya et al., 2005). Maurya and colleagues have applied these ideas to the reduction of signaling network models. In one approach, they use genetic algorithms to identify set of parameters, which are good candidates to explain a certain set of experimental data. Subsequently, they perform multiparametric sensitivity analysis to rank the parameters in accordance with their importance. Finally, on the basis of this ranking and in an iterative manner, they 'knock out' parameters and check if the resulting reduced model can still describe the experimental data. Using this approach, they could reduce a detailed model describing the GTPase-cycle signaling module from 48 to 17 reactions.

As an alternative method, they also applied mixed-integer non-linear-optimization technics – also a methodology developed mainly in chemical engineering – to the same problem, reducing the same model to 14 reactions (Maurya et al., 2005). The advantage of this approach is that topology and parameters are simultaneously determined, and thus the method is an order of magnitude faster.

Using similar principles but with a different goal, Maurya et al. developed a promising approach to set up models with a minimal number of state variables and parameters (Maurya et al., 2006). The rationale is to start with a simple model and add more complexity (via more complex kinetic laws and adding new elements to the network structure), until the model can reproduce a certain behavior.

Their framework is related to the design of control systems and the authors propose an analogy to the steps followed there: (i) optimization of the kinetic parameters = tuning of the controller parameters (ii) modification of the kinetic rate law expressions = modification of the control laws and (iii) addition or deletion of components = updating of the control structure.

It may be possible that different model structures lead to the same characteristics observed experimentally. On such cases, the procedure has to be complemented by experimental design as those described in Section 3 to discriminate model structures.

The method was applied to construct a model of the MAP kinase cascade, The final model that was developed comprised 5 state variables and 16 parameters, and shows good agreement with a previously published model with almost 100 state variables (Maurya et al., 2006).

6. Application of control theory

It has been proposed that, with the increasing knowledge on the structure of biochemical networks and especially the knowledge on signal transduction and processing, the application of theoretical concepts from control theory should become more important. Not only the biological sciences will benefit from control theory, but also vice versa, systems biological research will lead to new challenges for control engineering (Sontag, 2004). This section summarizes some of these ideas and illustrates them with simple examples.

Some of the methods introduced in this section are based on linearized models, that is, the models are given with the structure shown in Eq. (2). If one is dealing with a non-linear system, it has to be linearized around a working point/setpoint to apply the corresponding methods. This leads to limitations since the response of the non-linear system corresponds to that of the linearized one only for the setpoint (and approximately for very close points, such as those achieved upon small inputs). However, for many applications in engineering sciences this approach has led to powerful control strategies and there are also examples showing that it can help to elucidate control principles in biological systems (Yi et al., 2000).

6.1. *Stability*

The most important task for a control engineer is to design a controller in such a way that the entire system, composed of the controller and the controlled system, is stable (expressed in simple words, a system is stable if, after a small perturbation, it returns to the original state). In principle, living systems should show a stable behavior (including attractors like limit cycles); otherwise, a component would accumulate to a point where it would lead to the collapse of the organism. Therefore, one would expect a simple behavior in this regard in biochemical reaction networks. However, in biochemical systems, due to their non-linearities, two or more steady-states can be observed. Furthermore, it results that these phenomena are used by nature to process signals; for example, a multistable system can lead to an irreversible switch behavior, which is used to take irreversible decisions, such as differentiation or cell fate (Laurent and Kellershohn, 1999; Xiong and Ferrell, 2003; Eissing et al., 2004). In general, not all of these steady-states can be observed experimentally since they can be unstable. The analysis of systems showing such an interesting behavior will help to better understand the design of the networks and also will open possibilities in which way such systems can be manipulated.

A basic principle used in control engineering is the analysis of an overall system from the properties of its single components. In control theory, the open loop composed of controller and controlled system is commonly analyzed and the behavior of the closed loop is calculated from properties of the single submodels. This is the classical way to design a controller that has to fulfill some requirements with respect to influences of disturbances or setpoint tracking.

Sontag and co-workers consider a special class of systems, monotone systems, and show that under some conditions these systems cannot oscillate nor show any chaotic behavior (Sontag, 2004). Monotone systems can be non-linear, but they are well-behaved in the mathematical sense (Sontag, 2005): if one modifies the stimulus (or other conditions), and the system starts to response, say, with a higher value than before, one would expect that, for a ‘simple’ system, the response would remain higher for the whole trajectory. Also, one would expect that, if an even higher value for the input is applied, the response would be even higher. Similar behavior would be expected for the initial values, as well. This intuitive property is what characterizes a monotone system.

Monotony can be assured (more specifically, strong input/output monotony; here only SISO systems are considered) if, for a system derived from a very general system defined as in Eq. (1), the signs of all elements in the Jacobian matrix J, and the derivatives $\partial f/\partial u$ and $\partial y/\partial c$, are sign definite and there is no negative feedback in the adjacency matrix associated to J (Angeli et al., 2004). This property is important in the context of biochemical system as it can be analyzed very easily, since information on the connections between the elements (e.g. arrows with +for activation and −for inhibition) can often be seen in cartoons in biological papers.

Multistablity is guaranteed for some ranges of feedback strengths, if the system in open loop is monotone and the steady-state response k (the so-called I/O characteristic) is sigmoidal (Angeli et al., 2004): if the output of system is connected via a monotone increasing function g to the input, i.e. the feedback is positive, the steady-states of the closed loop system can be calculated by considering the intersections between the functions k and g^{-1} (the inverse of the feedback characteristic g).

Fig. 10 shows the open and the closed loop system for lactose uptake (see Fig. 1) and a plot of the stimulus-response curve. This example also illustrates the fact that non-linearities in cellular systems are needed to establish a certain functionality (Tyson et al., 2003): in the example given above, the non-linearity is used to guarantee that the system avoids a frequent turn on and off if some fluctuations in the stimulus are present. Other examples for functionalities in cellular systems are switch-like signal amplification (e.g. the MAP kinase cascade; Huang and Ferrell, 1996), adaptation (e.g. bacterial chemotaxis, Alon et al., 1999) and oscillations during the cell cycle (Tyson et al., 2003).

Another approach, Chemical Reaction Network Theory, developed in the field of chemical engineering by Feinberg and colleagues (see, e.g. Feinberg, 1995), can be of great interest to explore stability in systems biology, since it allows to determine whether a certain (bio)chemical network can present multistationarity. The strength of this theory relies on its ability to provide assertions independently of specific parameters. Therefore, it helps to discard multistationarity for a certain biochemical structure (Conradi et al., 2005).

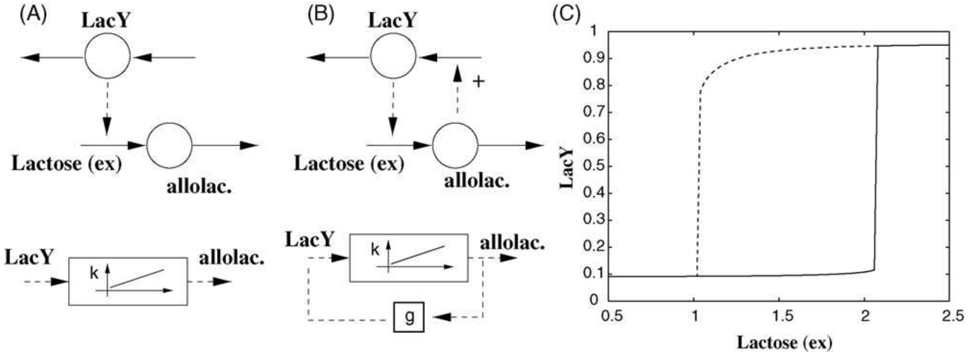


Fig. 10. (A) Open loop for the lactose uptake system. The concentration of the uptaking protein determines the concentration of allolactose. (B) Closed loop. (C) Simulation of the stimulus-response curve. A hysteresis can be detected, that is, the system possesses three steady-states – two stable and one unstable– for a broad range of stimuli.

6.2. *Integral feedback*

An interesting biochemical example for a well known technical feedback structure was discovered in the bacterial chemotaxis. The steady state output of the system (the phospho-rylated form of protein CheY) does not depend on the input – or better here, disturbance of the system – which is an attractant for the bacteria, e.g. a substrate, such as glucose. As shown by Doyle and co-workers, the system can be represented in such a way that an integral feedback can be detected (Yi et al., 2000).

Integral feedback is used frequently in technical control systems to regulate disturbances and to keep the system at a desired setpoint. As can be seen in Fig. 11, systems with integral feedback will respond very differently to either changes of the setpoint or disturbances.

In this context, it is a common approach to linearize the models to apply linear control theory. Furthermore, often Laplace transformations are used to convert the system of o.d.e.′s (a general form is given in Eq. (2)) to a system of algebraic equations. The transformed system now “works” in the frequency domain, that is, the system is analyzed with respect to amplitude amplification and phase shift of a given sinusoidal input function with frequency ω. Linear systems are characterized by the fact, that: (i) the frequency ω will be leveled off while a phase shift can occur and (ii) that the input amplitude will be altered (amplified or damped). The response of the system with respect to setpoint tracking or disturbance regulation can be described by separate transfer functions. These transfer functions are composed of properties of the controller and the controlled system but have to fulfill different tasks: new setpoints have to be reached very fast and precisely, while disturbances have to be eliminated quickly.

From the linear system with a single input u and a single output y, the transfer function P (i.e., an algebraic relationship between in- and output in the frequency domain) for the controlled system can be calculated:

$$Y = C(sI - A)^{-1}bU = PU. \tag{32}$$

With the transfer function for the controller R, output Y (output in the frequency domain) of the closed loop system (Fig. 11) can now be calculated by two parts G_1 and G_2, which are the transfer function for the setpoint tracking W and for the disturbance Z, respectively:

$$Y = G_1 W + G_2 Z \tag{33}$$

Both transfer functions are composed from the individual transfer functions of the controller and the plant. It can be shown that the transfer function

$$S = \frac{1}{1 + PR} \tag{34}$$

plays an important role for the dynamics of the closed loop. From an analysis of the closed loop, it results that if $S = 0$, then $G_2 = 0$ and thus the effect of Z dissapears. However, it can be shown that a conservation relationship is valid, if S is considered over all frequencies ω: the integral over all frequencies is zero. This is shown in Fig. 11 with a cellular network showing nearly an ideal adaptive behavior: enzyme *ez* catalyzes the synthesis

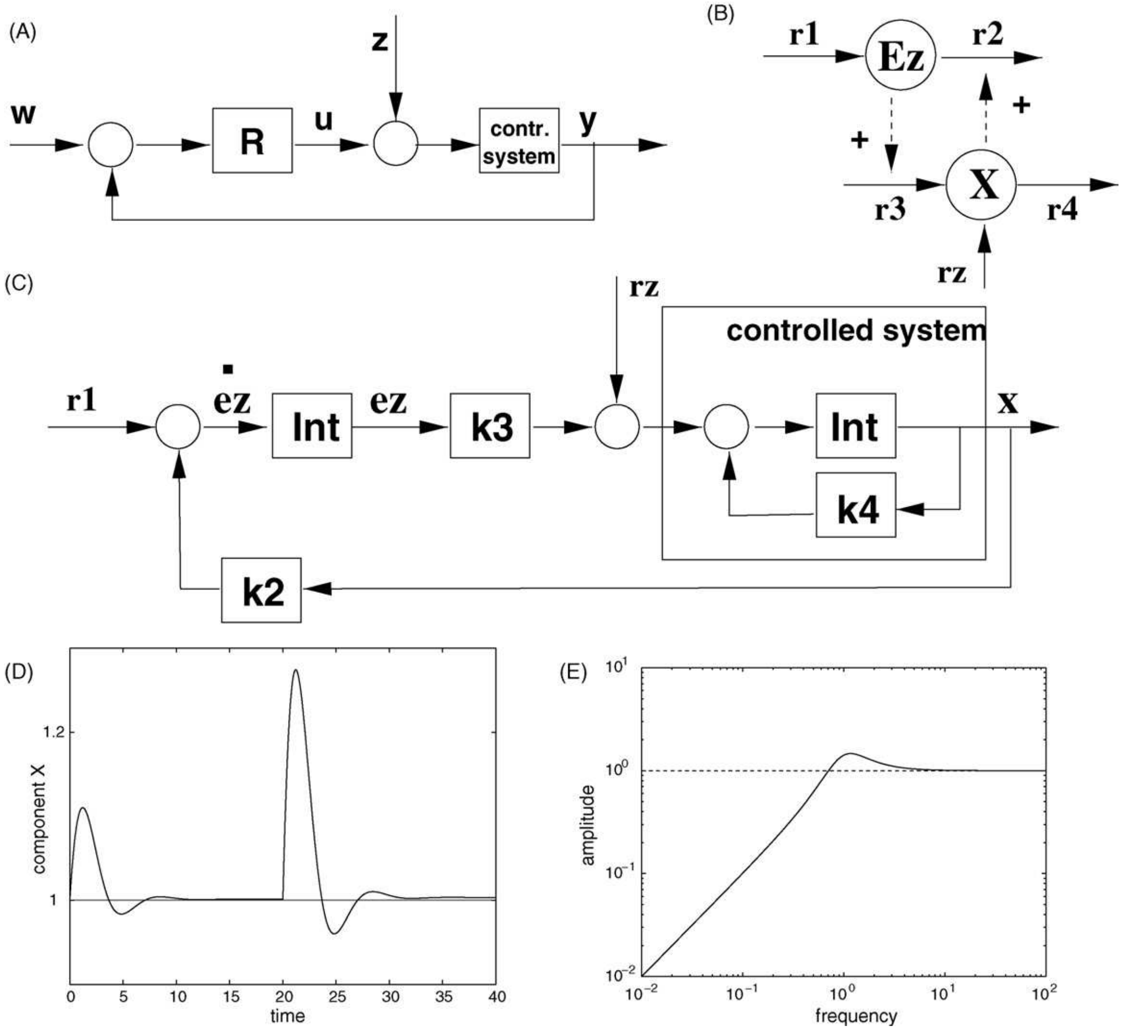


Fig. 11. (A) General scheme of a closed loop system with integral feedback. (B) Reaction network. (C) Scheme of the reaction network as a closed loop. **INT** is integration. When the degradation of *ez* works near saturation, the concentration of *ez* becomes independent from itself. (D) Simulation with two different values of r_z. For the second disturbance, the adaption is not perfect. (E) Plot of the magnitude of S over the frequency range. The area below and above the dashed line are equal.

of metabolite X while X feeds back to enzyme degradation of ez. According to the figure, and assuming a Michaelis-Menten kinetics for protein degradation, the non-linear o.d.e.′s for the system read:

$$\dot{e}z = r_1 - k_2 \frac{xez}{ez + K_e}, \qquad \dot{x} = k_3 ez - k_4 x + r_z \quad (35)$$

with rate r_z as disturbance on X. As can be seen in the Fig. 11, r_1 is the set point and the simple structure can be translated into a feedback control loop with integral feedback, if the degradation of the enzyme works in saturation, that is, the degradation of ez becomes independent of itself. This means that Eq. (35) is simplified to:

$$\dot{e}z = r_1 - k_2 x, \qquad \dot{x} = k_3 ez - k_4 x + r_z. \quad (36)$$

Depending on the strength of the disturbance, the systems adapt more or less very precisely (see Fig. 11). The Figure shows also a so-called Bode plot of $S(\omega)$, that is, the amplification for various frequency values.

Based on the conservation law, a part of the curve has to be larger than 1 (dashed line). Hence, depending on the input frequency, disturbances are damped for low frequencies and are amplified for higher values. This property has been described as the "robust but fragile" nature of cellular systems with integral feedback (Csete and Doyle, 2002). Since, in general, the input function is not a pure sinusoidal function, but, e.g. a step input, the expected time course response can be seen as a mixture of the response over all frequencies. In the simulation in Fig. 11, this can be seen in the time course of *X* that shows an undershoot.

If a system is able to perfectly balance the disturbance it has to possess a subsystem which itself can generate all disturbances. This "internal model principle" is known for a long time in control engineering. In Sontag (2003), it is shown in which way a system with a general structure can be decomposed in such a way. This will help to elucidate the structure of unknown networks that show adaptive behavior.

6.3. Robust control

A hallmark of models that describe dynamical processes in cells is that they show a high degree of uncertainty, since the knowledge is incomplete and the kinetic parameters are often unknown and have to be estimated from experimental data. In control theory, model and parameter uncertainties can be described in the "robust control" framework, see, e.g. Morari and Zafiriou (1988). It seems therefore natural to apply this concept for the analysis of cellular systems, as recently attempted in Kim et al. (2006). Kim and colleagues analyzed the dynamical behavior of Dictyostelium cells by means of robust control theory. The idea is to find parameter variations that destroy the oscillations. It results that the re-formulation of the given model into the robust control framework is difficult and that the calculation of the lower bounds seems to be impractical. However, the robust control approach is a promising approach to cover uncertainties in cellular systems.

7. Conclusions

Engineering sciences provide a bundle of computational tools and theoretical methods that are frequently applied in systems biology research. In this article, we summarize and illustrate recent advances in the field of mathematical modeling and model analysis from an engineering point of view.

A field closely related to systems biology is metabolic engineering, which was established in the eighties based on a quantitative description of biochemical processes together with the possibility to modify organisms by genetic alterations. The goal of metabolic engineering is the improvement of quality and quantity of interesting products in industrial applications. Therefore, the set up of models that are able to describe growth of microorganisms, substrate uptake and product formation was one of the central tasks in early years. With the possibility to modify strains and to introduce plasmids with genetic information of new proteins, the modeling of gene expression, plasmid stability and protein secretion gained importance. Clearly, the developments in metabolic engineering and systems biology will benefit from each other and synergistic effects can be expected (Nielsen and Olsson, 2002).

The elucidation of functionality provided by metabolic, genetic and signaling network structures based on a mathematical description is an important task, which we expect to be developed in the near future. Since the processes that must be described are characterized by uncertainties with respect to the components involved and the kinetic parameters, the set up of models and model analysis that cover such aspects will become very important. One promising way that is under investigation for linear systems is robust control analysis. This framework allows to include very general uncertainties or specific parameter uncertainties. A possible drawback is that the method "works" in the frequency domain and the applicability will probably not be easy.

The development of "simple" models may also be a useful approach. Although, big models comprising a high number of components and interactions between the components are useful, models that focus on crucial points like switches, etc., may make it easier to understand the functionality of a network. However, it is not a trivial task to extract the important players and interactions in a given network with many components, and new methods are necessary. Alternatively, one may start with the observed phenomena, such as oscillations or multiple steady-states: from non-linear dynamics, basic structures are known that are able to reproduce the desired behavior (Tyson et al., 2003); starting from these structures and in combination with the available biological knowledge, meaningful models can be set up, that can help for a better understanding of the observations.

A further crucial point in the near future will be the design of new experiments based on prior knowledge and experimental results. The goal will be to design informative experiments to elucidate the structure of the network and to improve the accuracy of the parameters. Methods proposed so far are based on the formulation with deterministic o.d.e.′s or steady-state equation systems where uncertainties are not included. Therefore, extended methods are required that take into account that the knowledge on some parts of the networks is limited.

Central to the methods described in this article is a focus on model set up and model analysis. However, the work of engineers in classical fields is also focused on problems of synthesis, like designing plants and controllers for technical applications. In biology, an analogous discipline is currently emerging: synthetic biology. Here, as in the case of systems biology, biology challenges engineering tools, and probably classical engineering approaches will have to be extended to cope with the complexity of cellular systems (Andrianantoandro et al., 2006).

Acknowledgments

The authors thank R. Schenkendorf for performing the calculations with the Kalman-Filter, and J. Huard, H. Conzelmann and S. Klamt for critical reading the manuscript, and D. Flockerzi for references on time-scale-based model reduction. J.S.R. acknowledges founding of the German Research Foundation (FOR521).

References

Alon, U., Surette, M.G., Barkai, N., Leibler, S., 1999. Robustness in bacterical chemo-taxis. Nature 397, 168–171, Letters to Editor.

Alves, R., Antunes, F., Salvador, A., 2006. Tools for kinetic modeling of biochemical networks. Nat. Biotechnol. 24, 667–672.

Andrianantoandro, E., Basil, S., Karig, D.K., Weiss, R., 2006. Synthetic biology: new engineering rules for an emerging discipline. Mol. Syst. Biol. 2, E1–E14.

Angeli, D., Jr Ferrell, J.E., Sontag, E.D., 2004. Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. Proc. Natl. Acad. Sci. U.S.A. 101 (7), 1822–1827.

Asprey, S.P., Macchietto, S., 2000. Statistical tools for optimal dynamic model building. Comput. Chem. Eng. 24 (2–7), 1261–1267.

Bettenbrock, K., Fischer, S., Kremling, A., Jahreis, K., Sauter, T., Gilles, E.D., 2006. A quantitative approach to catabolite repression in *Escherichia coli*. J. Biol. Chem. 281 (2578–2584).

Blinov, M.L., Faeder, J.R., Goldstein, B., Hlavacek, W.S., 2004. Bionetgen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. Bioinformatics 20 (17), 3289–3291.

Borisov, N.M., Markevich, N.I., Hoek, J.B., Kholodenko, B.N., 2005. Signaling through receptors and scaffolds: independent interactions reduce combinatorial complexity. Biophys. J. 89 (2), 951–966.

Box, G.E.P., Hill, W.J., 1967. Discrimination among mechanistic models. Technometrics 9, 57–71.

Brun, R., Reichert, P., Kiinsch, H.R., 2001. Practical identifiability analysis of large environmental simulation models. Water Resour. Res. 37 (4), 1015–1030.

Butcher, E.C., Berg, E.L., Kunkel, E.J., 2004. Systems biology in drug discovery. Nat. Biotechnol. 22 (10), 1253–1259.

Chen, B.H., Asprey, S.P., 2003. On the design of optimally informative dynamic experiments for model discrimination in multiresponse nonlinear situations. Ind. Eng. Chem. Res. 42 (7), 1379–1390.

Conradi, C., Saez-Rodriguez, J., Gilles, E.D., Raisch, J., 2005. Using chemical reaction network theory to discard a kinetic mechanism hypothesis. IEE Proc. Syst. Biol. 152 (4), 243–248.

Conzelmann, H., Saez-Rodriguez, J., Sauter, T., Bullinger, E., Allgöwer, F., Gilles, E.D., 2004. Reduction of mathematical models of signal transduction networks: simulation-based approach applied to EGF receptor signaling. Syst. Biol. 1 (1), 159–169.

Conzelmann, H., Saez-Rodriguez, J., Sauter, T., Kholodenko, B.N., Gilles, E.D., 2006. A domain-oriented approach to the reduction of combinatorial complexity in signal transduction networks. BMC Bioinformatics 7, 34.

Cooney, M.J., McDonald, K.A., 1995. Optimal dynamic experiments for bioreactor model discrimination. Appl. Microbiol. Biotechnol. 43, 826–837.

Csete, M.E., Doyle, J.C., March 2002. Reverse engineering of biological complexity. Science 295, 1664–1669 (review).

Cuellar, A.A., Lloyd, C.M., Nielsen, P.F., Bullivant, D.P., Nickerson, D.P., Hunter, P.J., 2003. An overview of CellML 1.1, a biological model description language. Simulation 79 (12), 740–747.

de Jong, H., Geiselmann, J., Hernandez, C., Page, M., 2003. Genetic network analyzer: qualitative simulation of genetic regulatory networks. Bioinformatics 19 (3), 336–344.

Dochain, D., 2003. State and parameter estimation in chemical and biochemical processes: a tutorial. J. Process Contr. 13, 801–818.

Edwards, K., Edagar, T.F., Manousiouthakis, V.I., 1998. Kinetic model reduction using genetic algorithms. Comput. Chem. Eng. 22, 239–246.

Eissing, T., Conzelmann, H., Gilles, E.D., Allgower, F., Bullinger, E., Scheurich, P., 2004. Bistability analyses of a caspase activation model for receptor-induced apoptosis. J. Biol. Chem. 279 (35), 36892–36897.

Ervadi-Radhakrishna, A., Voigt, E.O., 2005. Controllability of non-linear biochemical systems. Math. Biosci. 196, 99–123.

Faeder, J.R., Blinov, M.L., Goldstein, B., Hlavacek, W.S., 2005. Combinatorial complexity and dynamical restriction of network flows in signal transduction. Syst. Biol. 2 (1), 4–15.

Feinberg, M., 1995. The existence and uniqueness of steady states for a class of chemical reaction networks. Arch. Rational Mech. Anal. 132 (4), 311–370.

Fenichel, N., 1979. Geometric singular theory for ordinary differential equations. J. Di. Eqs. 31, 53–98.

Gadkar, K.G., Varner, J., Doyle III, F.J., 2005. Model identification of signal transduction networks from data using a state regulator problem. Syst. Biol. 2 (1), 17–30.

Gerdtzen, Z.P., P, D., Hu, W.-S., 2004. Non-linear reduction for kinetic models of metabolic reaction networks. Metab. Eng. 6, 140–154.

Gilles, E.D., 1998. Network theory for chemical processes. Chem. Eng. Technol. 21 (2), 121–132.

Ginkel, M., Kremling, A., Nutsch, T., Rehner, R., Gilles, E.D., 2003. Modular modeling of cellular systems with ProMoT/Diva. Bioinformatics 19 (9), 1169–1176.

Gonzalez, A., Naldi, A., Sanchez, L., Thieffry, D., Chaouiya, C., 2006. Ginsim: a software suite for the qualitative modelling, simulation and analysis of regulatory networks. Biosystems 84 (2), 91–100.

Hartwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W., 1999. From molecular to modular cell biology. Nature 402 (6761-supp), C47–C52.

Huang, C.F., Ferrell Jr., J.E., 1996. Ultrasensitivity in the mitogen-activated protein kinase cascade. Proc. Natl. Acad. Sci. U.S.A. 93 (19), 10078–10083.

Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19 (4), 524–531.

Ideker, T., Galitski, T., Hood, L., 2001. A new approach to decoding life: systems biology. Annu. Rev. Genomics Hum. Genet. 2 (3), 343–372.

Janes, K.A., Lauffenburger, D.A., 2006. A biological approach to computational models of proteomic networks. Curr. Opin. Chem. Biol. 10, 73–80.

Jong, H., 2002. Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol 9-1, 67–103.

Joshi, M., Seidel-Morgenstern, A., Kremling, A, 2006. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. Metab. Eng. 8 (5), 447–455.

Katare, S., Bhan, A., Caruthers, J.M., Delgass, W.N., Venkatasubramanian, V., 2004. A hybrid genetic algorithm for efficient parameter estimation of large kinetic models. Comput. Chem. Eng. 28, 2569–2581.

Kim, J., Bates, D.G., Postlethwaite, I., Ma, L., Iglesias, P.A., 2006. Robustness analysis of biochemical network models. Syst. Biol. 153 (3).

Kitano, H., 2002a. Computational systems biology. Nature 420, 206–210.

Kitano, H., 2002b. Systems biology: a brief overview. Science 295 (5560), 1662–1664.

Klamt, S., Saez-Rodriguez, J., Gilles, E.D., 2006. Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC Syst. Biol. 1 (2).

Kremling, A., Fischer, S., Gadkar, K., Doyle, F.J., Sauter, T., Bullinger, E., Allgöwer, F., Gilles, E.D., 2004a. A benchmark for methods in reverse engineering and model discrimination: problem formulation and solutions. Genome Res. 14 (9), 1773–1785.

Kremling, A., Fischer, S., Sauter, T., Bettenbrock, K., Gilles, E.D., 2004b. Time hierarchies in the *Escherichia coli* carbohydrate uptake and metabolism. Biosystems 73 (1), 57–71.

Kremling, A., Jahreis, K., Lengeler, J.W., Gilles, E.D., 2000. The organization of metabolic reaction networks: a signal-oriented approach to cellular models. Metab. Eng. 2 (3), 190–200.

Kutalik, Z., Cho, K.-H., Wolkenhauer, O., 2004. Optimal sampling time selection for parameter estimation in dynamic pathway modeling. Biosystems 75, 43–55.

Laurent, M., Kellershohn, N., 1999. Multistability: a major means of differentiation and evolution in biological systems. Trends Biochem. Sci. 24 (11), 418–422.

Ljung, L., 1999. System Identification: Theory for the User, second ed. Prentice Hall PTR.

Marquardt, W., 1996. Trends in computer aided modeling. Comput. Chem. Eng. 20, 591–609.

Mattsson, S.E., Elmqvist, H., Otter, M., 1998. Physical system modeling with modelica. Control Eng. Pract. 6, 501–510.

Maurya, M.R., Katare, S., Patkar, P.R., Rundell, A.E., Venkatasubramanian, V., 2006. A systematic framework for the design of reduced-order models for signal transduction pathways from a control theoretic perspective. Comput. Chem. Eng. 30, 437–452.

Maurya, M.R., Bornheimer, S.J., Venkatasubramanian, V., Subramaniam, S., 2005. Reduced-order modelling of biochemical networks: application to the GTpase-cycle signalling module. IEE Proc. Syst. Biol. 152 (4), 229–242.

Moles, C.G., Mendes, P., Banga, J.R., 2003. Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Res. 13, 2467–2474.

Morari, M., Zafiriou, E., 1988. Robust Process Control. Prentice Hall.

Munack, A., 1992. Some improvements in the identification of bioprocesses. In: Karim, M.N., Stephanopoulos, G. (Eds.), Modeling and Control of Biotechnical Processes, 1FAC Symposia Series. 1FAC, Pergamon Press, pp. 89–94.

Nielsen, J., Olsson, L., 2002. An expanded role for microbial physiology in metabolic engineering and functional genomics: moving towards systems biology. FEMS Yeast Res. 2, 175–181.

Noble, D., 2002. Modelling the heart: insights, failures and progress. Bioessays 24, 1155–1163.

Papin, J.A., Reed, J.L., Palsson, B.O., 2004. Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. Trends Biochem. Sci. 29 (12), 641–647.

Pawson, T., Nash, P., 2003. Assembly of cell regulatory systems through protein interaction domains. Science 300 (5618), 445–452.

Powers, J.M., Singh, S., Paolucci, S., 2002. On slow manifolds of chemically reactive systems. J. Chem. Phys. (4), 1482–1496.

Rodriguez-Fernandez, M., Mendes, P., Banga, J.R., 2006. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. Biosystems 83, 248–265.

Saez-Rodriguez, J., Conzelmann, H., Sauter, T., Kholodenko, B.N., Gilles, E.D., 2005a. Domain-oriented and modular approaches to the reduction of mathematical models of signaling networks. In: Kummer, U., Pahle, J., Surovtsova, I., Zobeley, J. (Eds.), 4th Workshop on Computation of Biochemical Pathways and Genetic Networks. Logos-Verlag Berlin, September, pp. 13–20.

Saez-Rodriguez, J., Kremling, A., Gilles, E.D., 2005b. Dissecting the puzzle of life: modularization of signal transduction networks. Comput. Chem. Eng. 29 (3), 619–629.

Saez-Rodriguez, J., Hammerle-Fickinger, A., Dalai, O., Klamt, S., Gilles, E.D., Conradi, C. On the multistability of signal transduction motifs, submitted.

Saez-Rodriguez, J., Kremling, A., Conzelmann, H., Bettenbrock, K., Gilles, E.D., 2004. Modular analysis of signal transduction networks. IEEE Control Syst. Mag. 24 (4), 35–52.

Saez-Rodriguez, J., Mirschel, S., Hemenway, R., Klamt, S., Gilles, E.D., Ginkel, M., 2006. Visual setup of logical models of signaling and regulatory networks with promot. BMC Bioinformatics 7 (1), 506.

Saraiya, U., Duca, K., North, C., 2005. Visualization of biological pathways: requirements analysis, systems evaluation and research agenda. Inf. Visual. 4 (3).

Sauro, H.M., Ingalls, B., 2004. Conservation analysis in biochemical networks: computational issues for software writers. Biophys. Chem. 109 (1), 1–15.

Schmidt, H., Jirstrand, M., Wolkenhauer, O., 2006. Information technology in systems biology. it—Inf. Technol. 48 (Jahrgang(3)).

Schoeberl, B., Eichler-Jonsson, C., Gilles, E.D., Muller, G., 2002. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nat. Biotechnol. 20 (4), 370–375.

Segel, L.A., Slemrod, M., 1989. The quasi-steady-state assumption: a case study in perturbation. SIAM Rev. 31 (3), 446–462.

Sontag, E.D., 2003. Adaption and regulation with signal detection implies internal model. Syst. Control Lett. 50, 119–126.

Sontag, E.D., 2004. Some new directions in control theory inspired by systems biology. Syst. Biol. 1 (1), 9–18.

Sontag, E.D., 2005. Molecular systems biology and control. Eur. J. Control 11, 396–435.

Sorger, P.K., 2005. A reductionist's systems biology: opinion. Curr. Opin. Cell Biol. 17, 9–11.

Stephanopoulos, G., Henning, G., Leone, H., 1990. MODEL.LA. A modeling language for process engineering. I. The formal framework. Comput. Chem. Eng. 14, 813–846.

Tyson, J.J., Chen, K.C., Novak, B., 2003. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr. Opin. Cell Biol. 5 (12), 221–231.

Unger, J., Kroener, A., Marquardt, W., 1995. Structural analysis of differential–algebraic equation system—theory and application. Comput. Chem. Eng. 19, 867–882.

Vacheva, I., Eils, R., 2006. Computational systems biology platforms. it—Inf. Technol. 48 (Jahrgang(3)).

Wend, H.D., 1993. Strukturelle Analyse linearer Regelsysteme. Oldenbourg.

Wiggins, S., 1994. Normally Hyperbolic Invariant Manifolds in Dynamical Systems. Springer Verlag, New York.

Wolf, D.M., Arkin, A.P., 2003. Motifs, modules and games in bacteria. Curr. Opin. Microbiol. 6 (2), 125–134.

Wolkenhauer, O., 2001. Systems biology: the reincarnation of systems theory applied in biology? Briefings Bioinform. 2 (3), 258–270.

Xiong, W., Ferrell Jr., J.E., 2003. A positive-feedback-based bistable 'memory module' that governs a cell fate decision. Nature 426 (6965), 460–465.

Yi, T.-M., Huang, Y., Simon, M.I., Doyle, J., 2000. Robust perfect adaption in bacterial chemotaxis through integral feedback control. PNAS 97 (9), 4649–4653.

Zagaris, A., Kaper, H.G., Kaper, T.J., 2004. Analysis of the computational signal perturbation reduction method for chemical kinetics. J. Nonlinear Sci. 14, 59–91.

Zobeley, J., Lebiedz, D., Kammerer, J., Ishmurzin, A., Kummer, U., 2005. A new time-dependent complexity reduction method for biochemical systems. Lecture Notes Comput. Sci. 3380, 90.

Towards Whole Cell "in Silico" Models for Cellular Systems: Model Set-up and Model Validation

Andreas Kremling, Katja Bettenbrock, Sophia Fischer, Martin Ginkel,
Thomas Sauter, and Ernst Dieter Gilles

Max-Planck-Institute Dynamics of Complex Technical Systems, Sandtorstr. 1,
39106 Magdeburg, Germany, `kre@mpi-magdeburg.mpg.de`

Abstract. Based on recent developments for new measurement technologies that enable researches to get quantitative information on intracellular processes, the set-up of very detailed models describing metabolism as well as regulatory networks becomes very popular. However, biochemical networks are rather complex including many feed-forward and feedback loops. In this contribution we propose an interdisciplinary approach including the computer based set-up of models and strategies to validate the models with apparent experiments. This approach will offer a new way to meaningful models that can be used to make simulation experiments analogous to real laboratory experiments. The approach is applied to the bacterium *Escherichia coli*: A mathematical model to describe carbon catabolite repression is developed and in part validated. The model is aggregated from functional units describing carbohydrate transport and degradation. These units are members of the *crp* modulon and are under control of a global signal transduction system which calculates the signals that turn on or off gene expression for the specific enzymes. Problems of parameter identification for whole cell models are discussed.

1 Introduction

Recent efforts for a better understanding of cellular systems have resulted in multidisciplinary research alliances (mainly in the US) where researchers from biology, informatics and systems engineering work together. The aim of these initiatives is to model complex biological systems in such a way that experiments can be performed with the help of a computer analogous to experiments in a real laboratory. Even biological working groups have recognized the need of frameworks for a quantitative description of cellular processes and the importance of integrating experimental and theoretical/computational approaches [3]. Central in the work of biologists is the definition of 'modules' or 'functional units' as a critical level of cellular organization. A concept stating that cellular metabolism is structured in functional units which could be used

L. Benvenuti, A. De Santis, and L. Farina (Eds.): Positive Systems, LNCIS 294, pp. 95–102, 2003.

in modeling has been proposed [8]. The concept is based on the definition of submodels with characteristic features. The submodels are implemented in the computer tool PROMOT [2] that provides a graphical user interface where the submodels can be chosen from the model library and can be connected to other submodels. Based on the concept a mathematical model considering functional units describing uptake and metabolism of a number of carbohydrates in *Escherichia coli* is set up and analyzed [7, 6]. A novel experimental approach using isogenic mutant strains, i.e. a number of strains derived from one wild-type with a clearly defined background, was used to determine yet unknown or uncertain parameters.

The intention of the contribution at hand is twofold: (i) It summarizes the current state of the model and describes the model extensions having taken place recently. (ii) Discuss the problems on the way to whole-cell-models. Up to now only a few models are available that describe parts of the metabolism, and, simultaneously, provide adequate data for model validation.

2 Modeling concept

Our approach is based on the analysis and the combination of the available knowledge on metabolism, signal transduction and cellular control with system-theoretical methods. The modeling procedure thus has to be based on the molecular structure of the functional units in such a way that a cellular unit is represented by an equivalent mathematical submodel. This modular approach is a new feature in the mathematical modeling procedure and guarantees a high transparency for biologists and engineers.

The basis of the framework is the definition of a complete set of elementary modeling objects. They should be disjunct with respect to the biological knowledge they comprise to prevent overlapping. The modeling process proceeds along two coordinates: a structural and a behavioral coordinate. The structural coordinate represents a progressive combination and linkage of elementary modeling objects to higher aggregated model structures. Higher aggregated model structures are called *functional units*. Modeling along the behavioral coordinate means that to each of the elementary modeling objects equations have to be assigned.

Functional units are defined according to three biological motivated criteria: (i.) A common physiological task. (ii.) A common genetic unit. The genes for all enzymes of a functional unit are organized in genetical units (operons, regulons and modulons) and/ or in a hierarchical structure. (iii.) A common signal transduction network. All elements of a functional unit are interconnected within a common signal transduction system. The signal flow across the unit border ("cross-talk" or "cross- regulation") is small compared to the information exchange within the unit, such that the coordinated response to a common stimulus ("stimulon") helps to identify the members of a unit.

3 Modeling environment ProMoT/Diva

PROMOT enables the use of object-oriented modeling techniques including encapsulation, aggregation, and inheritance. In PROMOT, dynamic models are built by aggregating structural and behavioral modeling entities. The modeling entities in PROMOT are organized in an object-oriented class hierarchy with multiple inheritance. This concept from computer science was adopted to allow a better organization of complex modeling libraries and flexible implementation of large scale models. Every entity in this hierarchy inherits all parts and attributes from its respective super-classes. With this method abstraction is possible and more general and reusable entities can be formed. PROMOT provides a special modeling language as well as a graphical user interface (GUI) for interactive modeling. The modeling tool, as well as the simulation environment, are developed under different Unix-derived operating systems, however the main platform is Linux. The kernel of the system is implemented as a modeling server in object-oriented Common Lisp (using the Common Lisp Object System CLOS).

The numerical analysis of the models is done with the simulation environment DIVA. Within DIVA many different numerical computations are possible, based on facilities to calculate the steady state and dynamic behavior of the model using non-linear equation solvers and integrators. For metabolic models 2 methods are of special interest: (i) Parameter analysis with respect to experimental data. (ii) Identification of parameters and model accuracy.

4 Model for carbohydrate uptake in *E. coli*

Figure 1 shows the modeling objects with relevant in- and outputs. The global signal transduction system comprises the phosphoenolpyruvate (PEP)-dependent: glucose phosphotransferase system (PTS), the synthesis of cAMP, and the interaction of the cAMP·Crp complex with the specific DNA binding sites. Besides its sensory function, the PTS is the main glucose uptake system. Uptake of glucose by other transport systems and uptake of lactose, galactose, glycerol, and sucrose[1] are described in separate functional units. For the bacterial physiology it is well known that the control of transcription initiation is the main control principle. Therefore control of post transcriptional and of translation processes are not modeled in detail here.

One main feature of the model is the hierarchical structure of regulatory network. Based on the analysis of molecular interactions of proteins with DNA binding sites a new approach to develop mathematical models describing gene expression is applied. Detection of hierarchical structures in metabolic networks can be used to decompose complex reaction schemes. This is achieved by assigning each regulator protein to one level in the hierarchy. Signals are

---

[1] *E. coli* is not able to grow on sucrose; therefore an engineered strain is used here.

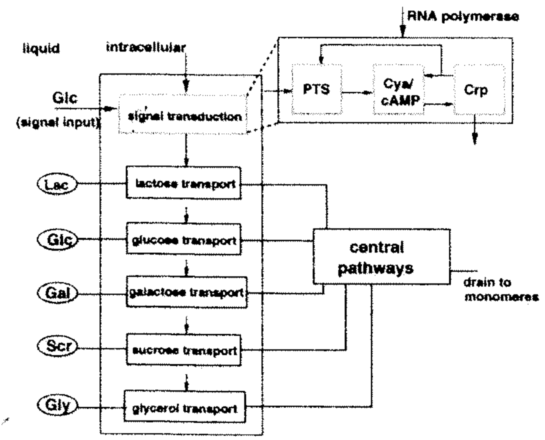


Fig. 1. The modules of the *crp*-modulon. The PTS is the sensory system. Protein EIIA and its phosphorylated form P~EIIA are the main outputs. The output signal from the Crp submodel describes the transcription efficiency of the genes and operons under control of the cAMP·Crp complex. Since there is no control of translation included in the model, the output is a measure for the rate of synthesis of the enzymes. A number of pathways are under control of the signal transduction pathway: glucose, lactose, galactose, sucrose and glycerol transport.

then transduced from the top level to the lower level, but not vice versa. The top level in the model is represented by the RNA polymerase, the second level by the global regulator Crp and the lowest level by e.g. the lactose repressor LacI. The overall comprises 63 states (ode's/ algebraic equations) and 251 parameters (see Table 1).

5 Model validation

Model validation is an essential part in modeling. In order to validate a model it is necessary to compare predictions given by the model with results from real experiments. The experiments have to be designed in a way that the measured data contain information about the different functional units included in the model. A strategy to identify parameters in very large models for cellular systems is still missing. Although the functional units are only weakly coupled to each-other, a number of problems arise during parameter identification:

- Up to now it is not clear if the available software is capable to solve all parameter fitting problems. In the present study we used maximal 10 experiments in one fit. A further problem is the finding of the most suitable state that should be measured to get the best information for the fit. From our modular approach results the idea using only one representative for

unit. However, the development of measurement methods for interesting metabolites/ proteins is still expensive.
- Although the units are only weakly connected, it is useful to analyze the interconnection, e.g. with a sensitivity analysis. However, this requires some starting values for the parameters and the states in the model. The chosen values might be far away from the real values and may lead to incorrect conclusions.
- Most experiments are performed with batch experiments. Here, the specific growth rate is maximal. The identification of Michaelis-Menten K_M values however requires low substrate concentration and therefore other growth rates are required. Low growth rates may lead to stress responses of the organism and the model is normally not able to describe this situation.
- Measurements of extracellular components are normally available and the uptake rate may be calculated. The simplest kinetic expression requires also information on the amount of enzyme. Since the amount may change during the experiment - this information is available seldom - parameters for the transport step can hardly be found.
- Incorporation of quantitative knowledge. Sometimes knowledge on the range of concentration of metabolites is available or, based on array data, knowledge that a gene have been expressed is measured.

A brief description of the theoretical background for parameter identification used so far is given in the following.

5.1 Parameter identification

To solve the equations the simulation environment DIVA was used. The integration algorithm DDASAC [1] has been chosen. To identify the model parameters the following approach is used: (i) Starting with parameters from literature, the model is analyzed with the method of Hearne [4], calculating a combination of parameters which have a maximal effect on the interesting states (states for which measured data are available). This sensitivity analysis gives a first impression on the sensitive parameters. (ii) Together with the measured data and the Fisher information matrix it was checked, if the sensitive parameters could be estimated. Applying a method introduced by [9] a set of parameters from the sensitive parameters were determined which could be estimated together with a given minimal variance γ. For the glucose/ lactose diauxic experiment, $m = 8$ states were measured (extracellular glucose and lactose, biomass and intracellular LacZ activity which is used as measure for LacZ concentration, galactose, acetate, cAMP in the medium, degree of phosphorylation of EIIA) and it can be expected that the parameters which can be estimated are related to the respective transport units. (iii) Parameter estimation: The whole model is given in the form

$$\dot{x} = f(x, p)\,, \tag{1}$$

with states x and parameter vector p. For a subset of the states $i = 1, m$ measurement data are available (z_{ik}) at time point t_k ($k = 1, N$). The aim of the parameter identification is to minimize the objective function $\Phi(p)$

$$\Phi(p) = \sum_{k=1}^{N} \sum_{i=1}^{m} (x_i(x_o, p, t_k) - z_{ik})^2 . \qquad (2)$$

To solve the optimization problem the SQP (**S**equential **Q**uadratic **P**rogramming) algorithm E04UPF from the NAG library was used.

5.2 Experimental approach

Published measurements dealing with diauxic experiments are often not well suited for the validation of mathematical models. The strains that have been used are not isogenic and measurements of different groups are difficult to compare. As the genetic background is often only poorly defined it is almost impossible to consider the genetic variations in model validation. In addition, the experimental setup is often not well documented or the design of the experiments is not useful for model validation.

A biological system can be characterized in different ways. One possibility is to stimulate the system by (i) changing the external conditions like growth medium, substrate or temperature, (ii) using different culture conditions like batch, continuous fermentations, deflection from steady state by a pulse, and transient conditions, (iii) by introducing a mutation and/or (iv) by alter the intracellular state of the cell e.g. by using different pre-culture conditions. Figure 2 summarizes all types of stimulation used in this study with respect to changes of the specific growth rate μ. In batch cultivations only information during growth with the maximal growth rate is obtained. Steady-state conditions are reached only after a long time period that may cause problems due to genetic alterations and due to substrate limiting conditions. If steady-state conditions are reached, pulse experiments can be performed to analyze very fast kinetics. As a wild-type strain LJ110, a well characterized derivative of the

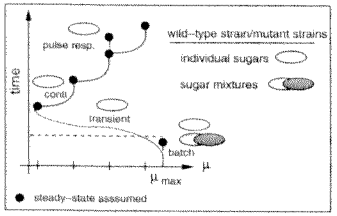


Fig. 2. Summary of experimental conditions used for parameter identification. See text for explanation.

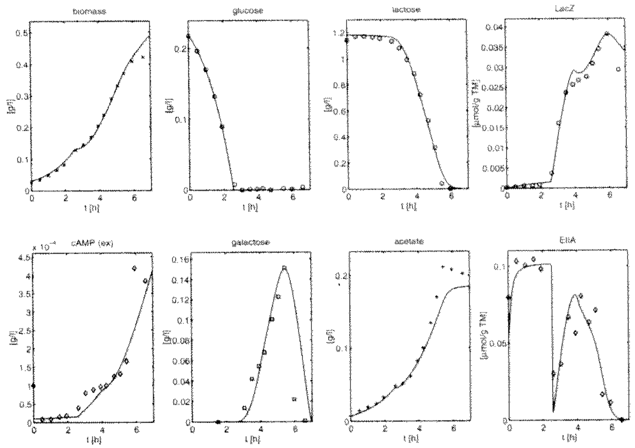


Fig. 3. Time course of measured states (solid) and experimental data (symbols) for a selected experiment with the wild-type strain LJ110 [5].

E. coli K-12 reference strain W3110 was chosen. Mutations were introduced in *cyaA*, *lacI*, *dgsA* and *ptsG*, i.e. in genes important in signal transduction influencing diauxic behavior. By characterizing the wild-type and these mutants with respect to growth on different carbohydrates and especially by recording time series of states during diauxic growth we were able to get enough measurements to estimate a relatively high number of parameters although few different states were measured.

5.3 Results

Based on the available measurements and the experiments performed a number of parameters could be estimated. Table 1 summarizes the findings for all functional units. Figure 3 shows exemplarily an experiment with the wild-type strain LJ110 when glucose and lactose are present in the medium.

6 Conclusion

The present study marks a starting point to set up whole cell models. A detailed model for carbohydrate uptake and metabolism with focus on the cellular control was developed and in part validated. Problems for parameter identification are the lack of consistent experimental data and the uncertainness about the choice of the measured quantity with respect of their importance for the fitting. Here, the development and application of new technologies like cDNA-arrays and proteomics will help to come to better solutions. A problem not addressed here is model structure identification. Current work focuses on the development of methods for experimental design if two or more different

Table 1. Summary of functional units, number of parameters and number of estimated parameters. About 20 different experiments are used for parameter fitting. [a] Parameters estimated with Metabolic Flux Analysis.

module name	param.	param. estimated	number of states	type
PTS (general)	21	9	9	ODE
PTS Glc	12	4	1	ODE
Cya	9	2	2	ODE
Crp	17	3	1	ODE
2nd Glc transporter	18	3	3	ODE
Lac transporter	16	7	4/2	ODE/ algebraic
Scr transporter	26	9	6	ODE
Gly transporter	24	5	5	ODE
Gal transporter	43	4	11/2	ODE/ algebraic
Catabolic reactions	51	11	8	ODE
Monomer synthesis	7	4^a+3	1	ODE
Liquid phase	7	5	8	ODE

References

1. M. Caracotsios and W. E. Stewart. Sensitivity analysis of initial value problems with mixed odes and algebraic equations. *Computers and Chemical Engineering*, 9(4):350–365, 1985.
2. M. Ginkel, A. Kremling, T. Nutsch, R. Rehner, and E. D. Gilles. Modular modeling of cellular systems with ProMoT/Diva. *Bioinformatics*, 2003. In press.
3. L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(Supp.):C47 – C52, 1999.
4. J. W. Hearne. Sensitivity analysis of parameter combinations. *Appl. Math. Modelling*, 9:106–108, 1985.
5. A. Kreming, K. Bettenbrock, S. Fischer, K. Jahreis, T. Sauter, and E.D. Gilles. Mathematical modeling of carbohydrate uptake systems in *Escherichia coli*: I. Growth under unlimited conditions. 2003. Submitted.
6. A. Kremling, K. Bettenbrock, B. Laube, K. Jahreis, J.W. Lengeler, and E.D. Gilles. The organization of metabolic reaction networks: III. Application for diauxic growth on glucose and lactose. *Metab. Eng.*, 3(4):362–379, 2001.
7. A. Kremling and E.D. Gilles. The organization of metabolic reaction networks: II. Signal processing in hierarchical structured functional units. *Metab. Eng.*, 3(2):138–150, 2001.
8. A. Kremling, K. Jahreis, J.W. Lengeler, and E.D. Gilles. The organization of metabolic reaction networks: A signal-oriented approach to cellular models. *Metab. Eng.*, 2(3):190–200, 2000.
9. C. Posten and A. Munack. On-line application of parameter estimation accuracy to biotechnical processes. In *Proceedings of the American Control Conference*, volume 3, pages 2181–2186, 1990.

Guaranteed Parameter Estimation for Cooperative Models

Michel Kieffer and Eric Walter

Laboratoire de Signaux et Systèmes – CNRS – Supélec – Université Paris-Sud
Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France,
`{kieffer,walter}@lss.supelec.fr`

Abstract. The parameters of cooperative models are estimated in a bounded-error context, *i.e.*, all uncertain quantities are assumed to be bounded, with known bounds. Guaranteed estimation is then the characterization of the set of *all* parameter vectors that are consistent with the model and experimental data, given these bounds. Interval techniques provide an approximate but guaranteed enclosure of this set. No parameter vector consistent with the experimental data and model structure can be missed, so this approach bypasses the structural identifiability study required by the usual approaches based on the local optimization of some cost function.

1 Introduction

This paper is about guaranteed estimation of the parameters of cooperative systems from experimental measurements. Estimation is performed in a bounded-error context, *i.e.*, all uncertain quantities (measurement noise, parameters to be estimated) are assumed to be bounded, with known bounds. In this context, parameter estimation may be formulated as finding the *set* of all parameter vectors that are consistent with the parametric model and experimental data, given the error bounds. Interval techniques provide an approximate but guaranteed enclosure of this set between two *subpavings*, *i.e.*, union of non-overlapping boxes. The approximation is guaranteed, as no consistent parameter vectors can be missed. Moreover, the precision of the approximation can be tuned by the user. With such techniques, no prior identifiability study is required. For example, a solution set consisting of two or more disconnected subsets may correspond to a model that is only locally identifiable. So far, this approach was mainly applied to models for which an analytical expression of the solution as a function of the parameters was available. This was because guaranteed integration of differential equations is very pessimistic when the parameter vector is only known to belong to some potentially large interval vector.

The purpose of this paper is to show that the concept of cooperativity makes it possible to extend the methodology to a very large class of

A Benchmark for Methods in Reverse Engineering and Model Discrimination: Problem Formulation and Solutions

Andreas Kremling,[1,5] Sophia Fischer,[1] Kapil Gadkar,[2] Francis J. Doyle,[2]
Thomas Sauter,[3] Eric Bullinger,[4] Frank Allgöwer,[4] and Ernst D. Gilles[1,3]

[1] *Systems Biology Group, Max-Planck-Institut für Dynamik komplexer technischer Systeme, 39106 Magdeburg, Germany;* [2] *Department of Chemical Engineering, University of California–Santa Barbara, Santa Barbara, California 93106, USA;* [3] *Institute for System Dynamics and Control Engineering and* [4] *Institute for Systems Theory in Engineering, University of Stuttgart, 70550 Stuttgart, Germany*

A benchmark problem is described for the reconstruction and analysis of biochemical networks given sampled experimental data. The growth of the organisms is described in a bioreactor in which one substrate is fed into the reactor with a given feed rate and feed concentration. Measurements for some intracellular components are provided representing a small biochemical network. Problems of reverse engineering, parameter estimation, and identifiability are addressed. The contribution mainly focuses on the problem of model discrimination. If two or more model variants describe the available experimental data, a new experiment must be designed to discriminate between the hypothetical models. For the problem presented, the feed rate and feed concentration of a bioreactor system are available as control inputs. To verify calculated input profiles an interactive Web site (http://www.sysbio.de/projects/benchmark/) is provided. Several solutions based on linear and nonlinear models are discussed.

The analysis of metabolic and regulatory pathways with mathematical models contributes to a better understanding of the behavior of metabolic processes (Kitano 2000). The setup of the structure of the model, that is, the stoichiometry of the biochemical reaction network, is mainly based on data from database systems or from literature. Recent efforts in measurement technologies like cDNA array data or 2D-gel electrophoresis (Ideker et al. 2001) will enable researchers to produce time courses of several substances from inside the cell. Given such data, a challenging task is to identify the underlying structure of the network ("reverse engineering") and—if two or more model structures are suited to describe the experimental data—to design new experiments that will allow discrimination between the model candidates. Further problems include identifiability of the model parameters, sensitivity of the parameters, and metabolic design (Stelling et al. 2001).

The main focus of work in the field of reverse engineering lies on the identification of genetic networks, that is, in which way transcription factors are connected to the respective genes. The methods used are based on a steady-state description (Tegner et al. 2003) or on Boolean networks (D'haesseleer et al. 2000; Repsilber et al. 2002). Using time-lagged-correlation matrices (Arkin and Ross 1995; Arkin et al. 1997) or genetic programming techniques (Koza et al. 2001), networks could also be reconstructed if time courses of selected state variables were available.

In contrast to the top-down approach represented by the reverse engineering techniques, the bottom-up approach starts with a mathematical model for genetic and metabolic networks based either on biochemical data from databases or on "cartoons" from literature. One major problem here is the estimation of uncertain or even unknown kinetic parameters, that is, the problem of parameter identification, that covers several tasks. (1) Identifiability: Simply speaking, identifiability is concerned with the following question. Given a particular model for a system and an input–output experiment, is it possible to uniquely determine the model parameters (Faller et al. 2003; Zak et al. 2003)? (2) Parameter estimation: Using optimization methods, a set of parameters is determined in such a way that the difference between the experimentally measured output and the predictive output of the mathematical model becomes minimal (Moles et al. 2003). (3) Finally, the accuracy of the parameters has to be calculated. This is normally done by determining the confidence limits of the estimated parameters (Faller et al. 2003; Swameye et al. 2003). To apply statistical methods for this purpose, a large amount of data is required. On the other hand, using the Fisher-Information-Matrix (see below), only a lower bound for the variances of the parameters can be obtained (Ljung 1999; Banga et al. 2002). This lower bound would be reached if the model equations were linear in the parameters, which is normally not the case. To overcome both problems, an alternative method, the bootstrap method (Press et al. 2002), could be applied.

If two or more model variants are available describing the same experimental observations, methods are available to design new experiments that allow us to discriminate between the variants. Early approaches are described in the literature (e.g., Box and Hill 1967; Munack 1992; Cooney and McDonald 1995). The key idea is to find an input profile that maximizes the difference of the outputs of the competing models. In a series of papers, Asprey and coworkers have developed methods to maximize the outputs of the system (Asprey and Macchietto 2000; Chen and Asprey 2003). This is achieved by using an extended weighting matrix including the variances of the measured state variables and the variances and the sensitivities of the parameters. In Chen and Asprey (2003), several methods for model discrimination are also reviewed.

Here, in silico experimental data for an organism growing in a chemostat as shown in Figure 1 are presented. For this purpose, a computer model was set up based on a fictive network struc-

---

[5] **Corresponding author.**
E-MAIL kremling@mpi-magdeburg.mpg.de; FAX 391-6110526.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1226004.

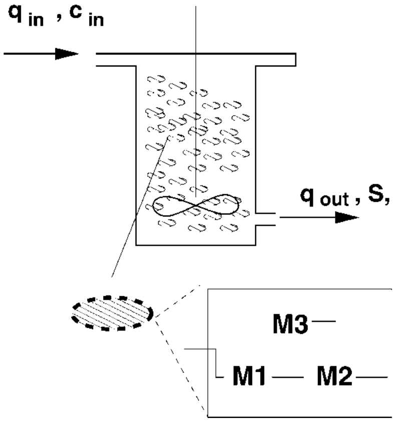


Figure 1 Scheme of the bioreactor. Inputs are flow rates q_{in}, q_{out}, and feed concentration c_{in}. Biomass is assumed to be homogeneously distributed in the reactor. The structure of the biochemical reaction network is unknown and must be identified.

ture. Parameters are chosen in such a way that a realistic behavior could be observed. After reaching a steady state, the flow rates q_{in}, q_{out} as well as the concentration of the substrate in the feed c_{in} are changed. Measurements are available for three metabolites, M1, M2, and M3, representing a small biochemical network of the organism, and for biomass B and substrate S. Because different algorithms for parameter estimation are already described in the literature (Moles et al. 2003), this contribution focuses on the accuracy of the parameters by comparing two methods for determining the variance of the parameters.

In the next section several problems are formulated to apply strategies in the field of reverse engineering and model discrimination. This paper focuses on different methods for model discrimination. For this purpose, two model variants are set up and parameters are estimated. The paper is written for the interested biological researcher and represents possibilities based on a system-theoretical approach. It will be shown that for the given problem it is not necessary to construct several mutant strains, which is often a time-consuming task, but instead, the application of system-theoretical methods using only control inputs available for a bioreactor system is sufficient to provide satisfactory results. Applications for these methods can be found frequently in the field of molecular and cell biology. Considering signal transduction pathways, open questions concern the mechanism of action of the stimulus, cross-talk phenomena, that is, the interaction of separated signal transduction units, and type of control, for example, control of activity or of synthesis of the components involved. Further applications are concerned with the choice of the correct kinetic description for a biochemical reaction (Asprey and Macchietto 2000) or with the distribution of metabolic fluxes in complex networks (Kremling et al. 2001).

METHODS

Benchmark Problem

Problem Formulation

Based on the measurement of components (intra- and extracellular) or expression data, the network structure has to be identified, that is, the interconnections between the given components have to be detected.

If two or more model variants can describe the available experimental data, the design of a new experiment is required to select the most feasible model structure. For larger submodels for cellular systems, measurements are not available for all state variables. Moreover, the development of new measurement techniques is very time consuming. Hence, strategies that require a lesser number of state variables to be measured and moreover strategies that identify these state variables are advantageous. To design a new experiment, inputs and outputs must be chosen in such a way that parameters can be identified. Furthermore, parameters can only be estimated with high accuracy if the control inputs direct them into sensitive regions.

The problem could also be used as a study in metabolic modeling for students to illustrate methods in model setup, model analysis, and experimental design.

Starting Conditions and Data Generation

Figure 2 shows time courses of metabolite concentrations *M*1, *M*2, and *M*3 as well as the time courses of biomass concentration *B* and substrate concentration *S*. The conditions during the chemostat experiment are summarized in Table 1. The molar mass for the substrate used is 342.3 g/mol. The initial conditions for biomass and substrate are 0.1 g/L and 2.0 g/L, respectively. The volume of the bioreactor was held constant at 1.0 L for the given time series (the maximal working volume of the reactor is V_{max} = 5.0 L).

Measurements are sampled every 2 h. To allow realistically complex behavior, the following procedure was used. A set of kinetic parameters was chosen for the (hidden) network. "Experimental data" (time profiles of substrate, biomass, and metabolites) were generated by simulation of this hidden network with the abovementioned initial conditions. With a random number *rand*, the absolute values of the state variables *x* were modified according to $\hat{x} = x(1 + rand)$, where *rand* is normally distributed with mean value $\bar{m} = 0$, and the standard deviation $\sigma = 0.1$.

With the information given so far, the problem of network identification can be solved.

For the problem of model discrimination, the following additional information can be used.

- Metabolite M1 is the first substance synthesized after uptake. The transport mechanism was identified as a Michaelis–Menten reaction law with the parameters given in Table 2.
- Substance M3 acts as an enzyme (E) converting metabolite M1 to M2. The reaction is irreversible, and the affinity (dissociation constant) of M1 was determined (Table 2).
- Degradation of M2 is also identified as a Michaelis–Menten reaction law with the parameters given in Table 2. It is assumed that flux from M2 is responsible for the entire biomass: M2 → biomass.
- The enzyme is subject to control (control of activity or control of synthesis).

To verify calculated input profiles an interactive Web site (http://www.sysbio.de/projects/benchmark/) is provided. The site offers the possibility to enter a vector of time points and corresponding values for the input profiles for q_{in}, q_{out}, and c_{in} as well as sampling time points (in *h*). Initial conditions for all state variables must also be given. Outputs are the time vector at the given sampling time points and a vector of all state variables with added random noise. The time series data are shown in several plots and can also be downloaded.

Model Formulation

Based on the information given above, equations are set up for the state variables. The equations for reactor volume, entire biomass concentration, and substrate concentration are formulated in a very general way:

$$\dot{V} = q_{in} - q_{out} \tag{1}$$

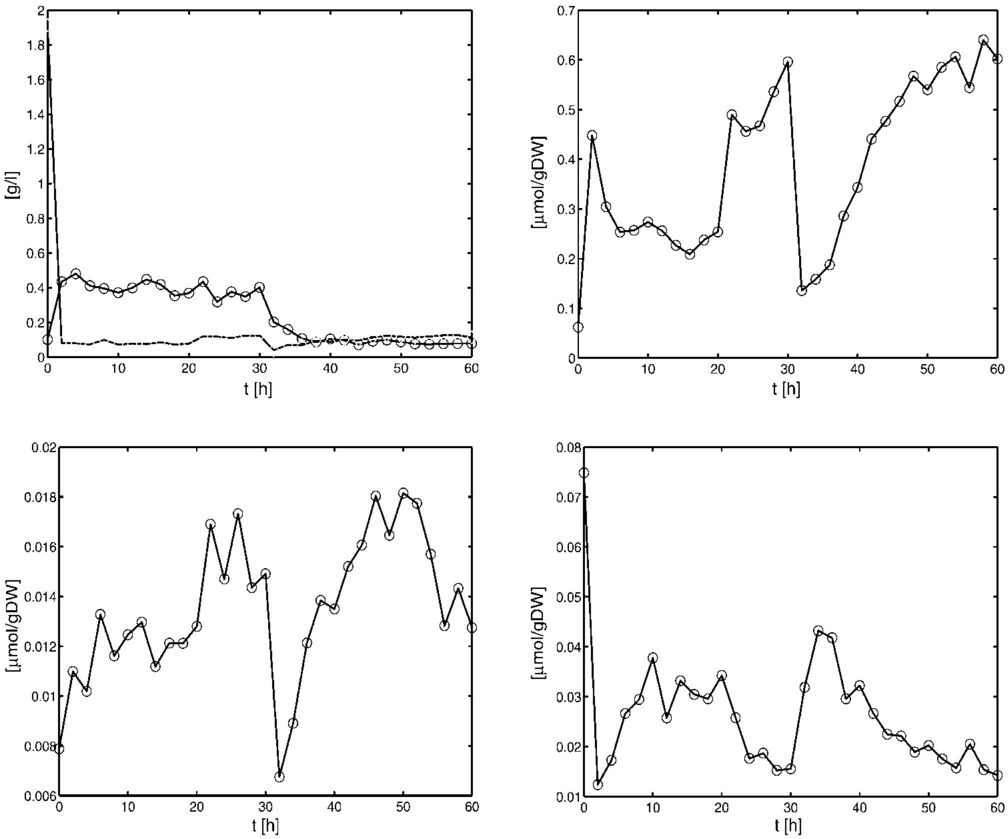


Figure 2 Time series data for biomass and substrate (*upper left*), for substance *M*1 (*upper right*), for substance *M*2 (*lower left*), and for substance *M*3 (*lower right*). Data were generated as described above. Numerical values of the data are given in the Appendix and can be downloaded from the Web site given in the problem formulation.

$$\dot{B} = \left(\mu - \frac{q_{in}}{V}\right) B \tag{2}$$

$$\dot{S} = q_{in}\,(c_{in} - S) - r_1\, Mw\, B\,, \tag{3}$$

where Mw is the molar mass of the substrate and r_1 is the uptake rate. A Michaelis–Menten kinetic rate law is used:

$$r_1 = r_{1max}\,\frac{S}{K_S + S}\,. \tag{4}$$

Based on the information given above, two possible model variants are formulated: Model A describes the conversion of M1 to M2 with a noncompetitive inhibition of the enzyme by M2:

$$r_{2A} = k_{2A}\, E\, \frac{M1}{K_{M1} + M1}\,\frac{K_{IA}}{K_{IA} + M2}\,, \tag{5}$$

where k_2 is the turnover number and K_{IA} the unknown affinity of the inhibitor M2 to the enzyme. Degradation of metabolite M2 is also described with a Michaelis–Menten kinetic rate law:

$$r_3 = r_{3max}\,\frac{M2}{K_{M2} + M2}\,. \tag{6}$$

Finally, enzyme synthesis is taken into account with a constant velocity:

$$r_{synA} = k_{synmaxA}\,. \tag{7}$$

In Model B, the control of enzyme synthesis instead of the control of enzyme activity is considered. Hence, equations 5 and 7 have to be modified. Now, for the enzymatic conversion of M1, a Michaelis–Menten kinetic rate law is assumed. For the enzyme synthesis, a formal kinetic rate law representing an inhibition is used:

$$r_{2B} = k_{2B}\, E\, \frac{M1}{K_{M1} + M1} \tag{8}$$

$$r_{synB} = k_{synmaxB}\,\frac{K_{IB}}{K_{IB} + M2}\,, \tag{9}$$

where K_{IB} represents inhibition of enzyme synthesis by M2.

The following system of equations for the concentrations *M*1, *M*2, and *E* is obtained for both models:

Table 1. Conditions During Continuous Culture Experiment

Time	Input
0–20h	q_{in} = 0.25 L/h q_{out} = 0.25 L/h c_{in} = 2.0 g/L
20–30h	q_{in} = 0.35 L/h q_{out} = 0.35 L/h c_{in} = 2.0 g/L
30–60h	q_{in} = 0.35 L/h q_{out} = 0.35 L/h c_{in} = 0.50 g/L

The volume of the bioreactor was held constant at 1.0 L.

$$\dot{M}1 = r_1 - r_2 - \mu M1 \tag{10}$$

$$\dot{M}2 = r_{2A/B} - r_3 - \mu M2 \tag{11}$$

$$\dot{E} = r_{synA/B} - \mu E\ . \tag{12}$$

The equations for the intracellular components also consider the dilution by growth represented by the specific growth rate μ. To describe the growth rate, it is assumed that part of the substrate taken up by the organisms is converted into biomass with a yield coefficient Y_{xs}The equation for μ is:

$$\mu = Y_{x/s} \cdot r_1\ . \tag{13}$$

With the vector of state variables $\mathbf{x} = [B, S, M1, M2, E]$, the vector of inputs $\mathbf{u} = [q_{in}, q_{out}, c_{in}]$, and the vector of model parameters $\mathbf{p}$, the model can now be written in the general form:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{p}, t)\ , \tag{14}$$

RESULTS

Estimation of Parameters and Confidence Intervals

Based on the experimental data and the given parameters, the following parameters have to be identified: Y_{xs}, $k_{2A/B}$, $k_{synmaxA/B}$, K_{IA}, and K_{IB}.

Parameter Estimation

Using a least-squares approach, the parameters should minimize the quadratic error between the simulations and the measured data. As the latter is only available at discrete time points $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$, the errors at each measurement time point are summed. The squared error is furthermore normalized by the standard deviation of the corresponding measurement noise σ_i and by the maximal measurement. Thus, less noisy signals are more weighted, and all measurements are brought to the same scale. This results in the following objective function that the optimal parameters should minimize:

$$J = \sum_{t \in \mathcal{T}} \sum_{i=1}^{M} \left(\frac{x_i(t) - \tilde{x}_i(t)}{\sigma_i \hat{x}_i} \right)^2, \text{ with } \hat{x}_i = \max_{t \in \mathcal{T}} x_i(t)\ , \tag{15}$$

where M is the number of states, x_i are the measured state variables, and $\tilde{x}_i$ the state variables of the models. The standard deviation of the noise is equal for all measurements, that is, $\sigma_i = 0.1 x_i$. Table 3 shows the resulting parameter values p_{opt} after a fit with the given experimental data. As the values of the objective functions attained for Model A and Model B differ only slightly, it is not clear which one of the models is better suited to fitting the benchmark problem.

Confidence Intervals

To estimate the confidence intervals of the parameters, two methods have been applied: local approximation by calculating the Fisher-Information-Matrix and a bootstrapping approach.

Table 2. **Kinetic Parameters for Synthesis of *M1*, Degradation of *M2*, and the Affinity of *M1* to Enzyme *E***

Synthesis of *M1* Values	$r_{\max} = 2.4 \times 10^4$ µmol/gDW h $K = 0.4437$ µmol/gDW
Affinity *M1* − *E* Value	$K = 12.2$ µmol/gDW
Degradation of *M2* Values	$r_{\max} = 3 \times 10^6$ µmol/gDW h $K = 10.0$ µmol/gDW

Table 3. **Identified Parameters of Both Models, Attained by Minimizing the Objective Function (15) Over the Benchmark Experiment**

Parameter	Model A	Model B
$Y_{X/S}$	6.968×10^{-5} g/µmol	7.031×10^{-5} g/µmol
K_{IA}	0.104 µmol/gDW	—
K_{IB}	—	0.166 µmol/gDW
k_2	5.988×10^6 L/h	5.559×10^6 L/h
k_{synmax}	7.2×10^{-3} µmol/gDW h	8.2×10^{-3} µmol/gDW h
Attained J	70	68

The Fisher-Information-Matrix is determined by the following equation:

$$\mathbf{F} = \sum_{t \in \mathcal{T}} \mathbf{S}^T \cdot \mathbf{MV}^{-1} \cdot \mathbf{S}\ , \tag{16}$$

where **MV** is the variance–covariance matrix of measurement errors and **S** is the sensitivity matrix:

$$\mathbf{S} = \begin{bmatrix} \frac{dx_1}{dp_1} & \frac{dx_1}{dp_2} & \cdots & \frac{dx_1}{dp_N} \\ \frac{dx_2}{dp_1} & \frac{dx_2}{dp_2} & \cdots & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{dx_M}{dp_1} & \cdots & \frac{dx_M}{dp_{N-1}} & \frac{dx_M}{dp_N} \end{bmatrix} \tag{17}$$

for a model with M considered states and N parameters. Because the state variables are time-dependent, the sensitivities are also time-dependent. A set of $M \cdot N$ differential equations has to be solved together with the M model equations (Varma et al. 1999):

$$\dot{\mathbf{S}} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \cdot \mathbf{S} + \frac{\partial \mathbf{f}}{\partial \mathbf{p}}\ . \tag{18}$$

Having solved the equations, the Fisher-Information-Matrix is calculated according to equation 16 by summing up all values over the time span. The Fisher-Information-Matrix is the inverse of the parameter estimation error covariance matrix of the best linear unbiased estimator (Posten and Munack 1990). The standard deviations of the parameters are therefore the square roots of the diagonal elements of $\mathbf{F}^{-1}$. They are, however, only lower bounds for the standard deviations, because the system is nonlinear in the parameters (Ljung 1999; Banga et al. 2002):

$$\sigma_i \geq \sqrt{F_{ii}^{-1}}\ . \tag{19}$$

The corresponding 95% confidence intervals can be approximated by two times the standard deviation (Press et al. 2002):

$$p_i - 2 \cdot \sigma_i \leq p_i^* \leq p_i + 2 \cdot \sigma_i \tag{20}$$

and are displayed in Figure 3 by solid lines. The figure shows relative confidence intervals Δp_i, that is, the confidence intervals have been normalized by the estimated parameters, given in Table 3. Thus, a value of 1 corresponds to the estimated parameter being equal to the optimal parameter value. For K_I, the calculated 95% confidence interval includes negative values, because a normal distribution was assumed, which is obviously not correct in this case.

The second approach estimates the "true" spreading of the parameters by repeating the parameter fitting to a large number of experiments, a so-called bootstrapping approach (Press et al.

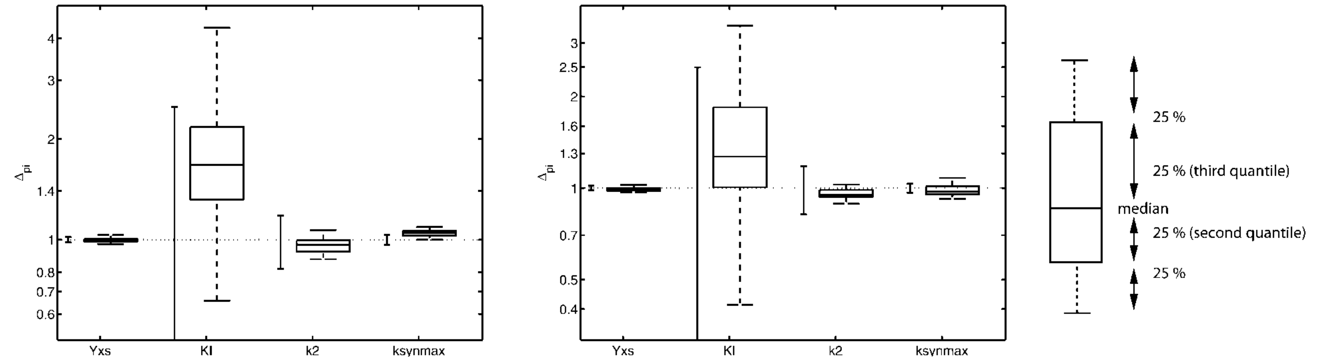


Figure 3 Parameter confidence intervals and box-plot. The parameter confidence intervals are shown normalized to the optimal values attained for the benchmark measurements. The 95% confidence interval based on the Fisher-Information-Matrix is depicted by the solid lines. The box-plot depicts the results of the bootstrap method. For both models, the estimated 95% interval for K_I includes negative values. Therefore, the whole interval is not depicted here.

2002). Here, 50 repeats were performed using the given Web site. In practice, such a large number of experiments would rarely be possible. Instead, "new experiments" can be generated by randomly picking a certain number of data points and moving them according to the uncertainty model of the corresponding measurement. The bootstrap approach estimates not only a mean and standard deviation of the parameter distribution, but also its shape. This can be visualized using a box-plot as depicted in Figure 3. A box-plot is a graphical representation of an ordered set of numbers. It depicts the median value by the central line. The median is the center value of a sorted list of data and is preferred to the mean as it is less sensitive to outliers in the data. The box shows where the central 50% of the values are, the so-called second and third quantiles. The vertical bars indicate how the remaining values are distributed. To eliminate the influence of outliers, the length of these bars is usually bounded. Here, 1.5 times the height of the box is used as maximal extension. The box-plot in Figure 3, for example, shows that the distribution is not symmetric, but that values larger than the median are spreading more than those below the median.

Clearly, the results of the two approaches differ quite substantially. This is due to nonlinear behavior of the system. Although the first approach (calculating **F**) assumes that the system is linear with respect to the parameters, the bootstrap approach is not based on a linearization. Its drawback is that the underlying experiment needs to be repeated several times. As high-throughput experiments become more common, bootstrap approaches might become more feasible in the future.

As expected, the estimation of parameter $Y_{X/S}$ yields almost identical values for both models (see Table 3). For the other parameters, the differences lie within the respective confidence intervals. Both models achieve a good agreement between the measurements and the simulated data, as observed from the attained objective functions in Table 3 and Figure 4. Discriminating between the two enzymatic hypotheses is therefore not possible.

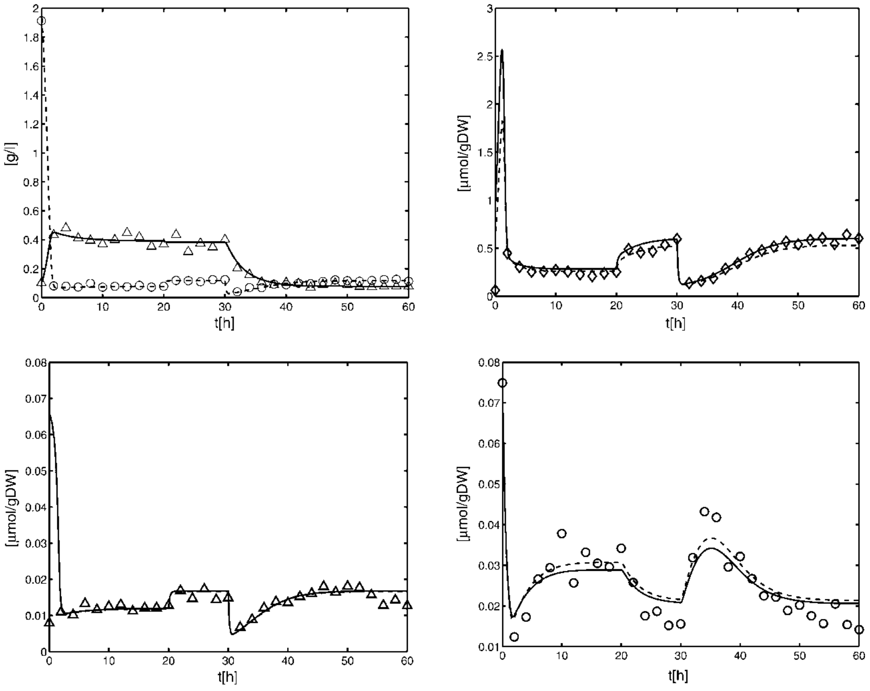


Figure 4 Benchmark data points (△, ○) versus the simulated time courses of Model A (solid) and Model B (dashed) using the parameters of Table 3. *Upper left:* X solid, S dashed; *upper right:* M1; *lower left:* M2; *lower right:* M3.

Solutions for Model Discrimination

In the following sections, different approaches to the model discrimination problem are discussed, and every approach suggests a new design experiment. All solutions presented here are based on the same structure of the model equations, as given in "Model Formulation" above.

Large Steps on the Inputs

The idea was to look for simple profiles of the manipulated variables, which can easily be implemented in a real world experiment. One simple possibility investigated here is applying large changes on the two inputs $q_{in} = q_{out}$ and c_{in}. This can result in an enhancement of small differences between the time curves calculated using the two tested models.

The strategy used in this section comprises (1) calculation of the steady state of four initial cases with low or high values of the feed concentration c_{in} and flow rates $q_{in} = q_{out}$; (2) simulation of 12 different step

Table 4. Experimental Procedure for Model Discrimination

Time (h)	c_{in} (g/L)	q_{in}, q_{out} (L/h)
0–24	2.0	0.4
24–60	0.1	0.4

experiments (four different initial conditions each with three different input changes: rate, concentration, and both) for both models; (3) fitting of the model parameters for each experiment and both models. Thus, 24 parameter sets are obtained, and the respective objective functions are calculated; (4) comparison of the resulting objective functions of Models A and B for each experiment. The objective functions for one experiment are different for both models if one model describes well the obtained data (low objective function) and the other does not (large objective function). Based on this comparison, the most discriminating experiment can be chosen. If the experimental data for the 12 versions of the second experiment would not be available, the parameter fitting step (3) was eliminated and the differences between both simulated time curves (using Models A and B) of every model state were used to identify the most discriminating experiment (4). Therefore, the model parameters based on the benchmark experiment would be used.

The most discriminating step that is suggested as a new experiment is summarized in Table 4. Starting in steady-state conditions with high flow rate and high feed concentration after 24 h, a change in the concentration is performed resulting in high-flow-rate and low-feed-concentration conditions. Several similarly discriminating cases were found but were not used in the following. For the rest of the possibilities, either poor fits to the Web site data and/or lower differences in the objective function were obtained (data not shown).

The new parameters for Model B are close to those attained by fitting only the benchmark experiment (see Table 5). The parameters of Model A, however, are quite different, in particular, K_I. The benchmark and the new experiment can be well fitted by Model B—see $M3$ in Figure 5. However, the Model A with the new parameter set is not any more able to fit $M3$ in the benchmark or the new experiment. Differences can be found all over the simulated time span, whereas the highest differences can be seen after the applied step (24 h) in the new experiment—see Figure 5. The time curves for biomass, substrate, and metabolites $M1$ and $M2$ show almost no differences between the two models. From the above, it can therefore be concluded that Model A can be discarded and that Model B describes the benchmark problem better with the proposed parameters. The control of the enzyme is realized by regulation of enzyme synthesis.

Linear Model Analysis—Analysis of the Phase Shift

The proposed solution is based on the linearized model. Regarding a steady-state solution ($\mathbf{x}^{ss}$) during continuous fermentation ($q_{in} = q_{out} = 0.25$ L/h, $c_{in} = 2.0$ g/L), the linearized model is given by:

$$\dot{\mathbf{x}} = \mathbf{J}\,\mathbf{x} + \mathbf{B}\,\mathbf{u}\,, \tag{21}$$

with the Jacobian

$$\mathbf{J} = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right|_{\mathbf{x}^{ss}} \text{ and } \mathbf{B} = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right|_{\mathbf{x}^{ss}}.$$

The input/output behavior of a linear system is characterized by two important observations: Stimulating the system with a given frequency w, the output shows the same frequency, but with a shift, named the phase shift, and amplified amplitude, named the gain. Linear dynamical model equations as given in equation 21 can be transformed to algebraic equations, called transfer functions, which can easily be handled.

For the proposed method, the gain and the phase shift for the transfer functions $G_{ij} = Y_i/U_j$ with outputs $y_1 = M1$, $y_2 = M2$, and $y_3 = E$ are analyzed. For Models A and B, all parameters are fixed except parameters K_{IA} and K_{IB}, respectively. The values for K_{IA} and K_{IB} are varied in the range $5 \times 10^{-3} < K_{I1/2} < 10.0$. Figure 6 shows the phase shift for input q and output $M1$. As can be seen, there exists a small frequency span where the two models display different phase shifts for all parameter combinations. Therefore, an experiment should be performed that forces the system with a distinct frequency inside the frequency window to see whether Model A or Model B is correct. To verify the approach, a frequency of $w = 0.5$ 1/h was chosen and phase shifts $-6.2 < \Delta\Phi_A < -23.16$, and $-23.43 < \Delta\Phi_B < -31.18$ for Models A and B, respectively, are expected. Figure 7 shows the time course of the input $q_{in} = q_{in}^0 + 0.1\,sin(wt)$ and the time course of $M1$ (data from the Web site). With the given data it was not possible to fit parameters K_{IA} or K_{IB} with high quality. However, for the solution provided, only the phase shift must be determined. The data were fitted with a second-order transfer function G:

$$G = \frac{1.54\,s^2 + 1.46\,s + 0.37}{s^2 + 0.67\,s + 0.06}\,. \tag{22}$$

The phase shift for the given frequency $w = 0.5$ is $\Delta\Phi = -28.38$, indicating that Model B is correct. Note that the linear model

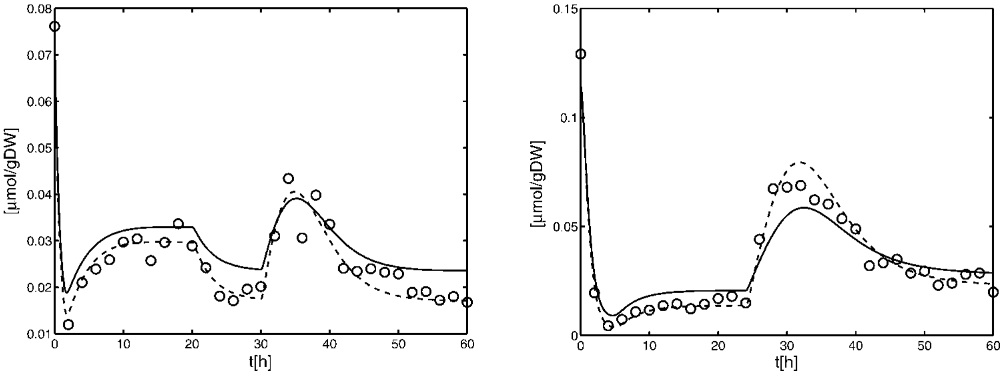


Figure 5 Data points of metabolite M3 (○) versus simulated time courses of Model A (solid) and Model B for the benchmark experiment (*left*) and the new experiment (*right*).

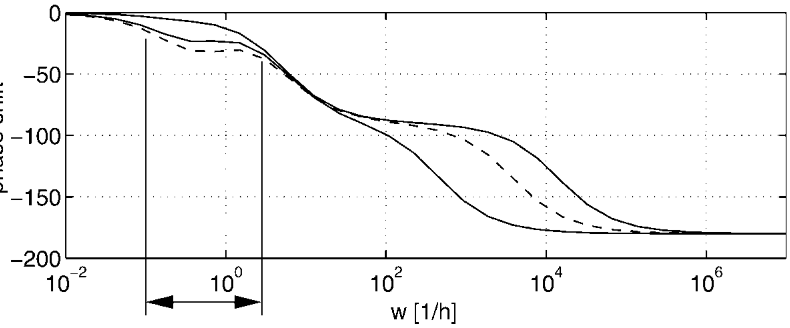


Figure 6 Phase shift for input q on output c_{M1}. Solid lines show maximal and minimal values for Model A, whereas dashed lines show minimal and maximal values for Model B varying parameters K_{IA} and K_{IB} between 5×10^{-3} and 10. For the small frequency span indicated by the vertical lines, the models are clearly separated (because of very small distances between the dashed lines, only one line can be seen).

with the correct parameters (but without noise) has a phase shift $\Delta\Phi = -30.66$ (see Appendix for the correct model).

Nonlinear Model Analysis

For the purpose of model discrimination, an experiment with an optimal input profile of the adjustable input variables (q_{in}, q_{out}, and c_{in}) has to be planned. For reasons of convenience, q_{in} and q_{out} are held equal here. The task can be formulated as the maximization of an objective function

$$\max_{u} = \int_{t_0}^{t_{end}} [\mathbf{\Delta x}^T(t)\mathbf{W}\mathbf{\Delta x}(t)]dt\,, \tag{23}$$

with $\mathbf{W}$ being a weighting matrix and $\mathbf{\Delta x}$ being the difference between the responses of the two competing Models A and B (indexes A and B are used further to point to the model variants).

Many different approaches for the choice of the weighting matrix can be found in the literature. It is obvious that weighting should be done if the interesting state variables are within different orders of magnitude. In this case, it is useful to use a diagonal weighting matrix with elements:

$$W_{ii} = \frac{1}{\left(\frac{x_{iA} + x_{iB}}{2}\right)^2}\,, \tag{24}$$

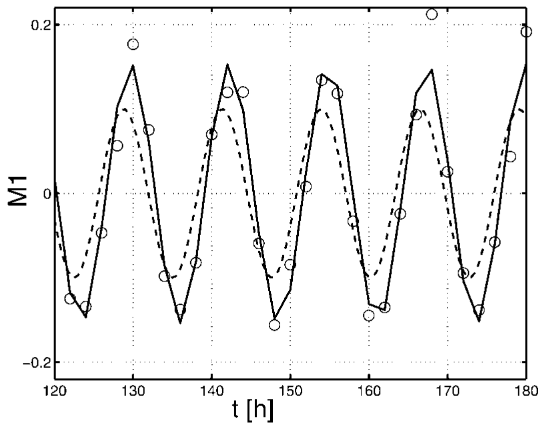


Figure 7 Time course of q_{in} (dashed), fitted (solid), and experimental values (circles) for $M1$; values are plotted minus mean values.

that is, to weight by the average of the two models. The objective function for a simple example with two state variables (x_1 and x_2) reads:

$$\max_{u} = \int_{t_0}^{t_{end}} \left[\frac{(\Delta x_1)^2}{\left(\frac{x_{1A} + x_{1B}}{2}\right)^2} + \frac{(\Delta x_2)^2}{\left(\frac{x_{2A} + x_{2B}}{2}\right)^2} \right] dt\,. \tag{25}$$

It is, however, also possible to include information about the measurement variances, the variances of the parameters of the model, and the sensitivity of these parameters with respect to the interesting state variables. This can be useful, because the values of the parameters may be uncertain. Buzzi Ferraris et al. (1984) and Chen and Asprey (2003) introduced such a strategy. The weighting matrix is formulated as follows:

$$\mathbf{W} = (\mathbf{MV} + \mathbf{VC}_A + \mathbf{VC}_B)^{-1}\,, \tag{26}$$

where $\mathbf{VC}$ is the variance–covariance matrix for model predictions:

$$\mathbf{VC} = \mathbf{S} \cdot \mathbf{PV} \cdot \mathbf{S}^T\,. \tag{27}$$

$\mathbf{PV}$ is the parameter estimation error variance–covariance matrix ($\mathbf{F}^{-1}$). It should be noticed that $\mathbf{PV}$ has to be approximated using the experiments carried out before, which in this case means only the benchmark experiment. Simplifying this approach by using only the diagonal elements of $\mathbf{MV}$, $\mathbf{PV}$, and $\mathbf{VC}$ clarifies its meaning: The squared model difference for one state variable is weighted by a sum given by its measurement variance, and the square of the sensitivity of each fitted parameter with respect to the state variable multiplied by the variance of the parameter. This means that the difference of a state variable contributes less to the objective function, if (1) the measurement error of that state variable is large and (2) the state variable in the designed experiment is very sensitive to parameters that could be estimated only with large errors using the experiment(s) carried out so far (here the benchmark experiment).

For a simple example with two state variables, the objective function looks now like this, if two parameters (index 1 and 2) are considered for each model:

$$\max_{u} = \int_{t_0}^{t_{end}} \left[\frac{(\Delta y_1)^2}{\begin{array}{c} MV_{11} + (S_{11,A}^2 \cdot PV_{11,A}) + (S_{12,A}^2 \cdot PV_{22,A}) \\ + (S_{11,B}^2 \cdot PV_{11,B}) + (S_{12,B}^2 \cdot PV_{22,B}) \end{array}} + \frac{(\Delta y_2)^2}{\begin{array}{c} MV_{22} + (S_{21,A}^2 \cdot PV_{11,A}) + (S_{22,A}^2 \cdot PV_{22,A}) \\ + (S_{21,B}^2 \cdot PV_{11,B}) + (S_{22,B}^2 \cdot PV_{22,B}) \end{array}} \right] dt\,, \tag{28}$$

Of course, this does not mean that $\mathbf{VC}$ has only diagonal elements (which would be mere chance), but that only the diagonal elements are considered in the approach.

This approach could help to avoid the case that an experiment is planned in which the model differences depend strongly on the value of parameters that are poorly fitted with the experiments carried out before. If the elements of $\mathbf{MV}$ are much larger than those of $\mathbf{VC}$ and the measurements have a similar standard variance (as in our case), it could be useful to use the following weighting matrix, that is, the simplified approach without consideration of the measurement variance:

$$\mathbf{W} = (\mathbf{VC}_A + \mathbf{VC}_B)^{-1}\,. \tag{29}$$

Table 5. Identified Parameters of Both Models, Attained by Minimizing the Objective Function (15) Over the New and the Benchmark Experiment

Parameter	Model A	Model B
$Y_{X/S}$	6.7×10^{-5} g/μmol	6.7×10^{-5} g/μmol
K_{IA}	4.8 μmol/gDW	—
K_{IB}	—	8.9×10^{-3} μmol/gDW
k_2	2.3×10^6 L/h	3.2×10^6 L/h
k_{synmax}	8.2×10^{-3} μmol/gDW h	1.8×10^{-2} μmol/gDW h

More interesting parameters, namely, the influence of the considered model state variables, the definition of the weighting matrix, and the influence of the optimization method, are analyzed and discussed. In this particular case study, the model structure is such that the biomass and the concentration of the substrate do not depend on the choice of the model. Therefore, only metabolites M1, M2, and M3 are of interest. Measurements in biological systems are, however, often very time consuming. Therefore, it is important to identify the state variables that have to be measured for model discrimination.

Both using the stochastic method and using the gradient-based optimization method may have advantages. With the stochastic method, one cannot be caught in local optima, whereas the gradient method leads to more exact results. Therefore, both methods will be compared. Equations for the concentrations of the state variables of both models are used as described in the section on “Model Formulation” above. In the case of the gradient-based method, the objective function is maximized using dynamic optimization offered by the DIVA simulation environment (Ginkel et al. 2003). In the case of the stochastic method, the “Optimized Step-Size Random Search” (OSSRS) algorithm developed by Sheela (1979) is used.

As a result of these considerations, optimization with several objective functions, differing in the weighting matrices used and the state variables or combinations of state variables considered, was performed with both optimization methods. For the calculations, the following conditions are fixed:

- Input moves are allowed every 10 h.
- The integration time is 60 h.
- The constraints used are given in Table 6. The biomass constraint ensures that washout is avoided. Moreover, there is enough biomass to be sampled out for the experimental measurements.
- The initial conditions for the state variables are chosen such that the steady-state values of both models are similar (stationary state with $q = 0.25$ L/h and $c_{in} = 2.0$ g/L).
- Parameter values for the models are as given in Table 3.

Table 7 summarizes the results obtained. A comparison between the values of the objective function can only be done for one approach, because it depends on the definition of $\mathbf{W}$. The differences in the values of the objective function are very small between stochastic and gradient-based method for nearly all cases, although the obtained input profiles differ strongly (data not shown). This hints of the existence of several local optima with very similar values of the objective function. In some cases (e.g., case 21) the gradient-based method was, however, stuck to local optima of very low quality. For cases 1–14 (see Table 7), only state variable $M1$ contributes significantly to the objective function. Therefore, equally high values are reached for all cases in which $M1$ was included in the objective function and much lower values are obtained for the cases in which $M1$ was not included. For cases 15–28, only $M2$ contributes significantly to the objective function. Only the optimal cases (boldface in Table 7) for each approach have been followed up further.

The following results are obtained from this first step: (1) the optimal input profiles differ strongly between the approaches and (2) none of the models can describe the experimental data with the set of parameters derived from the benchmark experiment. Figure 8 shows exemplarily in silico experimental and simulation data for case 23 ($\mathbf{W}$ as in equation 29, consideration of $M2$). Parameter fitting was therefore repeated in a second step with measurements from both experiments, the benchmark experiment and the new experiment, for the indicated cases.

After parameter estimation, Model A can be excluded in all cases, because the simulation of the enzyme does not fit the benchmark experiment. Figure 9 (left) shows this result exemplarily for case 1 (without weighting, consideration of only $M1$). The corresponding parameters can be found in Table 8. Exclusion of Model A could be verified by an F-test. The F-test uses the ratio of the standard deviations of two data sets and tests the null hypothesis that they are not significantly different. The standard deviations S of the residuals for the enzyme were calculated to be 1.7133×10^{-4} for Model A and 1.0104×10^{-5} for Model B. The level of significance was chosen to be $\alpha = 0.99$ and the data sets contained both 30 residuals.

$$F(30, 30)_{\alpha=0.99} = 2.3860 < \frac{S_A}{S_B} = 16.9567.$$

This means that the null hypothesis has to be rejected and the residuals of Model B have a significantly lower standard deviation than those of Model A. For the other weighting matrices, similar results were obtained (data not shown).

The findings of the proposed approach are discussed in the following: first, the focus is on the question of which model state variables have to be measured. Interestingly, the enzyme did not contribute significantly to the objective functions of all the approaches studied. The conclusion could have been, that it is not necessary to measure the enzyme. In the simulation results of the designed experiments, there are big differences in $M1$ and $M2$ between the two models, but both models can describe $M1$ and $M2$ after fitting. Therefore, without measurements of the enzyme, none of the models would have been able to discriminate between the two models after fitting.

The second question focuses on the weighting matrix that leads to the best results. All of the approaches could discriminate between the two models. It could, however, be seen as an advantage of the last approach (equation 29) that the simulation for the enzyme with Model A does additionally not fit measurements for the designed experiment (case 23; Fig. 9, right). This could again be verified by an F-test with a level of significance $\alpha = 0.99$. The standard deviations S of the residuals for the enzyme were calculated to be 0.0093 for Model A and

Table 6. Constraints Used for Finding Optimal Input Profiles

Constraints	Value
Minimum flow rate	0.05 L/h
Maximum flow rate	1.60 L/h
Minimum feed concentration	0.50 g/L
Maximum feed concentration	10.0 g/L
Minimum volume	1.00 L
Maximum volume	5.00 L
Minimum biomass concentration	0.05 g/L

Table 7. Summary of Results of Nonlinear Model Analysis

Approach	Case	State Variables	Optimization Method: Stochastic	Optimization Method: Gradient-Based
No weighting	1	**M1**	$\mathbf{3.7195 \times 10^{6}}$	$\mathbf{3.7342 \times 10^{6}}$
	2	*M2*	9.0224×10^{-5}	7.8716×10^{-5}
	3	*E*	3.6658×10^{-4}	8.015×10^{-7}
	4	*M1, M2*	3.7198×10^{6}	3.7342×10^{6}
W, unity matrix	5	*M1, E*	3.7198×10^{6}	3.7342×10^{6}
	6	*M2, E*	3.6658×10^{-4}	7.9476×10^{-5}
	7	*M1, M2, E*	3.7195×10^{6}	3.7367×10^{6}
Weighted by square of average	8	**M1**	**2.8287**	**2.8378**
	9	*M2*	0.0222	0.1201
	10	*E*	0.0476	0.0493
W as in equation 24	11	*M1, M2*	2.8394	2.8428
	12	*M1, E*	2.8685	2.8806
	13	*M2, E*	0.0686	0.0692
	14	*M1, M2, E*	2.8739	2.8857
Simplified Chen and Asprey	15	*M1*	1.356	3.6054
	16	*M2*	36.3359	2.1768
	17	*E*	2.6896	0.099
	18	*M1, M2*	36.4986	7.6538
W as in equation 26	19	*M1, E*	2.6897	2.5091
	20	*M2, E*	36.9019	2.4782
	21	**M1, M2, E**	**37.0645**	**7.763**
Simplified Chen and Asprey without measurement variance	22	*M1*	24.88	22.9863
	23	**M2**	$\mathbf{1.5598 \times 10^{11}}$	$\mathbf{1.4355 \times 10^{11}}$
	24	*E*	3.6623	2.6398
	25	*M1, M2*	1.5598×10^{11}	1.4355×10^{11}
W as in equation 29	26	*M1, E*	24.88	3.6468
	27	*M2, E*	1.5598×10^{11}	3.6207
	28	*M1, M2, E*	1.5598×10^{11}	1.4355×10^{11}

2.6202×10^{-4} for Model B. There were 30 measurements within the designed experiment, and

$$F(30, 30)_{\alpha=0.99} = 2.3860 < \frac{S_A}{S_B} = 35.4962 .$$

Figure 10 shows parameter confidence intervals for the following exemplary cases: (a) using only the benchmark experiment for parameter fitting, (b) using only the experiment case 1 (without weighting, consideration of $M1$), (c) using only the experiment case 23 (**W** as in equation 29, consideration of $M1$ and $M2$), and (d) using both the benchmark experiment and the experiment case 1. Each designed experiment leads to a reduction of the parameter confidence intervals, especially for parameters K_{IA} and K_{IB}, respectively. Case a shows by far the lowest values, lower than those obtained by using the two experiments in case d.

Third, the influence of the optimization method was analyzed. Both the stochastic and the gradient-based methods lead to similar results. Using the stochastic method ensures, however, that one is not stuck in a significantly suboptimal local optimum. On the other hand, the stochastic method is very time consuming.

DISCUSSION

A benchmark problem for reverse engineering, parameter identification, and model discrimination is presented. The focus of the investigation at hand lies on model discrimination. It is shown that for a problem that may arise in microbiology or cell biology, the application of system-theoretical methods allows one to come to satisfactory results without constructing several mutant strains. However, the application of the methods requires that the cellular system can be stimulated from outside. If a bioreactor system is available, the feed rate and the feed concentration may be used. For all methods, dynamical measurements, that is, time courses of interesting variables, are essential. Based on new measurement technologies like cDNA-arrays or proteomics, it is expected that such measurements are available in the near future. Clearly, the methods are general and do not depend on the special biochemical circuit under consideration.

Three methods for experimental design have been presented that were all able to discriminate between two model variants. Several parameters, namely, the influence of model state variables and control inputs, the definition of weighting matrices, and the influence of the optimization method were analyzed. In the case at hand, the problem is formulated in such a way that biomass and concentration of the substrate do not depend on the choice of the model. Only intracellular metabolites, M1, M2, and M3, are of interest. Measurements in biological systems are, however, often very time consuming. Therefore, it is important to identify the state variables that have to be measured for model discrimination. Given in silico experimental data, two model

Table 8. Identified Parameters of Both Models, Attained by Minimizing the Objective Function (15) Over the New Experiment Designed Without Weighting and the Benchmark Experiment

Parameter	Model A	Model B
K_{IA}	0.0138 µmol/gDW	—
K_{IB}	—	0.0136 µmol/gDW
k_2	22.4×10^{6} L/h	6.05×10^{6} L/h
K_{synmax}	0.00366 µmol/gDW h	0.0135 µmol/gDW h

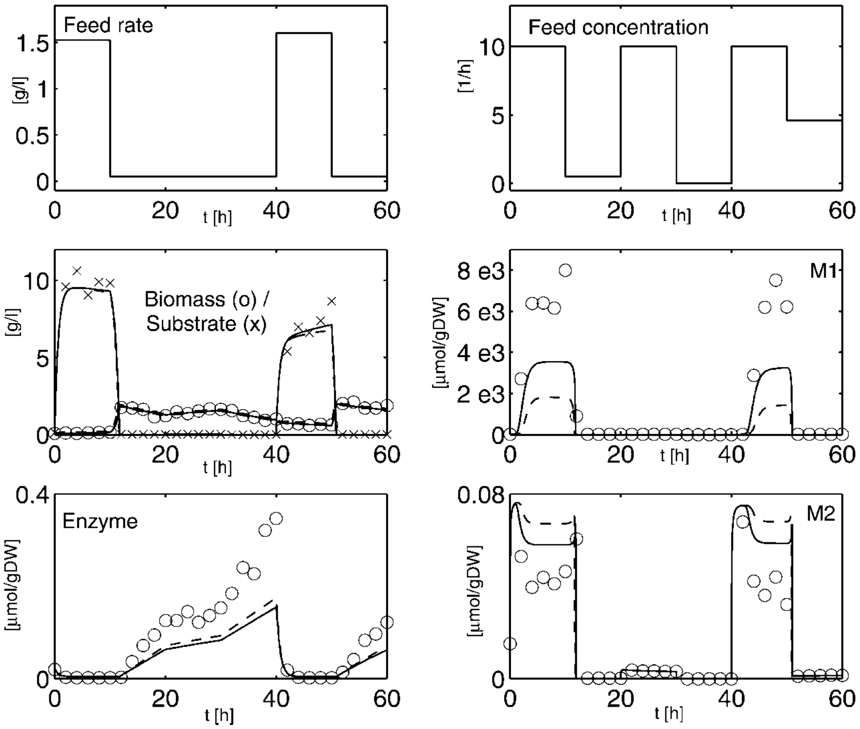


Figure 8 Optimal experiment, designed with **W** as in equation 29 (case 23). Optimal input profiles, in silico measurement results (circles), and results obtained with Model A (solid line) and Model B (dashed line) with the initial sets of parameters. Differences between the results of the two models are most notable in M1.

variants are formulated, and it was shown that both models are able to describe the given data.

Application of the three methods led to very different input profiles for inputs q and c_{in} in the experiment designed for model discrimination. The first approach focuses on the largest possible steps on the system inputs by starting from values representing the limits of meaningful inputs. Simulation runs have been carried out for the resulting 12 experimental versions. The experiment that led to the largest differences between the objective functions (equation 15) of the models has been chosen to be the new experiment. This method represents a very intuitive approach.

The section on "Linear Model Analysis" above provides a more "sophisticated" solution based on the phase shift of the linearized models. Using this approach, the phase shift of the output has to be determined given a calculated input frequency. The only input/output combination that could be used here was the pair q, c_{M1}. Drawbacks of the approach are generating such an input (needs a process control system) and the length of the experiment, because only the tuned system can be analyzed. Because the approach is based on linear models, the input signal should be small to stay within the linear range of the model. This leads to very small changes in the desired output that can be difficult to measure in a real-world experiment.

The third approach discriminates the models by bringing the states as apart as possible, however weighting the differences of the state variables. A method recently proposed by Chen and Asprey (2003) was simplified to clarify the weights used. The method calculates an input profile in such a way that the difference of a state variable contributes less to the objective function if the measurement error of that state variable is large and if the state variable in the designed experiment is very sensitive to parameters that could hardly be estimated using the benchmark experiment. Nonlinear optimization leads to very different input profiles, depending on the weighting matrices used and on the optimization method. One of these profiles represents also a form of large steps in the inputs (see Fig. 8), but the resulting differences in enzyme concentration are larger than those obtained by the first approach (cf. Figs. 5 and 9).

Common to all the methods is the observation that performing the newly designed experiment (here, with the interactive Web site) results in rather bad model predictions, if the in silico data are compared with the simulation. This is based on the large variance of the parameters determined in the initial experiment. Therefore, the parameters had to be identified again and Model A was excluded as a candidate model, because one state variable could not be fitted with both experiments. Interestingly, this state variable ($M3$) did not significantly contribute to the objective functions.

The model used for the Web interface is given in the Appendix. It is composed of both control of the enzyme activity and

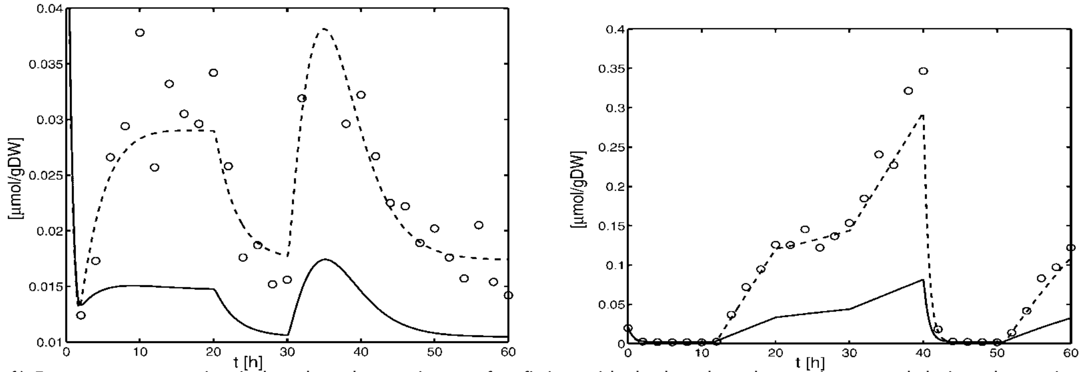


Figure 9 (*Left*) Enzyme concentration in benchmark experiment after fitting with the benchmark experiment and designed experiment for case 1. (*Right*) Enzyme concentration in the designed experiment (case 23) after fitting with the benchmark experiment and the designed experiment for case 23. In silico measurement results (circles) and results obtained with Model A (solid line) and Model B (dashed line). Results of Model A do not fit.

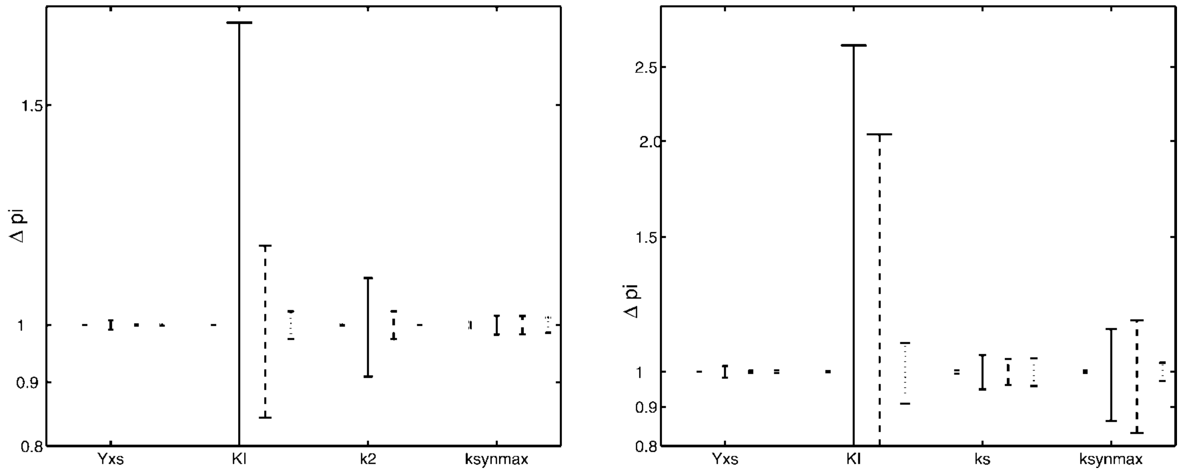


Figure 10 (*Left*) Parameter confidence intervals for Model A. (*Right*) Parameter confidence intervals for Model B. Four cases are compared. From *left* to *right* 1 (dot): experiment with **W** as in equation 29; 2 (solid) benchmark experiment; 3 (dashed) experiment obtained without weighting; 4 (dash-dot) benchmark experiment and experiment obtained without weighting.

control of enzyme synthesis. However, the influence of control of enzyme activity, represented by parameter K_{IA}, is very small. Therefore, the choice of Model B is the correct one. Comparing the parameters estimated in the first and third approaches and the correct parameters given in the Appendix (Table 9), the third approach gets better results. Moreover, the confidence region for the parameters is almost always smaller than for the benchmark experiment. For the second approach, the re-estimation of parameters is not necessary. However, one has to determine the phase shift for the frequency calculated that will last some time, because the system has to be tuned.

Based on our results, it is not possible to recommend one of these approaches. The application of one of these methods depends strongly on the possibilities to stimulate the system and to obtain measurements with high quality. The first method could be performed as a first initial experiment if there was little time to optimize the system. Comparing the stochastic versus the gradient-based optimization methods, the former leads to better results. However, the computational effort for this method is very high, as the calculation may last some days.

Another concern of this paper was the explanation and comparison of two methods for the determination of parameter accuracy. A very common method for this purpose is the approximation of parameter variances by use of the Fisher-Information-Matrix. The parameter variances obtained by this method represent, however, only lower bounds, that is, the actual variances will be larger. Furthermore, calculating the 95% confidence intervals as two times the standard deviations, as was done in this contribution, implies a normal distribution of the parameters. It is, therefore, not surprising that application of the bootstrapping approach, which does not have these drawbacks, leads to very different results (although the proportions between the parameters are similar). They represent the “true” spreading of the parameters. For the application of this method, either the possibility of repeating the experiment several times or the existence and application of an uncertainty model of the corresponding measurement are necessary. As high-throughput experiments become more common, bootstrapping approaches might become more feasible in the future.

Table 9. **Additional Parameters of the Correct Model**

Parameter	Value
$Y_{X/S}$	7.0×10^{-5} g/µmol
K_{IA}	0.01 µmol/gDW −
K_{IB}	10.0 µmol/gDW
k_2	6.0×10^{6} L/h
k_{synmax}	0.0168 µmol/gDW h

ACKNOWLEDGMENTS

This work was supported by the Alexander von Humboldt Foundation and National Science Foundation BES-0000961 (K.G. and F.J.D.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

APPENDIX

Measurement

time [h]	X [g/l]	S [g/l]
0	0.1088	1.9134
2.0000	0.4345	0.0805
4.0000	0.4811	0.0791
6.0000	0.4114	0.0734
8.0000	0.3956	0.0990
10.0000	0.3714	0.0724
12.0000	0.3995	0.0782
14.0000	0.4477	0.0752
16.0000	0.4190	0.0853
18.0000	0.3540	0.0725
20.0000	0.3690	0.0781
22.0000	0.4345	0.1195
24.0000	0.3183	0.1178
26.0000	0.3767	0.1099
28.0000	0.3489	0.1243
30.0000	0.4019	0.1249
32.0000	0.2023	0.0403
34.0000	0.1595	0.0703
36.0000	0.1068	0.0691
38.0000	0.0868	0.0933
40.0000	0.1047	0.0893
42.0000	0.0967	0.1000
44.0000	0.0714	0.0965
46.0000	0.0916	0.1122

48.0000	0.0992	0.1234
50.0000	0.0877	0.1180
52.0000	0.0766	0.1133
54.0000	0.0747	0.1196
56.0000	0.0769	0.1256
58.0000	0.0786	0.1269
60.0000	0.0781	0.1138
M1	M2	M3
0.0620	0.0079	0.0749
0.4479	0.0110	0.0124
0.3045	0.0102	0.0173
0.2534	0.0133	0.0266
0.2569	0.0116	0.0294
0.2736	0.0125	0.0378
0.2561	0.0130	0.0257
0.2268	0.0112	0.0332
0.2086	0.0121	0.0305
0.2375	0.0121	0.0296
0.2539	0.0128	0.0342
0.4895	0.0169	0.0258
0.4561	0.0147	0.0176
0.4673	0.0173	0.0187
0.5358	0.0144	0.0152
0.5961	0.0149	0.0156
0.1357	0.0067	0.0319
0.1584	0.0089	0.0432
0.1873	0.0121	0.0418
0.2860	0.0138	0.0296
0.3434	0.0135	0.0322
0.4408	0.0152	0.0267
time [h]	X [g/l]	S [g/l]
0.4767	0.0161	0.0225
0.5163	0.0180	0.0222
0.5675	0.0165	0.0189
0.5399	0.0181	0.0202
0.5851	0.0177	0.0176
0.6062	0.0157	0.0157
0.5443	0.0128	0.0205
0.6399	0.0143	0.0154
0.6020	0.0127	0.0142

The values of $M1$, $M2$, and $M3$ are in [µmol/gDW]. A file with the presented data can be downloaded from the Web site.

The Correct Model

The correct model is given by:

$$\dot{V} = q_{in} - q_{out} \tag{30}$$

$$\dot{B} = \left(\mu - \frac{q_{in}}{V} \right) B \tag{31}$$

$$\dot{S} = q_{in} \, (c_{in} - S) - r_1 \, Mw \, B \, . \tag{32}$$

For reaction rates r_1, r_2, and r_3 the following equations hold:

$$r_1 = r_{1max} \frac{S}{K_S + S} \tag{33}$$

$$r_2 = k_2 \, E \, \frac{M1}{K_{M1} + M1} \, \frac{K_{IA}}{K_{IA} + M2} \tag{34}$$

$$r_3 = r_{3max} \frac{M2}{K_{M2} + M2} \, . \tag{35}$$

Enzyme synthesis is taken into account with:

$$r_{synB} = k_{synmax} \frac{K_{IB}}{K_{IB} + M2} \, . \tag{36}$$

The following system of equations for the concentrations of $M1$, $M2$, and E is obtained for both models:

$$\dot{M}1 = r_1 - r_2 - \mu \, M1 \tag{37}$$

$$\dot{M}2 = r_2 - r_3 - \mu \, M2 \tag{38}$$

$$\dot{E} = r_{syn} - \mu \, E \, . \tag{39}$$

To describe the growth rate, it is assumed that part of the substrate taken up by the organisms is converted into biomass with a yield coefficient Y_{xs}. The equation for µ is:

$$\mu = Y_{xs} \cdot r_1 \, . \tag{40}$$

The correct parameters are summarized in Table 9.

REFERENCES

Arkin, A. and Ross, J. 1995. Statistical construction of chemical reaction mechanisms from measured time series. *J. Phys. Chem.* **99:** 970–979.

Arkin, A., Shen, P., and Ross, J. 1997. A test case of correlation metric construction of a reaction pathway from measurements. *Science* **277:** 1275–1279.

Asprey, S.P. and Macchietto, S. 2000. Statistical tools for optimal dynamic model building. *Comput. Chem. Eng.* **24:** 1261–1267.

Banga, J.R., Versyck, K.J., and Van Impe, J.F. 2002. Computation of optimal identification experiments for nonlinear dynamic process models: A stochastic global optimization approach. *Ind. Eng. Chem. Res.* **41:** 2425–2430.

Box, G.E.P. and Hill, W.J. 1967. Discrimination among mechanistic models. *Technometrics* **9:** 57–71.

Buzzi Ferraris, G., Forzatti, P., Emig, G., and Hofmann, H. 1984. Sequential experimental design for model discrimination in the case of multiple responses. *Chem. Eng. Sci.* **39:** 81–85.

Chen, H. and Asprey, S.P. 2003. On the design of optimally informative dynamic experiments for model discrimination in multiresponse nonlinear situations. *Ind. Eng. Chem. Res.* **42:** 1379–1390.

Cooney, M.J. and McDonald, K.A. 1995. Optimal dynamic experiments for bioreactor model discrimination. *Appl. Microbiol. Biotechnol.* **43:** 826–837.

D'haesseleer, P., Liang, S., and Somogyi, R. 2000. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **16:** 707–726.

Faller, D., Klingmüller, U., and Timmer, J. 2003. Simulation methods for optimal experimental design in systems biology. *Simulation* **79:** 717–725.

Ginkel, M., Kremling, A., Nutsch, T., Rehner, R., and Gilles, E.D. 2003. Modular modeling of cellular systems with ProMoT/Diva. *Bioinformatics* **19:** 1169–1176.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., and Hood, L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292:** 929–934.

Kitano, H. 2000. Perspectives on systems biology. *New Generation Computing* **18:** 199–216.

Koza, J.R., Mydlowec, W., Lanza, G., Yu, J., and Keane, M.A. 2001. Reverse engineering of metabolic pathways from observed data using genetic programming. In *Proceedings of the 6th Pacific Symposium on Biocomputing*, Hawaii, USA (eds. R.B. Altmann and A.K. Dunker), pp. 434–445. World Scientific Publishing Company.

Kremling, A., Bettenbrock, K., Laube, B., Jahreis, K., Lengeler, J.W., and Gilles, E.D. 2001. The organization of metabolic reaction networks: III. Application for diauxic growth on glucose and lactose. *Metab. Eng.* **3:** 362–379.

Ljung, L. 1999. *System Identification—Theory for the user*, 2nd ed. Prentice Hall PTR, Upper Saddle River, NJ.

Moles, G., Mendes, P., and Banga, J.R. 2003. Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Res.* **13:** 2467–2474.

Munack, A. 1992. Some improvements in the identification of bioprocesses. In *Modeling and control of biotechnical processes 1992*, IFAC Symposia series (eds. M.N. Karim and G. Stephanopoulos), pp. 89–94. IFAC, Pergamon Press, New York.

Posten, C. and Munack, A. 1990. On-line application of parameter estimation accuracy to biotechnical processes. In *Proceedings of the American Control Conference*, Vol. 3, pp. 2181–2186.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. 2002. *Numerical recipes in C: The art of scientific computing.* Cambridge University Press, Cambridge, UK.

Repsilber, D., Liljenström, H., and Andersson, S.G.E. 2002. Reverse engineering of regulatory networks: Simulation studies on a genetic algorithm approach for ranking hypotheses. *BioSystems* **66:** 31–41.

Sheela, B.V. 1979. Optimized step-size random search (OSSRS). *Computer Methods Appl. Mech. Engineer.* **19:** 99–106.
Stelling, J., Kremling, A., Ginkel, M., Bettenbrock, K., and Gilles, E.D. 2001. Towards a virtual biological laboratory. In *Foundations of systems biology* (ed. H. Kitano), Chap. 9, pp. 189–212. The MIT Press, Cambridge, MA.
Swameye, I., Miller, T.G., Timmer, J., Sandra, O., and Klingmüller, U. 2003. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc. Natl. Acad. Sci.* **100:** 1028–1033.
Tegner, J., Yeung, M.K.S., Hasty, J., and Collins, J.J. 2003. Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci.* **100:** 5944–5949.
Varma, A., Morbidelli, M., and Wu, H. 1999. *Parametric sensitivity in chemical systems.* Cambridge University Press, Cambridge, UK.
Zak, D.E., Gonye, G.E., Schwaber, J.S., and Doyle, F.J. 2003. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: Insights from an identifiability analysis of an in silico network. *Genome Res.* **13:** 2396–2405.

WEB SITE REFERENCES

http://www.sysbio.de/projects/benchmark/; Interactive Web site with online model.

Received January 27, 2003; accepted in revised form May 18, 2004.

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

BioSystems 73 (2004) 57–71

BioSystems

www.elsevier.com/locate/biosystems

Time hierarchies in the *Escherichia coli* carbohydrate uptake and metabolism

A. Kremling*, S. Fischer, T. Sauter, K. Bettenbrock, E.D. Gilles

Systems Biology Group, Max-Planck-Institut für Dynamik Komplexer Technischer Systeme, Sandtorstr. 1, 39106 Magdeburg, Germany

Received 23 April 2003; received in revised form 24 July 2003; accepted 2 September 2003

Abstract

The analysis of metabolic pathways with mathematical models contributes to the better understanding of the behavior of metabolic processes. This paper presents the analysis of a mathematical model for carbohydrate uptake and metabolism in *Escherichia coli*. It is shown that the dynamic processes cover a broad time span from some milliseconds to several hours. Based on this analysis the fast processes could be described with steady-state characteristic curves. A subsequent robustness analysis of the model parameters shows that the fast part of the system may act as a filter for the slow part of the system; the sensitivities of the fast system are conserved. From these findings it is concluded that the slow part of the system shows some robustness against changes in parameters of the fast subsystem, i.e. if a parameter shows no sensitivity for the fast part of the system, it will also show no sensitivity for the slow part of the system.

© 2003 Elsevier Ireland Ltd. All rights reserved.

Keywords: Phosphotransferase system; Time scale separation; Sensitivity analysis; Robustness; PEP/pyruvate ratio

1. Introduction

With developments in new measurement technologies, and therefore the availability of time courses for intracellular metabolites, the set up and validation of mathematical models for cellular systems (or parts of the metabolism) has become very popular. Detailed mathematical models promise a better understanding of the system under investigation, i.e. the models can be used for prediction and design and it might be possible to draw new conclusions in fields of application like biotechnology and medical science (Kremling et al., 2001b). These activities are summarized with the keyword “systems biology” and a number of projects mainly in the US (Agrawal, 1999, but also in Japan (Kitano, 2000)) have now begun.

In this contribution, we concentrate on a very important part of the bacterial regulatory system. The phosphotransferase system (PTS) is an uptake system for several carbohydrates in *Escherichia coli*. Besides this, it acts as a sensor and is involved in the control of uptake of a number of carbohydrates. For example, if glucose is present in the medium, the synthesis of many other C-source transport proteins and their corresponding catabolic enzymes is repressed. Since the PTS represents the start of the signal transduction pathway, the understanding of its dynamics is fundamental for the understanding of the whole pathway. Mathematical models for the PTS can be found in a number of contributions. Liao et al. (1996) present a simple approach covering all steps in one equation. Rohwer et al. (2000) discuss a very detailed model including all reaction steps. They analyze the steady-state behavior of the system and present results using metabolic control analysis (MCA). Recent studies take diffusion

---

* Corresponding author. Tel.: +49-391-6110-466.
E-mail address: kre@mpi-magdeburg.mpg.de (A. Kremling).

0303-2647/$ – see front matter © 2003 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2003.09.001

into account (Francke et al., 2002). In Kremling et al. (2001a) the reaction equations are divided into three modules: the first module describes the activities of the general PTS proteins EI and HPr, the second module describes the phosphotransfer from HPr to EIIA^{Crr} and the third module describes the actual transport step mediated by the glucose transporter EIICB^{Glc}.

Here, we have investigated carbohydrate uptake and metabolism on different time scales. The work was motivated by findings with laboratory experiments; a "pulse response" experiment revealed fast dynamics while diauxic growth on glucose plus lactose covered a broader time scale. Experiments were performed either with a genetically engineered sucrose positive strain or a wild type strain. The sucrose positive strain was used during the pulse response experiment. A mathematical model describing sucrose uptake and metabolism was introduced previously (Wang et al., 2001). Measurements of a pulse response experiment were used to identify parameters of the glycolysis and the phosphotransfer reactions of the PTS. The wild type strain was used to identify parameters involved in the control of glucose and lactose uptake. Since the work focuses on model analysis, the experimental findings are only summarized briefly.

For model analysis, only glucose uptake and metabolism are under consideration. A close relationship between time-scale hierarchies and robustness was pointed out in Rojdestvenski et al. (1999). Robustness is the insensitivity of a selected characteristic time course of a component, network function to sustain growth (Stelling et al., 2002, adaption precision (Barkai and Leibler, 1997)) with respect to changes of external or internal parameters (different environmental conditions, mutations, or altered kinetic parameters). Rojdestvenski et al. (1999) conclude that the decoupling of a system into two subsystems is necessary for robustness. Running on different time-scales is one of the characteristics that allow such decoupling. A robustness analysis of the overall system was performed to check the results of Rojdestvenski et al. (1999).

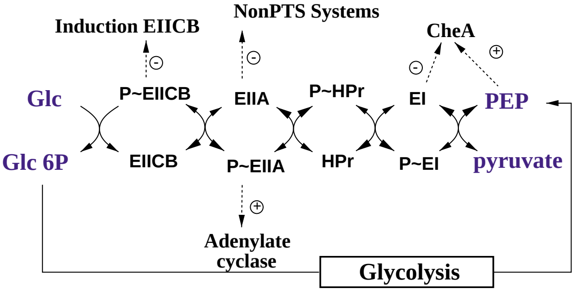


Fig. 1. Schematic representation of the Glc-PTS. Inputs are the entire concentrations of EI, HPr, EIIA, EIICB, PEP, pyruvate, and extracellular glucose. Important outputs are the phosphorylated and unphosphorylated forms of EIIA. These two conformations are measured in several experiments. Solid lines represent metabolic reactions and dashed lines signal outputs of the PTS.

2. *Escherichia coli* sugar uptake and metabolism

The PTS represents a transport and at the same time a signal transduction system responsible for carbon catabolite repression and inducer exclusion (Postma et al., 1993). In a set of five reactions, a phosphoryl group is transferred from phophoenolpyruvate (PEP) through two common intermediates, enzyme I (EI, gene *ptsI*) and the phosphohistidine carrier protein (HPr, gene *ptsH*), to the EII^{Glc}, and finally to the substrate (see Fig. 1 for the glucose PTS). EII^{Glc} consists of the soluble EIIA^{Crr} (gene *crr*, hereafter denoted as EIIA) and the membrane-bound transporter EIICB^{Glc} (gene *ptsG*) for glucose uptake. The Scr-PTS possesses a sucrose-specific membrane-bound transporter EIIBC^{Scr} (gene *scrA*), which also receives the phosphoryl group from EIIA. For further description of the PTS, see Postma et al. (1993). All proteins involved in this phosphorylation cascade act as signaling molecules, e.g. EI in chemotaxis, HPr in glycogen metabolism, and EIIA in inducer exclusion (by inhibition of lactose transport by the lactose permease LacY). The incoming phosphorylated form of the sugar is further metabolized: Glycolysis is the link between the transport reactions and their energy supply. Metabolism of glucose 6-phosphate during glycolysis results in two moles of PEP.

PTS-protein synthesis is under control of at least two regulators. While the cAMP·Crp complex acts as an activator (DeReuse and Danchin, 1988), Mlc is a repressor for *ptsG*, *ptsHI* and *crr* (Plumbridge, 1998). If glucose is present in the medium, EIICB^{Glc} is mainly in its dephosphorylated form. This form binds Mlc, and therefore prevents it from binding to the operator binding site (Tanaka et al., 2000; Lee et al., 2000).

3. Experimental results

The work was motivated by experiments performed in our laboratory. They were used as a basis for this study. Material and methods are according to Kremling et al. (2001a). The work was performed either with the wild type strain LJ110 (Zeppenfeld et al., 2000) or with LJ210. LJ210 (laboratory collection of K. Jahreis, Osnabrück) is a Scr$^+$ derivative of LJ110 that carries chromosomally the *scr* genes of pUR400 (Wohlieter et al., 1975; Schmid et al., 1982).

3.1. Pulse response experiment

To characterize the dynamic behavior of the glycolysis in interaction with the PTS, the response of the cells in steady-state to environmental disturbances was examined. In the experiments, we cultivated *E. coli* LJ210 in a continuous fermentation in a CSTR (type KLF2000, volume 2.0l, Bioengineering) with defined minimal medium as described in Kremling et al. (2001a). The disturbance of the culture in steady-state (dilution rate $D = 0.1\,\mathrm{h}^{-1}$, concentration of sucrose in the feed $c_{\mathrm{in}} = 17\,\mathrm{g/l}$) was performed by injection of a concentrated sucrose solution to a final concentration of 0.3 g/l. Resulting time courses of the fraction of unphosphorylated EIIA, and glycolysis metabolites (glucose 6-phosphate, fructose, fructose 6-phosphate, PEP, and pyruvate) were measured as described elsewhere (Ishizuka et al., 1993; Takahashi et al., 1998; Bergmeyer, 1979). The obtained trajectories are shown in Fig. 2. Note that during continuous culture, the steady-state condition $D = \mu$ holds true, where μ is the specific growth rate. Under the chosen experimental conditions, cells are sugar limited.

The added sucrose is consumed within 200 s. Gene expression can be neglected for this short time interval. The pulse can be followed in all measured glycolysis metabolites. The concentration of PEP decreases because of increased consumption for the transport process via the sucrose PTS. The important signaling component EIIA is totally dephosphorylated, and returns to the former steady-state after depletion of the added sucrose.

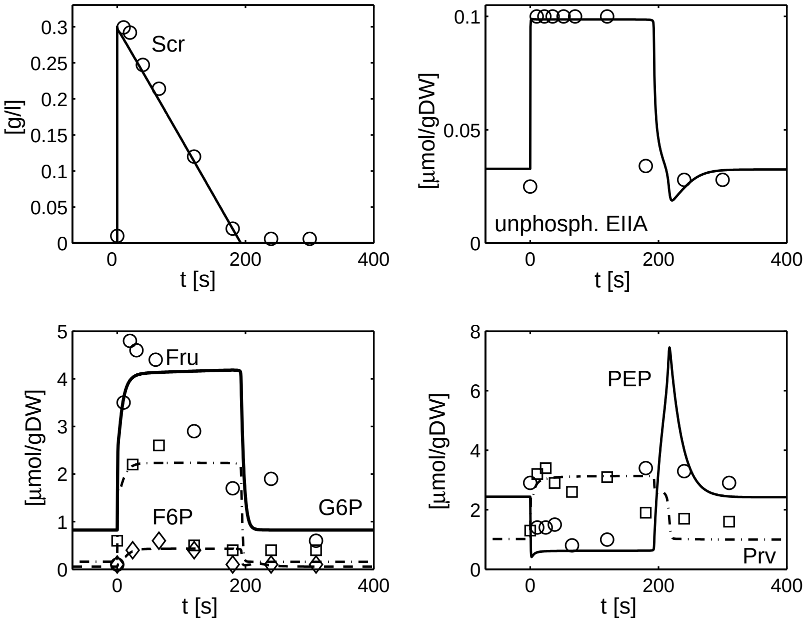


Fig. 2. Dynamic response of the PTS and glycolysis to a sucrose pulse disturbing a continuous culture in steady-state ($D = 0.1\,\mathrm{h}^{-1}$). The extracellular sucrose concentration was increased abruptly at $t = 0$ s to a final concentration of 0.3 g/l. Experimental results are represented by marks, simulation results by lines.

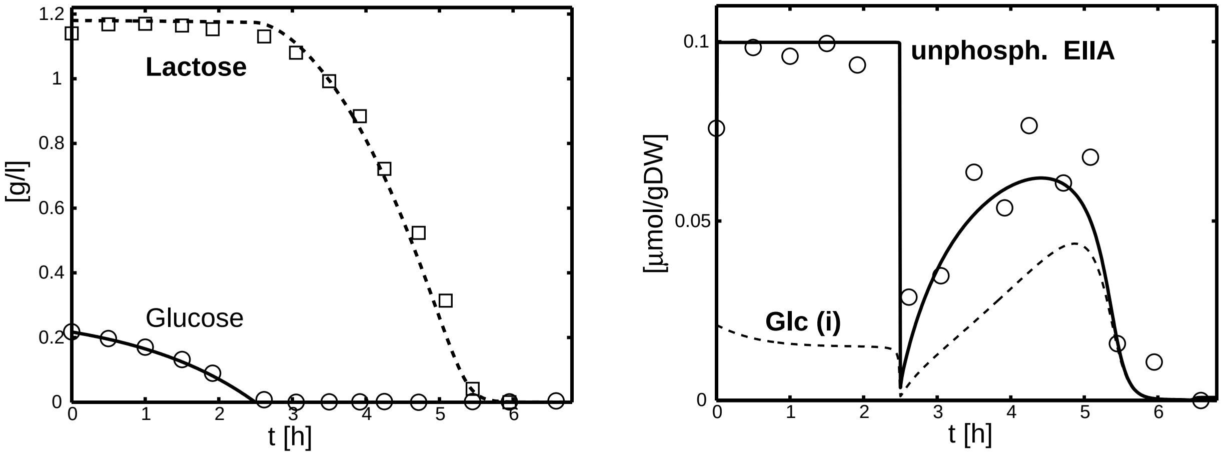


Fig. 3. Left: Time course of glucose and lactose in the medium during diauxic experiment. Right: Time course of the intracellular protein concentration of unphosphorylated EIIA and simulated intracellular glucose (dashed values were multiplied by factor 20). Experimental results are represented by marks, simulation results by lines.

3.2. Diauxic growth on two substrates

The growth of strain LJ110 in mixed cultures with glucose and lactose was characterized to analyze effects of gene expression on the dynamics of the PTS. Fig. 3 shows simulations and experimental results for extracellular glucose and lactose as well as for unphosphorylated EIIA when both carbohydrates are present in the medium at the beginning. In addition, model prediction for intracellular glucose is also shown.

As expected for the glucose phase, protein EIIA is mainly unphosphorylated. After the glucose is consumed, EIIA shifts very quickly to its phosphorylated form (representing fast dynamics) and subsequently becomes more and more unphosphorylated. This is probably because intracellular glucose may also be phosphorylated by the PTS. The observed dephosphorylation of EIIA can be interpreted as regulatory phenomenon; since EIIA is an uncompetitive inhibitor of the lactose permease LacY, the increase of unphosphorylated EIIA leads to a reduced lactose uptake rate. This prevents accumulation of glycolytic intermediates during high lactose uptake rates.

4. Model equations for the PTS and glycolysis

Model analysis will focus on glucose uptake and metabolism. Therefore, only the equations for the glucose PTS are discussed here.

4.1. Glucose PTS

The following reaction steps are incorporated into the model. Phosphoryl transfer from PEP to EI (dimer) is described by:

$$\mathrm{EI} + \mathrm{PEP} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{P} \sim \mathrm{EI} + \mathrm{Prv}, \tag{1}$$

where Prv stands for pyruvate. There is no evidence that dimerization of EI plays a role in the dynamics of the PTS for the investigations performed in this paper. The dimer is the most important conformation and is the only conformation considered here. Transfer of a phosphoryl group to HPr is described by:

$$\mathrm{P} \sim \mathrm{EI} + \mathrm{HPr} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{EI} + \mathrm{P} \sim \mathrm{HPr}. \tag{2}$$

Phosphoryl transfer from HPr to EIIA is described by:

$$\mathrm{P} \sim \mathrm{HPr} + \mathrm{EIIA} \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} \mathrm{HPr} + \mathrm{P} \sim \mathrm{EIIA}. \tag{3}$$

Since protein EIICB^{Glc} is membrane bound and since it is not clear if there is sequential binding of EIIA and the carbohydrate or a random binding (a carbohydrate molecule is able to bind to the unphosphorylated enzyme), a random kinetic rate law is used for the last two steps of the PTS; phosphoryl transfer from EIIA and the phosphorylation of the incoming carbohydrate (intracellular glucose is not included in the model, because it is assumed that the concentration is very low at all conditions used here):

$$P \sim EIIA + Glc_{ex} \overset{r_4}{\rightharpoonup} EIIA + Glc \sim P, \tag{4}$$

with the rate law r_4 for glucose is taken from Kremling et al. (2001a):

$$r_4 = \frac{k_4 c_{EIICB0} c_{P\sim EIIA} c_{Glc}}{(K_{EIIA} + c_{P\sim EIIA})(K_{Glc} + c_{Glc})}. \tag{5}$$

4.2. Glycolysis

To get a clear picture of the dynamics, a simplified model of glycolysis was used since the number of time constants is equal to the number of observables. The model comprises one step summarizing glycolytic reactions and drain into the monomers starting from glucose 6-phosphate, PEP, and pyruvate. Glycolysis in the simple model is described by:

$$G6P \overset{r_{gly}}{\rightharpoonup} 2PEP. \tag{6}$$

The pyruvate kinase reaction is as follows:

$$PEP \overset{r_{pyk}}{\rightharpoonup} Prv. \tag{7}$$

Furthermore, drain into monomer synthesis has been taken into account by the following reaction scheme:

$$\begin{aligned} &G6P \rightharpoonup^{r_{dr1}} \text{biosynthesis} \\ &PEP \rightharpoonup^{r_{dr2}} \text{biosynthesis} \\ &Prv \rightharpoonup^{r_{dr3}} \text{biosynthesis.} \end{aligned} \tag{8}$$

All reaction rates for the model described so far are summarized in Appendix A. As shown below, gene expression is responsible for very slow dynamics of the phosphorylated PTS components. For the analysis of the slow dynamics the lactose uptake system and its control by the cAMP·Crp complex are included from Kremling et al. (2001a). Since lactose is split into glucose and galactose, the intracellular glucose pool depends strongly on the concentration of the lactose permease and β-galactosidase. In the model, intracellular glucose can be phosphorylated by the PTS as well as by a glucokinase (gene *glk*).

4.3. Model parameters

A rough structure of the whole model is given in Fig. 4. The model differs from Kremling et al. (2001a) in the PTS equations and the simplified model for glycolysis. To estimate parameters from the presented data, a procedure used in Kremling et al. (2001a) was applied. (i) Starting with parameters from literature (Rohwer et al., 2000), a sensitivity analysis was performed to detect the most sensitive parameters. (ii) Together with the measured data and the Fisher information matrix (Ljung, 1999) it was determined, whether the sensitive parameters could be estimated, or not. Using a method introduced by Posten and Munack (1990) a set of parameters from the sensitive parameters that could be estimated together were determined. Parameters for the PTS and glycolysis are summarized in Table 1. Differences between the original model Kremling et al. (2001a) and the model presented here are negligible (data not shown).

The underlying reaction scheme for the PTS used here is simpler than that presented by Rohwer et al.

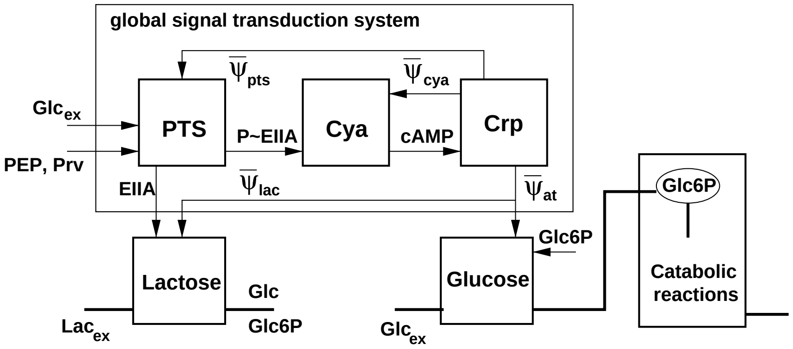


Fig. 4. Functional units in the *crpA*-modulon. Protein EIIA and its phosphorylated form P∼EIIA are the main output signals of the Glc-PTS. The output signals $\bar{\psi}$ from the CrpA submodel describe the transcription efficiency of the genes and operons under control of the cAMP·CrpA complex (Kremling and Gilles, 2001).

Table 1
Summary of the model parameters of the PTS and the simplified glycolysis

PTS		Glycolysis	
k_1	8.9E+6 gDW/μmol	k_{gly}	6500 μmol/gDW h
k_{-1}	5.9E+6 gDW/μmol h	K_{G6P}	1.5 μmol/gDW
k_2	2.8E+7 gDW/μmol h	k_{pyk}	1000.0 μmol/gDW h
k_{-2}	2.5E+7 gDW/μmol h	K_{PEP}	100.0 μmol/gDW
k_3	5.0E+6 gDW/μmol h	k_{g6p}	1500.0 1/h
k_{-3}	7.5E+6 gDW/μmol h	k_{pep}	200.0 1/h
k_4	9.17E+6 gDW/μmol h	k_{pyv}	25370 1/h
K_{EIIA}	0.0085 μmol/gDW		
K_{Glc}	0.0012 g/l		
c_{EI0}	0.012 μmol/gDW		
c_{HPr0}	0.12 μmol/gDW		
c_{EIIA0}	0.1 μmol/gDW		
c_{EIICB0}	0.003 μmol/gDW		

(2000). Model parameters were fitted using experimental results from Wang et al. (2001) and our own as yet unpublished experimental results. Therefore, the number of (uncertain) parameters was kept as low as possible. Since different strains and experimental conditions were used, one cannot expect the parameters to show a good agreement with parameters from literature. This is reflected, for example, by calculating the overall equilibrium constant K_{eq} for the first three reactions steps of the PTS:

$$K_{eq} = \frac{c_{P\sim EIIA}}{c_{EIIA}} \frac{c_{Prv}}{c_{PEP}}. \tag{9}$$

Here a value $K_{eq} = 1.13$ is obtained while the value from Rohwer et al. (2000) is $K_{eq} = 48.7$ and from Hogema et al. (1998) is $K_{eq} = 14.0$. The overall concentrations for the PTS proteins are fixed. For the experiment shown in Fig. 3, the value for c_{EIICB0} was taken from a simulation study with the model from Kremling et al. (2001a). The model takes into account that the synthesis of EIICB is under control of Mlc and the cAMP·Crp complex while the values for the other PTS proteins are taken as constants. In contrast, Rohwer et al. (2000) used constant concentrations of all PTS proteins.

5. Model analysis

Model analysis by means of theoretical tools is useful for a better understanding of the behavior of cellular systems. In silico, models are characterized by a large number of elements and interactions, i.e. the order of the system is rather high. However, systems running on different time-scales tend to couple the behavior of the fast modes to the slow modes. Algebraic equations can be used for the fast modes, if a steady-state is assumed. Furthermore, the influence of the parameters on the behavior of the system is important. Here, a sensitivity analysis is used to detect important model parameters.

5.1. Time-scale separation

First, we analyzed the time hierarchies for the conditions given during the “pulse response” experiment. Therefore, the eigenvalues and eigenvectors of a subsystem including PTS and glycolysis (observables EI, HPr, EIIA, G6P, PEP, and Prv) were calculated. To determine eigenvalues and eigenvectors a steady-state for $r = 1.1$ mmol/gDW h was chosen, representing the condition used in the experiment. The equations are linearized around the steady-state and the Jacobian J is obtained. A transformation of the old system observables $\underline{x}$ into new coordinates $\underline{z}$ with the inverse of the matrix of eigenvectors T^{-1}

$$\underline{z} = T^{-1}\underline{x}, \tag{10}$$

allows an analysis of the system in separate time windows characterized by the eigenvalues. Fig. 5 shows the entries of the rows of T^{-1} where the abscissa represents the observables of the original system. Plots A–C represent very fast processes. To detect the main components of one mode, the linearized system was stimulated by a step in the glucose concentration. Afterwards, the entries in the lines of T^{-1} must be multiplied with the concentrations of the respective observables. For line 1 in T^{-1}, representing the fastest mode, the main components are EI, Prv, HPr and PEP. These observables are involved in the first PTS reaction. In the second mode, the main components are HPr and EIIA, which are involved in the second PTS reaction. A clearer picture emerges when a new mode can be directly linked to one of the observables of the original system. This is the case for mode z_4, representing the dynamics of glucose 6-phosphate and mode z_5, representing the dynamics of pyruvate. The slowest mode, z_6, represents the dynamics of two original observables, namely PEP and glucose 6-phosphate.

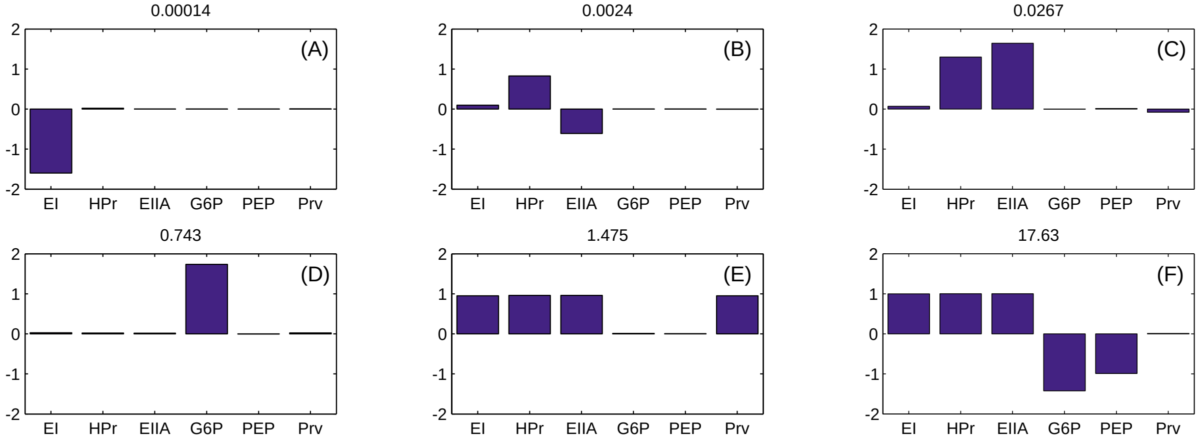


Fig. 5. Eigenvectors of the simplified reaction scheme for a glucose uptake rate of $r = 1.11\,\text{mmol/gDW h}$ (rows of the matrix T^{-1}). The headline of each plot gives the apparent time constant of the system in (s).

To verify the results for the group translocation through the PTS proteins EI, HPr, and EIIA, we simulated the first second(s) of the pulse response experiment with the nonlinear model. The time courses of the PTS proteins, glucose 6-phosphate, pyruvate and PEP as well as the time courses of the PTS rates (r_1, r_2, r_3, r_4) are shown in Fig. 6. Based on the estimated model parameters the rate of glucose uptake (r_4) varies with time. After a very fast increase it is not possible to maintain at this rate, since r_1, r_2, and r_3 are slower and the supply with phosphoryl groups could not be satisfied immediately, even though enough energy in form of PEP is available. This is in agreement with the simulation results of the intracellular concentrations of the PTS proteins; the concentration of phosphorylated EIIA shows a quick drop, because it transfers the phosphoryl groups to glucose. In contrast, the concentrations of phosphorylated EI and HPr drop slower. All PTS components reach a steady-state within 5 s. After the PEP pool is replenished by glycolytic reactions the uptake rate of the PTS rises again and reaches a steady-state after 30 s.

With respect to gene expression, PTS and glycolysis have much smaller time constants. The overall transport rate in this case reads

$$r = f(c_{\text{PEP}}, c_{\text{Prv}}, c_{\text{carbo}}, c_{\text{EIIBC0}}, c_{\text{EIIA0}}, c_{\text{EI0}}, c_{\text{HPr0}}), \tag{11}$$

which is the solution of the algebraic equation system, setting all time derivatives of the observables to zero. Fig. 7 summarizes the steady-state behavior of the system. The glycolytic metabolites—other than PEP—increase with increasing uptake rate while the fractions of the phosphorylated PTS components EIIA and EI decrease with an increasing uptake rate. The half maximal uptake rate was detected in the range of 0.23 mg/l (1.3 μM). The value is smaller than in other reports, perhaps reflecting in vivo conditions.

5.2. Robustness analysis

Rojdestvenski et al. (1999) stated, that the sensitivities of the fast part of the system are conserved when the dynamics of the remainder slow subsystem is analyzed. To check the assumption, a robustness analysis was performed. This was done by calculating changes of the time course of a selected observable with respect to changes in the kinetic parameters, i.e. by calculating the parameter sensitivities. This means that the slow subsystem shows some robustness against parameter changes, i.e. if a parameter shows no sensitivity for the fast subsystem, it will show no sensitivity for the slow subsystem.

Several methods are available to calculate and analyze parameter sensitivities defined by:

$$w_{ij} = \frac{\partial x_i}{\partial p_j}, \tag{12}$$

where x_i is a observable of the model and p_j is a model parameter. A very popular method is used in

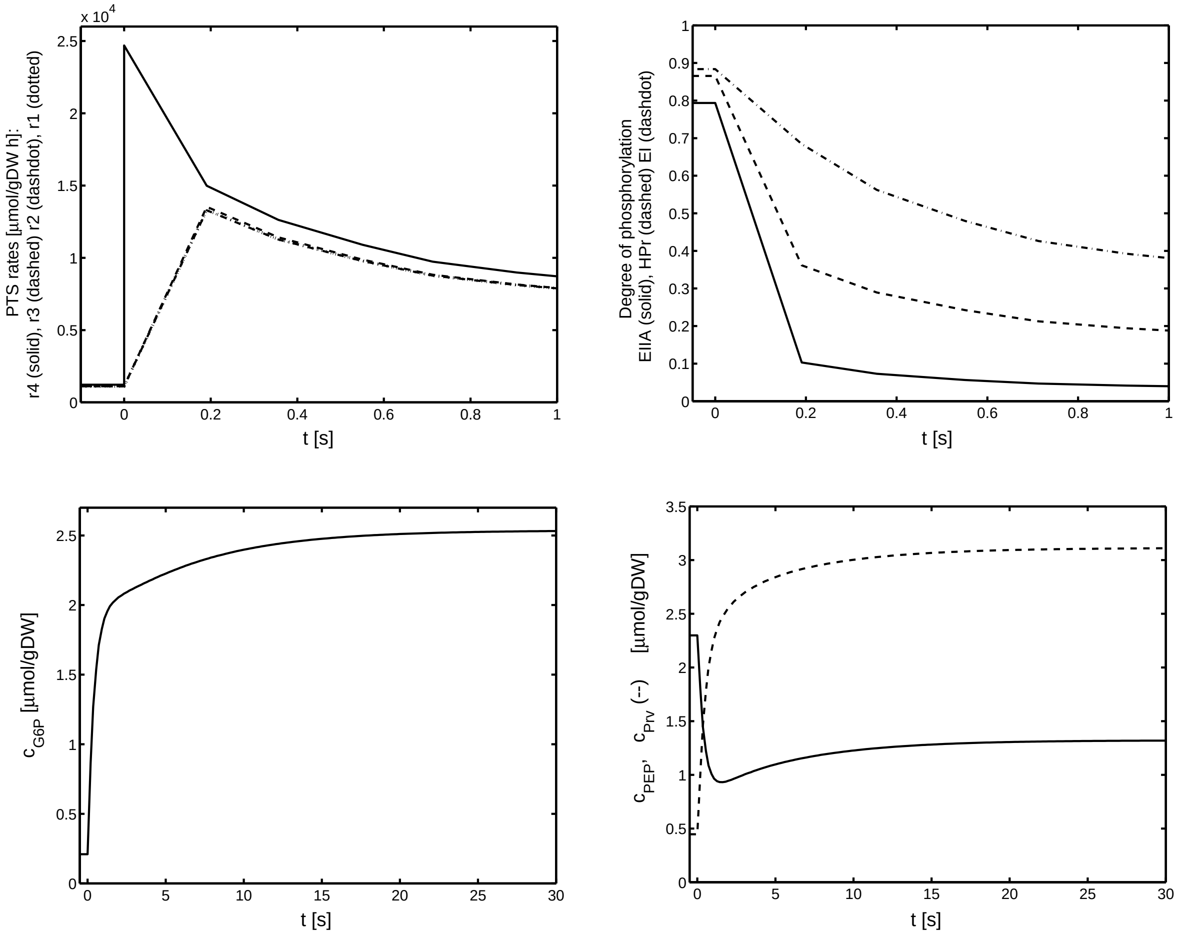


Fig. 6. Time course of the simulated PTS rates, the degree of phosphorylation of the PTS proteins, and selected intracellular metabolites (glucose 6-phosphate second row left, PEP, pyruvate second row right). For discussion of the dynamic behavior, see text.

Metabolic Control Analysis (e.g. see Heinrich and Schuster, 1996). The method uses several theorems to connect local sensitivities and to make a statement on global sensitivities and therefore relates the systemic behavior to local properties. Here, a method from Hearne (1985) developed for dynamic systems was applied. Hearne (1985) suggests finding the direction in which the overall parameter vector should be perturbed, so as to maximize the disturbance to the trajectories of interest. In contrast to other approaches, the method of Hearne (1985) integrates the sensitivities of selected parameters on selected observables. The method is also useful when looking for parameters that are able to alter specific observables for which measurements are available.

The method is based on the formulation of the following optimization problem:

$$\max \sum_{i=1}^{n} \sum \frac{\Delta x_i}{x_i} \Delta t, \tag{13}$$

with observables x_i and changes Δx_i of x_i due to changes $\Delta \boldsymbol{p}$ of parameter vector $\boldsymbol{p}$. The dimension of the system is n. The solution can be found by calculating the maximal eigenvalue and corresponding eigenvector of matrix $\boldsymbol{G}$ given in Appendix A. To calculate matrix $\boldsymbol{G}$, dynamic and algebraic equations for the sensitivities have to be solved. For a differential-algebra (DA) system with dynamic observables $\underline{x}$, algebraic observables $\underline{z}$, and parameter vector $\underline{p}$ of the general

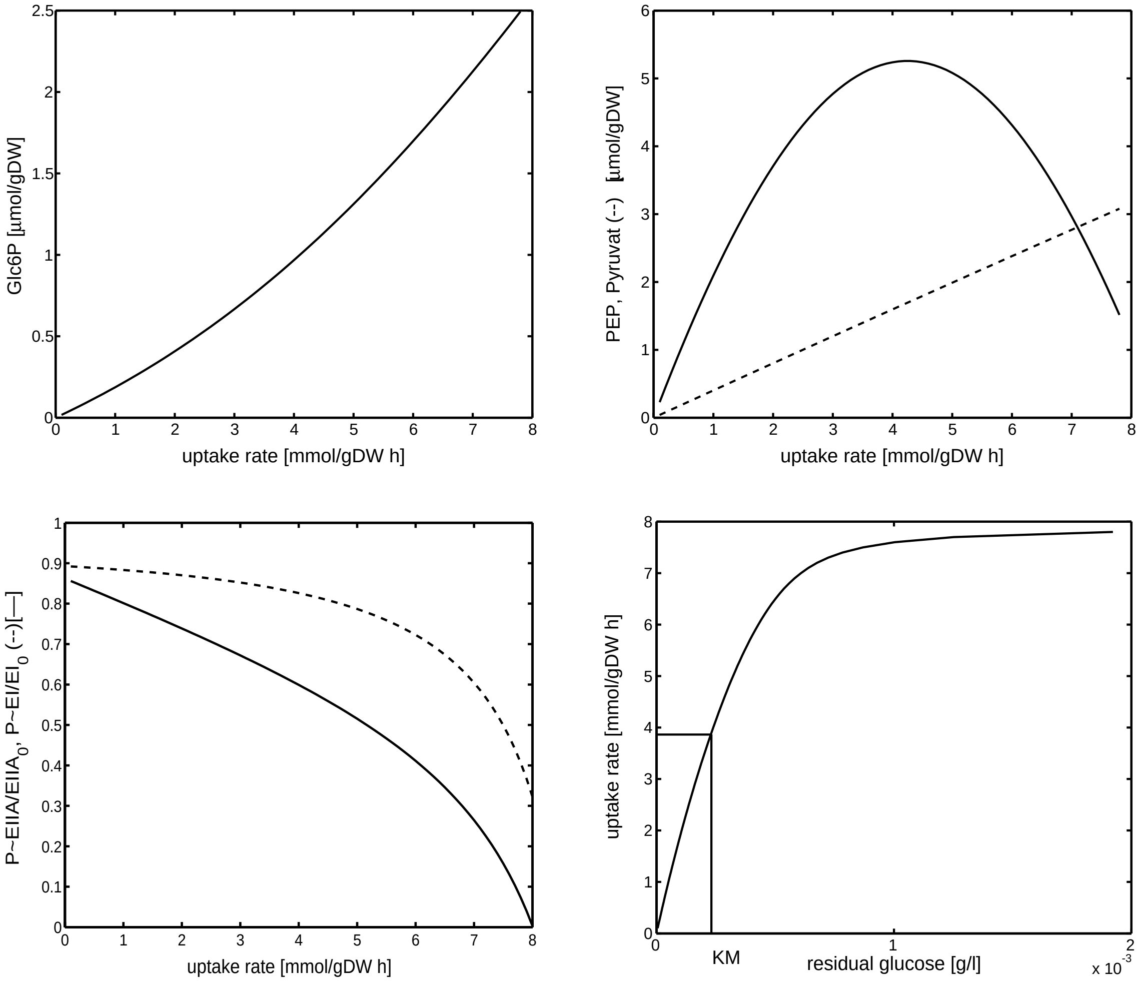


Fig. 7. Steady-state behavior of the PTS/glycolysis model. Glucose 6-phosphate (upper-left plot); PEP and pyruvate (upper-right plot: PEP solid, pyruvate dashed); fraction of phosphorylated EIIA and EI (lower-left plot: EIIA solid, EI dashed); uptake rate vs. residual glucose concentration (lower-right plot). The half maximal glucose uptake rate is achieved with 0.23 mg/l.

form:

$$\dot{\underline{x}} = \underline{f}(\underline{x}, \underline{z}, \underline{p}) \quad (14)$$

$$\underline{0} = \underline{g}(\underline{x}, \underline{z}, \underline{p}), \quad (15)$$

the matrix of the non-normalized sensitivities $W_x = (\mathrm{d}\underline{x}/\mathrm{d}\underline{p})$ and $W_z = (\mathrm{d}\underline{z}/\mathrm{d}\underline{p})$ can be calculated by the two following equations:

$$\dot{W}_x = \frac{\mathrm{d}\underline{f}}{\mathrm{d}\underline{p}} + \frac{\mathrm{d}\underline{f}}{\mathrm{d}\underline{x}} W_x + \frac{\mathrm{d}\underline{f}}{\mathrm{d}\underline{z}} W_z \quad (16)$$

$$0 = \frac{\mathrm{d}\underline{g}}{\mathrm{d}\underline{p}} + \frac{\mathrm{d}\underline{g}}{\mathrm{d}\underline{x}} W_x + \frac{\mathrm{d}\underline{g}}{\mathrm{d}\underline{z}} W_z. \quad (17)$$

For the analysis in this contribution the following strategy is used:

- The model is separated into two parts: a *fast* part and a remainder *slow* part. Based on the analysis above, observables and parameters are assigned to the respective parts.
- For the *fast* submodel, the sensitivity is calculated with the method of Hearne (1985), taking all observables and all parameters for the *fast* subsystem into account. This represents a “fingerprint” of the *fast* subsystem. The fingerprint is compared with the sensitivity of the same set of parameters, but now

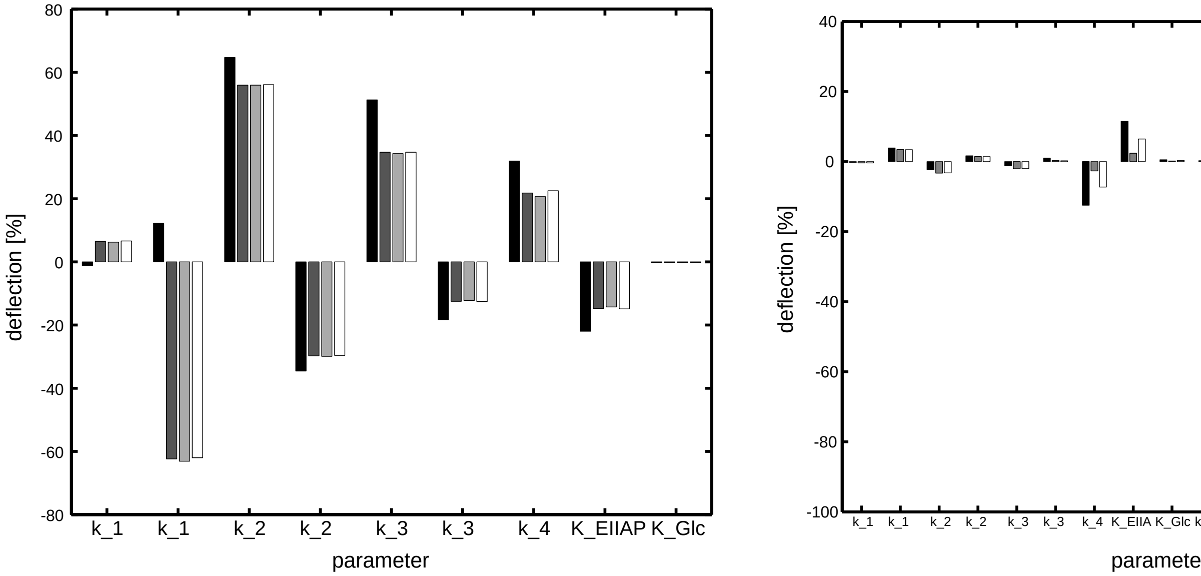


Fig. 8. Left-hand side: Case I: Fingerprint (black), glucose 6-phosphate (dark gray), PEP (light gray), pyruvate (white). With the excepti
vector of perturbation of fingerprint is comparable to the selected observables of the slow subsystem. Right-hand side: Case II: Fingerp
LacZ (white). For all parameters the sensitivity of the fast system (fingerprint) is conserved for selected slow observables.

with respect to selected observables of the *slow* part of the system.

5.2.1. Case study I: pulse experiment

The fast part of the system is represented by the observables of the PTS (EI, HPr, EIIA), while the slow part of the system comprises glycolytic reactions. For the pulse experiment gene expression can be neglected, due to the short period of time. For the slow part, observables glucose 6-phosphate (vector of perturbation is dark gray), PEP (light gray), and pyruvate (white) are selected (Fig. 8, left-hand side). Except for parameters k_1 and k_{-1}, the vectors are comparable, although the values for the fingerprint are slightly higher in all other cases.

5.2.2. Case study II: diauxic growth experiment

Here the fast subsystem is represented by the observables of the PTS (EI, HPr, EIIA) and the glycolysis (G6P, PEP, and Prv). The remainder system is the slow part. Fig. 8, right-hand side, summarizes the results. As representatives of the slow part, the entire biomass (vector of perturbation is dark gray) and the lactose splitting enzyme LacZ (white) are selected. For all observables the sensitivity of the slow system is conserved in the fast system. The sensitivity of glycolysis is the most important.

6. Discussion

A mathematical model for glucose uptake and metabolism is analyzed with respect to time hierarchies and robustness. Based on previous published results, simplified versions for the subsystems PTS and glycolysis are used during this study.

6.1. Dynamics

It was shown that the structure of the PTS with its successive phosphorylation steps with different dynamics can lead to temporally shifted dephosphorylation of the PTS components. A very common tool in engineering science is the analysis of the eigenvalues and eigenvectors of the linearized system. Thereby the system is transformed in new coordinates. Every coordinate reaches a steady-state with a time constant that is represented by the reciprocal of the absolute value of the respective eigenvalue. The analysis indicates that the three fastest modes reach a steady-state with time constants smaller than $\tau = 0.03$ s. It was not possible to assign one of these modes directly to one of the original observables. Palsson and Lightfoot (1984) showed that a mode may also represent an equilibrium condition. The simulation results in Fig. 6 show that this is not the case for the PTS rates. The steady-state condition at the beginning of the experiment is very close to the equilibrium (in the equilibrium all rates are zero) and shifts away after the glucose pulse. In contrast to the fast modes, the main components of the slow modes z_4 and z_5 could be assigned directly to one observable of the original system: glucose 6-phosphate and pyruvate, respectively. However, simulaton results show that the time constants cannot be detected with the nonlinear model. Although the concentration of glucose 6-phosphate rises very quickly, a steady-state is reached after 30 s that is much slower than expected from the linearized model.

Experimental data for the time course of PEP are presented in Hogema et al. (1998): the PEP concentration decreases by a factor of ≈30 within 15 s after stimulation and increases during a period of 2 min, which is slower than shown here. This might also be due to the different experimental conditions and strains used. The experiment shown in Fig. 2, which is the basis for the parameter estimation, starts under limiting conditions (the initial steady-state is reached after 2 days) while Hogema et al. (1998) performed the experiment with cells taken from the exponential growth phase. Considering the glucose uptake rate, we get the remarkable result that the maximal value is nearly 25 mmol/gDW h, showing the very high capacity of the transporter.

Based on the results above, the analysis of the linear model show that the observables of the model can be divided into two groups. One group representing PTS reactions and a second group representing glycolytic reactions. But, the linear model can give only hints on the time constants of the system. Simulation studies with the nonlinear model are therefore necessary. Including slower processes such as gene expression in the analysis, the findings indicate that the slow increase of the unphosphorylated form of EIIA during the diauxic growth experiment can be explained by a slow accumulation of intracellular glucose during lactose metabolism. As can be seen in Fig. 9, the

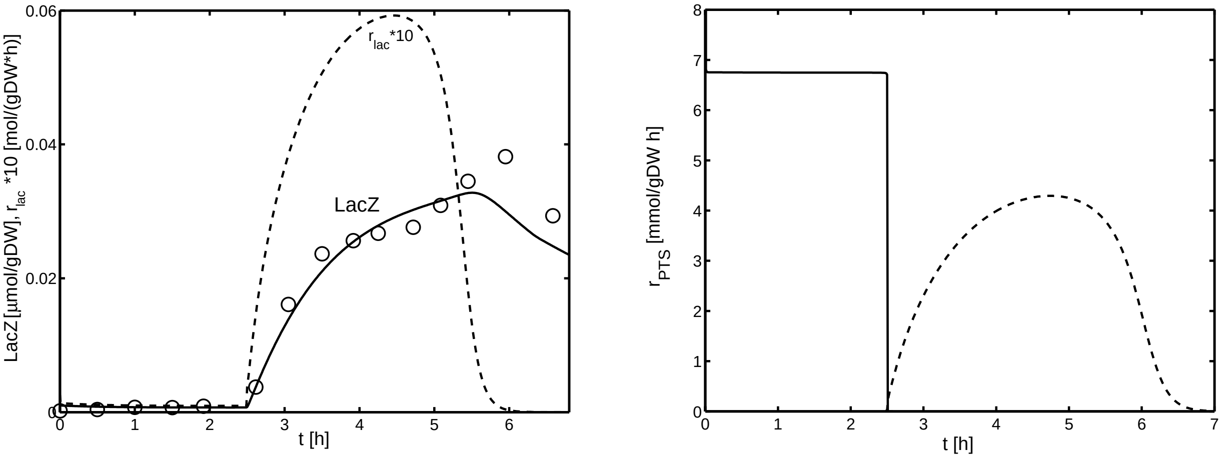


Fig. 9. Left: Time course of β-galactosidase (simulation and experimental results) and simulation of the rate of protein synthesis. Right: Uptake rate of glucose (solid), rate of phosphorylation of intracellular glucose (dashed).

induction of the *lac* operon lasts several hours, until a plateau is reached. Therefore, the rate through the PTS also rises slowly.

6.2. Steady-state characteristics

Regarding the steady-state characteristics, the degrees of phosphorylation of the PTS proteins EIIA and EI are nearly linear over a broad range (Fig. 7). For high uptake rates, the slope of the curve is decreasing, indicating a saturation behavior. This can also be seen in the plot of uptake rate versus glucose residual concentration. The values of PEP pass through a maximum at an uptake rate of 4 mmol/gDW h (close to the half maximal uptake rate), while glucose 6-phosphate and pyruvate show a linear dependency from the uptake rate. The initial values measured for PEP and pyruvate by Hogema et al. (1998) for glucose (6.8 and 2.2 μmol/gDW, respectively) are the same order of magnitude as those shown in the figure. The course of PEP reflects the very important role of this key metabolite in the entire system. For low uptake rates the PEP pool increases with increasing uptake rate, showing that PEP is available in sufficient amounts. For high uptake rates, the capacity of the glycolysis becomes more and more limiting, leading to decreasing PEP concentrations.

Since the degree of phosphorylation of protein EIIA is involved in the control of many carbohydrate transport systems, we calculated the degree of phosphorylation in dependency of the PEP/pyruvate ratio for the case where the PTS is active, and the case where the PTS is not active and a flux through glycolysis is enforced. When the PTS is not active the synthesis of pyruvate is realized only by the pyruvate kinase reaction. In Fig. 10 the pyruvate kinase flux was varied between 2 and 6 mmol/gDW h (the incoming flux to glucose 6-phosphate was fixed at 5 mmol/gDW h). As can be seen, the PEP/pyruvate ratio is decreasing with increasing pyruvate kinase flux. The right-hand side plot shows the shift of the degree of phosphorylation of EIIA in the cases where the PTS is active or not active. When the PTS is not active, from Eq. (9) the following equation will hold true for the degree of phosphorylation d_P:

$$d_\mathrm{P} = \frac{c_{\mathrm{P}\sim\mathrm{EIIA}}}{c_{\mathrm{EIIA0}}} = \frac{c_\mathrm{PEP}/c_\mathrm{Prv}}{1/K_\mathrm{eq} + c_\mathrm{PEP}/c_\mathrm{Prv}}, \quad (18)$$

showing that the curve does not depend on the overall amount of the PTS proteins. Note that Eq. (18) also represents the upper bound for the case that the PTS is active. This upper bound is reached if the concentration of the PTS proteins is increased.

The results of the model are in agreement with data in Hogema et al. (1998) (Fig.2 therein shows

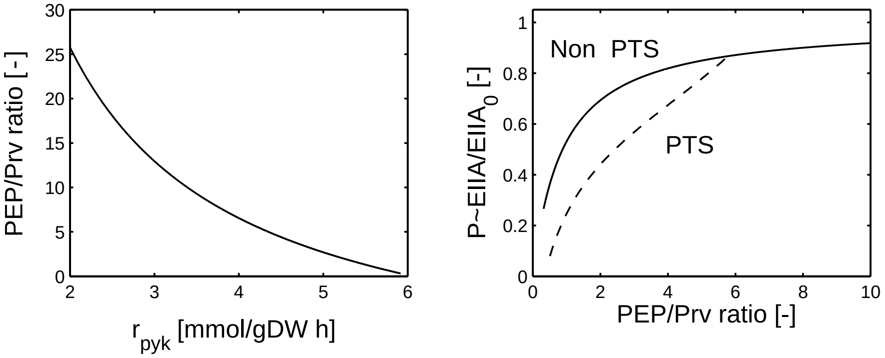


Fig. 10. Left: PEP/pyruvate ratio corresponding to the rate of pyruvate kinase when the PTS is not active. Right: Corresponding degree of phosphorylation for protein EIIA in the case that the PTS is not active (solid) and in the case where the PTS is active (dashed).

the degree of phosphorylation of EIIA for various carbon sources). Depending on the PEP/pyruvate ratio, the degree of phosphorylation is adjusted: For non-PTS sugars, e.g. glucose 6-phosphate, a low PEP/pyruvate ratio will also result in a low degree of phosphorylation, and therefore in low cAMP concentrations. If a non-PTS sugar is provided, the pyruvate kinase activity must be increased to establish the low PEP/pyruvate ratio. Data from literature indicates that the pyruvate kinase activity and concentration can be enhanced in a feed-forward loop by fructose 1,6-bis-phosphate (protein synthesis is under control of FruR, which is modified by fructose 1,6-bis-phosphate). Possibly, this results in a higher flux through pyruvate kinase if a high flux through glycolysis is possible, and therefore decreases the PEP/pyruvate ratio. Besides the pyruvate kinase activity other PEP metabolizing enzymes can also be expected to alter there activity in response to changes of the glycolytic flux and thereby also influence the PEP/pyruvate ratio. All findings indicate that for all sugars that feed into glycolysis, a high PEP/pyruvate ratio always points to a hunger situation, while a low PEP/pyruvate ratio signals a satisfactory situation.

6.3. Robustness

A new approach was introduced to analyze parameter sensitivities of slow and fast subsystems. The method calculates a vector of parameters that shows a maximal deflection of selected trajectories. These vectors are analyzed with respect to fast and slow observables of the model. The analysis of two experiments—pulse response and diauxic growth—reveals that some of the system parameters show a robust behavior: If a parameter shows no sensitivity in the fast system, it almost always show no sensitivity for slow observables. Therefore, it is concluded that the fast system may act as a kind of filter for the sensitivities. These findings may lead to an improvement during parameter identification for large systems composed of different units processing on different time scales. In the first example, the sensitivities are distributed on a number of parameters, showing that there is no single “bottleneck” in the PTS. For the second example discussed above (Case study II), it turns out that the glycolytic flux, represented by parameter k_{gly}, shows the highest sensitivity for the slow variables, e.g. biomass and LacZ. Since the slow processes reflect control of protein synthesis, e.g. via P $\sim$ EIIA, the importance of the glycolytic flux discussed already above and observed by Hogema et al. (1998) is confirmed from a theoretical point of view.

From research in microbiology, the PTS is designed as a sensor system. From our results and the cited experimental studies, it is concluded that it is a sensor system for external carbon source supply, but rather the glycolytic flux and the PEP/pyruvate ratio show a sensor function while the PTS acts more as a transmitter to process the signal to synthesize cAMP and finally activates Crp to start transcription.

Appendix A

A.1. Model equations for the simplified PTS/glycolysis model

The following rates are defined for reactions (1)–(3):

$$r_1 = k_1 c_{\mathrm{PEP}} c_{\mathrm{EI}} - k_{-1} c_{\mathrm{Prv}} (c_{\mathrm{EI0}} - c_{\mathrm{EI}}) \tag{A.1}$$

$$r_2 = k_2 c_{\mathrm{HPr}} (c_{\mathrm{EI0}} - c_{\mathrm{EI}}) - k_{-2} c_{\mathrm{EI}} (c_{\mathrm{HPr0}} - c_{\mathrm{HPr}}) \tag{A.2}$$

$$r_3 = k_3 c_{\mathrm{EIIA}} (c_{\mathrm{HPr0}} - c_{\mathrm{HPr}}) - k_{-3} c_{\mathrm{HPr}} (c_{\mathrm{EIIA0}} - c_{\mathrm{EIIA}}), \tag{A.3}$$

and the model equations read, together with Eq. (5):

$$\begin{aligned} \dot{c}_{\mathrm{EI}} &= -r_1 + r_2 \\ \dot{c}_{\mathrm{HPr}} &= -r_2 + r_3 \\ \dot{c}_{\mathrm{EIIA}} &= -r_3 + r_4. \end{aligned} \tag{A.4}$$

For glycolysis the following kinetics are used:

$$r_{\mathrm{gly}} = k_{\mathrm{gly}} \frac{c_{\mathrm{G6P}}}{K_{\mathrm{G6P}} + c_{\mathrm{G6P}}} \tag{A.5}$$

$$r_{\mathrm{pyk}} = k_{\mathrm{pyk}} \frac{c_{\mathrm{PEP}}}{K_{\mathrm{PEP}} + c_{\mathrm{PEP}}} \tag{A.6}$$

$$r_{\mathrm{dr1}} = k_{\mathrm{g6p}} c_{\mathrm{G6P}} \tag{A.7}$$

$$r_{\mathrm{dr2}} = k_{\mathrm{pep}} c_{\mathrm{PEP}} \tag{A.8}$$

$$r_{\mathrm{dr3}} = k_{\mathrm{prv}} c_{\mathrm{PRV}}, \tag{A.9}$$

and the equation system read:

$$\begin{aligned} \dot{c}_{\mathrm{G6P}} &= r_4 - r_{\mathrm{gly}} - r_{\mathrm{dr1}} \\ \dot{c}_{\mathrm{PEP}} &= -r_1 + 2 \times r_{\mathrm{gly}} - r_{\mathrm{pyk}} - r_{\mathrm{dr2}} \\ \dot{c}_{\mathrm{Prv}} &= r_1 + r_{\mathrm{pyk}} - r_{\mathrm{dr3}}. \end{aligned} \tag{A.10}$$

A.2. Method of Hearne (1985)

The method of Hearne (1985) uses the normalized sensitivities ω_{ij}

$$\omega_{ij} = w_{ij} \frac{p_j}{x_i} = \frac{\partial x_i}{\partial p_j} \frac{p_j}{x_i},$$

$$\boldsymbol{\Omega} = \begin{bmatrix} \omega_{11} & \cdots & \omega_{1m} \\ & \cdots & \\ \omega_{n1} & \cdots & \omega_{nm} \end{bmatrix}, \tag{A.11}$$

and the vector of sensitivities $\boldsymbol{s}$

$$\boldsymbol{s} = \left[\frac{\Delta p_1}{p_1} \cdots \right], \tag{A.12}$$

to formulate the optimization problem

$$\max \quad \boldsymbol{s}^{\mathrm{T}} \left(\sum \boldsymbol{\Omega}^{\mathrm{T}} \boldsymbol{\Omega} \Delta t \right) \boldsymbol{s}. \tag{A.13}$$

With scaling of the sensitivities

$$\boldsymbol{s}^{\mathrm{T}} \boldsymbol{s} = 1 \tag{A.14}$$

as additional constraint, the problem can be reformulated as an Euler–Lagrange equation with the solution

$$G\boldsymbol{s} = \lambda \boldsymbol{s}, \tag{A.15}$$

with

$$\boldsymbol{G} = \sum \boldsymbol{\Omega}^{\mathrm{T}} \boldsymbol{\Omega} \Delta\, t. \tag{A.16}$$

References

Agrawal, A., 1999. New institute to study systems biology. Nat. Biotechnol. 17, 743.

Barkai, N., Leibler, S., 1997. Robustness in simple biochemical networks. Nature 387.

Bergmeyer, H.U., 1979. Methoden der Enzymatischen Analyse. Verlag Chemie, Weinheim.

DeReuse, H., Danchin, A., 1988. The *ptsH*, *ptsI*, and *crr* genes of the *Escherichia coli* phosphoenolpyruvate-dependent phosphotransferase system—a complex operon with several modes of transcription. J. Bacteriol. 170, 3827–3837.

Francke, C., Westerhoff, H.V., Blom, J.G., Peletier, M.A., 2002. Flux control of the bacterial phosphoenolpyruvate: glucose phosphotransferase system and the effect of diffusion. Mol. Biol. Rep. 29, 21–26.

Hearne, J.W., 1985. Sensitivity analysis of parameter combinations. Appl. Math. Model. 9, 106–108.

Heinrich, R., Schuster, S., 1996. The regulation of cellular processes. Chapman & Hall.

Hogema, B.M., Arents, J.C., Bader, R., Eijkemanns, K., Yoshida, H., Takahashi, H., Aiba, H., Postma, P.W., 1998. Inducer exclusion in *Escherichia coli* by non-PTS substrates: the role of the PEP to pyruvate ratio in determining the phosphorylation state of enzyme IIA^{Glc}. Mol. Microbiol. 30, 487–498.

Ishizuka, H., Hanamura, A., Kunimura, T., Aiba, H., 1993. A lowered concentration of cAMP receptor protein caused by glucose is an important determinant for catabolite repression in *Escherichia coli*. Mol. Microbiol. 10, 341–350.

Kitano, H., 2000. Perspectives on systems biology. New Gen. Comput. 18 (3), 199–216.

Kremling, A., Gilles, E.D., 2001. The organization of metabolic reaction networks: II. Signal processing in hierarchical structured functional units. Metab. Eng. 3 (2), 138–150.

Kremling, A., Bettenbrock, K., Laube, B., Jahreis, K., Lengeler, J.W., Gilles, E.D., 2001a. The organization of metabolic reaction networks: III. Application for diauxic growth on glucose and lactose. Metab. Eng. 3 (4), 362–379.

Kremling, A., Sauter, T., Bullinger, E., Ederer, M., Allgöwer, F., Gilles, E.D., 2001b. Biosystems engineering: applying methods from systems theory to biological systems. In: Yi, T.-M., Hucka, M., Morohashi, M., Kitano, H. (Eds.), Proc. of the Second International Conference on Systems Biology, California Institute of Technology, Pasadena, CA, pp. 282–290.

Lee, S.J., Boos, W., Bouche, J.P., Plumbridge, J., 2000. Signal transduction between a membrane-bound transporter, PtsG, and a soluble transcription factor, Mlc, of *Escherichia coli*. EMBO J. 19, 5353–5361.

Liao, J.C., Hou, S.-Y., Chao, Y.-P., 1996. Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. Biotechnol. Bioeng. 52, 129–140.

Ljung, L., 1999. System Identification—Theory for the User, 2nd ed. Prentice-Hall PTR, Upper Saddle River, NJ.

Palsson, B.O., Lightfoot, E.N. 1984. Mathematial modeling of dynamics and control in metabolic networks: I. On Michaelis–Menten kinetics. J. Theor. Biol. 111, 273–302.

Plumbridge, J., 1998. Expression of *ptsG*, the gene for the major glucose pts transporter in *Escherichia coli*, is repressed by Mlc and induced by growth on glucose. Mol. Microbiol. 29 (4), 1053–1063.

Posten, C., Munack, A., 1990. On-line application of parameter estimation accuracy to biotechnical processes. In: Proceedings of the American Control Conference, vol. 3, pp. 2181–2186.

Postma, P.W., Lengeler, J.W., Jacobson, G.R., 1993. Phosphoenolpyruvat: carbohydrate phosphotransferase systems of bacteria. Microbiol. Rev. 57, 543–594.

Rohwer, J.M., Meadow, N.D., Roseman, S., Westerhoff, H.V., Postma, P.W., 2000. Understanding glucose transport by the bacterial phosphoenolpyruvate:glycose phosphotransferase system on the basis of kinetic measurements in vitro. J. Biol. Chem. 275, 34909–34921.

Rojdestvenski, I., Cottam, M., Park, Y.I., Öquist, G., 1999. Robustness and time-scale hierarchy in biological systems. BioSystems 50, 71–82.

Schmid, K., Schupfner, M., Schmitt, R., 1982. Plasmid mediated uptake and metabolism of sucrose by *Escherichia coli*. J. Bacteriol. 151, 68–75.

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., Gilles, E.D., 2002. Metabolic network structure determines key aspects of functionality and regulation. Nature 420, 190–193.

Takahashi, H., Inada, T., Postma, P., Aiba, H., 1998. CRP down-regulates adenylate cyclase activity by reducing the level of phosphorylated IIAGlc, the glucose-specific phosphotransferase protein, in *Escherichia coli*. Mol. Gen. Genet. 259, 317–326.

Tanaka, Y., Kimata, K., Aiba, H., 2000. A novel regulatory role of glucose transporter of *Escherichia coli*: membrane sequestration of a global repressor Mlc. EMBO J. 19, 5344–5352.

Wang, J., Gilles, E.D., Lengeler, J.W., Jahreis, K., 2001. Modeling of inducer exclusion and catabolite repression based on a PTS-dependent sucrose and non-PTS-dependent glycerol transport systems in *Escherichia coli* K-12 and its experimental verification. J. Biotechnol. 92, 133–158.

Wohlieter, J.A., Lazare, J.R., Snellings, N.J., Johnson, E.M., Syneki, R.M., Baron, R.S., 1975. Characterization of a transmissible genetic element from sucrose fermenting samonella strains. J. Bacteriol. 122, 401–406.

Zeppenfeld, T., Larisch, C., Lengeler, J.W., Jahreis, K., 2000. Glucose transporter mutants of *Escherichia coli* K-12 with changes in substrate recognition of the IICB^{Glc} and induction behavior of the *ptsG* gene. J. Bacteriol. 182, 4443–4452.

Workbench zur Modellbildung, Simulation und Analyse zellulärer Systeme

Workbench for Model Set Up, Simulation, and Analysis of Cellular Systems

Andreas Kremling, Steffen Klamt, Martin Ginkel, Ernst Dieter Gilles, Max-Planck-Institut für Dynamik komplexer technischer Systeme, Magdeburg

Zusammenfassung Die aktuelle Forschung in der molekularen Genetik und die Erfolge bei der Analyse von Genexpression und Proteinfunktion führen zu einer bisher unerreichten Fülle von Informationen über biologische Phänomene. Werkzeuge, die eine quantitative Beschreibung und Analyse ermöglichen, haben dabei eine entscheidende Bedeutung. Der Beitrag stellt Werkzeuge zur Modellerstellung, -simulation und -analyse vor, die bereits für eine Anzahl von biologischen Modellsystemen (Bakterien, Hefen) angewendet wurden. Das Werkzeug ProMoT dient zur automatischen Erstellung der Modellgleichungen, die anschliessend vom Gleichungslöser Diva numerisch untersucht werden können. Eine Analyse der Modellstruktur sowie die Berechnung von stationären Flüssen ist mit dem FluxAnalyzer möglich. ►►► **Summary** Current research in molecular genetics and success in developing methods for functional genomics and proteomics lead to an unrivaled knowledge on biological systems. Tools that allow for a quantitative description and analysis become therefore more and more important. The paper at hand introduces tools for model set up, dynamical simulation and analysis. The tools were successfully applied for a number of biological model systems. Tool ProMoT supports the automatical generation of model equations; the equations were afterwards solved numerically with Diva. To analyze the structure of and to calculate flux distributions in metabolic networks, the FluxAnalyzer was developed.

KEYWORDS J.3 [Life and Medical Science] Workbench, cellular systems

1 Einleitung

Die Erfolge der modernen molekularen Biologie bei der Analyse der genetischen Strukturen bei einer Vielzahl von Organismen hat die Bioinformatik zu einer sehr populären Wissenschaft gemacht. Heute stehen daher eine große Anzahl von Datenbanksystemen zur Verfügung, die umfangreiche Datenmengen speichern und strukturieren. Diese Datenbanken konzentrieren sich hauptsächlich auf Sequenzdaten, Stoffwechsel- und Regulationswege unterschiedlicher Mikroorganismen. Neue Messtechniken wie beispielsweise die Chip-Technologie und die Gel-Elektrophorese gestatten heute einen quantitativen Blick in die Zelle. Andere Techniken erlauben die Messung von intrazellulären Metaboliten in einem Zeitfenster von 2/100 Sekunden und gestatten daher die Analyse von sehr schnellen Dynamiken, wenn die Zellen entsprechend angeregt werden. Diese beiden Entwicklungen – verfügbares Wissen über die genetische Organisation von Zellen und die neuen Messtechniken – ebnen den Weg der Biologie von einer qualitativen zu einer quantitativen Wissenschaft. Stoffwechsel- und Regulationswege sind jedoch durch eine große Anzahl interagierender Komponenten gekennzeichnet. Um diese komplexen Systeme besser verstehen und womöglich auch Vorhersagen über das ganzheitliche Verhalten machen zu können, bedarf es aber weiterer Hilfsmittel. Eines dieser Hilfsmittel ist die detaillierte mathematische Beschreibung von zellulären Vorgängen. Mathematische Modelle stellen dann das Herzstück einer neuen Vorgehensweise bei der Analyse des Wachstums- und Produktbildungsverhaltens von zellulären Systemen dar. Dieser neue Forschungsansatz wird als *Systems Biology* oder manchmal auch als *Computational Biology* bezeichnet. Er zeichnet sich durch eine starke Kooperation zwischen Naturwissenschaftlern, Informatikern und Ingenieuren aus und versucht, durch eine starke Kopplung zwischen Experiment und Theorie einen Beitrag zum verbesserten Verständnis der in einer Zelle ablaufenden Vorgänge zu leisten.

In diesem Beitrag sollen einige Rechnerwerkzeuge vorgestellt werden, die in unserer Arbeitsgruppe entwickelt und eingesetzt werden. Diese betreffen die Bereiche Modellentwicklung, dynamische Simulation, Parameteridentifikation, dynamische Optimierung sowie die Analyse stationärer Modelle und zellulärer Netzwerke. Bild 1 gibt einen Überblick und zeigt, wie die einzelnen Werkzeuge gekoppelt sind.

Das Tool ProMoT dient dabei als Repräsentation des verfügbaren biologischen Wissens und als Hilfsmittel, um die komplexen Modelle zu erstellen. Bei beiden Aufgaben stützt es sich auf ein Modellierungskonzept, welches für die Modellierung von zellulären Systemen entwickelt wurde. Darauf wird im nächsten Abschnitt kurz eingegangen. Mit der Simulationsumgebung Diva besteht nicht nur die Möglichkeit, dynamische Simulationsstudien zu betreiben, sondern das Werkzeug, das sich durch eine sehr effiziente Numerik auszeichnet, kann zur Analyse von Sensitivitäten und zur Schätzung von unsicheren und unbekannten Parametern herangezogen werden. Wichtige Merkmale von ProMoT–Diva im Unterschied zu anderen Simulationswerkzeugen sind die graphische Benutzeroberfläche, der durchgehend modulare Aufbau der Modelle bei gleichzeitiger Transparenz des zugrundeliegenden Gleichungssystems sowie die sehr effiziente Numerik (siehe auch [1]). Das Werkzeug FluxAnalyzer dient schließlich der Analyse von stationären Modellen und zellulären Netzwerken. Es stellt eine ganze Reihe von Methoden, die aus den Bereichen *Metabolic Flux Analysis* und *Metabolic Pathway Analysis* bekannt sind, zur Verfügung.

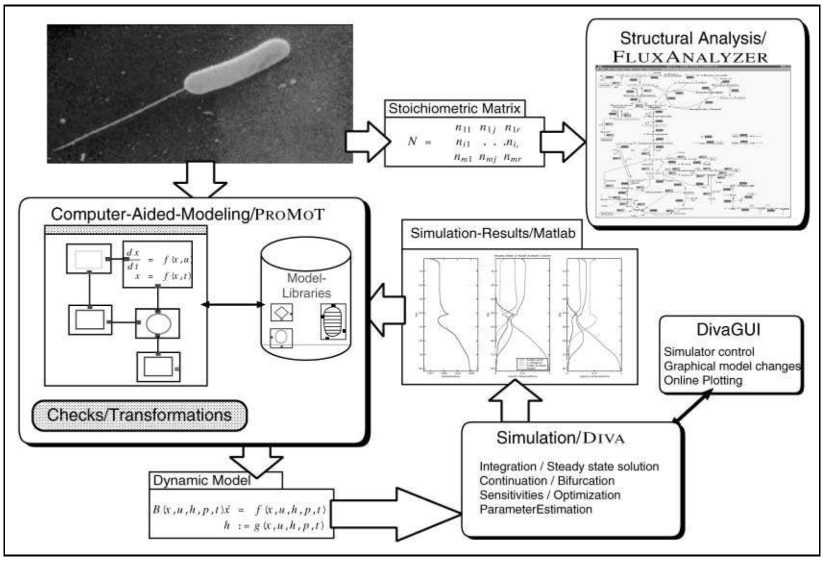


Bild 1 Überblick über die vorgestellten Werkzeuge. In ProMoT wird das verfügbare biologische Wissen strukturiert und so aufbereitet, dass mathematische Modelle aus vorgefertigten Teilmodellen erstellt werden können. Die Teilmodelle sind in einer umfangreichen Modellbausteinbibliothek abgelegt. Die Modelle können dann so abgespeichert werden, dass mit der Simulationsumgebung Diva Sensitivitäts- und Simulationsstudien durchgeführt werden können. Für die Analyse von stationären Modellen steht der FluxAnalyzer zur Verfügung. Dieser benötigt nur die stöchiometrischen Koeffizienten des biochemischen Reaktionsnetzwerkes.

2 Modellierungskonzept für zelluläre Systeme

An dieser Stelle sollen in aller Kürze Grundzüge eines Modellierungskonzeptes vorgestellt werden, welches eine strukturelle Dekomposition des zellulären Reaktionsnetzwerkes erlaubt [12]. Dabei wird davon ausgegangen, dass die Biophase in ihrem globalen Verhalten einer gemittelten Zelle entspricht. Eine zentrale Idee des Konzeptes ist es, dem Anwender Modellbausteine zur Verfügung zu stellen, die parametrisiert und mit anderen Modellbausteinen zu höher strukturierten Modellen – den Funktionseinheiten – verschaltet werden. Die Modellbausteine besitzen strukturelle Eigenschaften und verhaltensbeschreibende Eigenschaften. Die strukturellen Eigenschaften erlauben eine passende Verschaltung der Bausteine, während die Verhaltensbeschreibung den einzelnen Bausteinen eine mathematische Beschreibung zuordnet.

Bausteine, die einzelne Metabolite oder Proteine beschreiben, sowie Bausteine, die die biochemische Stoffumwandlung charakterisieren, werden als elementare Modellbausteine bezeichnet. Aus ihnen kann ein mathematisches Modell des gesamten Stoffwechsels aufgebaut werden. Zur vollständigen Beschreibung der in einer Zelle ablaufenden Prozesse sind jedoch noch weitere Grundbausteine notwendig: Zelluläre Systeme sind in der Lage, sich sehr schnell auf ändernde Umweltbedingungen einzustellen. Dies liegt zum einen an der Möglichkeit der Informationsverarbeitung, um einen äußeren Reiz – etwa in Form einer drastischen Veränderung der Substratkonzentration – in ein zelluläres Signal umzuwandeln. Dieses Signal wird weitergeleitet und verarbeitet, um eine zelluläre Antwort hervorzurufen. Die Prozesse der Signaltransduktion beruhen hauptsächlich auf Wechselwirkungen von Proteinen. Des Weiteren besitzt die Zelle eine hohe Anzahl von Steuer- und Regelkreise, die es erlauben, gewünschte Stoffwechselwege zu- oder abzuschalten und damit in optimaler Weise auf die neue Situation zu reagieren. Dies äußert sich beispielsweise in der Änderung der Syntheserate oder Aktivität der entsprechenden Stoffwechselenzyme. Zur Beschreibung dieser Prozesse wird zusätzlich die Modellklasse der Signalwandler benötigt.

Die Aggregation der elementaren Modellbausteine sollte nach vorgegebenen Regeln erfolgen. Beim vorgestellten Konzept werden dabei drei biologisch motivierte Kriterien herangezogen, die eine Gruppierung von elementaren Modellbausteinen zu Funktionseinheiten erlauben. Die Ableitung mehr formaler Kriterien ist Gegenstand laufender Forschungsarbeiten.

3 Modellerstellung mit ProMoT

3.1 Objektorientierte Modellerstellung

ProMoT [1] ist ein objektorientiertes, gleichungsbasiertes Modellierungswerkzeug. Es kann kontinuierliche und gemischt kontinuierlich-ereignisdiskrete Modelle für die Simulationsumgebung Diva erstellen. Die Modelle werden dabei aus einer abstrakten symbolischen Repräsentation im Modellierungswerkzeug in Fortran Unterprogramme überführt und übersetzt, wodurch eine sehr effiziente Simulation ermöglicht wird. ProMoT erlaubt dem Modellierer, objektorientierte Techniken zu benutzen. Modellbausteine in ProMoT sind Klassen, die einen lokalen Zustand kapseln, als Container für andere Bausteine dienen und durch Vererbung spezialisiert und erweitert werden können. Es werden strukturelle und verhaltensbeschreibende Modellbausteine unterschieden. Strukturell wird das Gesamtmodell in *Module* unterteilt, die bestimmten biologischen Einheiten auf verschiedenen Ebenen des Gesamtsystems entsprechen. Entsprechend dem vorgestellten Modellierungskonzept werden auf der untersten Ebene molekularbiologische Spezies und Reaktionen als Module beschrieben, aber auch höher strukturierte Funktionseinheiten wie Stoffwechselwege, Regulationssysteme und ganze Bioreaktoren werden durch Module repräsentiert. Module besitzen Schnittstellen, so genannte *Terminals*. Durch die Terminals können die gekapselten Module miteinander interagieren. Das lokale Verhalten eines Moduls wird durch Variablen, algebraische Gleichungen und gewöhnliche Differentialgleichungen beschrieben. Das Gesamtgleichungssystem kann im allgemeinen Fall ein differentialalgebraisches System mit einem Index kleiner gleich 1 sein. Einige Variablen der Module werden den Terminals zugeordnet und anschließend bei einer Verknüpfung der Terminals mit Variablen anderer Module durch Koppelgleichungen verbunden. Diskrete Zustandsänderungen werden mit Hilfe von Petrinetzen repräsentiert, deren Transitionen in Abhängigkeit der kontinuierlichen Variablen feuern. In der Systembiologie wird dies aber bislang wenig verwendet, daher soll hier nicht näher darauf eingegangen werden.

Modellbausteine in ProMoT sind in einer objektorientierten Klassenhierarchie mit multipler Vererbung organisiert. Dieses Konzept aus der Informatik wurde aufgegriffen, um komplexe Bibliotheken von Modellbausteinen flexibel zu gestalten und besser organisieren zu können. Für systembiologische Modellierungsprojekte wurde eine Bibliothek von Modulen erstellt, die sowohl elementare Modellelemente wie Stoffspeicher und Stoffwandler aber auch wiederholt vorkommende Funktionseinheiten zur Beschreibung der Genexpression und der Signaltransduktion modellieren. Eine direkte Anbindung an Datenbanken, um biochemische Daten direkt in Modelle umzuwandeln, ist z. Zt. nicht vorgesehen. Allerdings ist geplant, Modelle, die in SBML [4] vorliegen, einzulesen und Modellvarianten auch in SBML zu exportieren.

Die Module in ProMoT besitzen standardisierte Terminals, die eine universelle Verknüpfung der Bausteine ermöglichen. Die Terminals repräsentieren dabei Signale (Kon-

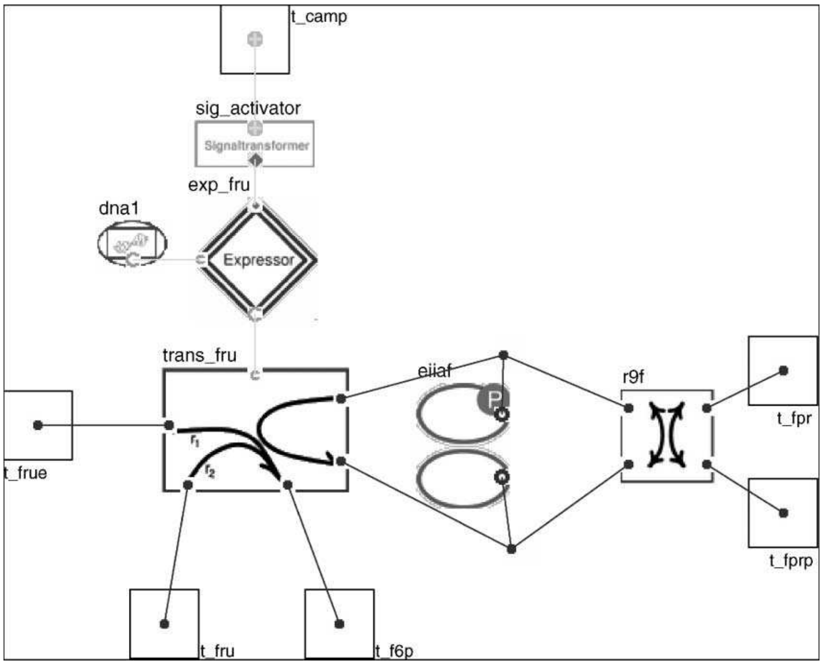


Bild 2 Modell eines Fruktosetransportweges von *E. coli* im graphischen Editor von ProMoT. Der Stoffwandler `trans_fu` wandelt sowohl externe als auch interne Fruktose (durch die Terminals `t_frue`, `t_fru`) in einer Phosphorylierungsreaktion in Fruktose-6P (`t_f6p`) um. Die dazu benötigte Energie wird duch die phosphorylierte Form des Enzyms `eiiaf` bereitgestellt. Die Synthese des Transportenzyms wird durch den Expressionsbaustein (durch die Doppelraute repräsentiert) beschrieben.

zentrationen oder Konzentrationsverhältnisse) oder Stoffflüsse (bidirektionaler Austausch einer Konzentration und einer Flussrate). Benutzer können neue Module mit Hilfe eines graphischen Editors oder textuell in der *Model Definition Language* (MDL) von ProMoT eingeben. Im graphischen Editor (siehe Bild 2) wird ein Flussdiagramm bearbeitet, in dem Module aus der geladenen Modellbibliothek durch Drag'n Drop aggregiert und miteinander verbunden werden. Mit Hilfe der Modelliersprache können spezielle eigene Gleichungsmodelle in elementaren Modulen implementiert werden. Dabei kann der Modellierer auf abstrakte Superklassen aus der Modellbibliothek zurückgreifen. MDL ist eine deklarative, objektorientierte Sprache, die eine Beschreibung der Modellelemente, aber keinen imperativen Code enthält. Sie wird vom Modellierungswerkzeug gelesen und geschrieben und wird auch als Datenformat zur Speicherung der Modellbibliotheken genutzt.

3.2 Implementation des Modellierungswerkzeuges

Ist der Aufbau eines Modells abgeschlossen, generiert ProMoT daraus simulationsfähigen Code. Dazu wird die Instanz einer Modulklasse erzeugt (siehe Bild 3), wobei alle aggregierten Bestandteile auf ihre Vollständigkeit (z. B. fehlende Referenzen, Startwerte für die Variablen) überprüft werden. Anschließend wird das resultierende Gleichungssystem für den Simulator erzeugt. Dabei werden alle in den Modulen definierten Gleichungen und Koppelgleichungen zusammengefasst und strukturell analysiert. Strukturelle Fehler des Gleichungssystems werden hierbei erkannt und dem Nutzer mit Hinweisen zur Fehlerbehebung signalisiert. Fehlerfreie Gleichungssysteme werden symbolisch transformiert, um die Effizienz des Simulationscodes zu steigern. Dabei werden einfache, explizite algebraische Gleichungen durch direkte Zuweisungen ersetzt und latente Variablen eliminiert, die vom Nutzer nicht zur Ausgabe markiert wurden. Dieser Schritt ist notwendig, weil die feinkörnige Strukturierung der Module zwar flexibleres Arbeiten bei der Modellerstellung erlaubt aber auch eine große Anzahl solcher Gleichungen und Variablen erzeugt. Das endgültige Gleichungssystem wird mit Hilfe eines Code-Generators in Fortran-Unterprogramme codiert und kann dann mit den numerischen Bibliotheken von Diva compiliert und gelinked werden.

ProMoT ist, wie in Bild 3 dargestellt, in zwei Teilen implementiert:

(1) einem Kernel, der alle Manipulationen der Modellelemente und der Dateien mit den MDL-Quelltexten durchführt und die Algorithmen zur Gleichungsmanipulation enthält und

(2) einer graphischen Oberfläche in Java, die verschiedene Ansichten auf die Wissensbasis von Modellelementen und das graphische Editieren von Modulen erlaubt.

Diese beiden Teile interagieren als Modell und View/Controller miteinander. Die Kommunikation der beiden Teile erfolgt über Corba.

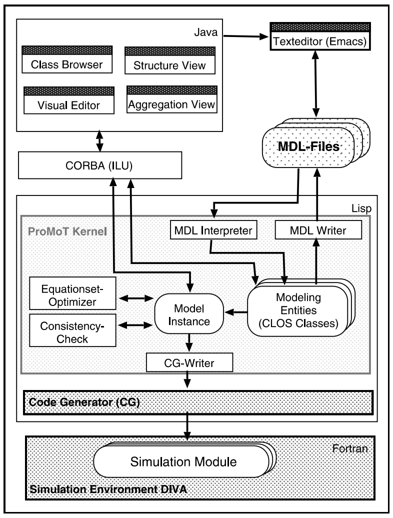


Bild 3 Softwarearchitektur von ProMoT: Der Kern ist in objektorientiertem Common Lisp implementiert, die graphische Oberfläche basiert auf JFC/Swing und interagiert nach einem MVC Pattern mit dem Kern. Die Kommunikation erfolgt über Corba.

4 Numerische Analyse des dynamischen Modells mit Diva

Eine numerische Analyse der mit ProMoT erstellten Modelle erfolgt mit der Simulationsumgebung Diva, die eine ganze Reihe von Methoden zur Berechnung stationärer Zustände und zeitlicher Verläufe aus den nichtlinearen Differentialgleichungen bietet. Die Gleichungen müssen dazu in der linear-impliziten Form

$$\boldsymbol{B}\,\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{p}, t) \tag{1}$$

vorliegen. Hierbei steht $\boldsymbol{x}$ für den Zustandsvektor, $\boldsymbol{p}$ für den Parametervektor, $\boldsymbol{y}$ für den Ausgangsvektor und t für die Simulationszeit. Matrix $\boldsymbol{B}$ ist die so genannte Descriptor-Matrix des Differential-Algebra-Systems. Gleichungen $\boldsymbol{f}$ repräsentieren die rechte Seite des gesamten Systems. Im Folgenden sollen zwei Aspekte besonders herausgestellt werden:

(1) Parameteranalyse auf der Basis verfügbarer Messdaten und

(2) Parameterschätzung.

4.1 Parameteranalyse und -sensitivitäten

Die Analyse von Sensitivitäten wird oft mit der *Metabolic Control Analysis* [3] in Verbindung gebracht. Diva benutzt die Parametersensitivitäten $w_{ij} = \partial x_i / \partial p_j$ zur Unterstützung der Parameterschätzung. Dazu wird in einem ersten Schritt eine Parameterkombination gesucht, die einen großen Einfluss auf ausgewählte Kurvenverläufe besitzt. Normalerweise werden hierzu nicht alle Zustandsgrößen ausgewählt, sondern nur solche, für die Messwerte zur Verfügung stehen. Das Verfahren von Hearne [2] berechnet einen Parametervektor, der die Trajektorien maximal aus der Ausgangslage auslenkt. Parameter, die in diesem Vektor eine hohe Gewichtung besitzen, zeigen eine große Sensitivität. Diva erlaubt auch die Berechnung von Sensitivitäten von Reaktionsraten bezüglich ausgewählter Parameter. Dies erlaubt im Besonderen auch die Ermittlung von Flusskontrollkoeffizienten, wie sie aus der *Metabolic Control Analysis* bekannt sind.

In einem zweiten Schritt ist nun zu prüfen, ob die sensitiven Parameter auf der Basis der verfügbaren Messdaten überhaupt geschätzt werden können. Dazu wird die Fisher-Informations-Matrix herangezogen [13]. Diese Matrix ist durch

$$\boldsymbol{F} = \sum_{k=1}^{N} \left[\boldsymbol{W}(t_k)^T \boldsymbol{C}(t_k)^{-1} \boldsymbol{W}(t_k) \right]$$

definiert, wobei $\boldsymbol{W}(t_k) = \partial \boldsymbol{x} / \partial \boldsymbol{p}$ die Matrix der Sensitivitäten ist, und $\boldsymbol{C}(t_k)$ die Kovarianzmatrix. Wendet man nun eine Methode an, die in [15] vorgeschlagen wurde, kann man Gruppen von Parametern bestimmen, die zusammen geschätzt werden können. Dazu ist eine Varianz γ der Parameter vorzugeben, die man im besten Fall erreichen will. Die Wahl von γ hängt von der Güte der Messdaten und den Anforderungen an die Parameterschätzung ab. Bei zellulären Systemen muss man aber davon ausgehen, dass die Parameter doch eine recht breite Streuung ($\pm 20\%$) besitzen.

4.2 Parameterschätzung

Sind sensitive Parameter bestimmt, die auch – basierend auf den Messdaten – geschätzt werden können, findet die eigentliche Parameterschätzung statt. Dabei ist man allerdings auf die einmal festgelegte Modellstruktur beschränkt. Die Messdaten z_{ik} liegen für ausgewählte Zustandsgrößen zu diskreten Zeitpunkten t_k vor. Das Ziel der Parameterschätzung ist die Minimierung der Zielfunktion $\Phi(\boldsymbol{p})$, die durch

$$\sum_{k=1}^{N} \sum_{i=1}^{n} w_{ik}^2 \left(\frac{x_i(\boldsymbol{x}_o, \boldsymbol{u}, \boldsymbol{p}, t_k) - z_{ik}}{z_i^{\max}} \right)^2$$

gegeben ist. Hierbei erlauben die Faktoren w_{ik} und $z_i^{\max}$ die Skalierung einzelner Messpunkte oder des gesamten Experimentes. Für die Optimierung stehen unter anderem eine SQP (***S**equential **Q**uadratic **P**rogramming*) Methode aus der NAG Bibliothek [14] zur Verfügung.

5 Struktur- und Stoffflussanalyse in biochemischen Netzwerken mit dem FluxAnalyzer

5.1 Grundlagen

Auf der Basis klassischer biochemischer Methoden und nun auch sequenzierter Genome lassen sich ganze Stoffwechselnetze für verschiedene Organismen rekonstruieren, die über Datenbanken wie KEGG [5] und MetaCyc [6] zugänglich sind. Die stöchiometrische Struktur metaboler Netzwerke kann daher als die am besten charakterisierte Datengrundlage für die Modellierung betrachtet werden. Viele Stoffwechselmodelle lassen sich (etwas vereinfacht zu Gleichung (1)) darstellen als

$$\dot{\boldsymbol{x}} = \boldsymbol{N}\boldsymbol{r}(\boldsymbol{x}, \boldsymbol{p}, t)\,. \qquad (2)$$

Hier steht Vektor $\boldsymbol{x}$ für die Metabolitkonzentrationen, $\boldsymbol{N}$ ist die stöchiometrische Matrix, die gerade die Netzstruktur konserviert (Zeilen: Metabolite, Spalten: die Reaktionen mit den stöchiometrischen Koeffizienten) und $\boldsymbol{r}$ ist der Reaktionsratenvektor. Letzterer wird durch eine Funktion beschrieben, die die Kinetik des Reaktionsmechanismus repräsentiert und von den Konzentrationen und kinetischen Parametern abhängt. Insbesondere sind die Parameter oft kaum bekannt. Aufgrund der Tatsache, dass die Metabolitenkonzentrationen in der Zelle approximativ im Fließgleichgewicht vorliegen, vereinfacht man Gleichung (2) zu:

$$0 = \dot{\boldsymbol{x}} = \boldsymbol{N}\boldsymbol{r}\,. \qquad (3)$$

Dieses Gleichungssystem ist Grundlage von strukturellen (topologischen, stöchiometrischen) und stationären Analysen in Stoffwechselnetzen und ist durch die Struktur der Matrix $\boldsymbol{N}$ charakterisiert. Die speziellen Methoden zur Analyse von Gleichung (3), die insbesondere der *Metabolic Flux Analysis* und *Metabolic Pathway Analysis* gewidmet sind, legen nahe, eine angepasste Modellierungs- und Visualisierungsumgebung bereitzustellen. Diese wurde mit dem FluxAnalyzer realisiert. Während Diva also für die dynamische Simulation eines in ProMoT erstellten Modells herangezogen wird – die Ausgabe der Ergebnisse erfolgt in Matlab – ist der FluxAnalyzer speziell für stöchiometrische Analysen eines biochemischen Netzwerkes entwickelt worden.

5.2 Aufbau des FluxAnalyzers

Der FluxAnalyzer [9] ist ein Paket für das kommerzielle Programm MatLab (Mathworks, Inc.) und profitiert dadurch von bereits implementierten algebraischen Funktionen, sowie durch die eleganten Möglichkeiten in MatLab, Benutzeroberflächen zu konstruieren. Für die Analyse eines beliebigen Reaktionsnetzes legt der Benutzer ein Netzwerk-Projekt an, das aus zwei Teilen besteht (Bild 4):

(1) *Abstrakte Netzstruktur:* Mittels Masken kann der Benutzer neue Netzwerkelemente (vom Typ

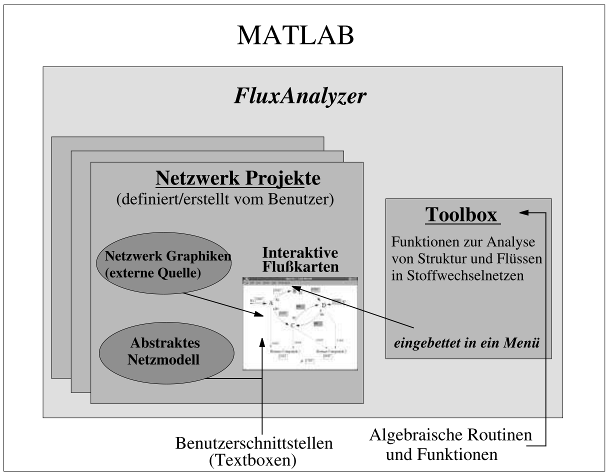

MATLAB
FluxAnalyzer
Netzwerk Projekte
(definiert/erstellt vom Benutzer)
Toolbox
Funktionen zur Analyse von Struktur und Flüssen in Stoffwechselnetzen
Netzwerk Graphiken (externe Quelle)
Interaktive Flußkarten
Abstraktes Netzmodell
eingebettet in ein Menü
Benutzerschnittstellen (Textboxen)
Algebraische Routinen und Funktionen

Bild 4 Aufbau des FluxAnalyzers.

„Metabolit“ oder „Reaktion“) deklarieren, die bestimmte Eigenschaften haben. So kann für jede Reaktion unter Anderem eine symbolische Reaktionsgleichung definiert – dies dient dem Aufbau der stöchiometrischen Matrix – und eine maximale und minimale Reaktionsrate angegeben werden. Alternativ kann eine Netzstruktur auch von ProMoT ex- und dann in den FluxAnalyzer importiert werden (Eine Darstellung der Ergebnisse des FluxAnalyzer in ProMoT ist nicht angedacht).

(2) *Interaktive Flusskarten (IFK):* Diese stellen das Kernkonzept der Interaktion und Visualisierung im FluxAnalyzer dar (siehe Beispiel in Bild 5).

Der Benutzer stellt selbstkreierte oder aus anderen Quellen (wie Datenbanken) erhältliche Netzwerkgrafiken zur Verfügung, die das Netz graphisch repräsentieren und als Hintergrund für die IFK dienen. Auf diesen Grafiken können dann Textboxen installiert werden, die jeweils einem Netzwerkelement zugeordnet sind. Beispielsweise kann an einem Reaktionspfad in der Grafik die zugehörige Textbox der entsprechenden Reaktion platziert werden, die dann die Reaktionsrate der Reaktion nach Berechnungen anzeigt. Ganz bewusst wurde hier auf Algorithmen zum automatischen Zeichnen des Netzgraphen verzichtet, da diese oftmals nicht den subjektiven Kriterien und Wünschen des Benutzers entsprechen. Außerdem sind so auch beliebige Annotationen möglich.

Jedes Netzwerk-Projekt kann vom Benutzer mit einer umfangreichen Sammlung an Algorithmen und Methoden für die metabole Fluss-, Struktur- und Pathwayanalyse untersucht werden. Diese können bequem in einem Menü gestartet und ohne detaillierte mathematische Kenntnisse ausgeführt werden. Die Ergebnisse werden direkt in den IFK in den Textboxen ausgegeben.

5.3 Funktionen zur Analyse von Stoffwechselnetzen

Metabolic Flux Analysis (Metabole Flussanalyse). Gleichung (3) ist normalerweise unterbestimmt, d. h. es können meistens keine stationären Raten im Netz berechnet werden. Oft können in Experimenten einige Raten gemessen werden (z. B. Aufnahme/Ausscheidung von Substraten/Produkten), die Gleichung (3) in ein inhomogenes Gleichungssystem überführen und mit denen dann zumindest einige unbekannte Raten bestimmt werden können. Die Eingabe gemessener und Ausgabe berechneter Raten erfolgt direkt in den Textboxen der IFK und

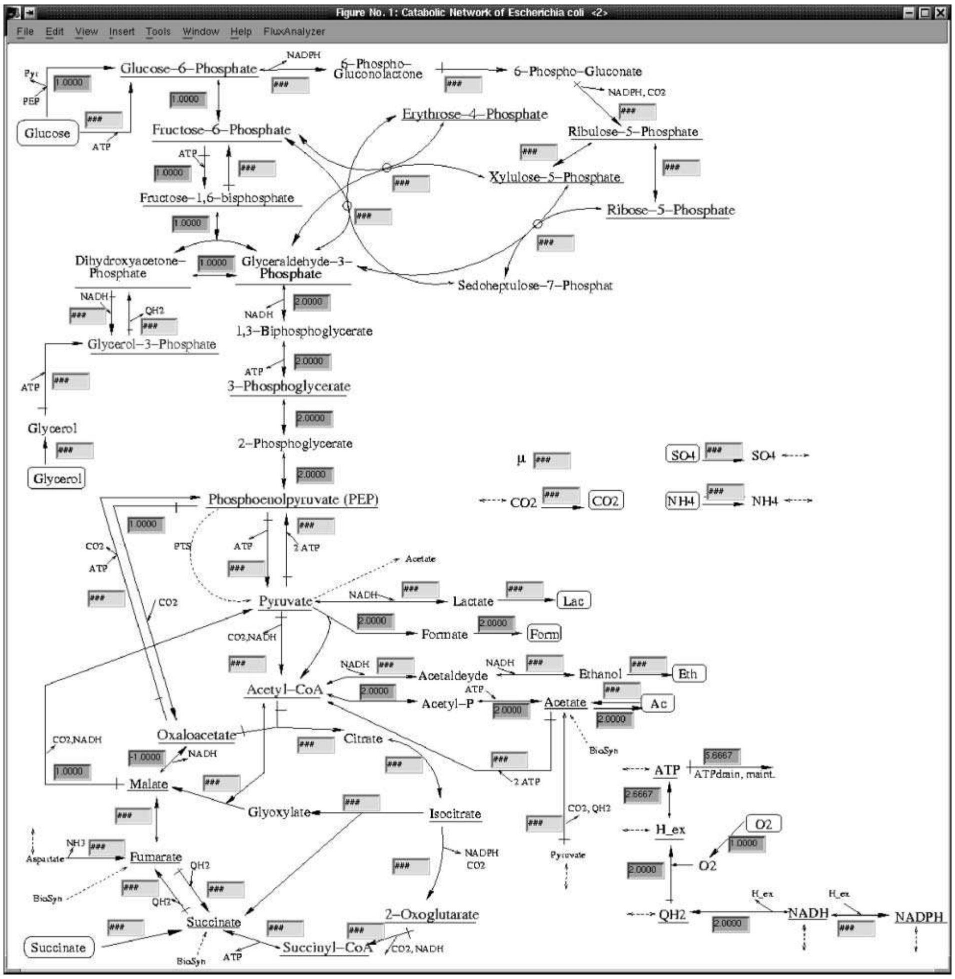


Bild 5 Interaktive Flusskarte, die in [17] eingesetzt wurde. Angezeigt wird gerade ein Elementarmodus (dunkle Boxen). In der Menüleiste befindet sich auch der Eintrag für den FluxAnalyzer.

wird farblich unterschiedlich markiert. Des Weiteren stehen Methoden für Konsistenzchecks – dies ist relevant bei redundanten Messungen –, zur Sensitivitätsanalyse und für den Vergleich unterschiedlicher Flussverteilungen zur Verfügung.

Optimierung von Flüssen Der Benutzer kann sich eine beliebige lineare Zielfunktion definieren (zum Beispiel um das Wachstum zu maximieren). Eine optimale Flussverteilung bezüglich dieser Zielfunktion kann berechnet und angezeigt werden.

Metabole Pathway Analyse auf Basis von Elementarmoden. Elementarmoden (EM) können als kleinste funktionale Teilnetze verstanden werden [16]. Die grundlegende Bedeutung der Elementarmoden ist erst in den letzten Jahren erkannt worden. Mit ihrer Hilfe können zum Beispiel „genetisch unabhängige“ Routen (Pathways) im Netz beschrieben, Flexibilität und Robustheit des Netzes bestimmt und die strukturelle Bedeutung einzelner Reaktionen bei unterschiedlichen Wachstumsbedingungen abgeschätzt werden [8; 17]. Die Berechnung von Elementarmoden in größeren Netzen ist aufgrund der kombinatorischen Komplexität eine diffizile Aufgabe [7]. Der FluxAnalyzer ermöglicht sowohl die effiziente Berechnung als auch eine eingehende statistische Untersuchung der berechneten Elementarmoden.

Andere strukturelle Eigenschaften wie Erhaltungsrelationen, strukturelle Inkonsistenzen (z. B. „Sackgassen“) und strukturelle Kopplungen können bestimmt werden. Des Weiteren kann die stöchiometrische Matrix exportiert oder graphisch ausgegeben werden. Die IFKs und die berechneten Daten können gespeichert werden.

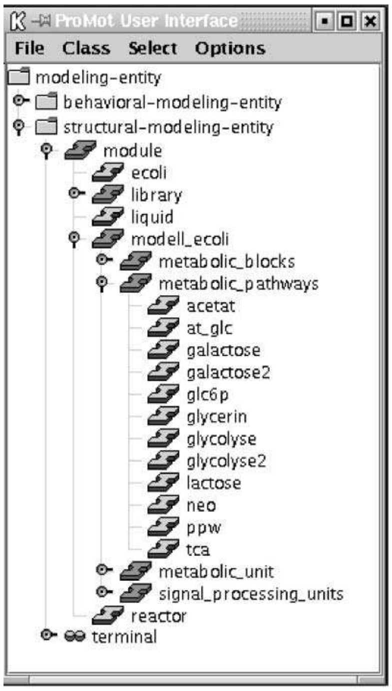
6 Modellorganismus Escherichia coli

Zur Beschreibung der Kohlenhydrataufnahme in *Escherichia coli* wurde ein detailliertes Modell erstellt [10; 11]. Wichtigstes Phänom, welches durch das Modell quantitativ richtig wiedergegeben wird, ist die so genannte ‚Kataboliten Repression‘. Kataboliten Repression meint die Fähigkeit des Zuckers, Glukose die Aufnahme einer ganzen Reihe von anderen Kohlenhydraten zu blockieren. In der Zelle wird diese Blockade durch eine Vielzahl von biochemischen Reaktionen, die ein komplexes Signaltransduktionssystem bilden, realisiert. Beteiligt sind die Aufnahmesysteme der Zucker, die als Sensoren fungieren, sowie Regulatorproteine am Ende der Kette, die direkt mit den entsprechenden Bindestellen auf der DNA interagieren und das Ablesen der entsprechenden Information hemmen oder aktivieren.

Um das Modell zu validieren, d. h. zu überprüfen, ob es die Realität richtig beschreibt, sind eine Reihe von Experimenten durchgeführt worden [10]. Diese Experimente wurden so geplant, dass das System aus unterschiedlichen Blickwinkeln betrachtet wird: Anregung der Kohlenhydrataufnahmesysteme durch Variation der Kohlenstoffquelle, Analyse der Signaltransduktionseinheiten durch Verwendung speziell konstruierter Mutantenstämme, die Defekte in der Signalweiterleitung besitzen, und Auflösung verschiedener Zeitfenster durch impulsförmige Anregung (die Experimente laufen innerhalb zwei Minuten ab) oder Batch-Versuche (die Experimente laufen über acht Stunden).

Das Modell ist vollständig in ProMoT implementiert. Bild 6 zeigt den Browser. Aufgeklappt sind die implementierten Stoffwechselwege. Unten im Bild ist beispielhaft der Modellbaustein „reactor“ gezeigt. Dieser umfasst die Flüssigphase und die biologische Phase. Durch Anklicken mit der Maus kann man sich die Bausteine dann detaillierter anschauen. Mit Hilfe des FluxAnalyzers wurde eine ausführliche Elementarmodenanalyse des Zentralstoffwechsel in *Escherichia coli* durchgeführt [17]. Durch die Auswertung von bis zu 500.000 Elementarmoden konnten wichtige Schlussfolgerungen für die strukturelle Funktionalität und Flexibilität in diesem metabolen Netz gezogen werden.

Mit Hilfe von ProMoT und Diva soll es langfristig möglich sein, Experimente am Rechner analog zu Laborexperimenten durchzuführen. Um die Funktionalität zu verbessern, ist geplant, Visualisierungstools einzubinden, um komplexe genetische und metabolische Netzwerke darstellen zu können. Anwendungsfelder des FluxAnalyzers finden sich vor allem in der Mikrobiologie, in der Biotechnologie und in der aufkommenden Systembiologie. Das Werkzeug wird bereits von drei industriellen Unterneh-



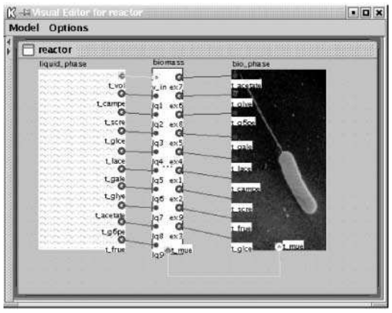


Bild 6 Oben: Browser von ProMoT mit den Modellbausteinen für das *E. coli* Modell. Aufgeklappt sind die implementierten Stoffwechselwege. Unten: Modellbaustein „Reactor“. Er beschreibt die Flüssigphase und die Biophase in einem Bioreaktor.

men und von mehreren Forschergruppen genutzt. Ausführlichere Beschreibungen findet man in [9] und auf der Webseite www.mpi-magdeburg.mpg.de/projects/fluxanalyzer.

Literatur

[1] M. Ginkel, A. Kremling, T. Nutsch, R. Rehner, and E. D. Gilles. Modular modeling of cellular systems with ProMoT/Diva. *Bioinformatics*, 19(9):1169–1176, 2003.

[2] J. W. Hearne. Sensitivity analysis of parameter combinations. *Appl. Math. Modelling*, 9:106–108, 1985.

[3] R. Heinrich and S. Schuster. *The regulation of cellular processes*. Chapman & Hall, 1996.

[4] M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, the rest of the SBML Forum: A.P. Arkin, B.J. Bornstein, D. Bray, A. Cornish-Bowden, A.A. Cuellar, S. Dronov, E.D. Gilles, M. Ginkel, V. Gor, I.I. Goryanin, W.J. Hedley, T.C. Hodgman, J.-H. Hofmeyr, P.J. Hunter, N.S. Juty, J.L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L.M. Loew, D. Lucio, P. Mendes, E. Minch, E.D. Mjolsness, Y. Nakayama, M.R. Nelson, P.F. Nielsen, T. Sakurada, J.C. Schaff, B.E. Shapiro, T.S. Shimizu, H.D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19:524–531, 2003.

[5] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28:27–30, 2000.

[6] P.D. Karp, M. Riley, M. Saier, I.T. Paulsen, J. Collado-Vides, S.M. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castro. The Ecocyc database. *Nucleic Acids Res.*, 30:56–58, 2002.

[7] S. Klamt and J. Stelling. Combinatorial complexity of pathway analysis in metabolic networks. *Molecular Biology Reports*, 29(1-2):233–236, 2002.

[8] S. Klamt and J. Stelling. Two approaches for metabolic pathway analysis? *Trends in Biotechnology*, 21:64–69, 2003.

[9] S. Klamt, J. Stelling, M. Ginkel, and E.D. Gilles. FluxAnalyzer: Exploring structure, pathways and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics*, 19:261–269, 2003.

[10] A. Kremling, K. Bettenbrock, B. Laube, K. Jahreis, J.W. Lengeler, and E.D. Gilles. The organization of metabolic reaction networks: III. Application for diauxic growth on glucose and lactose. *Metab. Eng.*, 3(4):362–379, 2001.

[11] A. Kremling and E.D. Gilles. The organization of metabolic reaction networks: II. Signal processing in hierarchical structured functional units. *Metab. Eng.*, 3(2):138–150, 2001.

[12] A. Kremling, K. Jahreis, J.W. Lengeler, and E.D. Gilles. The organization of metabolic reaction networks: A signal-oriented approach to cellular models. *Metab. Eng.*, 2(3):190–200, 2000.

[13] L. Ljung. *System Identification – Theory for the user*. Prentice Hall PTR, Upper Saddle River, New Jersey, second edition, 1999.

[14] J.J. Moré and S. J. Wright. *Optimization Software Guide*. SIAM, Philadelphia, USA, 1993.

[15] C. Posten and A. Munack. On-line application of parameter estimation accuracy to biotechnical processes. In *Proc. of the American Control Conf.*, vol. 3, pp. 2181–2186, 1990.

[16] S. Schuster, D. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat.Biotechnol.*, 18:326–332, 2000.

[17] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E.D. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420:190–193, 2002.

1

2

3

4

1 Dr. Andreas Kremling Geboren 1965 in Lahr/Schwarzwald, 1985–1992 Studium der Technischen Kybernetik an der Universität Stuttgart, 1993 Doktorand am ISR, Universität Stuttgart, 1998 Wissenschaftlicher Mitarbeiter am Max-Planck-Institut für Dynamik komplexer technischer Systeme (Fachgruppe Systembiologie)
Adresse: Max-Planck-Institut für Dynamik komplexer technischer Systeme, Sandtorstraße 1, 39106 Magdeburg, Tel: +49-0391-6110-466, Fax: +49-0391-6110-526,
E-Mail: kremling@mpi-magdeburg.mpg.de

2 Steffen Klamt Geboren am 04.02.1972 in Magdeburg, 1992–1998: Studium Angewandte Systemwissenschaft an der Universität Osnabrück, Abschluss als Diplom-Systemwissenschaftler April 1998, seit Mai 1998: Doktorand am Max-Planck-Institut für Dynamik komplexer technischer Systeme (Fachgruppe Systembiologie)
Adresse: Max-Planck-Institut für Dynamik komplexer technischer Systeme, Sandtorstraße 1, 39106, Magdeburg

3 Martin Ginkel Geboren am 09.02.1972 in Parchim (Meckl.), 1991–1997: Studium Informatik an der Universität Magdeburg, Abschluss als Diplominformatiker März 1997, seit 1998: Doktorand am Max-Planck-Institut für Dynamik komplexer technischer Systeme (Fachgruppe Systembiologie)
Adresse: Max-Planck-Institut für Dynamik komplexer technischer Systeme, Sandtorstraße 1, 39106 Magdeburg

4 Prof. Dr.-Ing. Dr. h.c. mult. Ernst Dieter Gilles Geboren am 16. Mai 1935 in St. Goarshausen, Studium der Elektrotechnik, TH Darmstadt (1954–1960), Promotion, TH Darmstadt (1964). Habilitation im Fach Regelungstechnik, TH Darmstadt (1966), Professor und Institutsdirektor am ISR der Universität Stuttgart (seit 1968), Direktor am Max-Planck-Institut für Dynamik komplexer technischer Systeme (seit 1997), Honorarprofessor Universität Magdeburg (1999)
Adresse: Max-Planck-Institut für Dynamik komplexer technischer Systeme, Sandtorstraße 1, 39106 Magdeburg

Available online at www.sciencedirect.com

SCIENCE DIRECT®

BioSystems 78 (2004) 23–37

BioSystems

www.elsevier.com/locate/biosystems

Analysis of two-component signal transduction by mathematical modeling using the *KdpD/KdpE* system of *Escherichia coli*

A. Kremling[a,]*, R. Heermann[b], F. Centler[c], K. Jung[b], E.D. Gilles[a]

[a] *Systems Biology Group, Max-Planck-Institut für Dynamik Komplexer Technischer Systeme, Sandtorstr. 1; 39106 Magdeburg, Germany*
[b] *Department Biologie I, Ludwig-Maximilians-Universität München, Bereich Mikrobiologie, 80638 München, Germany*
[c] *Bio Systems Analysis Group, Department of Mathematics and Computer Science, Jena Center for Bioinformatics (JCB), Friedrich-Schiller-University Jena, 07743 Jena, Germany*

Received 12 February 2004; received in revised form 16 June 2004; accepted 25 June 2004

Abstract

A mathematical model for the *KdpD/KdpE* two-component system is presented and its dynamical behavior is analyzed. *KdpD* and *KdpE* regulate expression of the *kdpFABC* operon encoding the high affinity K^+ uptake system *KdpFABC* of *Escherichia coli*. The model is validated in a two step procedure: (i) the elements of the signal transduction part are reconstructed in vitro. Experiments with the purified sensor kinase and response regulator in presence or absence of DNA fragments comprising the response regulator binding-site are performed. (ii) The mRNA and molecule number of *KdpFABC* are determined in vivo at various extracellular K^+ concentrations. Based on the identified parameters for the in vitro system it is shown, that different time hierarchies appear which are used for model reduction. Then the model is transformed in such a way that a singular perturbation problem is formulated. The analysis of the in vivo system shows that the model can be separated into two parts (submodels which are called functional units) that are connected only in a unidirectional way. Hereby one submodel represents signal transduction while the second submodel describes the gene expression.

© 2004 Elsevier Ireland Ltd. All rights reserved.

Keywords: Escherichia coli; Two-component signal transduction; Model reduction; Singular perturbation; In vivo dynamics

1. Introduction

Mathematical modeling and dynamical simulation become more and more important for the understanding of the complex behavior of metabolic and regulatory networks of cellular systems. Although there is a large qualitative knowledge, especially for bacteria, quantitative research is still scarce. Therefore, relative simple biological (sub-)systems must be analyzed to get more insight in the dynamics of intracellular processes. A number of such processes are related to the survival in a broad range of environmental conditions. Several parameters like the supply of different

---

* Corresponding author. Tel.: +49 391 6110 466; fax: +49 391 6110 526.
E-mail address: kremling@mpi-magdeburg.mpg.de (A. Kremling).

0303-2647/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2004.06.003

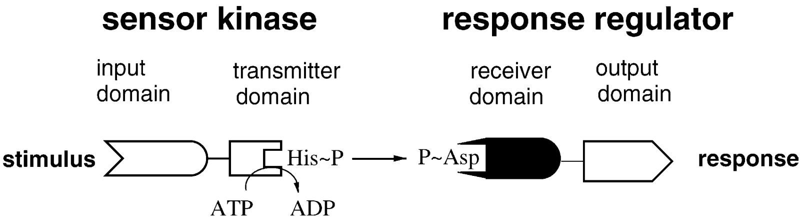


Fig. 1. General scheme of a two-component signal transduction system.

nutrients, the sudden presence of toxic substances, pH, temperature, O_2 concentration, osmolality, or different other factors can rapidly change. To survive, bacteria are forced to monitor their environment constantly and to adapt to changing conditions immediately. Therefore, bacteria have established special signal transduction systems to execute adaptive responses to changing environmental conditions. The simplest circuits consist of two protein components (two-component systems): a sensor kinase, often anchored in the cytoplasmic membrane, and a cytoplasmic response regulator that mediates an adaptive response, usually a change in gene expression (Fig. 1). Two-component systems are widespread in bacteria, archaea and plants. In *Escherichia coli*, 30 sensor kinases and 32 response regulators have been found. However, the number of two-component systems differs enormously in different bacteria, ranging from 0 in *Mycoplasma genitalium* to 80 in *Synechocystis sp.*, in which the corresponding genes account for nearly 2.5% of the genome (see Stock et al., 2000 for review). Table 1 shows a broad range of conditions and adaptive responses controlled by two-component systems in different bacteria.

Sensor kinases typically contain an N-terminal input domain which is connected via a linker to a C-terminal transmitter domain. Response regulators typically consist of a N-terminal receiver domain coupled to one ore more C-terminal output domains (Fig. 1). Upon perception of a stimulus, the input domain of the sensor kinase modulates the signaling activity of its transmitter domain, resulting in autophosphorylation of a highly conserved histidine residue with the γ-phosphoryl group of ATP. Then, the phosphoryl group is transferred to an aspartate residue of the response regulator receiver domain, resulting

Table 1
Sensor kinase/response regulator systems control various processes

Function/stress	Organism	System
Oxygen sensing	*E. coli*	ArcB/ArcA
Nitrate and nitrite respiration	*E. coli*	NarX/NarL, NarQ/NarP
Chemotaxis	Various	CheA/CheY
Nitrogen utilization	*E. coli*	NtrB/NtrC
K^+ supply	Various	*KdpD/KdpE*
Phosphate supply	Various	PhoR/PhoB; PhoQ/PhoP
Antibiotics	*Enterococcus faecium*	VanS/VanR
Osmolarity	*E. coli*, *Salmonella typhimurium*	EnvZ/OmpR
Gene transfer	*Bacillus subtilis*	ComP/ComA
Sporulation	*Bacillus subtilis*	KinB/Spo0F, KinA/Spo0F
Cell cycle	*Caulobacter crescentus*	CckA/CtrA
Photosynthetic apparatus	*Rhodobacter capsulatus*	RegB/RegA
Virulence	*Bordetella pertussis*	BvgS/BvgA
Quorum sensing	*Vibrio harveyi*	LuxN/LuxO; LuxQ/LuxO
Symbiosis	*Bradyrhizobium japonicum*	NodV/NodW
Development	*Myxococcus xanthus*	SasS/SasR

in an activation of the output domain(s) to trigger response. In most cases the response is an alteration in the transcription level of a special gene or gene cluster (see Bourret et al., 1991; Parkinson, 1993; Parkinson and Kofoid, 1992; Stock et al., 1990, 2000 for reviews).

This contribution deals with the mathematical description of a reaction mechanism representing a two-component system for the control of K^+ uptake in *E. coli*. The *KdpD/KdpE* system is one example of a typical two-component system, which regulates the expression of the *kdpFABC* operon encoding the high affinity K^+ transport system *KdpFABC* in *E. coli* most notably under K^+ limiting conditions (Walderhaug et al., 1992; Altendorf and Epstein, 1996; Jung and Altendorf, 2002). A model for two-component signal transduction was set up by (Fisher et al., 1996) earlier. They determined a number of reaction parameters for phosphotransfer from the sensor kinase VanS to the response regulators VanR and PhoB in *Enterococcus*. Models for other signal transduction systems in *E. coli* are described e.g. by Wong et al., 1997; Kremling et al., 2001; Van Dien and Keasling, 1998; Koh et al., 1998.

The strategy to set up and to analyze the mathematical model was as follows: First, the *KdpD/KdpE* signal transduction cascade was reconstructed in vitro (for details, see Appendix A). A first model was set up describing autophosphorylation of *KdpD*, transfer of the phosphoryl group between sensor kinase and response regulator, dephosphorylation of *KdpE* $\sim$ P, and binding of the response regulator to DNA fragments comprising the specific response regulator-binding site mentioned above. The model was validated by a set of experiments. In the second step, the overall in vivo system was analyzed. Since the *kdpFABCDE* regulon comprises the genes for the transporter as well as the sensor kinase/response regulator elements, an autocatalytic behavior was observed. The model was extended to describe mRNA and protein synthesis. The amount of mRNA and the number of transporter molecules inside the cell were determined experimentally. The mathematical model was used to analyze time scales of the stimulus response. To evaluate the quality of the model, steady-state values of the concentration of the K^+ uptake system for different stress conditions were calculated and compared with experimental data.

2. Model equations for the in vitro system

Although it is known that during enzymatic activities proteins form a number of temporary complexes, in this contribution a rather simple reaction mechanism was used to describe the two-component system (Fig. 2). Incorporating such temporary complexes increases the number of unknown parameters. Since for cellular systems measurements of the system components are difficult and only a subset of components can be measured, the model structure should also be as simple as possible to facilitate parameter identification (Saez-Rodriguez et al., 2004).

The reaction equations are:

Autophosphorylation :

$$\mathrm{ATP} + KdpD \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{ADP} + KdpD^p \qquad \text{(R1)}$$

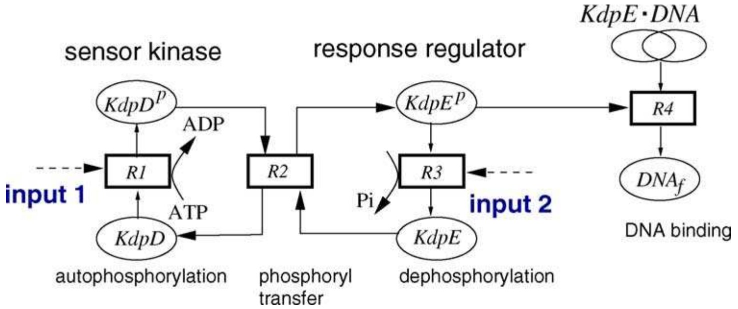


Fig. 2. The reaction mechanism for the *KdpD/KdpE* two-component system. The ellipses represent proteins. Two putative input loci are under consideration. Input 1 describes alterations of the kinase activity while input 2 describes the alterations of the phosphatase activity. The phosphorylated response regulator binds upstream of the *kdpFABC* promoter/operator region and in interaction with the RNA polymerase triggers *kdpFABC* expression. The reactions denoted as autophosphorylation, phosphoryl transfer, dephosphorylation and DNA binding are described in the text.

Phosphoryl group transfer :

$$KdpD^p + KdpE \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} KdpD + KdpE^p \tag{R2}$$

Dephosphorylation:

$$KdpE^p + KdpD \overset{k_3}{\rightarrow} KdpE + KdpD\,(+\mathrm{P}_i) \tag{R3}$$

DNA binding:

$$2\,KdpE^p + \mathrm{DNA}_f \underset{k_{-b}}{\overset{k_b}{\rightleftarrows}} KdpE - \mathrm{DNA} \tag{R4}$$

For the *KdpD/KdpE* system it is yet not known at which stage the stimulus enters the system and two possible input loci are under consideration (denoted in the following as input 1 and input 2). In this contribution only one stress condition is under investigation and therefore the stimulus is represented as a fixed parameter. In reaction R1 the stimulus (via input 1) enhances the kinase activity that results in autophosphorylation of the sensor kinase (state variables *KdpD*, $KdpD^p$) by ATP. In reaction R2 the phosphoryl group is transferred to the response-regulator (state variables *KdpE*, $KdpE^p$). *KdpE* ∼ P contains the active output domain. It is known that *KdpE* can be phosphorylated by alternative phosphor donors in the presence of an truncated form of *KdpD* (Heermann et al., 2003). However, this phosphorylation plays a minor role in the presence or absence of *KdpD*. Because of this uncertainty, phosphorylation of *KdpE* by acetyl phosphate and other low molecular weight phosphordonors was not included into the model. The dephosphorylation of *KdpE* ∼ P by the cognate sensor kinase *KdpD* is described in reaction R3 (controlled via input 2: decrease of dephosphorylation). It has been shown earlier that *KdpE* ∼ P dephosphorylation is only dependent on *KdpD* (Jung et al., 1997), so that other phosphatases are not considered in the model. Note that, although the stoichiometry of the back reaction of R2 and reaction R3 are similar, the underlying mechanisms are different: for R3 it is assumed that *KdpD* acts as an enzyme converting a substrate (*KdpE* ∼ P) into products (*KdpE* and P_i) while in R2 the phosphoryl group is transferred between the regulatory proteins. Hence, the reaction rate for R3 (parameter k_3) is not included in the balance equation for the sensor. However, high values of the phosphorylated sensor *KdpD* ∼ P results in low values for the free form of the sensor and therefore leads to a decreased rate of dephosphorylation. An intrinsic rate of dephosphorylation, i.e. a decay of the phosphorylated response regulator was not observed (Puppe et al., 1996). The activated response regulator forms a dimer and is then able to bind to the free DNA (state variable DNA_f) in reaction R4 to build a transcription complex (state variable $KdpE - \mathrm{DNA}$). Non-specific binding of the response regulator to other DNA binding sites could be neglected in the model (experimental data not shown).

The equations for the system are assorted in the following. For the total concentration of the kinase $KdpD_0$ and of the response regulator $KdpD_0$ the following equations hold:

$$KdpD_0 = KdpD + KdpD^p \tag{1}$$

$$KdpE_0 = KdpE + KdpE^p + 2KdpE - \mathrm{DNA} \tag{2}$$

And for the entire concentration of the DNA fragments DNA_0:

$$\mathrm{DNA}_0 = \mathrm{DNA}_f + KdpE - \mathrm{DNA} \tag{3}$$

The three remaining differential equations read:

$$\begin{aligned}\frac{\mathrm{d}KdpD^p}{\mathrm{d}t} &= -k_{-1}KdpD^p\,\mathrm{ADP} - k_2\,KdpD^p\,KdpD \\ &\quad + k_1\,KdpD\,\mathrm{ATP} + k_{-2}\,KdpD\,KdpE^p\end{aligned} \tag{4}$$

$$\begin{aligned}\frac{\mathrm{d}KdpE^p}{\mathrm{d}t} &= -k_{-2}\,KdpD\,KdpE^p - k_3\,KdpE^p\,KdpD \\ &\quad -2\,k_b\,KdpE^{p2}\,\mathrm{DNA}_f + k_2\,KdpD^p\,KdpE \\ &\quad + 2\,k_{-b}\,KdpE - \mathrm{DNA}\end{aligned} \tag{5}$$

$$\begin{aligned}&\frac{\mathrm{d}KdpE - \mathrm{DNA}}{\mathrm{d}t} \\ &\quad = -k_{-b}\,KdpE - \mathrm{DNA} + k_b KdpE^{p2}\,\mathrm{DNA}_f\end{aligned} \tag{6}$$

To compare situations in vivo and in vitro and to study the influence of the phosphoryl source, ATP is taken as a constant parameter value although it is consumed in minor amounts during the in vitro reaction.

2.1. Parameter estimation and results

The goal of parameter estimation is to find a set of parameters that can describe the experimental results. Here, it was based on the following experiments. The complete signal transduction cascade was reconstructed in vitro (see Appendix A for experimental details). Briefly, purified and reconstituted *KdpD*

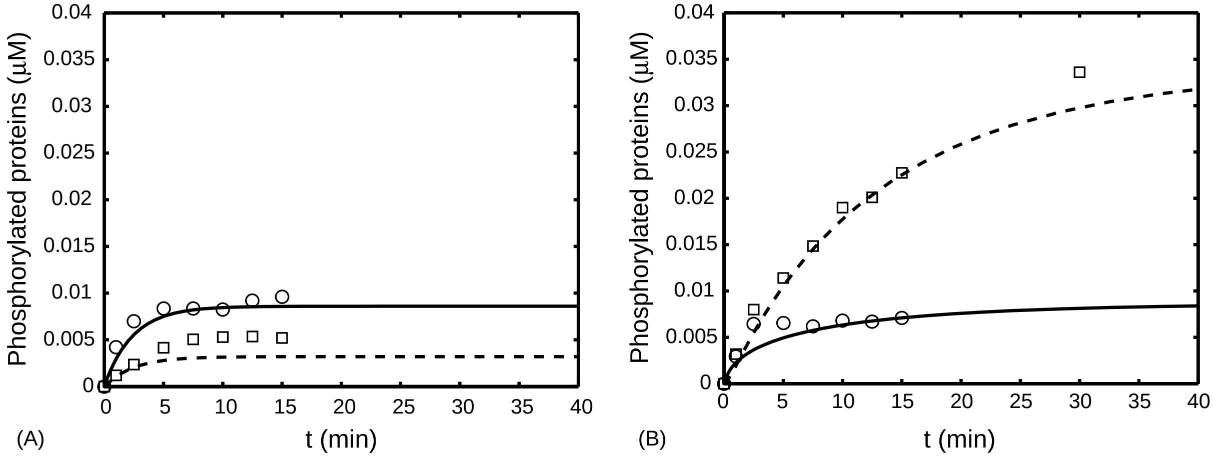


Fig. 3. Experimental and simulation results for the in vitro system. (A) Stress situation (0 mM K^+), no DNA fragments comprising the chromosomal *KdpE*-binding site were present. (B) Stress situation (0 mM K^+), in the presence of DNA fragments. Solid lines: simulation of phosphorylated *KdpD*, dashed lines: simulation of phosphorylated *KdpE*, circles: experimental data for phosphorylated *KdpD*, squares: experimental data for phosphorylated *KdpE*.

was mixed with purified *KdpE* (ratio 1:4). Then, an ATP/ADP mixture (ratio 12.5:1) was added, and the amounts of phosphorylated sensor kinase and response regulator were determined over time (Fig. 3). Two experiments were performed: one in the absence and one in the presence of DNA fragments comprising the response regulator-binding site. In both experiments phosphorylated *KdpD* was detectable albeit at very low amounts. In the presence of DNA fragments comprising the *KdpE*-binding site, the amount of phosphorylated *KdpE* was about 10 times higher as in the absence of these DNA fragments after 30 min.

Subsequently, we searched for one set of parameters that described these experimental results. Parameter values fitted with a least-square algorithm (MATLAB©) are summarized in Table 2. To have an impression on the quality of the fitted parameters, a sensitivity analysis was performed according to the method Hearne (Hearne, 1985). The method seeks a perturbed parameter vector to maximize the disturbance to the solution trajectory. Therefore, the method is useful for analyzing effects of a combination of parameter changes on the system. The parameters with the highest sensitivity were k_1, k_3, k_b and k_{-b}.

2.2. Model analysis and model reduction

2.2.1. Stationary behavior

In the described experiment a low degree of phosphorylation (<1%) was observed. However, the conditions used in the in vitro experiments did not reflect situations expected in vivo. Therefore, to reproduce intracellular conditions the influence of the ATP concentration on the steady-state values of the phosphorylated response regulator was under investigation. Fig. 4 shows steady-state values of the degree of phosphorylation $KdpE^p/KdpE_0$ in dependence on the ATP concentration. As can be seen, under in vivo conditions (>1.5 mM) a considerable higher degree of phosphorylation is expected. Such high ATP concentrations could not be adjusted in vitro, because by further addition of cold ATP, no measurement signal can be detected.

2.2.2. Singular perturbation problem

For the understanding of the overall behavior of cellular systems, mathematical models can be used to de-

Table 2
Parameter values for the in vitro data set

In vitro parameters	
k_1 = 0.0029 1/h μM	DNA_0 = 100 μM
k_{-1} = 0.00088 1/h μM	$KdpD_0$ = 1 μM
k_2 = 108 1/h μM	$KdpE_0$ = 4 μM
k_{-2} = 1080 1/h μM	ATP = 100 μM
k_b = 5400 1/h μM^2	ADP = 8 μM
k_{-b} = 360 1/h	
k_3 = 90 1/h μM	(0 mM K^+/no DNA)
k_3 = 90 1/h μM	(0 mM K^+/DNA)

Parameters were estimated with a least square algorithm.

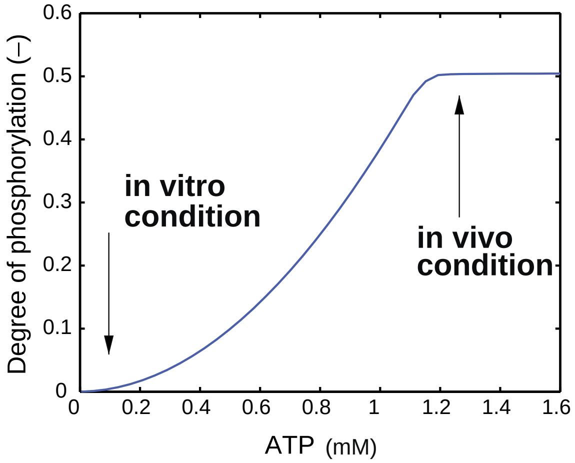


Fig. 4. Steady-state behavior of the system. The degree of phosphorylation is defined as $KdpE^p/KdpE_0$. For in vivo conditions the degree of phosphorylation is much higher than for the conditions used in vitro.

tect new (“emergent”) properties. A first step in model analysis is model reduction, i.e. to come to a “simpler” description of the system. Here, the number of independent states (the order of the system) is reduced by coupling two (or more) states by algebraic equations. One possibility to obtain algebraic equations is to analyze the time hierarchies of the system and to regard fast modes—in comparison to the chosen time window—as in steady-state.

To illustrate the approach, the scheme is simplified. Only two reactions are considered. Reaction (A1) describes phosphorylation and dephosphorylation of a response regulator rp from a general source s (constant entity) and reaction (A2) describes the interaction of rp with the DNA binding-site:

$$\text{synthesis:} \qquad s \underset{k_-}{\overset{k}{\rightleftarrows}} rp \tag{A1}$$

$$\text{DNA–binding:} \qquad rp \underset{k_{-b}}{\overset{k_b}{\rightleftarrows}} rd \tag{A2}$$

Scaling the O.D.E.s with $dt = 1/k\, d\tau$, and rearranging leads to the system:

$$rp' = s - K\, rp - \frac{k_b}{k}\,(rp - K_b\, rd) \tag{7}$$

$$rd' = \frac{k_b}{k}\,(rp - K_b\, rd)\,, \tag{8}$$

with $K = k_-/k$ and $K_b = k_{-b}/k_b$. Introducing a new state $r = rp + rd$ leads to a singular perturbation problem with $\epsilon = k/k_b$:

$$r' = s - K\,(r - rd) \tag{9}$$

$$\epsilon\, rd' = r - (1 + K_b) rd. \tag{10}$$

For very small ϵ it is allowed to perform a model reduction with $\epsilon = 0$. The system (A1,A2) can be rewritten with one O.D.E. for r:

$$r' = s - K(r - rd) \tag{11}$$

and one algebraic equation for rd:

$$rd = \frac{r}{1 + K_b}. \tag{12}$$

For the overall original model (R1)–(R4), scaling was performed with $dt = 1/k_2\, KdpE_0\, d\tau$ and based on the fitted set of parameter $\epsilon = k_2/k_b\, KdpE_0 \approx 5 \times 10^{-3}$. Model reduction for the original system is then equivalent to the assumption that reaction (R4) is in rapid equilibrium:

$$KdpE^{p2}\, \mathrm{DNA}_f = K_b\, KdpE - \mathrm{DNA}. \tag{13}$$

Thereby, the system is reduced to two O.D.E.s for $KdpD^p$ and $KdpE^p$, two algebraic equations to calculate the free amount of DNA (DNA_f), and unbound response regulators ($KdpEs^p_f$) and two algebraic equations for the total amount of sensor kinase and response regulator:

$$\begin{aligned}\frac{d\, KdpD^p}{dt} &= -k_{-1}\, KdpD^p\, \mathrm{ADP} - k_2\, KdpD^p\, KdpE \\ &\quad + k_1\, KdpD\, \mathrm{ATP} + k_{-2}\, KdpD\, KdpE^p_f\end{aligned} \tag{14}$$

$$\begin{aligned}\frac{d\, KdpE^p}{dt} &= -k_{-2}\, KdpD\, KdpE^p_f - k_3\, KdpE^p_f\, KdpD \\ &\quad + k_2\, KdpD^p\, KdpE\end{aligned} \tag{15}$$

$$KdpE^p = KdpE^p_f + 2\frac{KdpE^{p2}_f\, \mathrm{DNA}_f}{K_b} \tag{16}$$

$$\mathrm{DNA}_0 = \mathrm{DNA}_f + \frac{KdpE_f^{p2}\,\mathrm{DNA}_f}{K_b}, \tag{17}$$

where $K_b = k_{-b}/k_b$ is the binding affinity of $KdpE_f^p$ to the binding site DNA_f. The two algebraic equations for the total amount of sensor kinase and response regulator read:

$$KdpD_0 = KdpD + KdpD^p \tag{18}$$

$$KdpE_0 = KdpE + KdpE^p. \tag{19}$$

Differences during simulation experiments between the two models could hardly be detected (data not shown).

3. Model equations for the in vivo system

To describe the overall system in vivo, the model was extended with equations for the mRNA (state variable RNA)

$$\text{RNA: (nucleotides)} \overset{r_{tr}}{\rightarrow} \text{RNA} \overset{k_z}{\rightarrow} \text{degradation} \tag{20}$$

and the dynamical equations for the proteins *KdpFABC* ($KdpF$), total *KdpD* ($KdpD_0$), and total *KdpE* ($KdpE_0$)

$$KdpFABC\text{: (amino acids)} \overset{r_{tl1}}{\rightarrow} KdpF \overset{k_d}{\rightarrow} \text{degradation} \tag{21}$$

$$KdpD\text{: (amino acids)} \overset{r_{tl2}}{\rightarrow} KdpD_0 \overset{k_d}{\rightarrow} \text{degradation} \tag{22}$$

$$KdpE\text{: (amino acids)} \overset{r_{tl3}}{\rightarrow} KdpE_0 \overset{k_d}{\rightarrow} \text{degradation} \tag{23}$$

The organisation of the *kdpFABCDE* regulon is shown in Fig. 5. Under *kdpFABC*-inducing conditions a transcript *kdpFABC* is formed, and probably by a read-through effect the transcription of *kdpDE* is also enhanced. Indeed, under K^+-limiting growth conditions, increased amounts of *KdpD* and *KdpE* are detectable (unpublished information). This is taken into account in equations (22) and (23). The existence of a putative terminator was analyzed but could not be detected (data not shown). However, different molecule numbers of the transport system, sensor kinase and response regulator are expected although they are co-regulated. This is considered by different values for the translation efficiency.

Since it is assumed that the concentration of nucleotides and amino acids are not limiting, the rate laws r_{tr} and r_{tl_i} do not depend on the monomer concentration. To describe transcription efficiency based on the interaction of a number of response regulators a new method that was introduced previously was used to calculate the rate of mRNA synthesis r_{tr} (see Section 3.1). To describe translation efficiency the following approach is used:

$$r_{tl_i} = k_{tl_i}\,\mathrm{RNA}, \tag{24}$$

while for protein degradation a first order law with parameter k_d is used.

3.1. Brief summary of the modeling approach to describe transcription initiation

The method is based on the hierarchical structure of the regulatory network and calculates the transcription efficiency by neglecting unimportant interactions between regulator proteins and DNA-binding

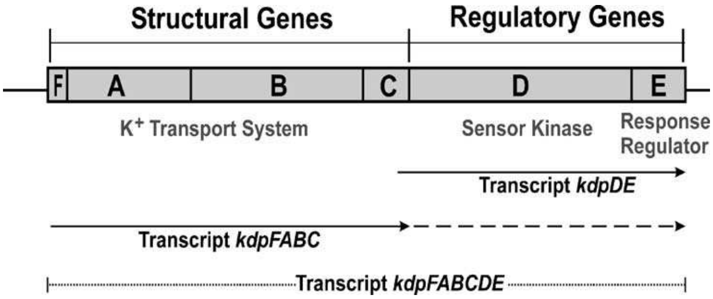


Fig. 5. Schematic representation of the *kdpFABCDE* regulon. The *kdpDE* operon is transcribed from its own promoter. Under *kdpFABC*-inducing conditions a transcript *kdpFABC* is formed, and the transcription of *kdpDE* is enhanced, probably by a read-through effect.

sites (Kremling and Gilles, 2001). Since the RNA polymerase is essential for transcription, it represents the cellular or top level while other regulator proteins have a special function (or are more specific) in metabolism, they are e.g. activators or inhibitors for the expression of specific genes. These regulator proteins are therefore assigned to further levels in the hierarchy. The hierarchical model structure allows a signal transduction from the top to the lowest level but not vice versa. Therefore, some interactions of the proteins are neglected which leads to a simpler model structure in comparison to a complete model, including all interactions.

The model used here assumes that the amount of RNA polymerase and the concentration of the σ factor are constant entities. Therefore, a basic activity of the RNA polymerase is considered by state ψ (the value is fixed). The interaction of the regulator with the DNA-binding sites enhances RNA polymerase activity (state $\bar{\psi}$). The equations to derive an expression for $\bar{\psi}$ are given in the Appendix A. The following equation will hold for the rate of mRNA synthesis r_{tr}:

$$r_{tr} = k_{tr}\, \bar{\psi}\, \mathrm{DNA}_0. \tag{25}$$

The rate of transcription is proportional to the RNA polymerase activity and the number of templates.

The O.D.E.s for the in vivo system read:

$$\frac{d\mathrm{RNA}}{\mathrm{d}t} = k_{tr}\, \bar{\psi}\, \mathrm{DNA}_0 - (k_z + \mu)\, \mathrm{RNA} \tag{26}$$

$$\frac{\mathrm{d}KdpD_0}{\mathrm{d}t} = k_{tl2}\mathrm{RNA} - (k_d + \mu)S_0 \tag{27}$$

$$\frac{\mathrm{d}KdpE_0}{\mathrm{d}t} = k_{tl3}\, \mathrm{RNA} - (k_d + \mu)R_0 \tag{28}$$

$$\frac{\mathrm{dKdpF}}{\mathrm{d}t} = k_{tl1}\, \mathrm{RNA} - (k_d + \mu)F, \tag{29}$$

where μ is the specific growth rate during the experiment.

The concentration of the mRNA was determined by Northern blot analysis. Therefore, an additional equation for mRNA^m (state variable RNA^m) was used to relate the measured quantity to the mRNA concentration.

$$\mathrm{RNA}^m = k_m\, \mathrm{RNA} \tag{30}$$

Parameter k_m was estimated based on data on mRNA synthesis from *E. coli* (Bremer and Dennis, 1987).

The dynamical equations for $KdpD^p$ and $KdpE^p$ read:

$$\begin{aligned}\frac{\mathrm{d}KdpD^p}{\mathrm{d}t} &= -k_{-1}\, KdpD^p\, \mathrm{ADP} - k_2\, KdpD^p\, KdpE \\ &\quad + k_1\, KdpD\, \mathrm{ATP} + k_{-2}\, KdpD\, KdpE^p\end{aligned} \tag{31}$$

$$\begin{aligned}\frac{\mathrm{d}KdpE^p}{\mathrm{d}t} &= -k_{-2}\, KdpD\, KdpE^p - k_3\, KdpE^p\, KdpD \\ &\quad + k_2\, KdpD^p\, KdpE\end{aligned} \tag{32}$$

Note that Eqs. (14) and (15) are similar to Eqs. (31) and (32) except for using $KdpE^p$ instead $KdpE^p_f$ on the right side. Explanation in Section 4.

3.2. Parameter estimation and results

The measured time course for the mRNA concentration shows an interesting unexpected dynamic (Fig. 6). After reaching a maximum at 10 min mRNA decreases until a new steady-state is reached after approximatly 40 min. Since the current structure of the model is not able to describe the observation, a hypothesis was formulated to describe the measured data: the decrease of the mRNA concentration could only be explained when a reset is assumed. Since the model does not include the participation of a functional transporter in any way it is assumed here, that the transport of K$^+$ ions mediated by *KdpFABC* counteracts the stimulus. Since no information on possible detailed mechanisms is available, an empirical black-box approach is used. Fig. 7 shows an extended scheme used to describe the mRNA dynamics.

For the black box the following mathematical expression is used: assuming that the stimulus enters the system by the dephosphorylation of the response regulator (parameter k_3), an increase of the intracellular K$^+$ concentration mediated by the K$^+$ uptake system *KdpFABC* should increase this parameter with increasing transporter concentration resulting in a dephosphorylation of *KdpE*. This is described by the following equation:

$$k_3 = k_h \frac{KdpF}{KdpF + K_h}, \tag{33}$$

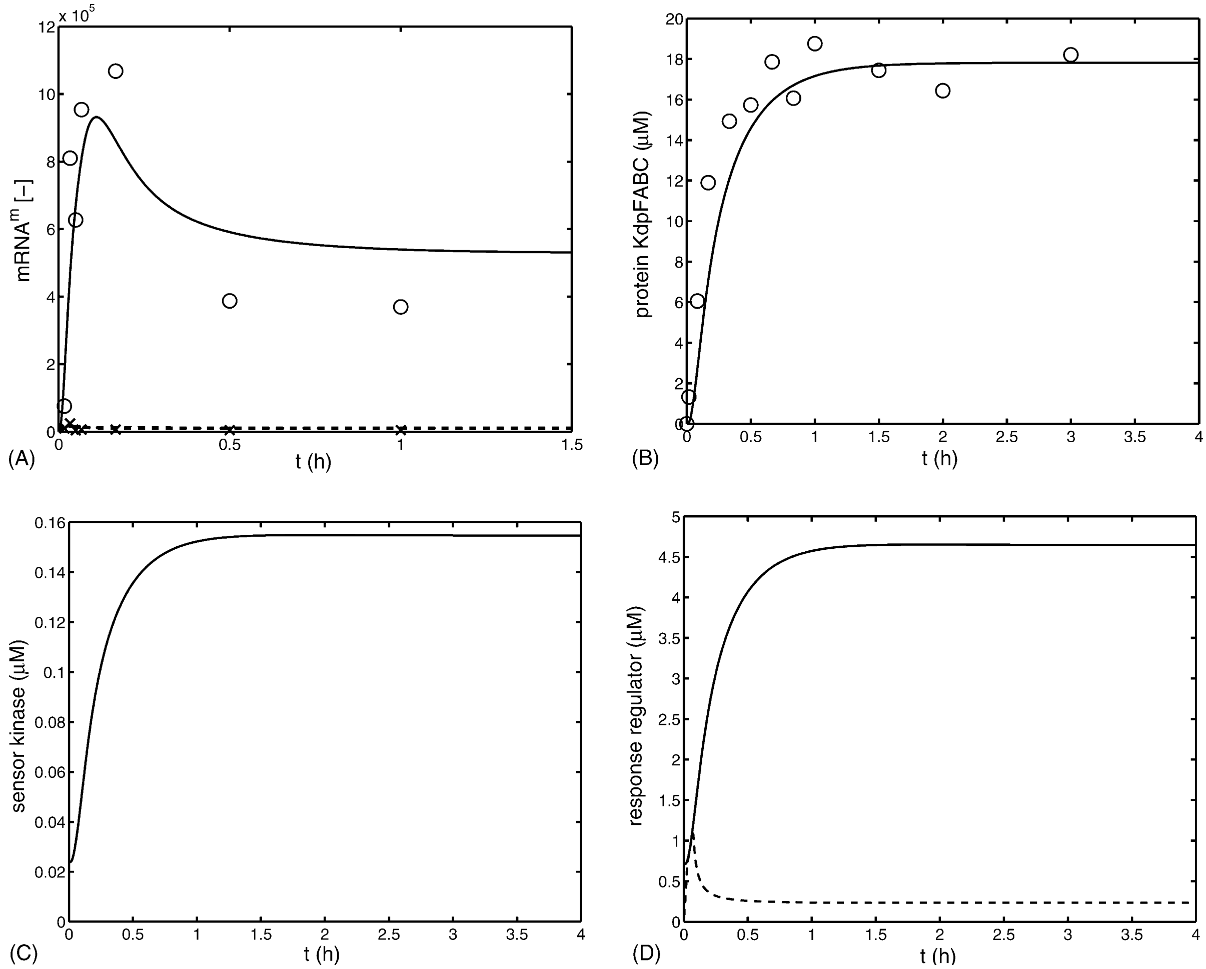


Fig. 6. Experimental and simulation results of the extended in vivo system. (A) Time course of mRNA^m. Circles represent an experiment under K$^+$ limiting conditions, solid line represents the simulation result; crosses represent an experiment during uninduced conditions, dashed line represents the simulation result. (B) Time course of protein *KdpFABC*. Solid lines: simulation results under K$^+$-limiting conditions, circles: experimental data under K$^+$-limiting conditions. (C) Time course of simulated sensor kinase. (D) Time course of simulated response regulator. Solid line represents the entire concentration of the response regulator, dashed line represents the phosphorylated response regulator.

where k_h and K_h represent adjustable parameters. The second putative input via the phosphorylation of the sensor kinase is not considered further. Parameter k_1 is taken as a constant.

Based on the available in vivo experimental data, parameters for the expression velocity could be determined by least-square parameter fit (MATLAB$^{©}$). However, simulation studies with the in vitro parameters lead to unrealistic results. Therefore, parameters of autophosphorylation and phosphoryl transfer were estimated, while the parameters for DNA binding were fixed to the values estimated for the in vitro system. For the in vivo system measurements for the sensor kinase and response regulator were not available and estimation of all parameters of the two-component unit seems not feasible. Parameters k_{-1} and k_{-2} are therfore taken out of the list of parameters that were estimated while they are fixed to empirical values. Since data are only available for the amount of protein *KdpFABC* the following assumption is made:

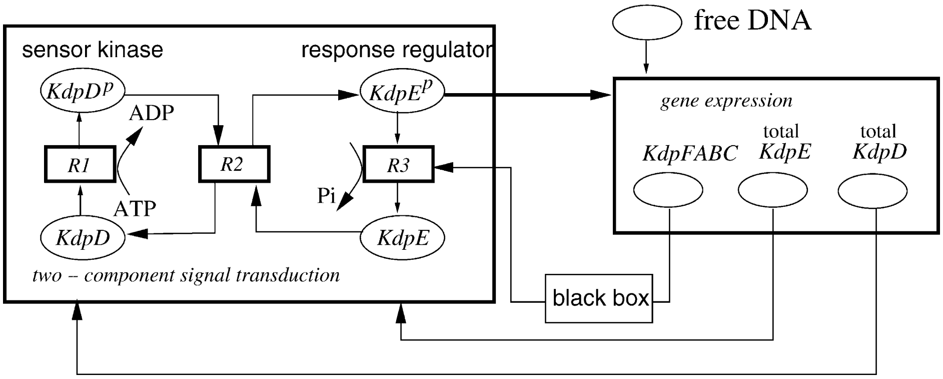


Fig. 7. The extended reaction mechanism for the two-component system. The model describes the expression of the proteins *KdpFABC*, *KdpD* and *KdpE*. The total amount of sensor kinase and response regulator are now further inputs in the two-component module. Due to unexpected dynamics of the mRNA a black box model is introduced describing a possible feedback from the transporter to the two-component module.

the number of molecules of *KdpE* have a fixed ratio $ra = 0.26$ to the number of molecules of *KdpFABC*, i.e. the concentration of the response regulator is assumed to be *ra* times the concentration of *KdpFABC*. The ratio between sensor kinase and response regulator is 1/30 ((Polarek et al., 1992) & unpublished own results).

In Fig. 6 simulation and experimental results are shown for a K^+ concentration $c_{K^+} = 0.02$ mM in the medium (parameters in Table 3, the values for ATP and ADP are chosen to be 2 mM and 0.2 mM, respectively). The time course of the mRNA and protein *KdpFABC* is reproduced with the model very well. To calculate the basal activity of the promoter, represented by pa-

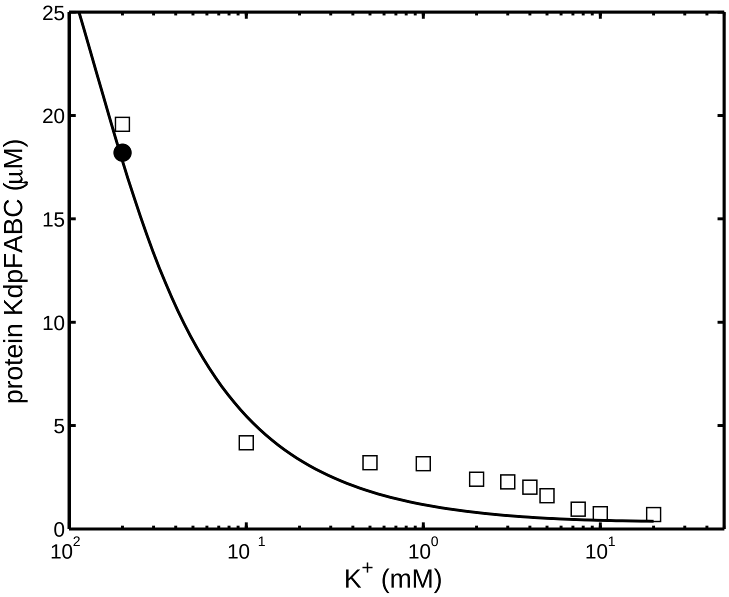


Fig. 8. Experimental data and simulation prediction for *KdpFABC* under different stress situations. Plotted are steady-state values against K^+ concentration. Squares: experimental data of steady-state values, closed circle: final value for *KdpFABC* taken from Fig. 6, solid line represents simulation results.

Table 3
Parameters for the in vivo data set

In vivo parameters	
$k_1 = 248.39$ 1/h μM	$\mathrm{DNA}_0 = 0.0054$ μM[b]
$k_{-1} = 10^{-4}$ 1/h μM[a]	$\psi = 0.0017$
$k_2 = 5789.3$ 1/h μM	$\alpha = 0.008$
$k_{-2} = 0.041$ 1/h μM[a]	ATP = 2 mM[b]
$K_b = 0.0667$ $\mu\mathrm{M}^2$[d]	ADP = 0.2 mM[b]
$k_{tr} = 2000$ 1/h[e]	$k_z = 28.87$ 1/h
$k_{tl1} = 1150$ 1/h	$k_d = 1.51$ 1/h
$k_{tl2} = 10$ 1/h	$k_{tl3} = 300$ 1/h
$k_h = 9982.5$ 1/h μM	$K_h = 0.04$ μM[a]
$k_m = 6.97 \cdot 10^9$	μ = 0.5 1/h[c]

Parameters were estimated with a least-square algorithm ([a]empirical values, [b]intracellular concentrations, [c]experimental, [d]taken from the in vitro experiment, [e]taken from (Bremer and Dennis, 1987)).

rameter ψ in the model, the mRNA was measured also during uninduced conditions (see also Fig. 6 left hand side). The value obtained, $\psi = 0.0017$, is comparable to a value obtained for the lac operon (Kremling et al., 2001).

To evaluate the quality of the model, the steady-state concentration of the uptake system *KdpFABC* was determined experimentally and compared to simulation predictions under different stress conditions. In the model the value for k_h is increased proportional to the increase of the stimulus concentration c_{K+} according to the formula:

$$k_h = k_h^0 \frac{c_{K+}}{c_{K+}^0} \tag{34}$$

where k_h^0, c_{K+}^0 represent the conditions given for the experiment in Fig. 6.

Fig. 8 compares the experimental and simulation results. A remarkable similarity of both curves was observed. Increasing the K^+ concentration in the medium to 1 mM results in a shut off of gene transcription while for higher K^+ concentrations a minor change of the steady-state values was observed.

4. Discussion

In this contribution a complete signal transduction pathway in *E. coli* starting from the sensory element to the cellular response is under investigation. For several reasons the Kdp system was chosen as an ideal test system:

- the signal transduction pathway is very short, it comprises only two elements;
- the signal transduction pathway could be reconstructed in vitro and is therefore accessible for measurements under two conditions (in the presence and absence of DNA fragments);
- the cellular response, i.e. the number of molecules of *KdpFABC* could be measured in vivo as well as the amount of *kdpFABC* mRNA under different stress conditions.

A mathematical model was set up to describe the experimental results. The model has a rather simple structure: temporary complexes between ATP and sensor kinase or between sensor kinase and response regulator were neglected to reduce the number of unknown or uncertain parameters. Parameters were estimated based on a number of experiments. For most experiments a good agreement between simulation and measurement data was achieved. For concentrations of K^+ between 0.5 mM $< c_{K+}^0 <$ 5 mM the residuals are larger. This is based on the fact, that these measurements were not used in the fitting procedure described above.

The in vitro experiments have shown a very low degree of phosphorylation. This might be due to two reasons: the ATP concentration used in the experiments didn't match intracellular conditions. In Fig. 4 the degree of phosphorylation is extrapolated to represent intracellular conditions. A degree of phosphorylation of nearly 50% could be achieved. On the other hand, as can be seen in Fig. 3 the amount of DNA shifts the equilibrium to the phosphorylated component. However, in an intracellular environment the number of binding sites for the response regulator is rather low and therefore could not result in a higher degree of phosphorylation.

Model reduction is always a powerful tool to reduce the degree of freedom in dynamical systems. Here, we applied the singular perturbation approach to show that different time hierarchies appear in the system under investigation. The analysis led to a system with O.D.E.s coupled with (implicit) algebraic equations, i.e. the dynamics of the binding of the response regulator is a “mirror” of the dynamics of the autophosphorylation and phosphoryl transfer.

In Kremling et al. (2000) we proposed a general framework to decompose metabolic and regulatory networks into functional units. Functional units are representing submodels with limited autonomy. One of the aspects under investigation considers signal transduction and describes the process of transcription initiation in Kremling et al. (2000) described with a modeling object called "coordinator" with an input/output relation. The aim of the decomposition is the definition of a set of submodels with fixed attributes which are the basis for a computer tool that support the modeler in setting up complex models (Kremling et al., 2001). Based on the analysis of the proposed mathematical model for the two-component system a reduced model comprising algebraic equations and O.D.E.s was developed. The reduced model consists out of two submodels which are connected in both directions in the in vitro case: submodel 1 describes the signal transduction to activate the response regulator while submodel 2 describes the interaction of the activated response regulator with the DNA control sequence. Analyzing the system in vivo shows however, that both units can be linked in a one-way direction. This is based on the fact that the in vivo amount of the DNA-binding site is very small in comparison to the DNA amount used in vitro. This leads to the observation that the concentration of $KdpE^p$ and of $KdpE_f^p$ in the reduced model are nearly equal, i.e. only a minor number of molecules is bound to the DNA. A decomposition of a more complex signal transduction pathway for catabolite repression was shown in (Kremling et al., 2000). However, in this paper the decomposition was based on biologically motivated criteria. Here, we show that there is also a theoretical basis that will allow the separation in different submodels.

Since the sensor kinase and the response regulator are auto-controlled, the dynamics of mRNA synthesis strongly depends on the initial conditions. Values used are in the expected range described in the literature (≈ 0 molecules sensor kinase, 300 molecules response regulator (Polarek et al., 1992) & unpublished own results). The low number of molecules makes it necessary to use a stochastic modeling approach. This was done and the results were compared with the deterministic approach. Differences could hardly be detected (data not shown). A good agreement between the experimental results and the simulated time course of the state variables was achieved by assuming a feedback from the protein *KdpFABC* to the input. In this way the model can be used for an experimental design to evaluate this hypothesis. For this purpose a mutant strain, defective in the uptake of K^+ via the KdpFABC system will be used. When the hypothesis is correct a monotone increase of the mRNA concentration is expected.

Parameter values estimated in the in vitro case led to unrealistic results in the in vivo case. This is mainly based on the fact that in vitro the phosphorylation degree of *KdpE* was very low (0.5%). Parameters used for in vitro case therefore lead to a very slow accumulation of the mRNA in contrast to the observed experimental results shown in Fig. 6. Hence, some of the parameters had to be identified again. For the parameter fit, two experiments with low and high K^+ concentration were used. Simulation results of the response regulator at a low K^+ concentration showed that that degree of phosphorylation reaches nearly 100% at the very beginning of the experiment (Fig. 6), and reaches a steady value of approximately 5%. A high degree of phosphorylation seems to be necessary for a quick response to environmental conditions. Based on the results, a steady-state characteristic curve was predicted with the model and finally compared with experimental results (Fig. 8). Here, also a good agreement was achieved.

Two-component signal transduction is one of the important mechanisms for bacteria to sense their environment and to respond to altered conditions. The present study gives some insights into the dynamics of such systems and can be used as starting conditions for other systems.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft [JU 270/4-1 (K.J.)] and the Fonds der Chemischen Industrie. We thank Mechthild Krabusch for excellent technical assistance and Britta Hasemeier for providing data for Fig. 6.

Appendix A

A.1. Additional equations for the in vivo system

The proposed method to calculate the transcription efficiency $\bar{\psi}$ is based on two algebraic equations

(16,17). They are extended in the following way:

$$KdpE^p = KdpE^p{}_f + 2\,\frac{KdpE^{p2}_f\,\mathrm{DNA}_f}{K_b}\left(1+\frac{1}{\alpha\,K}\right) \tag{A.1}$$

$$\mathrm{DNA}_0 = \mathrm{DNA}_f\left(1+\frac{1}{K}\right) + \frac{KdpE^{p2}_f\,\mathrm{DNA}_f}{K_b}\left(1+\frac{1}{\alpha\,K}\right), \tag{A.2}$$

where parameter $K = (1-\psi)/\psi$ represents the basal activity of the RNA polymerase and parameter α is an amplification factor. Transcription can occur if RNA polymerase is active alone or together with the activator. Hence, the fraction $\bar{\psi}$ of occupied promoter is given by:

$$\bar{\psi} = \frac{\mathrm{DNA}_f}{K\cdot\mathrm{DNA}_0}\left(1+\frac{KdpE^{p2}_f}{\alpha\,K_b}\right). \tag{A.3}$$

A.2. Reconstruction of the KdpD/KdpE signal transduction cascade in vitro

We reconstructed the complete signal transduction cascade of the *KdpD/KdpE* system in vitro. Purified *KdpD* (Jung et al., 1997) in proteoliposomes and purified *KdpE* (Heermann et al., 2003) in a ratio of 1 to 4 μM were mixed with 100 μM DNA comprising the DNA-binding site of *KdpE* (Sugiura et al., 1992) in buffer (50 mM Tris/HCl pH, 7.5, 10% glycerol (v/v), 0.5 M NaCl, 2 mM dithiotreitol). The double-stranded DNA fragments comprising the *KdpE*-binding sites were obtained by annealing of two complementary oligonucleotides. The upper strand sequence (from 5′ to 3′) has the following sequence: 5′-CATTTTTATACTTTTTTTACACCCCGCCCG-3′. The reaction was started by addition of 100 μM [γ-^{32}P]ATP (0.476 Ci/mmol), 8 μM ADP (ratio of 1 to 12.5, reflecting the ratio in living cells), and 110 μM $MgCl_2$. At the times indicated, samples were taken and the reactions were stopped by addition of an equal volume of 2× concentrated sodium dodecyl sulfate (SDS) sample buffer (Laemmli, 1970). In each case, samples were immediately subjected to SDS polyacrylamide gel electrophoresis (Laemmli, 1970). Gels were dried, the amount of radio-labeled proteins was detected by exposure of the gels to a phosphor screen, and the images were analyzed with a PhosphorImager SI system (Molecular Dynamics) using [γ-^{32}P]ATP as a standard. In parallel, the phosphorylation degree of *KdpE* ∼ P was determined in a gel-free detection system. This method consists of direct spotting of the phosphorylated sample on nitrocellulose after removal of *KdpD* ∼ P (ultracentrifugation) and [γ-^{32}]P ATP (gel filtration).

A.3. Quantification of the produced KdpFABC complex

Expression of *kdpFABC* was measured at the translational level by quantitative Western blot analysis. *E. coli* K-12 cells were grown at 37 °C in phosphate-buffered minimal medium (Epstein and Kim, 1971) containing 10 mM K^+ until the mid-exponential phase, filtered and resuspended in pre-warmed medium of lower K^+ concentration (0.02 mM K^+), and harvested at the indicated time. To measure the amount of *KdpFABC* complex in steady state, cells were grown in minimal media containing the indicated K^+ concentrations, and harvested at an absorbance of ≈1.0 at 600 nm. Cells were resuspended in SDS sample buffer and subjected to SDS–polyacrylamide gel electrophoresis (Laemmli, 1970). Quantification of *KdpFABC* was basically performed following the protocol developed for lactose permease (Sun et al., 1996). Briefly, proteins were electro-blotted to a nitrocellulose membrane. Blots were then blocked with 5% (w/v) bovine serum albumin (BSA) in 10 mM Tris/HCl (pH 7.5)/0.15 M NaCl (buffer A) for 1 h. Anti-KdpB antibody was added at a final dilution of 1:5000, and incubation was continued for 1 h. After washing with buffer A, ^{125}I-protein A (Amersham Biosciences) was added at a final dilution of 1:5000, and incubation was continued for 1 h. After washing thoroughly, the membrane was exposed to a phosphor screen. Known amounts of purified *KdpFABC* complex were used to obtain a standard curve. The amount of *KdpFABC* complex was then quantified using the PhosphorImager SI system (Molecular Dynamics) by comparison to the standard curve.

A.4. Quantification of kdpFABC mRNA

For quantification of the produced *kdpFABC* mRNA under K^+-limiting conditions, *E. coli* K-12 cells were grown at 37 °C in phosphate-buffered minimal medium (Epstein and Kim, 1971) containing 10 mM K^+ until the mid-exponential phase, filtered, and subsequently resuspended in medium of lower K^+ concentration (0.02 mM K^+) or the same medium as before (10 mM K^+). At the indicated times, cells were harvested and the RNA was prepared according to (Aiba et al., 1981). For quantitative Northern blot analysis, 20 μg of RNA from each sample was separated by electrophoresis in 1.2% (w/v) agarose-1.1 M formaldehyde gels in MOPS (morpholinepropanesulfonic acid) buffer. Equal loading of samples onto the gel was verified by ethidium bromide staining of the rRNA in a separate gel. RNA was transferred to Hybond-N nylon membrane (Amersham Biosciences) by upward capillary action. Hybridization was performed following a standard protocol (Sambrock et al., 1989) using γ-^{32}P-radio-labeled dCTP PCR fragments as specific probes for kdpA (nt 1009 to 1794). Radioactivity was quantified with the PhosphorImager SI (Molecular Dynamics).

References

Aiba, H., Adhya, S., de Crombrugghe, B., 1981. Evidence for two functional gal promoters in intact *E. coli* cells. J. Biol. Chem.

Altendorf, K., Epstein, W., 1996. The KdpATPase of *Escherichia coli*. In: Dalbey, R.E. (Eds.), Advances in Cell and Molecular Biology of Membranes and Organelles. JAI Press, vol. 5, Greenwich, London, pp. 401–418.

Bourret, R.B., Borkovich, K.A., Simon, M.I., 1991. Signal transduction pathways involving protein phosphorylation in prokaryotes. Annu. Rev. Biochem. 60, 401–441.

Bremer, H., Dennis, P.P., 1987. Modulation of chemical composition and other parameters of the cell by growth rate. In: Neidhardt, F.C. (Ed.), (Editor in Chief) *Escherichia coli* and *Salmonella typhimurium*. ASM Press, Washington, DC, pp. 1527–1542.

Van Dien, S.J., Keasling, J.D., 1998. A dynamic model of the *Escherichia coli* phosphate-starvation response. J. Theor. Biol. 190, 37–49.

Epstein, W., Kim, B.S., 1971. Potassium transport loci in *Escherichia coli* K-12. J. Bacteriol. 108, 639–644.

Fisher, S.L., Kim, S.-K.K., Wanner, B.L., Walsh, C.T., 1996. Kinetic comparison of the specifity of the vanomycin resistance kinase VanS for two response regulators. VanR and PhoB. Biochemistry 35, 4732–4740.

Hearne, J.W., 1985. Sensitivity analysis of parameter combinations. Appl. Math. Modelling 9, 106–108.

Heermann, R., Altendorf, K., Jung, K., 2003. The N-terminal input domain of the sensor kinase *KdpD* of *Escherichia coli* stabilizes the interaction between the cognate response regulator *KdpE* and the corresponding DNA-binding site. J. Biol. Chem. 278 (51), 51277–51284.

Jung, K., Altendorf, K., 2002. Towards an understanding of the molecular mechanisms of stimulus perception and signal transduction by the *KdpD/KdpE* system of *Escherichia coli*. J. Mol. Microbiol. Biotechnol. 4, 223–228.

Jung, K., Tjaden, B., Altendorf, K., 1997. Purification, reconstitution, and characterization of *KdpD*, the turgor sensor of *Escherichia coli*. J. Biol. Chem. 272, 10847–10852.

Koh, B.T., Tan, R.B.H., Yap, M.G.S., 1998. Genetically structured mathematical modeling of trp attenuator mechanism. Biotechnol. Bioeng. 58, 502–509.

Kremling, A., Bettenbrock, K., Laube, B., Jahreis, K., Lengeler, J.W., Gilles, E.D., 2001. The organization of metabolic reaction networks. Part III. Application for diauxic growth on glucose and lactose. Metab. Eng. 3 (4), 362–379.

Kremling, A., Gilles, E.D., 2001. The organization of metabolic reaction networks. Part II. Signal processing in hierarchical structured functional units. Metab. Eng. 3 (2), 138–150.

Kremling, A., Jahreis, K., Lengeler, J.W., Gilles, E.D., 2000. The organization of metabolic reaction networks: a signal-oriented approach to cellular models. Metab. Eng. 2 (3), 190–200.

Laemmli, U.K., 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227, 680–685.

Parkinson, J.S., 1993. Signal transduction schemes of bacteria. Cell 73 (5), 857–871.

Parkinson, J.S., Kofoid, E.C., 1992. Communication modules in bacterial signaling proteins. Annu. Rev. Genet. 26, 71–112.

Polarek, J.W., Williams, G., Epstein, W., 1992. The products of the *kdpDE* operon are required for expression of the Kdp ATPase of *Escherichia coli*. J. Bacteriol. 174 (7), 2145–2151.

Puppe, W., Jung, K., Lucassen, M., Altendorf, K., 1996. Characterization of truncated forms of the *KdpD* protein, the sensor kinase of the K^+-translocating Kdp system of *Escherichia coli*. J. Biol. Chem. 271, 25027–25034.

Saez-Rodriguez, J., Kremling, A., Gilles, E.D., Dissecting the puzzle of life: Modularization of signal transduction networks. Comput. Chem. Eng., 2004. in press.

Sambrock, J., Fritsch, E.F., Maniatis, T., Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1989.

Stock, A.M., Robinson, V.L., Goudreau, P.N., 2000. Two-component signal transduction. Annu. Rev. Biochem. 69, 183–215.

Stock, J.B., Stock, A.M., Mottonen, J.M., 1990. Signal transduction in bacteria. Nature 344, 395–400.

Sugiura, A., Nakashima, K., Tanaka, K., Mizuno, T., 1992. Clarification of the structural and functional features of the osmoregulated kdp operon of *Escherichia coli*. Mol. Microbiol. 6, 1769–1776.

Sun, J., Wu, J., Carrasco, N., Kaback, H.R., 1996. Identification of the epitope for monoclonal antibody 4B1 which uncouples lactose and proton translocation in the lactose permease of *Escherichia coli*. Biochemistry 35, 990–998.

Walderhaug, M.O., Polarek, J.W., Voelkner, P., Daniel, J.M., Hesse, J.E., Altendorf, K., Epstein, W., 1992. *KdpD* and *KdpE*, proteins that control expression of the *kdpABC* operon, are members of the two-component sensor-effector class regulators. J. Bacteriol. 174, 2152–2159.

Wong, P., Gladney, S., Keasling, J.D., 1997. Mathematical model of the lac operon: Inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. Biotechnol. Prog. 13, 132–143.

Metabolic networks: Biology meets Engineering sciences

A. Kremling, J. Stelling, K. Bettenbrock, S. Fischer and E.D. Gilles

Corresponding author: E. D. Gilles
Systems biology group
Max-Planck-Institut für Dynamik komplexer technischer Systeme
Sandtorstr. 1; 39106 Magdeburg, Germany
Tel.: +49-0391-6110-451; Email: gilles@mpi-magdeburg.mpg.de

Abstract

A hallmark of systems biology is the interdisciplinary approach to the complexity of biological systems, in which mathematical modeling constitutes an important part. Here, we use the example of sugar metabolism in the simple bacterium *Escherichia coli* and its associated control to illustrate the process of model development. Even for this well–characterized biological system, a close interaction between experimentation and theoretical analysis revealed novel, unexpected features. Additionally, the example shows how concepts from engineering sciences can facilitate the formal investigation of biological networks. More generally, we argue that analogies between complex biological and technical systems such as modular structures and common design principles provide crystallization points for fruitful research in both domains.

(111 words)

1 Systems Biology: An Interdisciplinary Approach

For the past 30 years, it has been characteristic for biology to be qualitative and descriptive, directed towards the understanding of the molecular detail. However, for the understanding of complex system properties like optimal control, adaptation and memory, both the systems components and their interactions have to be considered. Primarily the new 'omic technologies now make the complete determination of biological systems a realistic goal (Selinger et al., 2003). As a result, biology moves from the focus on few components to the study of networks of molecular interactions that give rise to complex physiological functions (Alm and Arkin, 2003). Systems biology adopts this holistic view on biological function. However, several characteristics distinguish it from, and extend bioinformatics approaches to network analysis. A hallmark of many cellular networks, such as the intricate networks in cellular regulation, is that they respond dynamically towards extra- and intracellular conditions and signals. Only by means of a quantitative description of the systems' constituents and interactions, the resulting behavior can be understood in terms of the quantitative dynamics.

Achieving this goal furthermore requires a theory-based approach to the complexity not understandable by intuition alone. Mathematical modeling of complex biological systems plays a central role in systems biology because it allows for a formalized treatment of biological networks in the computer, using tools from mathematics and systems sciences (Tyson et al., 2001; Kitano, 2002b). Ideally, mathematical modeling requires and entails a precise representation of the knowledge on the system, and of hypotheses for unknown mechanisms. It allows one to apply formal methods of analysis. Mainly these two characteristics are expected to lead to a deepened understanding of the biological systems under consideration (Endy and Brent, 2001; Gilman and Arkin, 2002). Consequently, the efforts directed towards a quantitative, system-level understanding in biology rely on an interdisciplinary approach combining concepts from biology, information sciences and systems engineering. A central objective of systems biology is finally to develop virtual representations of cells and organisms. These representations should allow for computer experiments similar to experiments with real biological systems. Thereby the way for a predictive biology can be paved, which will enhance, for instance, the understanding and the treatment of human diseases (Stelling et al., 2001; Kitano, 2002a). There are already some examples of systems biological approaches that successfully couple experimental and theoretical approaches. They cover a broad spectrum of organisms and systems (www.siliconcell.net). The analysis of bacterial chemotaxis can be regarded as a paradigm of such an approach. The extensive experimental and theoretical analysis has helped substantially in the understanding of the system (Barkai and

Leibler, 1997). Currently, however, the knowledge on virtually any biological system does not permit to detail a complete list of parts, interactions and mechanisms, on which 'true' mathematical representations could be built. Instead, despite considerable progress in high–throughput experimentation, the resulting networks are still incomplete and bear inaccuracies (von Mering et al., 2002). Under these circumstance, an often encountered argument is that theoretical analysis should await an – in some sense – complete biological knowledge before becoming meaningful. We and others, however, argue that only an iterative cycle of experimentation and theory will be able to fulfill the promises of systems biology. Experiments generate data and hypotheses, subsequent mathematical modeling allows to assess the compatibility of both, and to derive novel or alternative explanations that can be evaluated in new experiments (Stelling et al., 2001; Kitano, 2002b).

'Traditional' biology integrates new findings into cartoons of pathways or regulatory networks, or uses new knowledge to revise these representations. Similarly, mathematical models are 'work in progress' (Lee et al., 2003). In this process, however, unbiased predictions from formal representations can reveal unexpected properties of, or critical components in biological systems as in a recent experimental and theoretical analysis of the Wnt signaling pathway (Lee et al., 2003). In another case, mathematical modeling suggested a bistable trigger as a core element of cell cycle regulation a long time before an experimental confirmation of the mechanism was obtained (Novak and Tyson, 1993; Pomerening et al., 2003; Sha et al., 2003). Here, we use the control of sugar uptake in the simple bacterium *Escherichia coli* to show that an iterative cycle of experimentation and model development can yield deeper insight into apparently well–understood systems. In particular, our background in engineering sciences provides concepts and methods for this study. We will focus the discussion of recent developments and future challenges in systems biology on potential (further) contributions that engineering could make to understand complex biological systems.

2 Model set-up

Starting point of every model developing procedure is the biological knowledge available from literature or text books. For *Escherichia coli*, knowledge on metabolism as well as for genetic regulation is rich and especially the lactose operon and its control has been investigated for a long time. Starting with the pioneering work of Jacques Monod who proposed the concept of defining operons as a sequence of genes that were expressed in a coordinated manner, current research in molecular biology has revealed a number of further strategies of cellular systems to adapt very efficiently to alterations of environmental conditions. Here, we used the lactose metabolism, i.e. the uptake of lactose and its break down to precursors, as an origin for the model set up presented. In successive steps we extend the model to cope with further environmental situations like different carbon sources to show how the individual pathways are organized to fulfill their physiological task and how the cells arrange the interaction of different pathways on a higher level of control. This approach differs from previous studies and modeling efforts on the PTS mainly in that it aims to an understanding of the interactions of genetic regulation and metabolism. Previous approaches mainly delt with small subsystems covering either only metabolic reactions or only genetic regulation. A very seminal example is the work of Rohwer et al.(Rohwer et al., 2000) who set up a detailed kinetic model of the PTS phosphorylation chain. This model gives interesting insights into the effects of complex formation, molecular crowding and flux response coefficients of these reactions but as the system is uncoupled from metabolism and genetic regulation it is not suitable for the understanding of the coupling of both levels.

2.1 Environment – the liquid phase

Considering a bio-reactor, the environment of the cells is described with the concentration of the carbon source S in the liquid phase. Since below, the focus is on the cellular interior, the overall biomass X is taken as the macroscopic variable. Growth of the biomass is coupled to the uptake of the carbon sources via yield coefficient Y and uptake rate r . The uptake rates are functions of concentration of substrate in the bio-reactor and the concentration of the respective transport system which is located in the cytoplasmic membrane. For one substrate the respective equations read for a batch process:

$$\begin{aligned} \dot{X} &= \mu \cdot X \\ \dot{S} &= -r \cdot mg \cdot X \end{aligned} \tag{1}$$

with the specific growth rate μ is given by $\mu = Y * r$ and mg is the molecular weight of the carbon source (fluxes are given in [μmol/g DW h] and concentrations in the liquid phase are given in [g/l]). The equations in (1) are very general and are widely used in bioprocess engineering, since they describe the overall behavior of the biomass and the substrate in a simple manner. To describe the uptake reactions in a more detailed way, biological knowledge on the individual pathways has to be incorporated. As an example, the lactose pathway is considered in the following.

2.2 Lactose pathway

Lactose is taken up via the lactose permease LacY (gene *lacY*). The permease works as a symporter, i.e. for every molecule lactose that is taken up, some molecules of H^+ is also taken up from the medium. Intracellular lactose is split into glucose and galactose by the β-galactosidase enzyme LacZ (gene *lacZ*). One by-product of this reaction is allolactose, the natural inducer of the lactose operon. If allolactose is present inside the cell, it deactivates the lactose repressor LacI, which blocks the binding of the RNA polymerase and therefore prevents the synthesis of the mRNA. A further transcription factor, Crp, which is activated by cAMP, activates the transcription of the operon. As can be seen in Figure 1 the lactose pathway represents a loop with positive return. The more allolactose is present, the more protein can be synthesized. With increasing amounts of the respective enzymes, allolactose is also degraded faster and a steady-state can be reached. From the scheme, it becomes also clear that the initial conditions for all components could not be zero, if the system should be inducible. If lactose is not present in the medium, a few molecules are necessarily available in each cell and will allow induction by lactose. In Figure 1 the modeling objects for the lactose pathway are shown. For the enzymatic reactions simple Michaelis-Menten type kinetics are used. To describe the transcription efficiency, a reasonable approach is the choice of the fraction of free promoter binding sites with respect to all available promoter binding sites for the lactose operon. In comparison to the model equation system (1) the uptake of the carbohydrate is described more realistic since the synthesis of the transport system is included, which leads to a short delay of uptake.

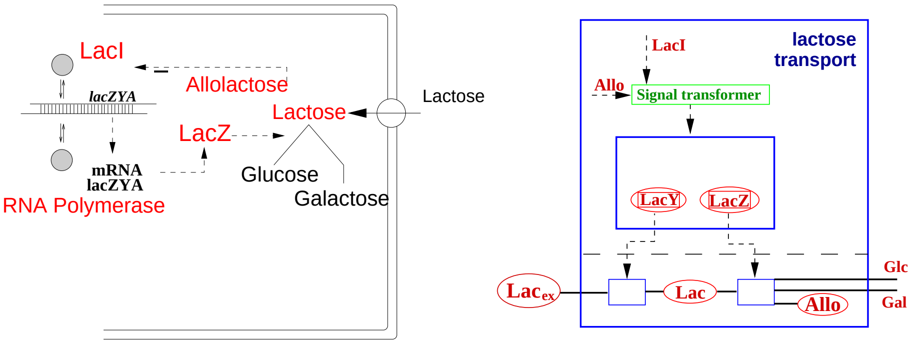


Figure 1: Lactose uptake and metabolism. Left: Schematic diagram of lactose induction. Right: Representation of the sub-model with modeling objects.

2.3 Glucose uptake

To extend the scheme for a further carbohydrate, here glucose, knowledge on the transport system on the metabolic and genetic level is incorporated. A very important experimental observation is diauxic growth, if glucose and lactose are provided at the same time in the bio-reactor. Therefore, the model must be set up in such a way, that this behavior is reproduced. Starting with metabolism, besides the uptake reactions (Figure 2), which comprise four proteins, glycolytic reactions are also included, since the energy for the transport comes from phosphoenolpyruvate (PEP). Glucose is taken up by the phosphoenolpyruvate dependent glucose phosphotransferase system (PTS). In a sequence of five steps, the high energy bond is translocated to the incoming substrate that appears in its phosphorylated form inside the cell. Connecting both pathways only on the metabolic level, does not lead to the required behavior in a simulation study. Therefore, knowledge on the genetic level of control has to be included also.

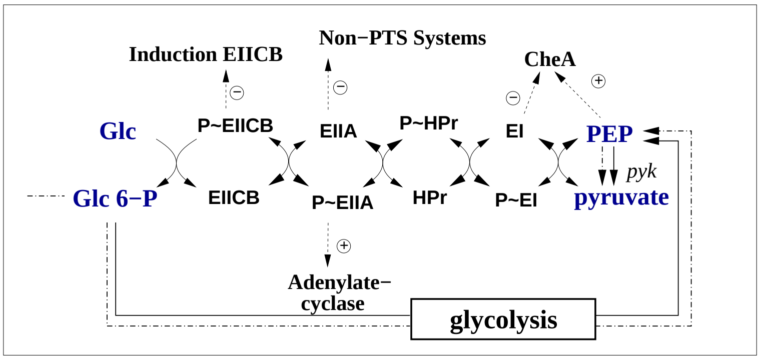


Figure 2: Schematic representation of the glucose PTS. Inputs are the entire concentrations of EI, HPr, EIIA, EIICB, PEP, pyruvate, and extracellular glucose. Important outputs are the phosphorylated and unphosphorylated forms of EIIA. These two conformations are measured in several experiments. Solid lines represent metabolic reactions and dashed lines signal outputs of the PTS. Dash-dot lines represents metabolic flux in case of no PTS transport system. Since PEP is also converted by the pyruvate kinase reaction (gene *pyk*) to pyruvate, the degree of phosphorylation is strongly influenced by PEP and pyruvate, even if the PTS is not active (after Kremling et al. (2004)).

Initially, knowledge on genetic regulation was restricted to a cAMP·Crp dependent induction of the gene *ptsG* which codes for the actual transport system EIICB^{Glc}. Transcription factor Crp is called a global transcription factor since it is involved in the expression of nearly 200 genes. C Since the lactose operon is also under control of cAMP·Crp the question arose, in which way the local control by LacI and the global control via Crp have to be modeled. Years ago, Lee and Bailey (1984ab) proposed a method where the transcription efficiency η is proportional to the fraction ψ_P of occupied promoters. The influence of an inhibitor, e.g. a repressor, blocking the promoter is taken into account with $1 - \psi_R$ which represents the free sites. Activators are taken into account by parameter α in the term $(1 + \alpha\,\psi_A)$. For the transcription efficiency the following equation holds according to the method of Lee and Bailey

$$\eta \;= \psi_P\;(1\;-\;\psi_R)\;(1\;+\;\alpha\,\psi_A)\,. \tag{2}$$

The proposed method is limited to the consideration of single operons. To be more flexible and to allow model extensions in a very simple way, we proposed a new method with focus on the hierarchical set up of the genetic regulation (Kremling and Gilles, 2001). For this method, the transcription factors are assigned to different levels in the hierarchy. The lowest level is represented by individual pathways, e.g. the lactose repressor LacI which is involved only in lactose metabolism. The second level is represented by global transcription factors, e.g.

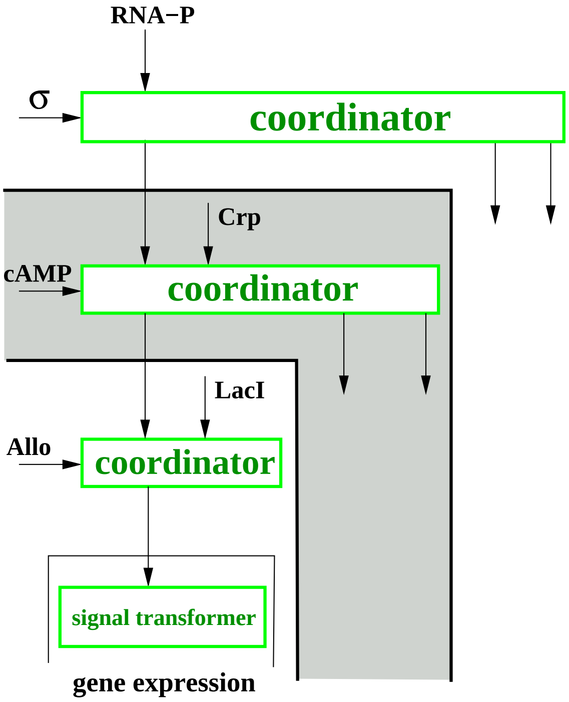


Figure 3: Hierarchical set up of the genetic regulation network. Signals are transduced from the top level to the lower level but not vice versa. The lowest level represents individual pathways, the second level represents global transcription factors while the highest level is reserved for the RNA polymerase.

Crp, which control a number of pathways. The highest level is reserved for the RNA polymerase which is involved in nearly all transcription processes (Figure 3).

As far as we have described the details of metabolism and genetic control, the picture is not yet completed and simulation results does not show a diauxic growth behavior. The missing link between both pathways is the interaction between the PTS, here output protein EIIA^{Glc} (gene *crr*) and (i) the lactose permease and (ii) the activation of the cAMP generating enzyme adenylate cyclase CyaA. In the following, both effects are analyzed in a detailed way (Figure 4). Protein EIIA^{Glc} is expected to be in either of two states: phosphorylated or unphosphorylated. It is known that unphosphorylated EIIA^{Glc} is able to inhibit the lactose permease (as well as some more enzymes in different carbohydrate uptake pathways). This is referred to as "inducer exclusion" since it prevents the entry of the substrates. On the other hand the phosphorylated form of EIIA^{Glc} is able to activate CyaA and therefore activates the synthesis of cAMP. However, the degree of phosphorylation of EIIA^{Glc} depends on a number of input variables as can be seen in Figure 2. In the case that the PTS is not active, the degree of phosphorylation depends only on the concentration of PEP and pyruvate (Kremling et al., 2004). Different PEP and pyruvate concentrations resulting in different EIIA^{Glc} phosphorylation states have already been demonstrated for varying growth substrates indicating that this imput is the most important one (Hogema et al., 1998). Another major input is the dephosphorylation of the PTS proteins by incoming substrates. Model analysis by dynamical simulation studies with the proposed model structure gives interesting insights in the dynamics of the intracellular components. We started with a batch experiment where glucose and lactose are provided from the beginning. Figure 5 shows the time course of selected state variables (all model equations and parameters are summarized in (Kremling et al., 2001)). As expected glucose is taken up while lactose is not. After the run out of glucose,

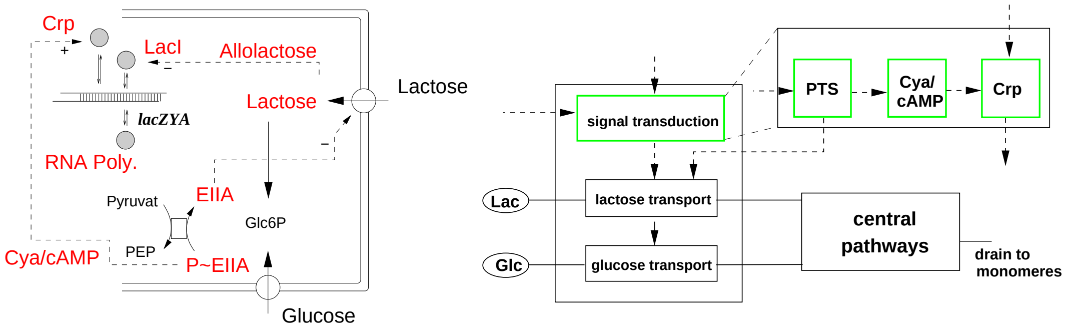


Figure 4: Interaction between the PTS and the lactose pathway. Left: Schematic diagram. Right: Modeling objects.

the PTS protein EIIA^{G}lc shows a very quick switch from the unphosphorylated to the phosphorylated form. This abolishes the inhibition of the lactose permease and furthermore leads to an activation of gene expression by the cAMP·Crp complex. cAMP is very low during the glucose uptake and rises in the lactose phase as a consequence of the degree of phosphorylation of EIIA^{G}lc. For the simulation it was assumed that some molecules of EIICB^{Glc} were available from the beginning. Since the promoter of *ptsG* has a high basal activity, the concentration during the glucose phase remains nearly constant. However, in the lactose phase, the concentration of EIICB^{Glc} rises due to the higher cAMP levels. Since glucose is no longer available for uptake and growth, the further synthesis of EIICB^{Glc} seemed not to be meaningful. In fact, during the time period when the model was developed genetic research revealed that, a so far unknown transcription factor, Mlc (also called DgsA), is involved in the specific control of EIICB^{Glc} (Plumbridge, 1998). The repressor is active if no glucose is present in the medium and leads to a shut off of gene expression during growth on lactose. Since the detailed mechanism was unclear, a simple model for repression of the *ptsG* gene (Kremling et al., 2001) shows a satisfactory behavior. This can also be seen in Figure 5.

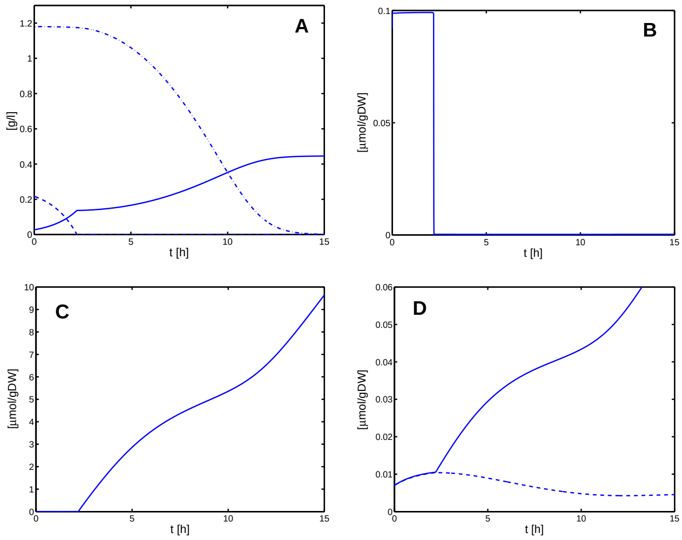


Figure 5: Simulation results of selected state variables. **A** Biomass (solid), extracellular glucose (dashed) and lactose (dash-dot). **B** EIIA. **C** cAMP. **D** EIICB^{Glc}. Comparison of two model variants. Repression of EIICB^{Glc} is not included (solid) and included with a simple model (dashed).

2.4 More detailed description of regulatory phenomena

Comparing the simulation results of the proposed model with experimental data, we noticed that the model was not able to reproduce the intracellular dynamics: (i) In (Inada et al., 1996) the time course of intracellular cAMP was measured. In an experiment using glucose and lactose as carbon sources, cAMP shows an adaptive behavior, i. e. after a steep rise at the end of the glucose uptake phase, the concentration of cAMP goes back to the values observed in the glucose phase. To reproduce this behavior, we included knowledge on the genetic control of the proteins involved in the signal transduction pathway, Cya and Crp, respectively. While Cya is negatively controlled by the cAMP·Crp complex, Crp is auto-controlled. The proposed mechanism is rather complex: For low cAMP·Crp concentrations transcription is inhibited while for larger concentrations an activation is proposed (Hanamura and Aiba, 1992). In Figure 6 the impact of the model extension is shown. Now, the qualitative behavior is reproduced correctly. (ii) The second model extension focuses on the kinetics of the inducer exclusion.

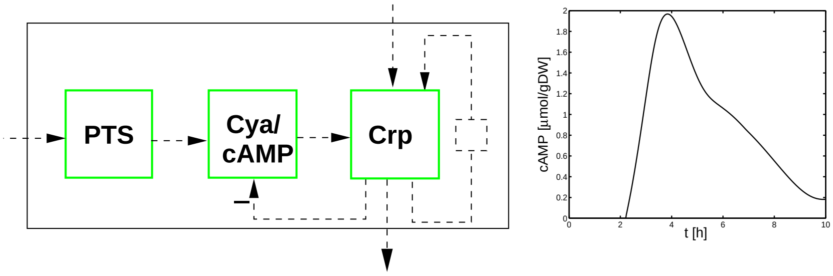


Figure 6: Model extension and simulation results of cAMP. The feedback loop to the adenylate cyclase Cya leads to an adaptive behavior. Crp is auto-controlled. This is indicated by the dashed box.

A common approach for modeling enzymatic kinetics is to use Michaelis-Menten type kinetics. A mechanism to describe inhibition extends the simple Michaelis-Menten equation by additional factors. A widely accepted assumption hereby is, that the amount of inhibitor (normally a metabolite) does not change significantly during binding at the enzyme since the concentration of the enzyme is much lower than the concentration of the metabolite. In (Rohwer et al., 1998) an interesting experiment is described where it is shown that the proportion of the concentration of enzyme and inhibitor is near one, depending on the experimental design used. In our model protein EIIA^{Glc} interacts with the lactose pathway. Interestingly, inhibition occurs only, if lactose is present in the medium. To include these facts into the model, the inhibitor EIIA^{Glc} (unphosphorylated) is assumed to be in two conformations that are in equilibrium:

$$EIIA^f \quad \overset{K}{\rightleftharpoons} \quad EIIA \cdot LacY \cdot Lac_{ex} \tag{3}$$

with $EIIA^f$ being the free form, $EIIA \cdot LacY \cdot Lac_{ex}$ being a ternary complex of EIIA, lactose permease LacY and external lactose, and K being the overall binding affinity. In the model equations for the PTS, only the free form is used as the driving potential.

(iii) Own measurements during the glucose/lactose diauxie experiment revealed some interesting dynamics of the degree of phosphorylation of protein EIIA^{Glc} during the second growth phase. As shown in Figure 5 EIIA^{Glc} is in its phosphorylated form during growth on lactose. Our experimental observation, however, indicates a slow rise of the unphosphorylated form for two hours and afterwards a slow decrease. It was speculated that the splitting of intracellular lactose into galactose and glucose and subsequent phosphorylation of intracellular glucose in glucose 6-phosphate is involved in the dephosphorylation of EIIA^{Glc}. As sketched in Figure 7 glucose has two possibilities to get phosphorylated: on the one hand, a glucokinase phosphorylates intracellular glucose with ATP or, on the other hand, intracellular glucose gets phosphorylated by the PTS. In the former case, protein EIIA^{Glc} remains in the phosphorylated state while in the second case, EIIA^{Glc} gets more and more dephosphorylated depending on the accumulation of intracellular glucose.

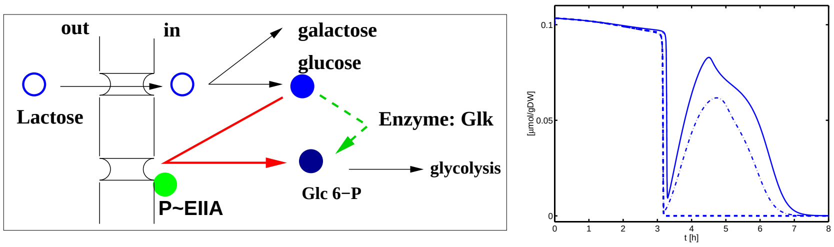


Figure 7: Model extension and simulation results of EIIA^{Glc}. During the second growth phase, phosphorylation by either the glucokinase or the PTS is possible. The simulation on the right side show the expected results: Flux only via the PTS (solid line), only via the glucokinase (dashed line) or a mixture form both possibilities (dash-dot line).

2.5 Regulation by Mlc

The simulation shown above indicates that the PTS phosphorylates the intracellular glucose. To further verify this hypothesis, experiments with mutant strains were designed that differ only in one gene of interest (isogenic mutants). Strain Glk$^-$ misses the glucokinase enzyme while strain Mlc$^-$ misses the specific repressor for EIICB^{Glc} and the other PTS proteins HPr and EI. Simulation results show that the Mlc$^-$ strain should show lower values of EIIA^{Glc} in the lactose phase since higher levels of the PTS proteins are expected to phosphorylate EIIA^{Glc} in a more efficient way. In Figure 8, the dynamics of protein EIIA^{Glc} (unphosphorylated) is shown. A good agreement between the simulation results and the experimental data is observed. After fitting the parameters of the model, all experiments could be described with a single set of parameters (publication in preparation). Note, that for parameter identification experimental data for other state variables like biomass, extracellular substrates, extracellular cAMP and LacZ was used. Figure 9 shows the time course of these state variables for the wild type strain during the batch experiment.

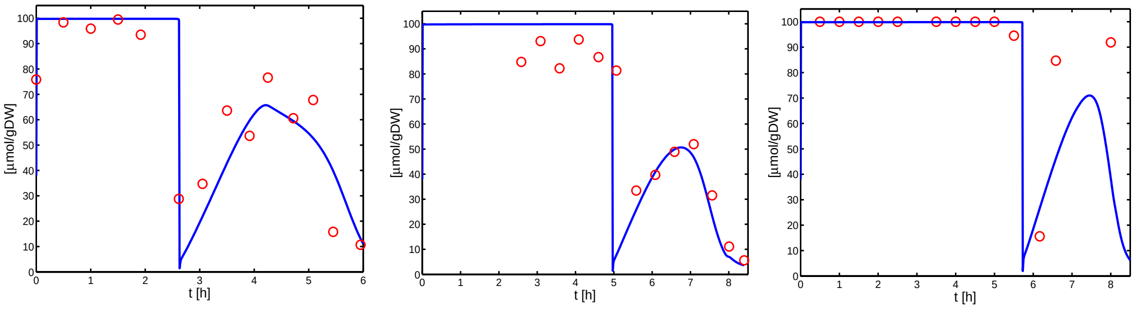


Figure 8: Simulation (solid line) and experimental data (circles) for three batch experiments using wild type strain (left), Mlc$^-$ strain (middle) and Glk$^-$ (right). The model was fitted to the data; the experiments could be described with one set of parameters (publication in preparation).

2.6 Model analysis – implications for diauxic growth

The model described so far was extended step by step by incorporating pathways for additional carbohydrates. The current version is able to describe the uptake of six sugars, glucose, lactose, galactose, glycerol, glucose 6-phosphate, and sucrose (*E. coli* wild type strain is not able to grow on sucrose; therefore a mutant strain with a sucrose PTS was constructed and analyzed (Kremling et al., 2004)). To

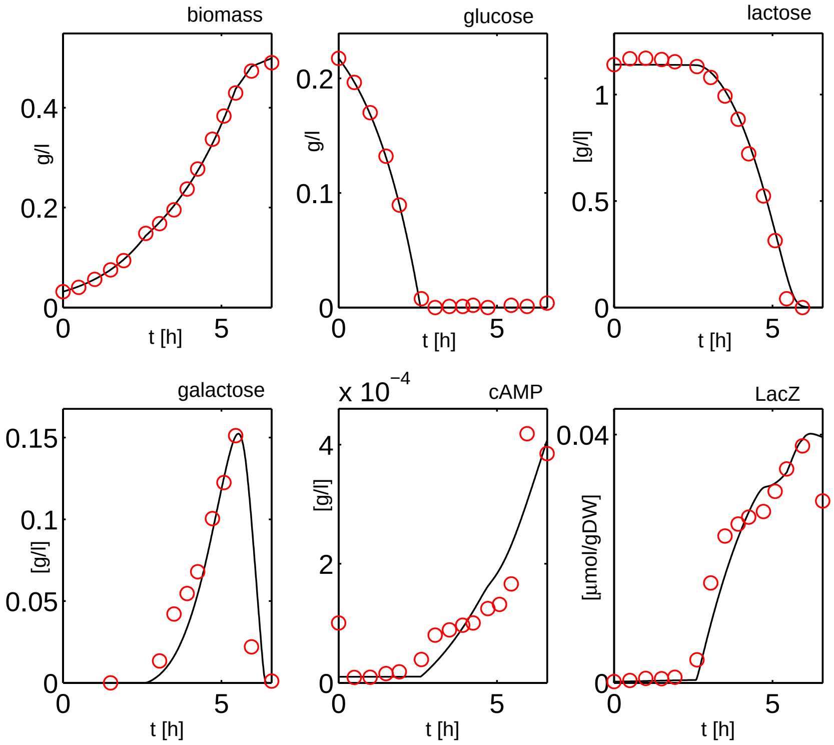


Figure 9: Simulation (solid line) and experimental data (circles) for the wild type strain during the batch experiment. Glucose is taken up right from the beginning while lactose is taken up in the second growth phase. Interestingly, galactose, a product from the LacZ reaction is excreted in the medium at the beginning of the second growth phase. When lactose runs out, *E. coli* uses galactose as additional carbohydrate source. cAMP is also excreted in large amount during the second growth phase.

fit the parameters, experiments under different environmental conditions, experiments with mutant strains, and experiments with different pre-culture conditions were performed (publication in preparation). The model has 50 state variables and needs 300 parameters. For nearly all parameters values were found in literature. Based on a sensitivity and parameter analysis, 60 parameters could be estimated from the experimental data.

The key elements of the model are summarized for a PTS carbohydrate and a non PTS carbohydrate in Figure 10. The transport system are normally under dual control. Besides a carbohydrate specific control by repressors like LacI, GalR, or GlyR, most systems are under control of the global regulator Crp thereby depending on the degree of phosphorylation of the PTS protein EIIA^{Glc}. In this case the advantages of a systems biological approach become obvious. Because of the wealth of important and interacting regulations, metabolite concentrations

Table 1: Summary of functional units, number of parameters and number of estimated parameters. About 20 different experiments are used for parameter fitting. [a] Parameters estimated with Metabolic Flux Analysis.

module name	param.	param. estimated	number of states	type
PTS (general)	21	9	9	ODE
PTS Glc	12	4	1	ODE
Cya	9	2	2	ODE
Crp	17	3	1	ODE
2nd Glc transporter	18	3	3	ODE
Lac transporter	16	7	4/2	ODE/ algebraic
Scr transporter	26	9	6	ODE
Gly transporter	24	5	5	ODE
Gal transporter	43	4	11/2	ODE/ algebraic
Catabolic reactions	51	11	8	ODE
Monomer synthesis	7	4^a+3	1	ODE
Liquid phase	7	5	8	ODE

and protein states, only a quantitative systems oriented approach will help in the understanding and will be able to identify the abilities of the system. It can also help to identify some general properties of the system.

Diauxic growth is observed for a number of couples of carbohydrates. With the model at hand and the simplified scheme in Figure 10 some general conclusions could be drawn: There is no unique control circuit that leads to diauxic behavior. Rather, diauxic growth is the result of a number of different control schemes and kinetic parameter constellation. So, a PTS sugar does not repress the uptake of a non PTS sugar in general. Own measurements with glucose and glucose 6-phosphate (a non PTS sugar with an uptake system that is also under control of the cAMP·Crp complex) show that the uptake of glucose is repressed while glucose 6-phosphate is taken up immediately. Measurement of the synthesis of the glucose transporter EIICB^{Glc} by LacZ fusion revealed that EIICB^{Glc} is no more synthesized although glucose is present in the medium. In (Morita et al., 2003) it is speculated that high concentrations of glycolysis intermediates like glucose 6-phosphate or fructose 6-phosphate may be involved in the down regulation of the EIICB^{Glc} messenger RNA. With the model, simulation studies can be done to verify the hypothesis. Figure 11 show simulation results for glucose 6-phosphate uptake. If it is assumed that high levels of intracellular glucose 6-phosphate is able to inhibit the synthesis of protein EIICB^{Glc} the uptake of glucose is inhibited in the first growth phase. Since glucose 6-phosphate does not accumulate any longer, transporter EIICB^{Glc} can be synthesized again for uptake of glucose.

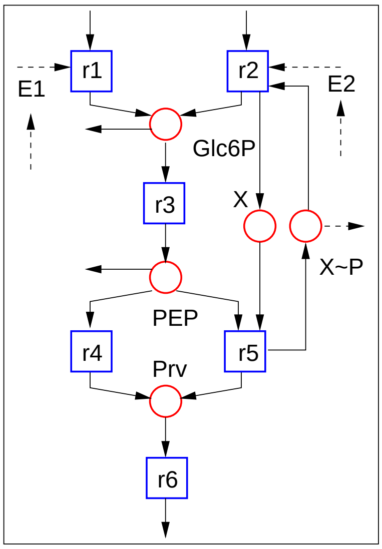


Figure 10: Key elements of the carbohydrate uptake systems. r1 represents the uptake system for a non PTS sugar, r2, together with r5 the uptake system for a PTS sugar, X and $X{\sim}P$ represent a PTS protein, r3 glycolysis, r4 the pyruvate kinase reaction, and r6 the drain of pyruvate (Prv). $E1$ and $E2$ are the respective proteins for the transporter. Both are subject to control. Most of the carbohydrate transport systems are controlled by Crp that is activated by cAMP. In the scheme this is represented by the signal arrow coming from a PTS representative (X$\sim P$).

3 Recent Developments and Future Challenges

The example of modeling carbohydrate uptake in *E coli* showed that close interactions between experimentation and theoretical analysis may yield novel insight into an 'old' biological system. Apparently, for less well characterized cellular systems, the question of how to best organize these interactions is of even more relevance. Besides discussing recent developments and challenges in this aspect of systems biology, we will broaden our view to more general principles of organization and function. In all cases, engineering sciences offer concepts and methods that can help in understanding biology. We will draw on analogies between complex biological and technical systems to illustrate this point.

3.1 Experimentation and Theory

The characterization of network components and interactions in qualitative and quantitative terms is a prerequisite for an integrated understanding as well as for realistic mathematical models of biological systems (Kitano, 2002b). For in-

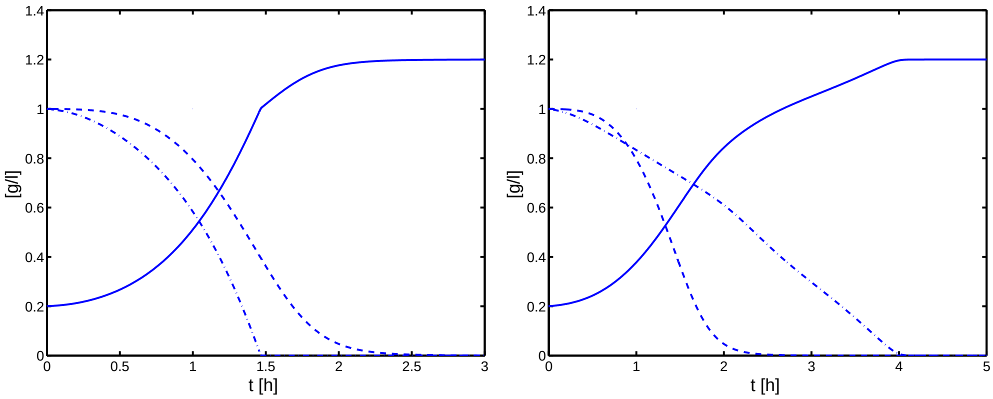


Figure 11: Simulation study for glucose 6-phosphate uptake. Left: In the uncontrolled system both sugars are taken up in parallel. Right: Assuming a feedback loop from intracellular glucose 6-phosphate to the synthesis of the glucose uptake system, the uptake of glucose is inhibited. Biomass (solid line), glucose (dash-dot), glucose 6-phosphate (dashed).

stance, determining all interactions between the components in an organism of low complexity such as *E. coli* has been estimated to require between 50 and 40,000 microarray experiments (Selinger et al., 2003). Hence, optimizing the way in which these experiments are conducted holds great promises for the efficiency of systems approaches. In experimental biology, educated guesses in 'traditional' hypothesis–driven research and, more recently, comprehensive studies using, for instance, systematic gene knock–outs prevail. Systems engineering offers a large body of theory for the identification problem (Ljung, 1999) that can be employed to assess the information content of experimental data (as for the estimation of parameters in our *E. coli* example), and to suggest efficient strategies for generating quantitative data. For instance, a recent study applied tools from systems sciences to an artificial gene network in order to analyze the effect of (inherent) stochastic fluctuations and (purpose–driven) input perturbations on the identification of model parameters (Zak et al., 2003). We believe that, besides specific predictions leading to new experiments once a mathematical model is available, systematic investigations of this type can result in more general guidelines for experimental strategies to quantitatively characterize biological networks.

Many biological systems of interest, however, are not yet amenable to this approach relying on detailed mathematical models, for instance, owing to an incomplete and / or inaccurate knowledge on components and interactions. There, the challenge is to derive the system's working principles from the observable behavior. This reverse–engineering usually entails the discrimination between a large number of hypothetical mechanisms to infer the causal relationships. For the analysis of gene networks, for instance, several theoretical approaches to the problem have been suggested. They range from boolean networks that consider only the 'on' and 'off' states of genes to detailed dynamical models (D'haeseleer et al., 2000). A great challenge for the future obviously is to assess the relative power of the methods, and their data requirements. More generally, however, it will be crucial to find a common basis for theoretical approaches at different resolution that are currently incompatible with each other. Only such a unification will allow for the desired gradual transition from coarse to very detailed representations of complex biological networks, depending on the knowledge on the system and the specific interest of the investigator (Ideker and Lauffenburger, 2003). Here, for instance, general systems theory provides a theoretical framework (Willems, 1991) that could be built upon. In brief, it regards systems (and models as their representation) as functional entities that simply map a set of inputs to a set of outputs. As such, it enables a general treatment of models similar to the ideas outlined in (Selinger et al., 2003).

3.2 Modules and Hierarchies

One parallel between biological and technical systems is particularly striking and can greatly facilitate the systems biology approach: It is increasingly accepted that both types of systems are composed of semi-autonomous modules that perform a specific function. Biological modules acting as switches, triggers, amplifiers, and other functional units are paralleled by similar devices in, for instance, electrical and control engineering (Hartwell et al., 1999; Nurse, 2003). Modularity in general, and these analogies in particular, have at least three important implications for our ability to understand integrated biological systems: (i) they allow for the decomposition of complex networks into manageable units, which can later be re–assembled to obtain the whole picture, (ii) corresponding modular concepts for mathematical modeling and formal analysis facilitate theoretical investigations in systems biology as illustrated by our *E. coli* example, and (iii) it will be possible to draw on the large repertoire of methods and insights from engineering sciences by elaborating common operating principles of prototypical technical and biological (sub)systems (Stelling et al., 2001; Csete and Doyle, 2002).

A major current challenge for elucidating and exploiting modularity in biology, however, is to find objective criteria for the demarcation of modules. Several approaches have been suggested in the literature, for instance, regarding the dissection of complex metabolic networks into simpler modules (Schuster et al., 1993). Most intuitively, functional units can be characterized as performing a common physiological task and belonging to the same genetic unit and / or signal processing entity (Kremling et al., 2000). Yet, similar to the delineation of pathways from complex interaction maps in traditional biology, in many cases these ‘soft’ criteria prevent an unambiguous assignment of modules. Methods from graph theory that analyze the components (nodes) and their interactions (links) in networks yielded statistically overrepresented ‘motifs’ in transcriptional networks (Shen-Orr et al., 2002). A particular functionality could be assigned to some of these recurring small networks of interactions through more detailed dynamic analysis. For instance, a three-gene circuit termed the ‘feed–forward motif’ specifically either accelerates or delays transcriptional responses (Mangan and Alon, 2003). These analyses are confined to small patterns of interactions, and their role in the larger system is unclear at present. At a larger scale, graph–theoretical approaches revealed a hierarchical ordering of modules for the genome–wide metabolic network of *E. coli* (Ravasz et al., 2002). However, graph models may be too coarse to reflect biological functionality (Arita, 2004).

Furthermore, concepts exist that explicitly take function into account from the beginning. For instance, metabolic pathway analysis identifies the smallest functional units in metabolism, but these units are usually overlapping (Rohwer et al., 1996; Schuster et al., 2000). The search for co–regulated genes in libraries on gene expression data obtained by microarrays showed common patterns of hierarchical modularity in different organisms, yet the resolution of individual modules is influenced by adjustable parameters of the analysis method (Bergmann et al., 2004). Finally, a recent proposal concerns the demarcation of modules based on a criterion from systems theory, namely the absence of retro–activity (Saez-Rodriguez et al., 2004). In summary, thus, albeit a multitude of methods to analyze modularity in biological systems exists, their caveats do not allow to conclusively specify modules – or to prove their existence. Apparent next steps could consist in, for instance, a systematic comparison of the analysis results for a model system. It will be tempting to develop hybrid approaches taking into account multiple criteria for delineating modules. In addition, a hierarchical structure of biological networks raises important, largely unaddressed questions on the role of hierarchies in the co–ordination of cellular functions. Modularity and hierarchies open new directions for the multi–level analysis of biological systems, for which, for instance, electric circuit engineering provides suitable paradigms (Nurse, 2003). Not only systems biology, but also engineering theory will benefit from analogies between biological and technical systems.

3.3 Functions and Design Principles

The notion of function is a common denominator of biological and engineered systems. In contrast, physical systems may show equally complex networks, resulting in complex behavior. However, they arise without purpose, and are not driven by evolution or voluntary engineering as for the first two classes of systems (Hartwell et al., 1999). The crucial point here is that, to perform similar functions, biological and synthetic systems use similar design principles. Negative feedback, for instance, serves to maintain homeostasis in both domains. Consequently, translation of engineering principles into the realm of biology will have a major impact on understanding the structure and function of complex biological systems (Csete and Doyle, 2002). At a detailed level, two directions of future research appear obvious. As for perfect adaptation in bacterial chemotaxis, mapping a complicated biological network to a well–known engineering principle – integral feedback control in this case – may explain the observed behavior (Yi et al., 2000). Conversely, necessary conditions for achieving a particular function in engineered systems can guide detailed investigations in biology. For example, methods from control theory were recently employed to provide an analytical method for deciding whether positive feedback in biology leads to bistable switching (Angeli et al., 2004); when such a behavior is observed *in vivo*, 'missing links' in the assumed circuit diagram could, hence, be identified. This kind of studies, however, is only at the beginning. Important avenues of future research will be to examine control–theoretical concepts such as (structural) identifiability and controllability with respect to their applicability to biological systems.

Systems biology and engineering alike are presumably most challenged by the need to understand and / or to optimize highly integrated systems with a large number of interacting components. In both domains, robustness, that is, resistance to perturbations and failures constitutes a prominent design goal. Some ingredients for achieving this property such as feedback control, modularity, and hierarchies are known in engineering, and engineered systems were highly optimized in this regard (Csete and Doyle, 2002). However, it seems reasonable to assume that evolution in biology came up with more efficient and / or alternative solution to the problem. Hence, in our opinion, analyzing the design principles of biology in this respect will prove beneficial both for systems biology and for engineering. Model–based analyses of metabolic networks in bacteria already revealed parts of the control logic: whereas control at the level of fluxes ensures optimal growth for each particular situation the organism encounters (Ibarra et al., 2002), the control of metabolic gene expression seems to trade–off the efficiency in this situation, and the organism's flexibility to respond to environmental changes (Stelling et al., 2002). Although seemingly being at completely different levels of abstraction, the search for design principles profoundly feeds back on the interactions between experimentation and theory. For instance, the insight into metabolic control was obtained by using the structure of metabolic networks alone. Hence, theoretical investigations may help to decipher information from well–known properties, and to indicate less rewarding, in addition to new directions of experimental research.

4 Conclusions

Biological complexity is the substrate for the emerging field of systems biology, with the aim of an integrated understanding of complex biological systems as its driving force. Beyond this, however, we believe that a main characteristic of the systems biology approach is its interdisciplinary nature that combines methods and concepts from biology, information sciences and engineering. In particular, mathematical modeling of biological systems will serve to achieve the goals of systems biology, and to help establishing a more quantitative biology. As our example of sugar uptake in *E. coli* and its control showed, a close interaction between experimental biology and computational analysis is able to establish quantitative and predictive mathematical models. Such models can, for instance, be employed to reveal inconsistencies in the current knowledge on a system, assess the explanatory power of alternative hypotheses, and ultimately suggest new experiments that verify or falsify the model predictions. We believe that this iterative cycle, combining experimentation and theory will be essential for the success of systems biology.

A major current challenge, thus, is to increase the efficiency of the interactions. In this case, as for other fields that warrant more intense research, engineering can provide well-established theoretical concepts. Analogies between complex biological and technical systems are obvious, for instance, their modular and hierarchical structure, the notion of functions that have been optimized, and the underlying general design principles. Future research in these fields can be anticipated to yield operating principles that will increase our comprehension of how complex systems in generals are designed and perform. Moreover, for biology, such design principles will guide detailed investigations of specific biological systems. For engineering, they can provide new paradigms (or revive old ones), for instance, regarding the efficient control of integrated technical systems. A major obstacle on the way to gain these potential benefits from systems biology, however, is the still existing 'clash of civilizations' (Huntington, 1993) between the sciences in biology and engineering. Finding a common language and educating a new breed of scientists that are familiar with both fields (Lazebnik, 2002), thus, should be a central objective of all initiatives in systems biology.

References

E. Alm and A.P. Arkin. Biological networks. *Curr. Opin. Struct. Biol.*, 13: 193–202, 2003.

D. Angeli, J.E. Ferrell, Jr., and E.D. Sontag. Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. *Proc. Natl. Acad. Sci. U.S.A.*, 101:1822–27, 2004.

M. Arita. The metabolic world of *Escherichia coli* is not small. *Proc. Natl. Acad. Sci. U.S.A.*, 101:1543–47, 2004.

N. Barkai and S. Leibler. Robustness in simple biochemical networks. *Nature*, 387, 1997.

S. Bergmann, J. Ihmels, and N. Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.*, 2:E9, 2004.

M.E. Csete and J.C. Doyle. Reverse engineering of biological complexity. *Science*, 295:1664–69, 2002.

D. Endy and R. Brent. Modelling cellular behaviour. *Nature*, 409:391–95, 2001.

A. Gilman and A.P. Arkin. Genetic "code": representations and dynamical models of genetic components and networks. *Annu. Rev. Genomics Hum. Genet.*, 3:341–69, 2002.

A. Hanamura and H. Aiba. A new aspect of transcriptional control of the *Escherichia coli crp* gen: positive autoregulation. *Molecular Microbiology*, 6: 2489–2497, 1992.

L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray. From molecular to modular cell biology. *Nature*, 402 (Supp.):C47–C52, 1999.

B. M. Hogema, J. C. Arents, R. Bader, K. Eijkemanns, H. Yoshida, H. Takahashi, H. Aiba, and P. W. Postma. Inducer exclusion in *Escherichia coli* by non-PTS substrates: the role of the PEP to pyruvate ratio in determining the phosphorylation state of enzyme IIA^{Glc}. *Mol. Microbiol.*, 30:487–498, 1998.

S.P. Huntington. The clash of civilizations. *Foreign Affairs*, 72:22–28, 1993.

R.U. Ibarra, J.S. Edwards, and B.O. Palsson. *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420: 186–89, 2002.

T. Ideker and D. Lauffenburger. Building with a scaffold: emerging strategies for high- and low-level cellular modeling. *Trends Biotechnol.*, 21:255–62, 2003.

T. Inada, K. Kimata, and H. Aiba. Mechanism responsible for glucose-lactose diauxie in *Escherichia coli*: challenge to the camp model. *Genes Cells*, 1: 293–301, 1996.

H. Kitano. Computational systems biology. *Nature*, 420:206–10, 2002a.

H. Kitano. Systems biology: a brief overview. *Science*, 295:1662–64, 2002b.

A. Kremling, K. Bettenbrock, B. Laube, K. Jahreis, J.W. Lengeler, and E.D. Gilles. The organization of metabolic reaction networks: III. Application for diauxic growth on glucose and lactose. *Metab. Eng.*, 3(4):362–379, 2001.

A. Kremling, S. Fischer, T. Sauter, K. Bettenbrock, and E. D. Gilles. Time hierarchies in the *Escherichia coli* carbohydrate uptake and metabolism. *BioSystems*, 73(1):57–71, 2004.

A. Kremling and E.D. Gilles. The organization of metabolic reaction networks: II. Signal processing in hierarchical structured functional units. *Metab. Eng.*, 3(2):138–150, 2001.

A. Kremling, K. Jahreis, J.W. Lengeler, and E.D. Gilles. The organization of metabolic reaction networks: A signal-oriented approach to cellular models. *Metab. Eng.*, 2(3):190–200, 2000.

Y. Lazebnik. Can a biologist fix a radio? - Or what I learned while studying apoptosis. *Cancer Cell*, 2:179–82, 2002.

E. Lee, A. Salic, R. Kruger, R. Heinrich, and M.W. Kirschner. The roles of APC and axin derived from experimental and theoretical analysis of the Wnt pathway. *PLoS Biol.*, 1:E10, 2003.

S. B. Lee and J. E. Bailey. Genetically structured models for *lac* promotor-operator function in the *Escherichia coli* chromosome and in multicopy plasmids: *lac* operator function. *Biotechnology and Bioengineering*, 26:1372–1382, 1984a.

S. B. Lee and J. E. Bailey. Genetically structured models for *lac* promotor-operator function in the *Escherichia coli* chromosome and in multicopy plasmids: *lac* promotor function. *Biotechnology and Bioengineering*, 26:1383–1389, 1984b.

L. Ljung. *System identification : theory for the user.* Prentice Hall PTR, Upper Saddle River, NJ, 2nd edition, 1999.

S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. U.S.A.*, 100:11980–85, 2003.

P. D'haeseleer, S. Liang, and R. Somogy. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16:707–26, 2000.

C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.

T. Morita, W. El-Kazzar, Y. Tanaka, T. Inada, and H. Aiba. Accumulation of glucose 6-phosphate or fructose 6-phosphate is responsible for destabilization of glucose transporter mRNA in *Escherichia coli*. *J. Biol. Chem.*, 278(18): 15608–15614, 2003.

B. Novak and J.J. Tyson. Numerical analysis of a comprehensive model of M-phase control in Xenopus oocyte extracts and intact embryos. *J. Cell Sci.*, 106: 1153–68, 1993.

P. Nurse. Understanding cells. *Nature*, 424:883, 2003.

J. Plumbridge. Expression of *ptsG*, the gene for the major glucose pts transporter in *Escherichia coli*, is repressed by Mlc and induced by growth on glucose. *Molecular Microbiology*, 29(4):1053–1063, 1998.

J.R. Pomerening, E.D. Sontag, and J.E. Ferrell Jr. Building a cell cycle oscillator: hysteresis and bistability in the activation of Cdc2. *Nat. Cell Biol.*, 5:346–51, 2003.

E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–55, 2002.

J. M. Rohwer, R. Bader, H. V. Westerhoff, and P. W. Postma. Limits to inducer exclusion: Inhibition of the bacterial phosphotransferase system by glycerol kinase. *Molecular Microbiology*, 29:641–652, 1998.

J. M. Rohwer, N. D. Meadow, S. Roseman, H. V. Westerhoff, and P. W. Postma. Understanding glucose tranport by the bacterial phosphoenolpyruvate:glycose phosphotransferase system on the basis of kinetic measurements in vitro. *J. Biol. Chem.*, 275:34909–34921, 2000.

J. M. Rohwer, S. Schuster, and H.V. Westerhoff. How to recognize monofunctional units in a metabolic system. *Journal of Theoretical Biology*, 179:214–228, 1996.

J. Saez-Rodriguez, A. Kremling, and E. D. Gilles. Dissecting the puzzle of life: Modularization of signal transduction networks. *Computers & Chemical Engineering*, 2004. Accepted.

S. Schuster, D.A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, 18:326–32, 2000.

S. Schuster, D. Kahn, and H. V. Westerhoff. Modular analysis of the control of complex metabolic pathways. *Biophys Chem.*, 48:1–17, 1993.

D.W. Selinger, M.A. Wright, and G.M. Church. On the complete determination of biological systems. *Trends Biotechnol.*, 21:251–54, 2003.

W. Sha, J. Moore, K. Chen, A.D. Lassaletta, C.S. Yi, J.J. Tyson, and J.C. Sible. Hysteresis drives cell-cycle transitions in *Xenopus laevis* egg extracts. *Proc. Natl. Acad. Sci. U.S.A.*, 100, 2003.

S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, 31(1):64–68, 2002.

J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E.D. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420:190–93, 2002.

J. Stelling, A. Kremling, M. Ginkel, K. Bettenbrock, and E.D. Gilles. Towards a Virtual Biological Laboratory. In H. Kitano, editor, *Foundations of Systems Biology*, pages 189–212. MIT Press, Cambridge, MA, 2001.

J.J. Tyson, K. Chen, and B. Novak. Network dynamics and cell physiology. *Nat. Rev. Mol. Cell Biol.*, 2:908–16, 2001.

J.C. Willems. Paradigms and puzzles in the theory of dynamical systems. *IEEE Transac. Automat. Control*, 36(3):259–94, 1991.

T.-M. Yi, Y. Huang, M.I. Simon, and J. Doyle. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc. Natl. Acad. Sci. U.S.A.*, 97(9):4649–53, 2000.

D.E. Zak, G.E. Gonye, J.S. Schwaber, and F.J. Doyle III. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: Insights from an identifiability analysis of an in silico network. *Genome Res.*, 13:2396–405, 2003.

BMC Systems Biology

Research article

Open Access

Analysis of global control of *Escherichia coli* carbohydrate uptake

Andreas Kremling*, Katja Bettenbrock and Ernst Dieter Gilles

Address: Max-Planck-Institut Magdeburg, Systems Biology, Sandtorstr. 1, 39106 Magdeburg, Germany

Email: Andreas Kremling* - kremling@mpi-magdeburg.mpg.de; Katja Bettenbrock - bettenbrock@mpi-magdeburg.mpg.de;
Ernst Dieter Gilles - gilles@mpi-magdeburg.mpg.de

* Corresponding author

Published: 13 September 2007

BMC Systems Biology 2007, **1**:42 doi:10.1186/1752-0509-1-42

Received: 8 May 2007
Accepted: 13 September 2007

This article is available from: http://www.biomedcentral.com/1752-0509/1/42

© 2007 Kremling et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Global control influences the regulation of many individual subsystems by superimposed regulator proteins. A prominent example is the control of carbohydrate uptake systems by the transcription factor Crp in *Escherichia coli*. A detailed understanding of the coordination of the control of individual transporters offers possibilities to explore the potential of microorganisms e.g. in biotechnology.

Results: An o.d.e. based mathematical model is presented that maps a physiological parameter – the specific growth rate – to the sensor of the signal transduction unit, here a component of the bacterial phosphotransferase system (PTS), namely EIIA^{Crr}. The model describes the relation between the growth rate and the degree of phosphorylation of EIIA ^{crr} for a number of carbohydrates by a distinctive response curve, that differentiates between PTS transported carbohydrates and non-PTS carbohydrates. With only a small number of kinetic parameters, the model is able to describe a broad range of experimental steady-state and dynamical conditions.

Conclusion: The steady-state characteristic presented shows a relationship between the growth rate and the output of the sensor system PTS. The glycolytic flux that is measured by this sensor is a good indicator to represent the nutritional status of the cell.

Background

Mathematical models of cellular systems describing metabolism, signal transduction and gene expression are becoming more and more important for the understanding of the underlying molecular processes. Since the earliest work to elucidate the molecular nature of regulatory structures by J. Monod, the knowledge of the detailed interactions between the components that are responsible for carbohydrate uptake in *Escherichia coli* is steadily increasing. Although current research on individual uptake systems like glucose still reveals new players that maybe play a role in local control [1], the knowledge of individual uptake systems is rich and is used as a basis to set up mathematical models to describe and analyze the properties of the control circuits. E.g. for the lactose uptake system in *E. coli*, it was shown that the autocatalytic action of inducer allolactose is responsible for the existence of multi-stationarity [2]. Such nonlinear properties of sub-networks are often described and assigned to a certain functionality of the system. The understanding of how different stimuli of the same type – in this study carbohydrates – are sensed by the cells and how these different signals are processed is still lacking. Here, we used experimental data published by our group [3,4] to elucidate and characterize such a global control circuit, that is, a regulatory scheme, that senses a physiological parameter like the specific growth rate and maps it to the degree of phosphorylation of the intracellular component EIIA^{Crr}. EIIA^{Crr} is a component of the phosphoenolpyruvate (PEP): carbohydrate phosphotransferase system (PTS). The PTS is not only a transport system for a number of carbohydrates but also acts as a sensory system. Sensor elements like the PTS can be seen as logic elements that process external stimuli into intracellular signals. High fluxes through the glycolysis, corresponding to high growth rates result in a low degree of phosphorylation of EIIA^{Crr}. At first view, this is surprising, since, assuming a linear reaction chain, high fluxes result in high pool concentrations, based on the (normally) monotone dependency of the reaction rate on the substrate concentration. The PTS together with the glycolysis can now be seen as an element that allows a transformation of high fluxes into a low pool concentration. This is not only due to the existence of two complementary pools like EIIA^{Crr} and its phosphorylated form, but as we will show, depends strongly on the flux distribution at the PEP node. High fluxes through the glycolysis result in low values of the phosphorylated form of EIIA^{Crr} while low fluxes indicate a hunger situation and the global transcription factor cAMP · Crp is activated.

Interestingly, the relationship between growth rate and degree of phosphorylation of EIIA^{Crr} could be seen in various growth situations of the wild type strain growing on single substrates like glucose, lactose, and glycerol and for growth on mixtures of substrates, and of a PtsG deletion mutant strain missing *ptsG*, a gene that is central for glucose transport.

Carbohydrate uptake by E. coli

The PTS of *E. coli* consist of two common cytoplasmatic proteins, EI (enzymeI) and HPr (histidine containing protein), as well as of an array of carbohydrate-specific EII (enzymeII) complexes. E.g. for glucose uptake, a phosphoryl group is transferred from phosphoenolpyruvate (PEP) through EI, HPr, EIIA^{Crr}, PtsG (also known as EIICB^{Glc}, that is the membranstanding transport protein) and finally to the substrate. Since all components of the PTS, depending on their phosphorylation status, can interact with various key regulator proteins the output of the PTS is represented by the degree of phosphorylation of the proteins involved in phosphoryl group transfer.

Figure 1 gives a rough sketch on the components that influence the degree of phosphorylation of protein EIIA^{Crr}: (i) Metabolic fluxes through the glycolysis. Extracellular glucose is taken up by PtsG and enters into the cell as glucose 6-phosphate. Other carbohydrates enter glycolysis at the same node (e.g. galactose and lactose) or at other nodes (e. g. glycerol at triose phosphate). The carbohydrates are further metabolized by glycolytic reactions. At node PEP, the flux is subdivided. One part is converted to

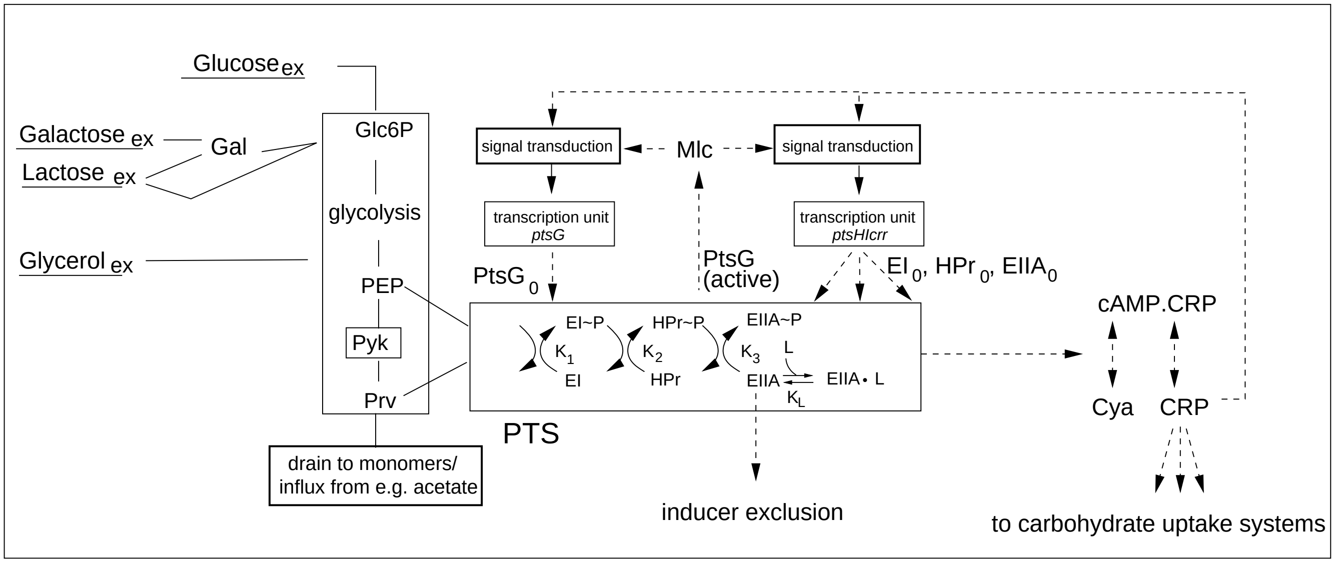


Figure 1
A rough scheme of the interactions of the PTS. The degree of phosphorylation of the PTS proteins is influenced by the flux through glycolysis and the overall concentration of the proteins. The respective genes are subject to transcriptional control by several transcriptions factors, e.g. Mlc and Crp and post-transcriptional control (not shown). The degree of phosphorylation of EIIA^{Crr} is furthermore influenced by interactions with other proteins (L) during inducer exclusion. In case of a PTS sugar, the phosphoryl group from EIIA^{Crr} is transferred to the transported sugar. E.g. glucose appears as glucose 6-phosphate inside the cell.

pyruvate by pyruvate kinase while the remainder part is converted to pyruvate by the PTS. Other fluxes from or to PEP or pyruvate are marginal and hence are not considered in this study. Fluxes from e.g. acetate uptake enter gluconeogenesis via pyruvate or from TCA. (ii) Overall concentration of the PTS proteins. The expression of the *pts* genes (*ptsHIcrr*, *pts*G) is subject to control by various regulators, with Mlc and Crp being the most important ones. Mlc is a repressor that is active if no glucose is present in the medium. A possible mechanism of the interaction between Mlc and EIICB^{Crr} is described in [1]. Crp is a global regulator that is involved in the regulation of a number of genes; it is activated by cAMP. cAMP is synthesized from ATP by adenylate cyclase (Cya). Both proteins, Crp and Cya, are subject to control by the cAMP · Crp complex itself. Recent investigations indicate that the *pts*G transcript is subject to post-transcriptional control by a small RNA (sRNA) regulator SgrS which is induced at different stress conditions, e.g. glucose-phosphate stress. This stress occurs when cells accumulate glucose 6-phosphate or the glucose analog a-methyl-glucoside 6-phosphate and leads to the degradation of PtsG mRNA [5-7]. (iii) Another parameter that determines the degree of phosphorylation of protein EIIA^{Crr} is the overall equilibrium constant K_{pts} that links the PEP/pyruvate ratio to the degree of phosphorylation. Figure 1 considers a general case where the phosphoryl group is transferred from PEP to EIIA^{Crr}. Furthermore, EIIA^{Crr} is considered to exist in a free form and in a form bound to a protein L involved in carbohydrate transport or metabolism (lactose permease, glycerol kinase). Then, the equilibrium constant K_{pts} can be determined as:

$$K_{pts} = \frac{K_1 \cdot K_2 \cdot K_3 \cdot (L + K_L)}{K_L} \qquad (1)$$

with K_1, K_2, K_3, K_L being the respective equilibrium constants from the single reactions shown in Figure 1. If EIIA^{Crr} is bound to lactose permease or glycerol kinase, it acts as an inhibitor that prevents uptake and/or metabolism of the substrate, an effect that is called inducer exclusion.

The intention of this contribution is to develop a model with a small number of state variables and parameters to work out the basic principles for the understanding of the sensor function. Nearly all parameters could be determined from experiments (for material and methods, [see Additional file 1]). The core of the model describes the mapping of the specific growth characteristics represented by the carbohydrate uptake rates to the degree of phosphorylation of the PTS component EIIA^{Crr}. The kinetic properties of the sensor which at the same time is a transport system are characterized and the output of the sensor is mapped to the rate of synthesis of genes that are under control of transcription factor cAMP · Crp. In this way, a closed loop is established that precisely adjusts the respective transport protein to maintain the incoming flux. The results are used to predict the transient behavior during glucose/glucose 6-phosphate diauxic growth and glucose/lactose diauxic growth. Finally, we also show that the approach can be generalized for other main growth substrates like acetate. In the end, a comparison with a corresponding detailed model on catabolite repression [3] is performed.

Results and discussion

Sensor characteristics

First, the steady-state properties of the core system, comprising glycolytic and PTS reactions, are analyzed. Predictions with the model are performed and compared with experimental data. Based on the molecular details, two situations are considered (Figure 2). *Case A* considers growth on glycolytic substrates, that is, carbohydrates that feed into glycolysis. This includes growth on PTS and on non-PTS substrates. E.g. glucose enters the cell by a PTS as glucose 6-phosphate, while lactose is a non-PTS substrate. Intracellular lactose is split into glucose and galactose by LacZ. The resulting intracellular glucose is phosphorylated by PtsG and/or by glukokinase. Galactose, too, is further metabolized and both enter via glucose 6-phosphate into glycolysis. In case of lactose, EIIA^{Crr}mediates inducer exclusion by binding to lactose permease. This alters the overall equilibrium constant as described above.

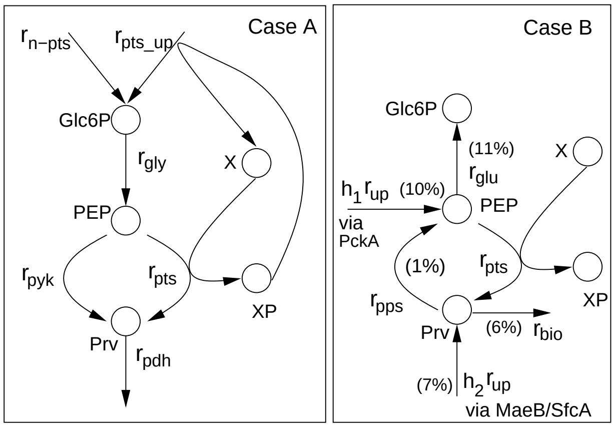


Figure 2
Reactions schemes that describe the fluxes through glycolysis and the PEP/pyruvate node. Left: *Case A*. Growth on glycolytic PTS substrate and non-PTS substrates. State variable *X* represents all PTS components. Right: *Case B*. Growth on gluconeogenetic substrates. Values in parenthesis indicate the flux distribution during growth on acetate [10] in % of the acetate uptake. Main routes to PEP and pyruvate are via PckA (PEP carboxykinase) and MaeB/SfcA (malate dehydrogenase).

The scheme simplifies the biological knowledge on metabolism and gene expression by lumping together reactions and components. In case A, carbohydrate uptake is represented by reactions r_{pts_up} for a PTS carbohydrate and $r_{n\text{-}pts}$ for a non-PTS carbohydrate. Glycolysis is simply represented by metabolite *Glc6P*. The flux at node *PEP* is subdivided into r_{pyk} for pyruvate kinase and r_{pts}. The drain from pyruvate (*Prv*) to other parts of the central metabolism is represented by r_{pdh}. Since other fluxes from or to PEP and pyruvate are rather marginal they are not considered in the model. Proteins EI, HPr, and EIIA^{Crr} of the PTS are represented by only one component *X* that exists either in the unphosphorylated form *X* or in the phosphorylated form *XP*. In case of a PTS carbohydrate, *XP* is used for transport via r_{pts_up}.

Case B considers gluconeogenetic substrates which feed into TCA or into other central metabolites below the PEP/pyruvate branch. Here, PEP and pyruvate are produced by a number of different reactions, e.g. from the TCA or via Acetyl CoA. Among these, PEP synthase (Pps) is active converting pyruvate directly to PEP. For substrates that enter TCA, two pathways are known that connect TCA and glycolysis: PckA (PEP carboxykinase) connects oxaloacetate and PEP, MaeB/SfcA (malate dehydrogenase) connect malate and pyruvate. These fluxes are represented by $h_1 r_{up}$ and $h_2 r_{up}$, respectively, with h_1 and h_2 are numbers between zero and one, representing a fraction of the uptake rate r_{up}. In a number of subsequent gluconeogenetic reaction steps (r_{glu}), PEP is then converted to glucose 6-phosphate.

Based on the knowledge presented so far, a simplified model structure is suggested that is able to simulate the different cases proposed above.

Glycolytic substrates

As was shown in a previous study [8], the metabolic part of the considered network reaches the steady-state very fast. Therefore, the steady-state assumption will be used as a starting point for model analysis. For *G6P*, *PEP*, *Prv* and the protein that represents the PTS, *XP*, the following equations that describe the dynamics are obtained from the scheme:

$$\dot{G6P} = r_{n-pts} + r_{pts\_up} - r_{gly} \tag{2}$$

$$\dot{PEP} = 2r_{gly} - r_{pyk} - r_{pts} \tag{3}$$

$$\dot{Prv} = r_{pts} + r_{pyk} - r_{pdh} \tag{4}$$

$$\dot{XP} = r_{pts} - r_{pts\_up} \tag{5}$$

where $r_{n\text{-}pts}$ and r_{pts_up} are the systems inputs and are related by the yield coefficients to the specific growth rate. *XP* is the system output. The following conditions will hold for the defined rates in steady-state:

$$r_{pts} = r_{pts\_up} \tag{6}$$

$$r_{gly} = r_{n\text{-}pts} + r_{pts\_up} \tag{7}$$

$$r_{pdh} = 2\,(r_{n\text{-}pts} + r_{pts\_up}) \tag{8}$$

$$r_{pyk} = 2\,r_{n\text{-}pts} + r_{pts\_up} \tag{9}$$

The kinetics for the rate laws are kept as simple as possible to describe the experimental data. The rate laws are assumed as follows:

$$r_{gly} = k_{gly}\,G6P \tag{10}$$

$$r_{pdh} = k_{pdh}\,Prv \tag{11}$$

$$r_{pts} = k_{pts}(PEP(X_0 - XP) - K_{pts}\,Prv\,XP) \tag{12}$$

$$r_{pyk} = k_{pyk}\,PEP\,f\,(PEP, \ldots), \tag{13}$$

with X_0 is the overall concentration of the PTS protein. The focus of the analysis will be on the branch point at PEP. To elucidate the correct choice of the kinetic rate law for the pyruvate kinase reaction, function *f* is introduced that represents different model variants. Function *f* depends on *PEP* but may also depend on different metabolites in the network.

The steady-state concentrations can be derived from the equations above:

$$G6P = \frac{r_{n-pts} + r_{pts\_up}}{k_{gly}} \tag{14}$$

$$\mathrm{Prv} = 2\frac{\mathrm{r_{n-pts}} + \mathrm{r_{pts\_up}}}{\mathrm{k_{pdh}}} \tag{15}$$

$$PEP = \frac{2r_{n-pts} + r_{pts\_up}}{k_{pyk} f} \tag{16}$$

$$XP = k_{pdh} \frac{X_0 - \dfrac{r_{pts\_up}}{k_{pts} PEP}}{k_{pdh} + 2K_{pts} k_{pyk} f \dfrac{r_{n-pts} + r_{pts\_up}}{2r_{n-pts} + r_{pts\_up}}} \tag{17}$$

The steady-state equation for *PEP* is given in implicit form since it depends on function *f*. In the following, growth situations on non-PTS and PTS sugars are considered separately.

Equation (17) for the non-PTS case reads

$$XP = k_{pdh} \frac{X_0}{k_{pdh} + K_{pts} k_{pyk} f} \qquad (18)$$

As can be seen cleary, the choice of *f* has a strong influence on the steady-state characteristics: Assuming $f = 1$, that is, the pyruvate kinase reaction is modeled as a first order reaction, *XP* is constant and independent from the uptake rate. This could not be observed in the experiments (see below). Assuming a Michaelis-Menten kinetics, that is, $f = \frac{1}{K + PEP}$, the steady-state concentration of *PEP* can be calculated via Equation (16):

$$PEP = \frac{2r_{n-pts}}{k_{pyk} f} = \frac{2r_{n-pts}}{k_{pyk} \frac{1}{K + PEP}} \qquad (19)$$

$$\Rightarrow \quad PEP = \frac{2Kr_{n-pts}}{k_{pyk} - 2r_{n-pts}} \qquad (20)$$

Since k_{pyk}, in this case, is the maximal reaction rate of r_{pyk}, *PEP* is an increasing monotone function in dependency on the uptake rate r_{n-pts}. Interestingly, this leads to values for *XP* that increase for increasing uptake rates. This result is again contradictory to the observed experimental results.

Equation (17) for PTS substrates reads:

$$XP = k_{pdh} \frac{X_0 - \frac{r_{pts\_up}}{k_{pts} PEP}}{k_{pdh} + 2K_{pts} k_{pyk} f} \qquad (21)$$

Differences for PTS and non-PTS substrates can be seen in the numerator that is always smaller in case of growth on PTS substrates. Since the denominator is always larger than in the case of non-PTS substrates, the curve of the PTS substrates will always be below the curves for non-PTS substrates.

To describe the available experimental data for growth on PTS and non-PTS substrates (Table 2 in [Additional file 1]), parameters were estimated by a least square approach. A reasonable fit could be obtained with

$$f = f_1(G6P) \cdot f_2(PEP) = G6P^n \cdot PEP^m. \qquad (22)$$

Since the pyruvate kinase in *E. coli* is a tetramer that needs activation from a glycolytic metabolite (in *E. coli* PykF is strongly activated by fructose 1,6-bis-phosphate, that is not included in the model, but is represented by glucose 6-phosphate instead), values for $n > 1$, $m \geq 1$ are analyzed. Equation (1) relates the overall PTS constant K_{pts} to individual reactions steps. Since measurements of proteins that influence K_{pts} are not available, K_{pts} represents a mean value for different situations considered in the experiments. For parameter identification 31 data points are considered, values $n = 2$, $m = 1$ are fixed and values for K_{pts} and X_0 are taken from literature (Table 6 in the [Additional file 1]); so, four parameters are estimated: k_{gly}, k_{pyk}, k_{pts}, and k_{pdh}.

The standard deviation $\hat{\sigma}$ of the measured data for the degree of phosphorylation of EIIA^{Crr} is estimated with the degree of freedom $df = 31$ (data points) -4 (parameters):

$$\hat{\sigma} = \sqrt{\frac{\sum (XP_i - XP_{m_i})^2}{df}} = 0.013. \qquad (23)$$

Based on the maximal value X_0 this corresponds to 13%. Figure 3 shows the results of the parameter estimation. Parameter values and confidence regions are summarized in Table 6 in [Additional file 1]. Dashed lines mark a 95% confidence band of the simulation based on the linearized system (linearized with respect to the parameters; for details [Additional file 1]).

A robustness analysis was performed as described earlier [8]. Instead of presenting individual sensitivities, a ranking of all sensitivities

$$w_i = \frac{\partial XP}{\partial p_i} \cdot \frac{p_i}{XP} \qquad (24)$$

based on the sensitivity matrix *W* with

$$W = \sum_j (w_i^T w_i) \,|_j \qquad (25)$$

with *j* is the index of the simulated data points was calculated. Together with a constraint, considering the deflection of the parameters $\underline{\Delta p}$

$$\underline{\Delta p}^T \cdot \underline{\Delta p} = 1 \qquad (26)$$

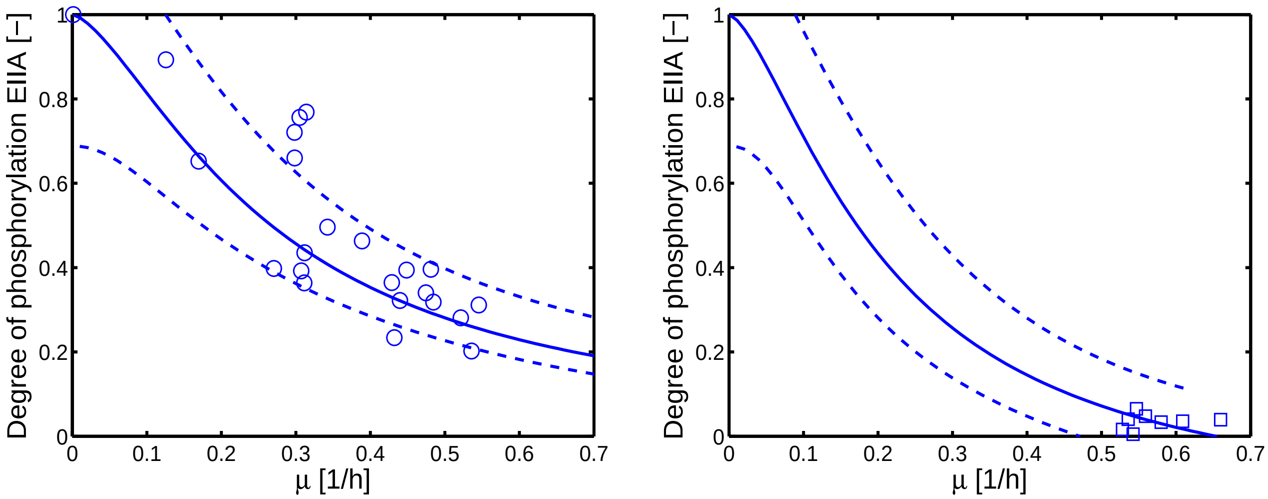


Figure 3
Course of the degree of phosphorylation of *X* in dependence on the growth rate. Since most experiments are performed with lactose and glucose, the specific growth rate can be converted with (nearly) the same yield coefficient (on a molar basis) into an uptake rate. Left: Growth on non-PTS carbohydrates (lactose and glycerol, wild type and mutant strain BKG47); Right: Growth on PTS carbohydrates (glucose, wild type). Dashed lines indicate a 95% confidence interval based on the simulated and the experimental data. The calculation is based on a linearization around the estimated parameters, therefore, it is not exptected that all the data can be found in-between the two limits.

the maximal deviation of the trajectories can be calculated by the eigenvectors and eigenvalues of matrix W [9]. The eigenvector corresponding to the maximal eigenvalue is $\underline{\Delta p}^{max}$. The parameter vector that leads to the maximal deviation is calculated then by $\underline{p}$ $(1 + \underline{\Delta p}^{max})$. Figure 4 summarizes the results. Four of the parameters are related to enzyme concentrations (X_0, k_{gly}, k_{pyk}, k_{pdh}) while the others are kinetic parameters of the PTS reaction (k_{pts}, K_{pts}) and the pyruvate kinase reaction (m, n). Interestingly, in the kinetic expression f of the pyruvate kinase parameter n describing the influence of the feed-forward control (activation of the pyruvate kinase by glucose 6-phosphate) shows maximal sensitivity in both cases. In general, the amount of enzyme has a bigger influence than the kinetic parameters. This will allow the cell to adjust the degree of phosphorylation by genetic control.

Gluconeogenetic substrates

For gluconeogentic substrates the scheme according to Figure 2, case B, is considered. The o.d.e's are:

$$\dot{PEP} = h_1 r_{up} + r_{pps} - r_{pts} - r_{glu} \tag{27}$$

$$\dot{Prv} = h_2 r_{up} + r_{pts} - r_{pps} - r_{bio} \tag{28}$$

$$\dot{XP} = r_{pts}. \tag{29}$$

Rate r_{up} is the system input. Rate r_{bio} is the flux from pyruvate to biosynthesis and r_{glu} is the rate of gluconeogenesis:

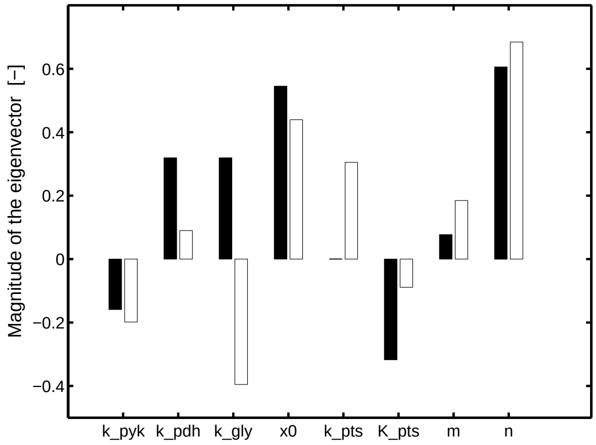


Figure 4
Results of the sensitivity analysis. Black bars indicate a non-PTS substrate while white bars indicate a PTS substrate. The size of the bars represent the level of the eigenvector of the sensitivity matrix *W* that correspond to the maximal eigenvalue. Note, that for the non-PTS case, parameter k_{pts} has zero sensitivity (r_{pts} is zero in this case).

$$r_{bio} = k_{bio} P_{rv} \tag{30}$$

$$r_{glu} = k_{glu} PEP \tag{31}$$

For the rate r_{pps} the following simple approach is used:

$$r_{pps} = k_{pps} Prv\ g(Prv, \ldots) \tag{32}$$

with function g representing the influence of pyruvate and possible effectors. Together with parameters h_1, h_2 and k_{pps} the rates are adjusted in such a way that data from a flux distribution [10] can be described.The percentage fluxes can be found in Figure 2. The steady-state equation for XP can be rewritten as:

$$XP = \frac{X_0}{1 + K_{pts}\,\dfrac{Prv}{PEP}}. \tag{33}$$

Simulation studies lead to the conclusion that function g should depend on PEP that acts as an inhibitor of PEP synthase. Otherwise, the degree of phosphorylation increases with increasing uptake rate which seems, also in this case, not to be meaningful. Indeed, literature research revealed that PEP synthase is negatively regulated by PEP [11]. Function g used is:

$$g = g_1(PEP) \cdot g_2(Prv) = \frac{1}{PEP^2} \cdot Prv, \tag{34}$$

taking into account that Pps is a dimer with two possible binding sites. A simulation study for different values of acetate uptake/growth rates are shown in Figure 5; data are taken from Table 3 [Additional file 1]. Another interesting observation where the PEP/pyruvate ratio may be involved was reported by the group of Liao [12]. They analyzed a wild type strain and a *pps* mutant strain when glucose and acetate are provided in the medium. They showed that the missing Pps protein has no influence on the general physiology but shows a significant influence on the transition time from growth on glucose to growth on acetate. In this case the degree of phosphorylation is a constant value:

$$XP = \frac{X_0}{1 + K_{pts}\,\dfrac{h_2 k_{glu}}{h_1 k_{bio}}}. \tag{35}$$

Liao and colleagues observed a drastic increase of the lag phase on acetate in the mutant strain during glucose/acetate diauxic growth. Our simple model predicts, that the degree of phosphorylation is a bit smaller than the values in the wild type strain. This confirms that Pps has nearly no influence on physiological parameters like the growth rate.

Model predictions

With the model developed so far, model predictions can be performed. Two cases are considered: the PEP/pyruvate ratio and growth on different single carbon sources.

PEP/pyruvate ratio

The PEP/pyruvate ratio could be predicted in dependency on the growth rate. Experimental data were taken from [4] and compared to the simulation results (Figure 6). As can be seen, the prediction fits to the data well.

Growth on single carbohydrates

To confirm that the model presented here is able to describe the sensor system for a number of carbohydrates, experimental data from batch experiments with different PTS and non-PTS carbohydrates were performed and compared with the model calculations [4]. As can be seen in Figure 7 the experimental results are in good agreement for a number of substrates. Except for N-acetyl-glucosamine, the measured data points fit well to the prediction. Note, that most of the PTS sugars use the phosphoryl group from HPr to transport the carbohydrate into the cell. If this is included in the calculation, the degree of phosphorylation of $EIIA^{Crr}$ d_{EIIA} depends on the fraction of phosphorylated HPr d_{HPr}:

$$d_{EIIA} = \frac{EIIA^P}{EIIA_0} = \frac{d_{HPr}}{d_{HPr} + K_3(1 - d_{HPr})} \tag{36}$$

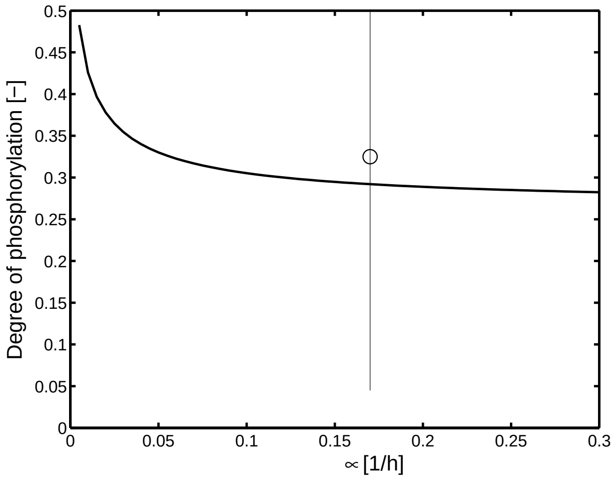


Figure 5

Degree of phosphorylation in dependence on the growth on acetate. Measurements are available for four experiments with nearly identical growth rate (errors bar is given for the four experiments). and the mean value is plotted.

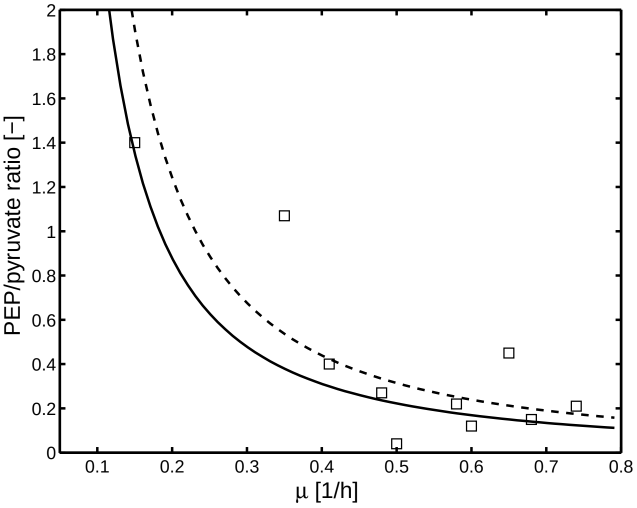


Figure 6
Measured and simulated relationship between the PEP/pyruvate ratio and the specific growth rate. Solid line is for PTS carbohydrates and the dashed line for non-PTS carbohydrates. Symbols respresent measured values (Table 3 in [Additional file 1] [4]).

Since the value for K_3, the equilibrium constant for the phosphoryl transfer HPr to EIIA^{Crr} is approximately 1 [3,13,14], values of d_{EIIA} and d_{HPr} are nearly equal. Therefore, in the model, state variable X can be used to represent HPr as well as EIIA^{Crr}.

Transcription efficiency and sensor kinetics

In order to set up a closed loop, further modules have to be characterized. First, the influence of phosphorylated EIIA^{Crr} on transcription efficiency is analyzed, afterwards the kinetics of the PTS transport system is investigated.

Transcription efficiency

Experiments to determine the influence of the degree of phosphorylation of EIIA^{Crr} on the transcription efficiency were performed with the cAMP · Crp independent promoter $scrK_P$ and the cAMP · Crp dependent promoter $scrY_P$ [4]. As can be seen in Figure 8, the activity of the cAMP · Crp independent promoter does not vary with the degree of phosphorylation of EIIA^{Crr} while the cAMP · Crp dependent promoter shows a sigmoidal behavior in the range below 0.6. From the data, a sigmoidal function g_T could be determined that maps the degree of phosphorylation of EIIA^{Crr} to the rate of protein synthesis:

$$g_T = k_b + k_{syn}\frac{XP^6}{XP^6 + K^6}. \qquad (37)$$

Unexpectedly, the Hill coefficient is high (n = 6) indicating a high sensitivity in a narrow range of the input.

Sensor kinetics

Experiments to determine the apparent K_M value of the PTS transporter for different PTS carbohydrates are reported in a number of publications [15]. In [4] experimental data determining the phosphorylation levels near

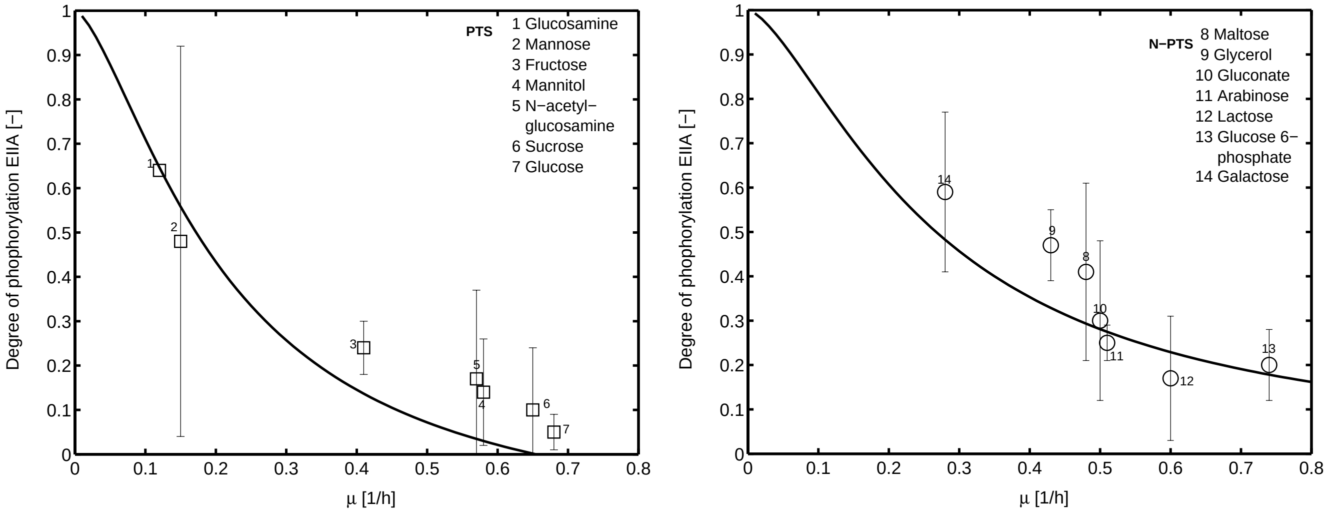


Figure 7
Experimental data showing the relationship between the specific growth rate μ and the degree of phosphorylation of EIIA^{Crr} for a number of different experiments performed with single carbohydrates. Left: PTS carbohydrates as indicated in the legend. Right: non-PTS carbohydrates as indicated in the legend. Samples are taken in the mid-log phase. Error bars indicate a 95% confidence interval.

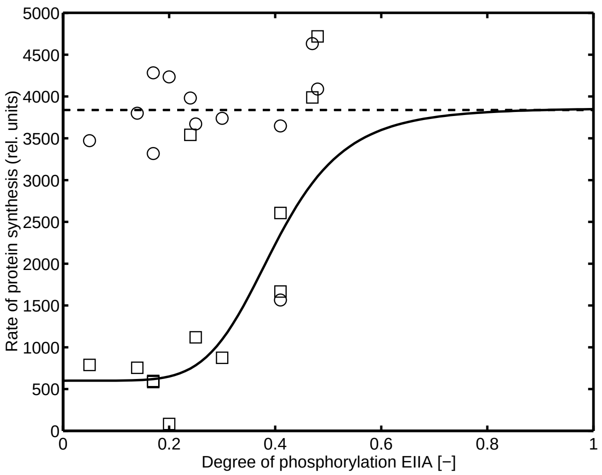


Figure 8
Relationship between the degree of phosphorylation of EIIA^{Crr}and promoter activity of *scrY* (squares, dependent from transcription factor Crp) and *scrK* (circles, independent from transcription factor Crp) [4]. The activity of the reporter protein is taken as a measure for protein synthesis. To calculate the transcription efficiency the raw data are multiplied with the specific growth rate μ. (Data from Table 3 in [Additional file 1] [4]).

these critical substrate concentrations are taken during continuous bioreactor experiments. During the starting phase of the continuous bioreactor experiments, the carbohydrate concentration drops until it becomes limiting. This decrease is much slower than it is in batch experiments, allowing for a better resolution of data in the low carbohydrate concentration ranges. Experiments were performed with the PTS substrates glucose and mannitol, having similar K_M values as determined in transport assays. To determine the kinetic parameters of the PTS, a two-substrate kinetics of the form

$$r_{pts\_up} = k_{pts\cdot up} E_{Glc} \frac{Glc\ XP}{(Glc + K_{glc})(XP + K_{EIIAP})} \tag{38}$$

with enzyme concentration E_{Glc}, turnover number $k_{pts\cdot up}$ and binding constants K_i is used. The parameters are determined from the dynamical experiments and are compared with the experimental data (for experimental data, see Tables in [Additional file 1]; for parameter values, see Table 6 in [Additional file 1]). Figure 9 shows the relationship between the measured residual carbohydrate concentrations during the bioreactor experiments and the measured degree of phosphorylation of protein EIIA^{Crr} together with a simulation result.

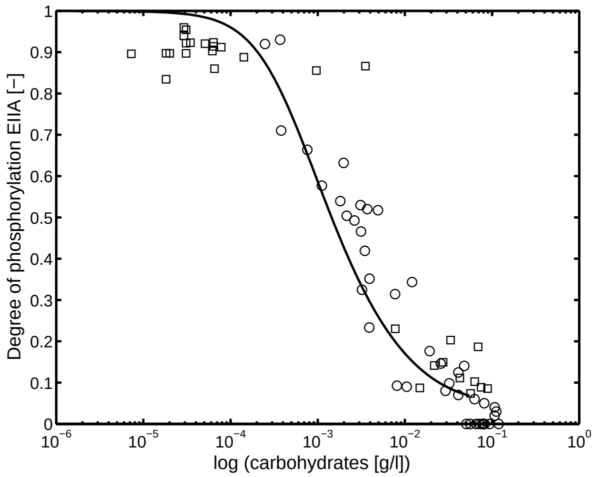


Figure 9
Measured and simulated relationship between residual carbohydrate concentrations during a continuous bioreactor experiments and the measured degree of phosphorylation of protein EIIA^{Crr}. Values are given for glucose (circles) and mannitol (squares). See Table 4 and 5 in [Additional file 1].

Closed loop dynamics and application to diauxic growth

Finally, a model with a closed loop, comprising the core model and individual uptake systems is set up. The model is applied to a complex growth situation, namely growth with a mixture of two substrates. Simulation studies for growth on mixtures of glucose/glucose 6-phosphate and of glucose/lactose are described.

Growth on glucose/glucose 6-phosphate

Glucose 6-phosphate represents an interesting growth substrate. This sugar-phosphate is taken up into the cell via the inorganic phosphate antiporter, UhpT [16] and can afterwards enter into glycolysis without further modification. A mixture of glucose and glucose 6-phosphate is a very interesting case because expression of both proteins depends on the cAMP·CRP complex. UhpT has been shown to influence cAMP levels in the cell [17]. It was concluded that neither glucose 6-phosphate nor another metabolite of glycolysis was directly involved in this effect but rather the flux through UhpT itself [17]. These results are confirmed by additional studies analyzing the effect of glucose 6-phosphate uptake on the degree of EIIA^{Crr} phosphorylation and the amount of cAMP [18]. In addition, it was shown that high intracellular Glc6P levels lead to the degradation of the *ptsG* mRNA [6,7] via the small regulatory RNA, SgrS [5] and hence to reduced concentrations of PtsG.

Carbohydrate transporters are inducible, that is, the enzymes are synthesized only if the respective substrate is present in the medium. To take this into account the rate of synthesis depends on Equation (37) and a second term g_I^{glc} and g_I^{g6p}, respectively, that describes induction. Although, it is known that high levels of glucose 6-phosphate influence glucose uptake, in the model, no interaction of glucose 6-phosphate as inhibitor of the transporter is included, since quantitative data are hardly available (see discussion on this topic in [17]). However, as can be seen in Figure 10 a strong inhibition of glucose uptake and concomitant, a decrease of the amount of the PtsG transporter is observed during glucose 6-phosphate uptake. To match these unexpected experimental data, an influence of the glucose 6-phosphate uptake system (E_{G6p}) on the rate of synthesis r_{syn}^{glc} of the glucose transporter is formulated as a "black box" model, function g_B (Equation (43)). This is done to account for possible effects on *ptsG* mRNA stability. The model introduced so far is complemented with equations for the substrates *Glc6P*, *Glc*, biomass *B*, and kinetics for the glucose 6-phosphate uptake. The additional equations are

:

$$\dot{B} = \mu\ B \tag{39}$$

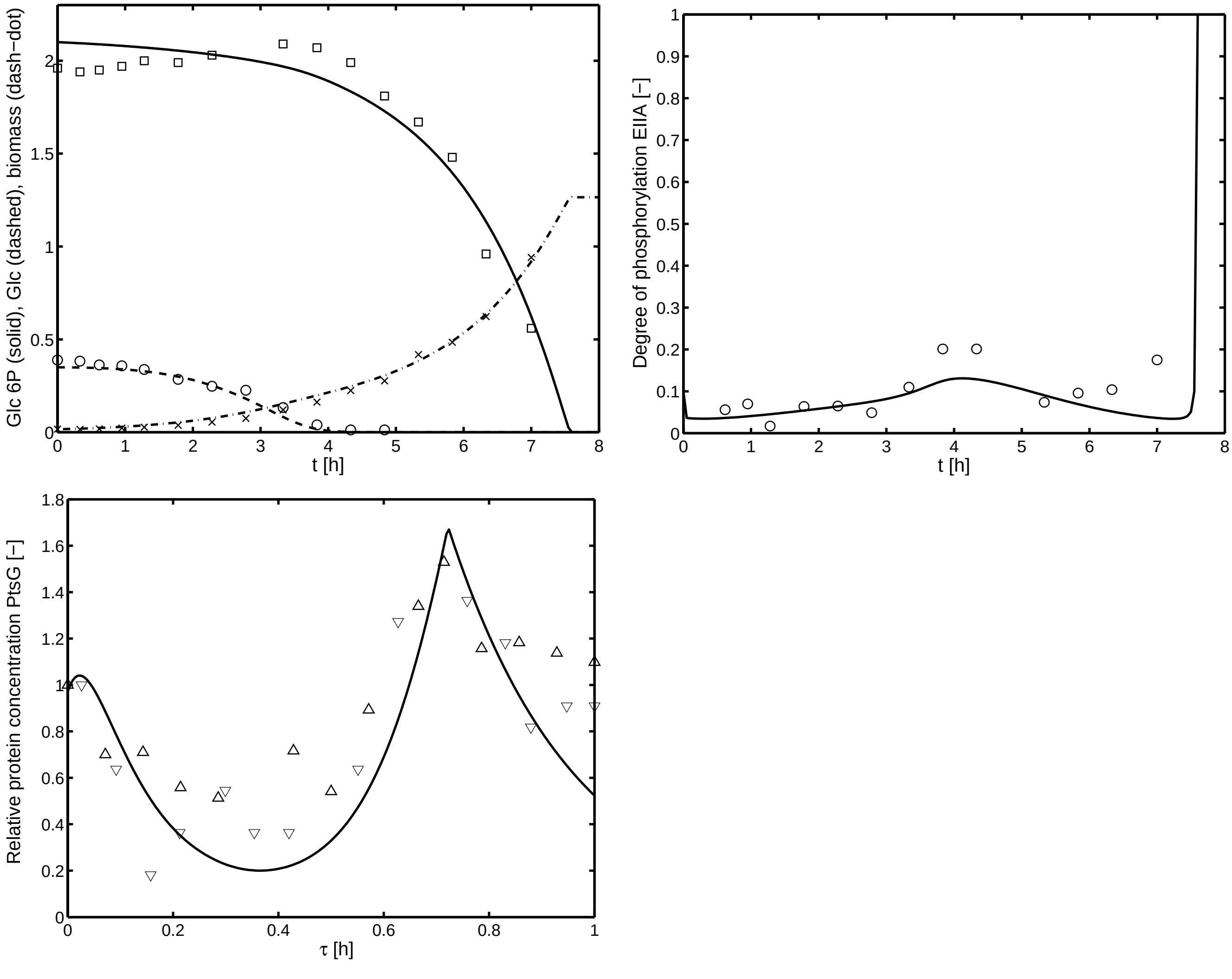


Figure 10

Glucose 6phosphate/glucose diauxic growth. Top: Time course of glucose 6-phosphate, glucose and biomass. Middle: Time course of the degree of phosphorylation of EIIA^{Crr}. Bottom: Time course of the activity of the glucose transporter monitored by a reporter gene fusion (strain LZ110). Shown are two different experiments (symbols ∇ for experiment 1 and Δ for experiment 2); here, the time was scaled to the maximal time of the experiment (7 h and 7.6 h).

$$\dot{Glc6P} = -r_{n-pts}B = -k_{g6p}E_{g6p}\frac{Glc6P}{Glc6P + K_{g6p}}B \quad (40)$$

$$\dot{Glc} = -r_{pts\_up}B \quad (41)$$

$$\begin{aligned}\dot{E}_{G6p} &= r_{syn}^{g6p} - (\mu + k_d)E_{g6p} = k_1 g_I^{g6p} g_T - (\mu + k_d)E_{G6p}\\ &= k_1\frac{r_{n-pts}}{r_{n-pts} + K_1}g_T - \mu\, E_{G6p}\end{aligned} \quad (42)$$

$$\begin{aligned}\dot{E}_{Glc} &= r_{syn}^{glc} - (\mu + k_d)E_{Glc} = k_2 g_I^{glc} g_B\, g_T - (\mu + k_d)E_{Glc}\\ &= k_2\frac{r_{pts}}{r_{pts} + K_2}\frac{K_I}{K_I + E_{G6p}}g_T - (\mu + k_d)E_{Glc},\end{aligned} \quad (43)$$

with the specific growth rate μ that is calculated with the yield coefficients Y_{g6p} and Y_{glc} in dependence on the substrate uptake:

$$\mu = Y_{g6p}\, r_{n\text{-}pts} + Y_{glc}\, r_{pts\_up}. \quad (44)$$

Parameters k_1 and k_2 are scaling factors and g_T is taken from Equation (37).

In the simulation (Figure 10), only the parameters for the glucose 6-phosphate uptake and the inhibition of the glucose transporter PtsG by the glucose 6-phosphate transporter UhpT are fitted while all other parameters are kept as described in the previous sections. Therefore, the time course of the degree of phosphorylation of EIIA^{Crr} is a prediction based on previous results. The time course of the substrates in the medium hints to an inhibition of glucose uptake during glucose 6-phosphate uptake. After consumption of glucose 6-phosphate, the growth rate slows down which results in a small increase of the degree of phosphorylation. During subsequent growth on glucose, the degree of phosphorylation of EIIA^{Crr} is again very low. For the experiment shown in Figure 10 the course of the glucose transporter was not measured. Therefore, the right plot of Figure 10 shows data from experiments with slightly different initial conditions. To compare the results, the time of the simulation experiment and the time of the wet experiment are scaled. The time course of the glucose transporter indicates that indeed the rate of gene expression is under control and is inhibited during growth on glucose 6-phosphate.

As described above, the cause for the down-regulation of PtsG is not clear. To check the intracellular levels of glycolytic metabolites, a simulation is performed that compare the experiment shown in Figure 10 with a model variant where no interaction between the two transporters is assumed ($K_I >> E_{G6p}$). As can be seen in Figure 11 the time course of glucose 6-phosphate and PEP are nearly equal in both experiments, indicating that these metabolites are hardly involved in the *pts*G mRNA degradation.

Growth on glucose/lactose

Finally, we simulated a diauxic growth experiment with glucose and lactose already introduced in [3] with the reduced model introduced here. In the reduced model, gene expression is modeled with the characteristic curve for the relationship of the degree of phosphorylation of EIIA^{Crr} on cAMP · Crp dependent promoters. The equation for the lactose in the medium, lactose transporter E_{Lac} and for the transporter kinetics read:

$$\dot{Lac} = -r_{lac}\,B = -k_{lac}\,E_{lac}\frac{Lac}{K_{lac} + Lac(1 + \frac{X0 - XP}{K_{IEIIA}\,X0})}B \quad (45)$$

$$\begin{aligned}\dot{E}_{Lac} &= r_{syn}^{lac} - (\mu + k_d)\,E_{lac} = k_3\, g_I^{lac} g_T - (\mu + k_d)E_{Lac}\\ &= k_3\frac{r_{lac}}{r_{lac} + K_3}g_T - (\mu + k_d)\,E_{Lac}\end{aligned} \quad (46)$$

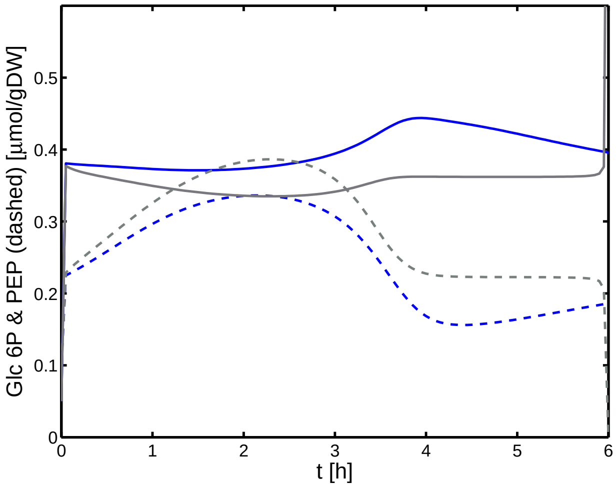


Figure 11
Simulation results for glucose 6-phosphate (solid lines) and PEP (dashed lines). The simulation compares two cases: the glucose transporter PtsG is under control of the glucose 6-phosphate transporter UhpT (black lines, corresponding to the simulation in Figure 9) or not (grey lines). Both simulation results in comparable concentrations of the two metabolites.

As can be seen in the simulation in Figure 12, the time course of LacZ (right plot) can describe the experimental data, however, with less accuracy than the detailed model presented in [3].

Comparison with a detailed model for catabolite repression

A very detailed model for catabolite repression was already introduced to describe a number of experiments under different conditions and with different strains [3]. However, especially for PTS uptake, only high growth rates were considered. Figure 13 compares the characteristic curve for the detailed model and the reduced model, introduced here and it can be seen, that, indeed, the detailed model fails to describe the experimental data for a broad range of the growth rate. The detailed model was also used to calculate a steady-state relationship for the transcription efficiency. As can be seen in the plot, again, for low growth rates, the detailed model fails to describe the experimental data.

Conclusion

The paper presents evidence that a sensitive metabolic regulation at the PEP/pyruvate node results in a relationship between the phosphorylation state of EIIA^{Crr}, an element of the sensory system PTS, and the specific growth rate μ. Under a variety of experimental conditions with a wild type strain and a mutant strain this relationship could be verified over a broad range of the growth rates, revealing the signaling and kinetic characteristics of the sensor. For

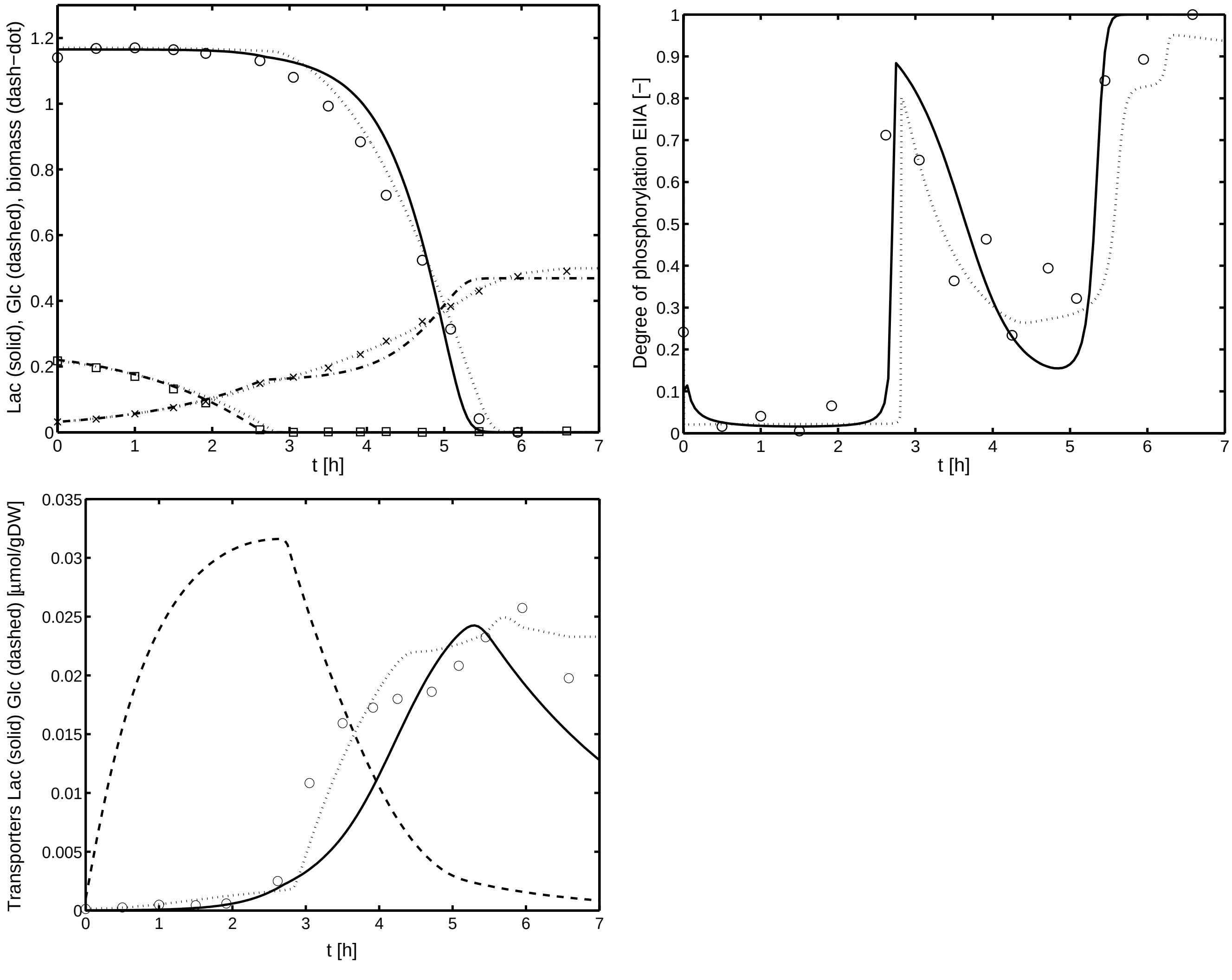


Figure 12

Glucose/lactose diauxic growth. Top: Time course of lactose, glucose and biomass. Middle: Time course of the degree of phosphorylation of EIIA^{Crr}. Bottom: Time course of the activity of LacZ Dotted line are simulations with the original model [3] while solid lines are simulations with the new model.

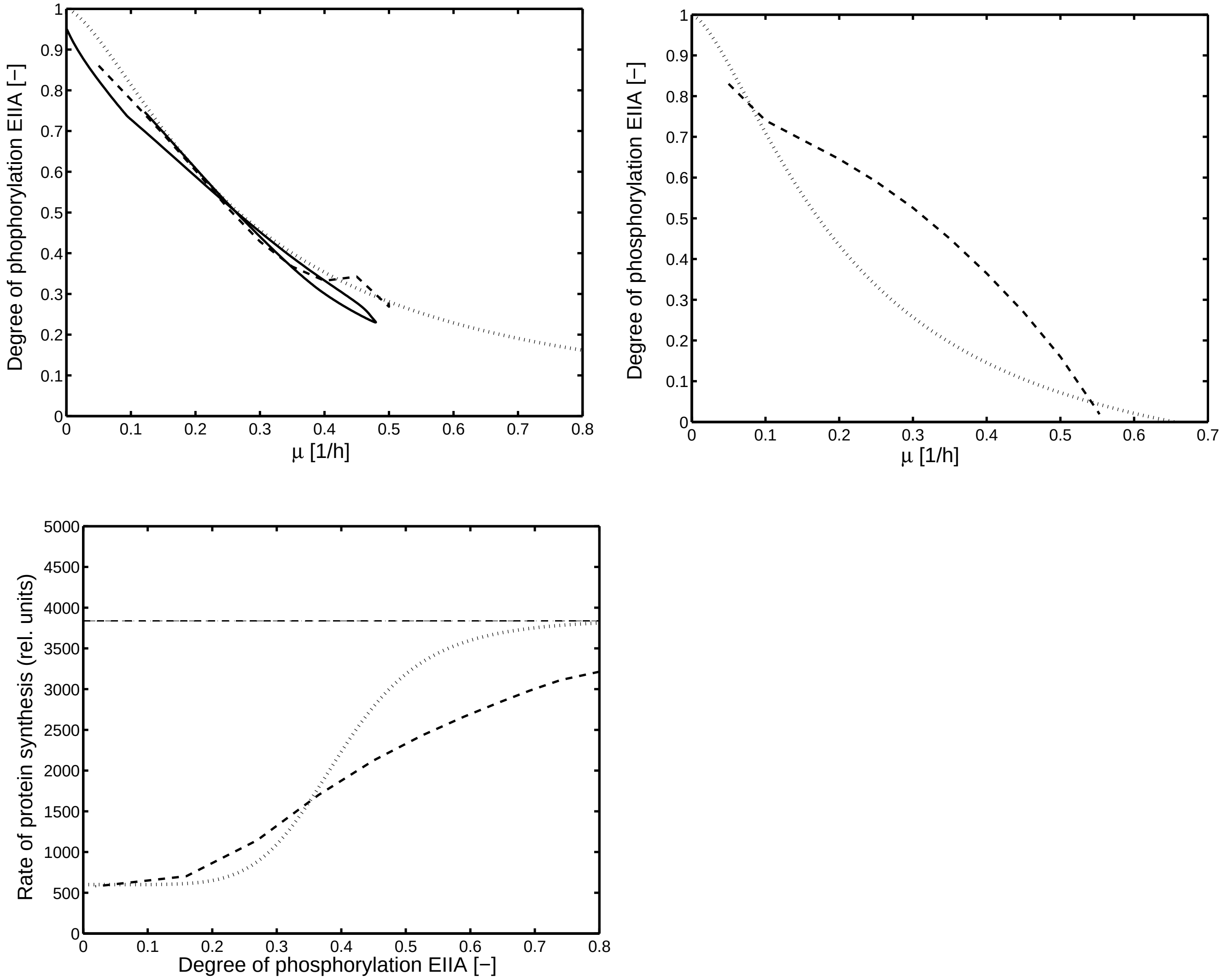


Figure 13

Comparison of simulation results of the proposed model with a more detailed model [3]. Top: Characteristic curve for non-PTS substrates. Solid line: Simulation of a batch experiment. (Figure 5 in the supplement in [3]). Values for the second growth phase, that is, growth on lactose are plotted. Dashed line: Simulation of a continuous fermentation with different values of the dilution rate with the model in [3]. Dotted line: Results with the proposed model. Middle: Characteristic curve for PTS substrates. Dashed line: Simulation of a continuous fermentation with different values of the dilution rate with the model in [3]. Dotted line: Results with the proposed model. Bottom: Characteristic curve to describe the relationship between the degree of phosphorylation of EIIA^{Crr} and the rate of protein synthesis. Dashed line: Simulation of a continuous fermentation with different values of the dilution rate with the model in [3]. Dotted line: Results with the proposed model.

the analysis of the system, a mathematical model with a small number of state variables (Table 1) was set up and based on an initial set of experimental data, model predictions were performed.

Several kinetic properties determine the degree of phosphorylation of the PTS protein EIIA^{Crr}. According to this study, the choice of the rate law for the pyruvate kinase is the most important one. While all other kinetic rate laws can be described with simple mass action rate laws, the pyruvate kinase has to be described with a power law kinetics. However, this choice is only true for a certain set of experimental conditions; considering only growth of the wild type on glucose, a simple rate law, as suggested by [19], is capable to describe experimental data. Based on a systems biology approach that considers different operational modi of the system and a directed stimulation of the system with respect to these modi, the present study shows that the simple rate law is not longer able to describe all experimental data. The core model comprises

Table 1: Summary of the state variables of the model

State variable	Comment
B	biomass
Glc6P	extracellular glucose 6-phosphate
Glc	extracellular glucose
Lac	extracellular lactose
G6P	glucose 6-phosphate; represents the metabolites in the upper part of the glycolysis
PEP	phosphoenolpyruvate
Prv	pyruvate
XP	represents the phosphorylated form of the PTS proteins (EI, HPr, EIIA^{Crr})
E_{G6P}	represents uptake system for glucose 6-phosphate (UhpT)
E_{Glc}	represents uptake system for glucose (PtsG)
E_{Lac}	represents uptake system for lactose (LacY)

four reactions for glycolysis, pyruvate kinase, PTS, and drain to monomers. Parameters were determined by fitting experimental data from a wild type strain and a PtsG mutant strain. To deconstruct the results, a robustness analysis was performed that ranks the parameters according to the influence on the degree of phosphorylation of EIIA^{Crr} in dependence on the growth rate. As expected, the biggest influence for both operational modi shows parameter n, that represents the influence of the feed-forward control of glucose 6-phosphate on pyruvate kinase. Furthermore, the overall concentration X_0 of enzyme EIIA^{Crr} has a big influence while the concentration of the other enzymes represented by k_{gly}, k_{pyk}, and k_{pdh} is moderate and comparable to the influence of the remaining kinetic parameters k_{pts} and K_{pts}.

The feed-forward loop is a special motive (a regulatory pattern that is more present than others) described in detail for genetic systems [20]. Here, we found that this motive is essential for the transformation of a high incoming flux (high growth rate) into a low PEP/pyruvate ratio. To verify this, the internal metabolites PEP and pyruvate are measured. Since the errors for the procedure of the PEP and pyruvate measurement are rather high [4], the data shown in Figure 6 should be interpreted rather as a trend and not as quantitative measurements. Although measurements for small growth rates are not available, the PEP/pyruvate ratio could be predicted very well for growth rates in the range between 0.15 1/h and 0.7 1/h.

In engineering science, sensor or measurement systems are designed in such a way that they don't influence the system that is measured. This is called "free of retroactivity". Considering the PTS operational mode in comparison to the non-PTS mode the difference of the curves is due to the transport activity of the PTS. Hence, the sensor PTS is not free of retroactivity; however, for small growth rates, indicating a severe stress situation, the difference between the PTS mode and the non-PTS mode is negligible.

As representative of gluconeogenetic substrates, growth on acetate was considered. The fluxes are adjusted in such a way that a flux distribution published previously, is matched. Measurements of the degree of phosphorylation of EIIA^{Crr} are in good agreement with the predicted values. The results also confirm that the Pps enzyme has only marginal influence on growth on acetate as described by [12]. However, the observation that a Pps mutant strain that grows simultaneously on glucose and acetate shows an extended lag phase could not be explained with model set up in this study.

The transcription efficiency according to Equation (37) revealed that the Hill-coefficient $n = 6$ is rather high. This might be due to several reasons: although the signal transduction pathway starting from EI and ending with Crp is rather short, several components and processes are involved. First cAMP is generated by the adenylate cyclase (Cya); second cAMP interacts with Crp to activate the transcription factor. Furthermore, transcription of Cya is also under control of Crp leading to a feedback loop. Since the kinetics of the individual steps are not yet characterized, the rather high Hill-coefficient can be seen as an overall measure of the sensitivity of the system. The kinetics determined are used to simulate the two dynamical experiments and a good agreement between the simulation data and the experimental data could be observed. This shows that not only the steady-state behavior can be reproduced well but also the dynamics of the sensor/actuator system.

The simplified scheme is used to analyze the growth behavior and the dynamics of *Escherichia coli* during growth on glucose/glucose 6-phosphate and on glucose/lactose. The model has to be extended to describe the kinetics of the transporters and the kinetics of gene expression for the relevant transporters. Since experimental data that characterize the K_{Glc} value for glucose can be found in the literature, the respective value K_{EIIA} for the degree of phosphorylation was determined by a simulation experiment with a random bi-bi double substrate kinetics, Equation (38), and experimental data from [4]. Parameters k_{max} and K_{EIIA} are determined by a least-square fit.

Growth on glucose/glucose 6-phosphate reveals the interesting observation that the concentration of the glucose transporter decreased during growth on glucose 6-phosphate. To match the experimental data, an inhibitory effect of the glucose 6-phosphate transporter UhpT on the glucose transporter PtsG was assumed and described with a simple kinetics. Previous studies revealed that the *ptsG* mRNA is under control by SgrS, a small RNA. It was shown that high levels of intracellular glucose 6-phosphate or fructose 6-phosphate lead to *ptsG* mRNA degradation [6,7]. Here, the model can be used to calculate the intracellular levels of glucose 6-phosphate and PEP in model variants with and without control of PtsG. As shown in Figure 11, no difference could be detected, indicating that the interaction between the two transporters is based on the activity of the glucose 6-phosphate transporter as suggested in [17]. Note, that to describe the time course of PtsG in Figure 10, three factors, namely the inhibition of PtsG by UhpT, induction of *ptsG* and global control of PtsG synthesis by Crp were taken into account and have to be adjusted very precisely.

A comparison with a detailed model for catabolite repression justifies the set up of the new model. Altough validated under different experimental conditons, the detailed model fails to describe growth on PTS carbohydrates on a broad range of the growth rate.

The approach is based on the development of a model with a minimal number of parameters that are necessary to describe the observations. Although some of the parameters have no defined mechanistic interpretation such models will facilitate the procedure of parameter analysis and estimation. The model is capable to simulate a broad range of experimental conditions and is suited for further studies on control systems on *E. coli* since it can be easily extended to describe other regulatory systems.

Methods

For simulation of the algebraic system, solving the o.d.e. system, and parameter estimation MATLAB was used. Files to simulate the system with MATLAB and the experimental data can be found on a website [21]. For the experimental data, see the [Additional file 1] and a further manuscript from our group [4].

Competing interests

The author(s) declares that there are no competing interests.

Authors' contributions

AK performs modeling, model analysis and parameter estimation. KB performs the experiments. EDG conceived of the study, and participated in its design and coordination. All authors read and approved the manuscript.

Additional material

Additional file 1

Supplementary Information. The Supplementary Information includes the description of the experimental data and the values for the kinetic parameters of the mathematical model.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-1-42-S1.pdf]

Acknowledgements

AK and KB are funded by the FORSYS initiative from the German Federal Ministry of Education and Research (BMBF).

References

1. Plumbridge J: **Expression of *ptsG*, the gene for the major glucose PTS transporter in *Escherichia coli*, is repressed by Mlc and induced by growth on glucose.** *Mol Microbiol* 1998, **29(4):**1053-1063.
2. Ozbudak E, Thattai M, Lim H, Shraiman B, van Oudenaarden A: **Multistability in the lactose utilization network of *Escherichia coli*.** *Nature* 2004, **427:**737-740.
3. Bettenbrock K, Fischer S, Kremling A, Jahreis K, Sauter T, Gilles ED: **A quantitative approach to catabolite repression in *Escherichia coli*.** *J Biol Chem* 2006, **281:**2578-2584.
4. Bettenbrock K, Sauter T, Jahreis K, Kremling A, Lengeler JW, Gilles ED: **Analysis of the Correlation between Growth Rate, EIIA^{Crr} (EIIA^{Glc}) Phosphorylation Levels and Intracellular cAMP Levels in *Escherichia coli* K-12.** *J. Bacteriol* 2007, **189:**6891-6900. [Accepted].
5. Vanderpool CK: **Physiological consequences of small RNA-mediated regulation of glucose-phosphate stress.** *Current Opinion in Microbiology* 2007, **10:**146-151.
6. Morita T, El-Kazzar W, Tanaka Y, Inada T, Aiba H: **Accumulation of glucose 6-phosphate or fructose 6-phosphate is responsible for destabilization of glucose transporter mRNA in *Escherichia coli*.** *J Biol Chem* 2003, **278(18):**15608-15614.
7. Morita T, Kawamoto H, Mizota T, Inada T, Aiba H: **Enolase in the RNA degradosome plays a crucial role in the rapid decay of glucose transporter mRNA in the response to phosphosugar stress in *Escherichia coli* .** *Mol Microbiol* 2004, **54:**1063-1075.
8. Kremling A, Fischer S, Sauter T, Bettenbrock K, Gilles ED: **Time hierarchies in the *Escherichia coli* carbohydrate uptake and metabolism.** *BioSystems* 2004, **73:**57-71.
9. Hearne JW: **Sensitivity analysis of parameter combinations.** *Appl Math Modelling* 1985, **9:**106-108.
10. Zhao J, Shimizu K: **Metabolic flux analysis of *Escherichia coli* K12 grown on ^{13}C-labeled acetate and glucose using GC-MS and powerful flux calculation method.** *J Biotech* 2003, **101:**101-117.
11. Chulavatnatol M, Atkinson DE: **Phosphoenolpyruvate Synthase from *Escherichia coli*.** *J Biol Chem* 1973, **248:**2712-2715.
12. Kao KC, Tran LM, Liao JC: **A global regulatory role of gluconeogenic genes in *Escherichia coli* revealed by transcriptome network analysis.** *J Biol Chem* 2005, **280(43):**36079-36087.
13. Sauter T, Gilles E: **Modeling and experimental validation of the signal transduction via the *Escherichia coli* sucrose phosphotransferase system.** *J Biotechnol* 2004, **110(2):**181-199.
14. Rohwer JM, Meadow ND, Roseman S, Westerhoff HV, Postma PW: **Understanding glucose tranport by the bacterial phosphoenolpyruvate:glycose phosphotransferase system on the basis of kinetic measurements in vitro.** *J Biol Chem* 2000, **275:**34909-34921.
15. Postma PW, Lengeler JW, Jacobson GR: **Phosphoenolpyruvate: carbohydrate phosphotransferase systems of bacteria.** *Microbiol Rev* 1993, **57(3):**543-594.
16. Maloney PC, Ambudkar SV, Anatharam V, Sonna LA, Varadhachary A: **Anion-exchange mechanisms in bacteria.** *Microbiol Rev* 1990, **54:**1-17.

17. Dumay V, Danchin A, Crasnier M: **Regulation of *Escherichia coli* adenylate cyclase activity during hexose phosphate transport.** *Microbiol* 1996, **142:**575-583.
18. Hogema BM, Arents JC, Bader R, Eijkemanns K, Yoshida H, Takahashi H, Aiba H, Postma PW: **Inducer exclusion in *Escherichia coli* by non-PTS substrates: the role of the PEP to pyruvate ratio in determining the phosphorylation state of enzyme IIA^{Glc}.** *Mol Microbiol* 1998, **30:**487-498.
19. Kremling A, Bettenbrock K, Laube B, Jahreis K, Lengeler J, Gilles E: **The organization of metabolic reaction networks: III. Application for diauxic growth on glucose and lactose.** *Metab Eng* 2001, **3(4):**362-379.
20. Mangan S, Zaslaver A, Alon U: **The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks.** *J Mol Biol* 2003, **334:**197-204.
21. **E. coli global control: model and experimental data** [http://www.mpi-magdeburg.mpg.de/people/kre/global_control/]

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."
Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

Bioinformatics Advance Access published January 10, 2008

A feed-forward loop guarantees robust behavior in *Escherichia coli* carbohydrate uptake

A. Kremling[*], K. Bettenbrock, and E.D. Gilles

Systems Biology Group, Max-Planck-Institute for Dynamics of Complex Technical Systems,
Magdeburg, Germany

Associate Editor: Prof. Thomas Lengauer

ABSTRACT

Motivation: In *E. coli* the phosphoenolpyruvate: carbohydrate phosphotransferase system acts like a sensory element which is able to measure the flux through glycolysis. Since the output of the sensor, the phosphorylated form of protein EIIA, is connected to the activity of the global transcription factor Crp, the kinetic and structural properties of the system are important for the understanding of the overall cellular behavior.

Results: A family of mathematical models is presented, varying with respect to their degree of complexity (number of reactions that are taken into account, number of parameters) that show a structurally and quantitatively robust behavior. The models describe a set of experimental data that relates the output of the sensor to the specific growth rate. A central element that is responsible for the structural robustness is a feed-forward loop in the glycolysis, namely the activation of the pyruvate kinase reaction by a metabolite of the upper part of the glycolysis. The robustness is shown for variations of the measured data as well as for variations of the parameters.

Availability: MATLAB files for model simulations are available on http://www.mpi-magdeburg.mpg.de/people/kre/robust/
A short description of the files provided on this site can be found in the Supporting information.

Contact: kremling@mpi-magdeburg.mpg.de

1 INTRODUCTION

In recent years, the set-up of mathematical models for cellular systems that describe metabolism, signal transduction and gene expression has become very popular and will lead to a better understanding of the underlying molecular processes. The knowledge on the detailed interactions between the components that are responsible for carbohydrate uptake in *Escherichia coli* is steadily increasing and current research on individual uptake systems like for glucose uptake via the phosphoenolpyruvate (PEP): carbohydrate phosphotransferase system (PTS) reveals new players that maybe play a role in control of these systems (Plumbridge, 1998). However, the knowledge on individual uptake systems is already rich and is used as a basis to set up mathematical models to describe and analyze the properties of the control circuits (e.g. see (Bettenbrock et al., 2006; Mahadevan et al., 2002; Santillan and Mackey, 2004)).

In previous reports (Bettenbrock et al., 2006; Kremling et al., 2007; Bettenbrock et al., 2007) we analyzed in detail a signal transduction pathway that senses the metabolic state of *E. coli* during carbohydrate uptake and processes the signal to activate Crp. Crp is a global transcription factor involved in the expression of a large number of genes, responsible for carbohydrate uptake and chemotaxis. A key element in this process is the PTS shown in Figure 1.

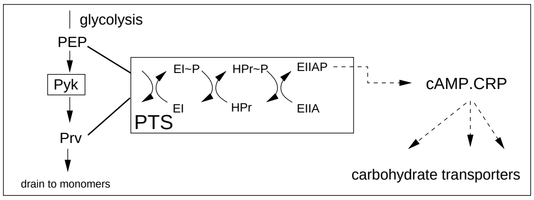


Fig. 1. Scheme of the PTS in *E.coli*. It senses the flux through glycolysis (shown here) and is also responsible for e.g. glucose uptake (not shown). The output, phosphorylated EIIA, activates the synthesis of cAMP which again activates transcription factor Crp. In turn, the cAMP·Crp complex is involved in transcription regulation of most of the carbohydrate transporters and the PTS proteins. Pyk is the pyruvate kinase.

The PTS is a transport and a sensory system at the same time. In a sequence of four reactions a phosphoryl group is transferred from metabolite PEP to protein EIIA^{Crr} (EIIA is used further in the text), the output of the sensory system. E.g. in case of glucose, the phosphoryl group is afterwards transferred the the actual transport protein EIICB^{Glc} and then to the incoming sugar. Interestingly, a relationship between the specific growth rate μ and degree of phosphorylation of EIIA could be seen in various growth situations of the wild type strain growing on single substrates like glucose, lactose, or glycerol, and also for growth on mixtures of substrates (see Hogema et al. (1998) and Bettenbrock et al. (2007)).

Often sensor elements can be regarded as logic elements that process external stimuli into intracellular signals. As an example, a NOT element with high input will result in a low response of the output. Circuits representing different logic elements are mainly found in signaling cascades of higher cells. Here, we present a logic element that can be found in bacterial metabolism: High fluxes through the glycolysis, corresponding to high growth rates result in a low degree of phosphorylation of EIIA. This is surprising, since, assuming a linear reaction chain, high fluxes result in high concentrations of the metabolites in the pathways, based on the (normally) monotone dependency of the reaction rate from the substrate concentrations. The PTS together with the glycolysis can now be seen as an element that allows the transformation of high fluxes into a low metabolite concentrations.

---

[*] to whom correspondence should be addressed

© The Author (2008). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Robustness is the insensitivity of a selected characteristic (time course of a component, steady-state characteristics, network function to sustain growth (Stelling et al., 2002), adaption precision (Barkai and Leibler, 1997)) with respect to changes of external or internal perturbations (different environmental conditions, mutations, altered kinetic parameters, or altered model structures). For the contribution at hand, we define a set of structural and quantitative robust mathematical models as models that fulfill the following conditions:

i) the models are quantitative, that is, they describe experimental data (time course data or steady-state characteristics) representing a cellular function with a given accuracy;

ii) the models of the family differ in the number of components, the number of reactions, the number of regulatory pattern and/or in the choice of the kinetic expressions for the reaction rates.

The paper presents a set of models that describe the dependency of the PTS output phosphorylated EIIA from the specific growth rate μ that is represented by the uptake rate. Often modelers are confronted with the argument that a model can reproduce any experimental data if the parameters are fitted properly. In this contribution, we show that only those models meet the above requirements (and are therefore members of the family) that show a certain structural motif, a regulatory pattern that is more frequent than others. In *E. coli*, the pyruvate kinase reaction is activated by a metabolite in the upper part of glycolysis. This activation represents a feed-forward loop. Motifs for genetic networks have been discussed frequently in the past and it turns out that the feed-forward loop is one of the most common structures in the *E. coli* transcriptional network (Mangan et al., 2003). Here, we found that this motif is essential for the robustness of the transformation of a high incoming flux (high growth rate) into a low concentrations of phosphorylated EIIA. The correct adjustment of the degree of phosphorylation of EIIA in dependence on the glycolytic fluxes is a necessity to survive: if a carbohydrate is running out (low fluxes), the cells has to synthesize proteins for other energy sources. This can only be realized if a transcription factor, here, Crp is activated. Therefore, from a physiological point of view, the correct detection of the flux distribution needs a robust network structure. The feed-forward loop is just such a structural element that allows broad variations of the participating components (here, PEP and pyruvate) but guarantees that the glycolytic flux is correctly mapped to the sensor output, here the phosphorylated form of EIIA. The interpretation of a feed-forward loop as a a motif that guarantees robustness is a new aspect in the discussion on design principles of cellular systems.

2 RESULTS

Figure 2 shows a scheme of the biochemical network that is responsible for metabolism of carbohydrates. In general, substrates enter glycolysis at different nodes. The scheme in Figure 2, left side, considers substrates that feed into glucose 6-phosphate, the first glycolytic metabolite. The scheme covers the central reactions of carbohydrate metabolism. The state variables are $Glc6P$ (glucose 6-phosphate), TP (triose phosphate), PEP (phosphoenolpyruvate), Prv (pyruvate), and $EIIA$/ $EIIAP$. $EIIA$/ $EIIAP$ represent all the PTS proteins. Since the reactions of the PTS are very fast in comparison to glycolytic reactions or gene expression (Kremling et al., 2004), the individual reactions of the PTS can be lumped together and can be described by a single equilibrium constant K_{pts}. Rates r_{up_npts} and r_{up_pts} represent uptake of either a non-PTS sugar or of a PTS sugar, respectively. Rates $r_{ppp/bio}$, r_{ppp}, $r_{tca/bio}$ represent fluxes from glucose 6-phosphate into pentose phosphate pathway and biosynthesis, flux from pentose phosphate pathway back to glycolysis and drain to TCA and biosynthesis from PEP, respectively. Fluxes through glycolysis are represented by r_{gly}, r_{eno}, and r_{pyk} and it is assumed that the respective enzyme concentrations depend on the growth rate μ. Rate r_{pts} represents the rate through the PTS. The equations for the state variables are summarized in the Supporting information.

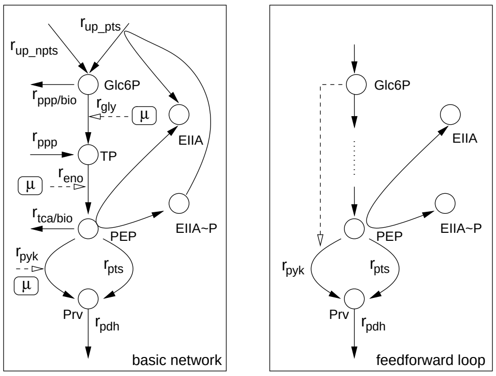


Fig. 2. Basic structure of the model variants. The PTS proteins are represented only by $EIIA$. Some reaction steps in the glycolysis are lumped since the drain flux to the monomers from the precursors are only marginal and have no influence on the results. PTS substrates and non-PTS substrates enter via glucose 6-phosphate. Drain to pentose phosphate pathway (ppp), biosynthesis (bio) and TCA (tca) are included as well as a dependency of the respective concentration of glycolytic enzymes from the growth rate (white arrows with μ symbol). In the models presented here, the feed-forward loop is realized by glucose 6-phosphate that activates pyruvate kinase, see also main text.

A feed-forward loop, the activation of the pyruvate kinase by a metabolite from the upper part of the glycolysis is described in the literature (Waygood and Sanwal, 1974) (right side of Figure 2). The activator of the pyruvate kinase is fructose 1,6-bis-phosphate. In the models introduced below, fructose 1,6-bis-phosphate is not included as a state variable. Therefore, it is replaced by glucose 6-phosphate. This is justified since in the upper part of the glycolysis the drain to anabolism is very small and it can be expected that the steady-state values do not differ very much between glucose 6-phosphate and fructose 1,6-bis-phosphate.

2.1 Sensory system

At first, the sensory system is considered and a relationship for the output of the PTS, phosphorylated EIIA, has to be derived. If the flux distribution at PEP is considered for the case that the PTS is present but not involved in uptake (e.g. growth on non-PTS sugars like glucose 6-phosphate, glycerol, etc.) the reaction rate of the reversible reaction r_{pts} has to be zero:

$$\begin{aligned} r_{pts} \quad &= \quad k_{pts}\ PEP\ EIIA\ - \\ & \quad\quad k_{pts}^{-}\ Prv\ EIIAP\ =\ 0 \end{aligned} \tag{1}$$

with reaction parameters k_{pts} and k_{pts}^{-}. Taking into account that the PTS proteins are either phosphorylated or not, the overall concentration $EIIA_0$ is introduced:

$$EIIA_0 \;=\; EIIA \;+\; EIIAP\,, \tag{2}$$

and Equation (1) is reorganized with respect to $EIIAP$:

$$EIIAP \;=\; \frac{EIIA_0}{1 \,+\, K_{pts}\,\frac{Prv}{PEP}} \;=$$

$$EIIA_0\;\frac{\frac{PEP}{Prv}}{K_{pts} \;+\; \frac{PEP}{Prv}} \tag{3}$$

with $K_{pts} \;=\; k_{pts}^{-}/k_{pts}$. In case of an active PTS, that is, the carbohydrate is taken up and phosphorylated by the PTS, the general structure of Equation (3) is also valid. The following steady-state equation will hold for PTS substrates:

$$r_{up\_pts} \;\;=\;\; r_{pts}\,, \tag{4}$$

and it follows

$$EIIAP \;=\; \frac{\frac{PEP}{Prv}\;EIIA_0 \;-\; \frac{r_{up}}{k_{pts}}}{K_{pts} \;+\; \frac{PEP}{Prv}}\,. \tag{5}$$

Equations (3) and (5) are the measurement equations for the system at hand and are central for the understanding of the experimental data. It is required that the PEP/pyruvate ratio decreases with increasing incoming fluxes (high growth rate) to guarantee a low output (Figure 3A). Measurements of steady-state values of phosphorylated EIIA during batch experiments with different carbon sources taken from Bettenbrock et al. (2007) confirm the approach. Equation (5) states that the degree of phosphorylation of EIIA is always smaller for PTS sugars than for non-PTS substrates and that the difference gets small for low uptake rates.

The PEP and pyruvate concentrations in a cell are difficult to measure. The techniques established for the measurement of metabolites generally generate data with high errors and reliable data about the variation of PEP and pyruvate concentrations with increasing growth rate are lacking. The PEP to pyruvate ratio in a cell can be measured indirectly via the phosphorylation level of EIIA and this ratio has been shown to decrease with increasing growth rate. The networks shown in Figure 2 lead to two different scenarios that are depicted in Figure 3: High growth rates require high fluxes and lead to increased fluxes through the pyruvate kinase. In a network without further control both, the PEP and the pyruvate concentrations have to increase with increasing growth rate (see Figure 3B) to match this requirement. In this case the PEP to pyruvate ratio is very sensitive to small fluctuations in the metabolite concentrations and is most probably difficult to control. On the contrary, in a network that is controlled via the feed-forward loop shown in Figure 2 (right plot) high fluxes can be realized by lowering the PEP concentration. In this case, the pyruvate concentration increases with increasing growth rate while the PEP concentration decreases. The high flux through the pyruvate kinase, in this case, is realized by the activation by glucose 6-phosphate. This scenario is much less fragile to small fluctuations or uncertainties in the metabolite concentrations, since the ratio of a decreasing metabolite (PEP) and an increasing metabolite (pyruvate) always decreases.

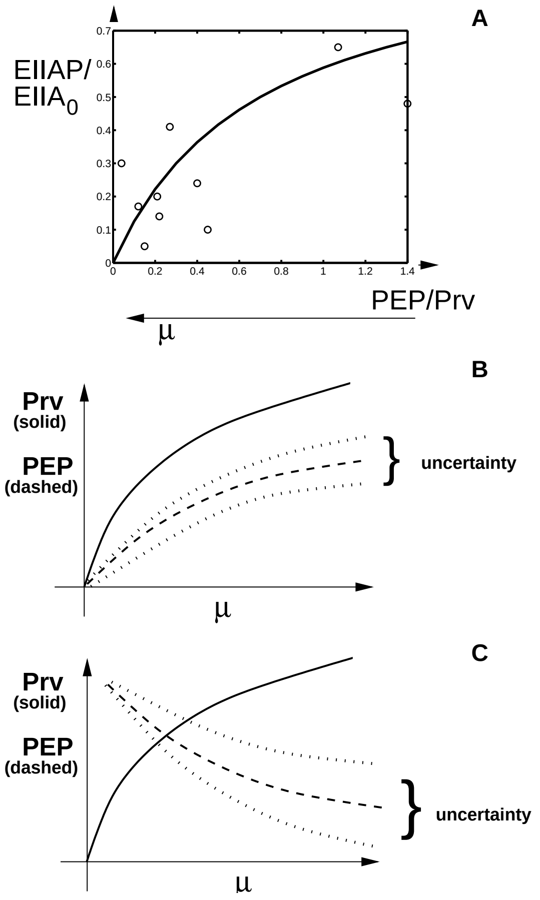


Fig. 3. Top: According to Equation (3), a high PEP/Prv ratio corresponds to high values for $EIIAP$. Measurements are taken from Bettenbrock et al. (2007), the solid line simulates Equation (3) with $K_{pts} = 0.7$. Middle: PEP and pyruvate (Prv) concentrations as functions of the growth rate. Sensitive structure since uncertainties lead to near equal characteristics. Bottom: PEP and Prv concentrations as functions of the growth rate in a controlled network. Robust structure since uncertainties does not change the PEP/Prv ratio very much.

In the following different model variants are introduced that correspond to two cases: uncontrolled network or controlled network with feed-forward loop.

2.2 Model variants

The model variants that are used for this study range form detailed to simple, but are all based on the available biological knowledge. The models differ (i) in the kinetic expression for r_{gly}, r_{pyk}, and r_{pdh}, (ii) in the dependency of the enzyme concentration on the growth rate, and (iii) in the incorporation of the drain into biosynthesis. With this approach, models of different complexity are generated and analyzed with respect to model verification by experimental data.

Model 1 considers all dependencies shown in Figure 2 (left plot). Fluxes $r_{ppp/bio}$, r_{ppp} and $r_{tca/bio}$ are 40%, 20% and 25% of the uptake rate (Holms, 1996). Levels of enzymes depend on the activity of the transcriptional/translational machinery and on the activity of transcription factors. This results in different concentrations of enzymes if the whole range of growth rates is considered. E.g. Seeto et al. (2004) report on the dependency of PtsG from the diluation rate during continuous cultivation. Experimental data for catalytic enzymes like in the glycolysis are, however, not available. To take into account possible dependencies, a simple relationship was included here: The dependency of scaled glycolytic enzyme concentrations e/e_0 from growth rate μ is as follows:

$$\frac{e}{e_0} = \frac{0.5\ \mu}{\mu + K_\mu} + 0.5\,. \tag{6}$$

The equation states that there is a basal level for a slow growth rate of 50% of the maximal level e_0. For higher growth rates, the enzyme concentration increases with increasing μ. Equation (6) represents only one possibility to take into account growth rate dependent enzyme levels. If experimental data will be available, appropriate functions can be used here instead the given one.

The models also differ in the choice of the kinetics for the glycolysis reaction r_{gly}, r_{pyk}, and r_{pdh}. Here, mass action law or Michaelis-Menten kinetics are used. Model 1 is the uncontrolled network. The variants of Model 1 that are used i the analysis are summarized in Table 1.

Model 2 represents the controlled network. In the controlled network, the pyruvate kinase reaction is controlled by a feed-forward loop. Different rate laws are used to describe the reaction rates; e.g. one choice for pyruvate kinase is based on recent publications that use a Monod-Wyman-Changeux kinetics (Bettenbrock et al., 2006; Chassagnole et al., 2002). Variants of Model 2 that are used for analysis are summarized in Table 1.

2.3 Parameter estimation and model assessment

The objective function to fit the parameters was formulated as a ordinary least square problem and a standard gradient based algorithm as it is provided by MATLAB was used for solving. In the Introduction we suggest to use a measure that allows to assess quantitatively the results of the parameter estimation. Since the measurement error of the measurements (experimental data used are described in detail in (Bettenbrock et al., 2006, 2007)) can hardly be determined we calculate the estimated standard deviation (or residual mean square (Montgomery et al., 2001)) $\hat{\sigma}$. Finally we relate $\hat{\sigma}$ on the overall concentration $EIIA_0$ to get a % value:

$$\sigma \;=\; 100\ \frac{\hat{\sigma}}{EIIA_0} \;=\; 100\ \frac{\sqrt{\frac{1}{N-n}\ \sum^{i} \epsilon_i^2}}{EIIA_0} \tag{7}$$

with residuals ϵ_i, N the number of data points and n the number of parameters that were used in the estimation. Experiments were performed with different substrates and substrate combinations under different conditions (Bettenbrock et al., 2006, 2007) and $N = 45$ data points are available. The number of estimated parameters is $n = 4$, that is, k_{gly}, k_{pyk}, k_{pdh}, and k_{pts} are estimated.

The model with the best fit is Model 2b, that is, the model with the highest complexity (Monod-Wyman-Changeux kinetics for the pyruvate kinase, drain fluxes to monomers as well as growth dependent enzyme concentrations are considered). In general, all variants of Model 2 have better results than the models of class Model 1, except for Model 1c. Interestingly, Model 2f with the simplest structure and the simplest rate laws but taking into account the feed-forward loop also reaches a good σ value. Figures 5 and 6 compare some of the model variants with experimental data.

Table 1. Summary model variants and kinetic expressions. Upper part: Columns describe rate of glycolysis r_{gly}, rate of pyruvate kinase r_{pyk}, pyruvate dehydrogenase r_{pdh}, growth dependent enzyme concentrations (E), drain to biosynthesis ($drain$) for the different model variants. †Monod-Wyman-Changeux kinetics: $r = \frac{k_{pyk}\ PEP\ (1+PEP/K_{PEP})^{\beta-1}\ (1+G6p/K_{G6p})^{\beta_2}}{K_{PEP}\ \left((1+PEP/K_{PEP})^{\beta}\ (1+G6p/K_{G6p})^{\beta_2} + L\right)}$; ‡MM: Michaelis Menten. The lower part specifies the kinetic expressions used.

	r_{gly}	r_{pyk}	r_{pdh}	E	$drain$
M1	g1	p1	d1	yes	yes
M1a	g1	p2	d1	yes	yes
M1b	g2	p2	d2	yes	yes
M1c	g2	p1	d1	yes	yes
M1d	g1	p1	d1	no	yes
M1e	g1	p2	d1	no	yes
M1f	g2	p2	d1	no	yes
M2	g1	p1*	d1	yes	yes
M2a	g1	p2*	d2	yes	yes
M2b	g1	p3	d1	yes	yes
M2c	g2	p1*	d1	yes	yes
M2d	g1	p1*	d1	no	yes
M2e	g1	p3	d1	no	yes
M2f	g1	p1*	d1	no	no

Kinetic expression		
g1	Mass action	$k_{gly}\, G6p$
g2	MM‡	$k_{gly}\ \frac{G6p}{G6p+K_{G6p}}$
p1	Power law	$k_{pyk}\, PEP^2$
p2	Power law	$k_{pyk}\, PEP^4$
p1*	Power law	$k_{pyk}\, PEP^2\ G6p^2$
p2*	Power law	$k_{pyk}\, PEP^4\ G6p^4$
p3	MWC†	
d1	Mass action	$k_{pyk}\, Prv$
d2	Power law	$k_{pyk}\, Prv^2$

2.4 Model analysis

In this section the sensitivity of the objective function σ is analyzed with respect to (i) variations of the measured data and (ii) with respect to model parameters. For the analysis one representative of the model class M1 (uncontrolled network) and one representative of model class M2 (controlled network) were considered. The models are comparable since their structure is the same except the feed-forward loop. It is expected that other model variants behave similarly than the representatives.

2.4.1 (i) Variations of the measured data Recently, we applied a statistical procedure – the bootstrap method (Efron and Tibshirani, 1993; DiCiccio and Efron, 1996) – to determine parameter uncertainties for nonlinear systems (Joshi et al., 2006). The method surmounts the theoretical limitations (e.g. the Fisher-Information matrix gives only a lower bound for the parameter variances in case of systems that are nonlinear in parameters) by assessing the uncertainties in statistics with data from finite samples. Like a

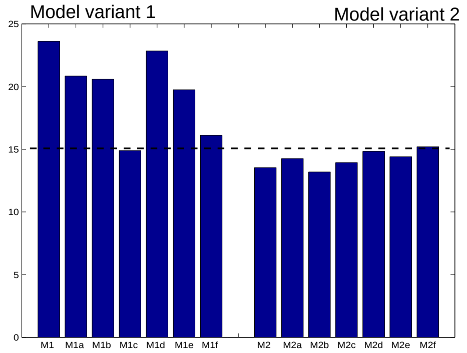


Fig. 4. Comparison of σ values according to Equation (7) for different model variants. Models of class 2 show for almost always a better behavior than model of class 1.

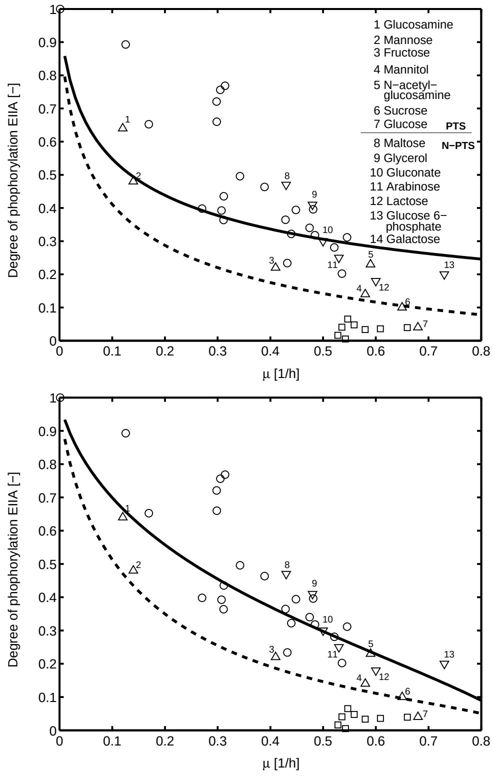


Fig. 5. Comparison of model variant M1a (top) and M1d with experimental data. Simulation: The solid line represents non-PTS sugars, the dashed line PTS sugars. Experimental data: ◯ non-PTS carbohydrates and ☐ PTS carbohydrates from Bettenbrock et al. (2006). ▽ non-PTS carbohydrates and △ PTS carbohydrates from Bettenbrock et al. (2007). Numbers indicate the different carbon sources as described in the legend.

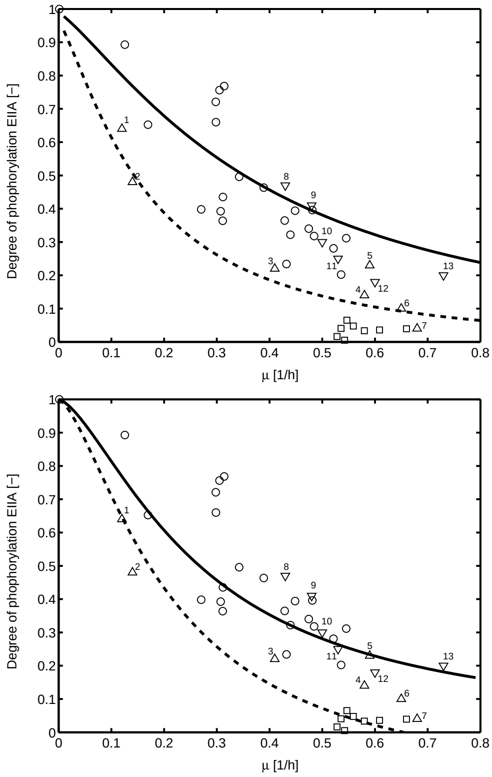


Fig. 6. Comparison of model variant M2b (top) and M2e with experimental data. The solid line represents non-PTS sugars, the dashed line PTS sugars. Same experimental data as in Figure 5.

Monte-Carlo method, the bootstrap method uses stochastic elements and repeated simulations to analyze the properties of the system under consideration.

Briefly, the analysis is performed in such a way, that an initial set of experimental data $\mathbf{S}$ is used as a data base. Performing parameter estimation result in a first set of parameters and σ value to assess the model quality. Due to measurement errors the repetition of the experiment leads to a slightly different set of data $\mathbf{S_1}$ and therefore to a different set of estimated parameters and σ value. The bootstrap approach now uses a large set of B-times replicated experimental data $\mathbf{S_1}, \mathbf{S_2}, \mathbf{S_3} \cdots \mathbf{S_B}$ to calculate statistical properties of the resulting distribution of the (re)-estimated set of parameters and σ values. Formal we look for

$$CR^{95}[\sigma] \;=\; f(\underline{Y_M}) \qquad (8)$$

with CR^{95} is a 95% confidence region of σ with respect to modification of the measured data $\underline{Y_M}$. The values for CR^{95} can directly be read off from the distributions shown in the following figures. Figure 7 shows the results of the bootstrapping with respect to measurement variations for model M1 and M2. For each of the 2000 runs, every single data point was modified by adding a random noise (normal distribution) of 2%. As can be seen, model variant M2d representing the controlled network shows a narrower distribution of the σ values.

2.4.2 (ii) Variations of parameters An analogous approach is applied to determine the influence of parameter variations $\underline{p}$ on the σ values:

$$CR^{95}[\sigma] \;=\; f(\underline{p})\,. \qquad (9)$$

To calculate $CR^{95}[\sigma]$, the model parameters that were used for parameter estimation (k_{gly}, k_{pdh}, k_{pts}, and k_{pyk}) are altered by adding a random number again from a normal distribution (10%),

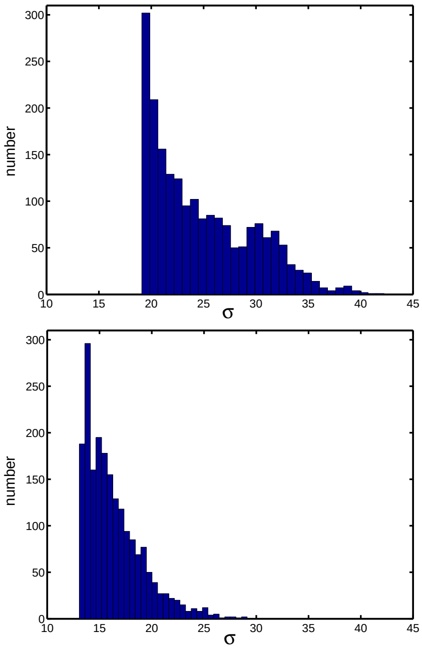


Fig. 7. Bootstrapping results for model M1 (top) and M2 (bottom). Histogram for 2000 runs for σ. The upper and lower values of the confidence region are 35.22%, 19.21% for M1 and 23.47%, 13.19% for M2, respectively.

simulating the system, and calculating σ with Equation (7). Figure 8 reveals that the model representing the controlled system shows a narrower distribution.

3 CONCLUSIONS

Robustness in cellular systems has been discussed very frequently in the last years. A prominent example is the bacterial chemotaxis where an integral feedback loop is responsible for precise adaptation of the system with respect to internal perturbations.

The signal flow, responsible for protein synthesis in the *E. coli* carbohydrate uptake network has been under investigation for a long time and several players have been described in the literature. Besides local control by carbohydrate specific regulators, the global regulator Crp is involved in transcription initiation for almost all genes in the network. Experimental data revealed a relationship between the specific growth rate of *E. coli* and the output of the sensor phosphorylated EIIA of the signaling pathway. The sensor measures the flux through the central pathways. This is realized by a network structure that transforms a high flux into a low response and can be seen therefore as a logic element with NOT function. A number of model variants are thinkable that allow a quantitative description of the experimental data. Therefore, a number of different models were set up and analyzed.

Model development is always a competition between a realistic description, that is based on the available knowledge, and a reduced or simplified description, that takes into account only the most important characteristics. In this contribution, the models presented

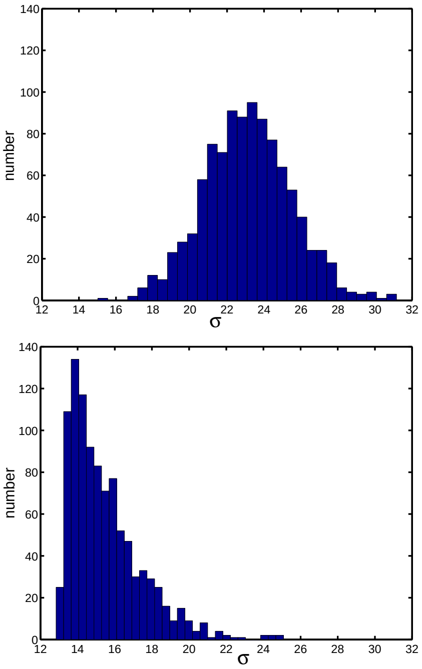


Fig. 8. Sensitivity of σ with respect to random parameter variations. Top: model M1d, bottom: model M2d. The upper and lower values of the confidence region are 27.81%, 18.45% for M1d and 20.42%, 13.25% for M2d, respectively.

show a different degree of complexity based on the kinetic expression for the single rates or the number of reactions that are taken into account. The idea behind this is to show that available experimental data can be described with good accuracy not only by a single model structure but by a whole class of models. Here, a number of different model structures, taking into account different flux distributions, kinetic expressions and possible dependencies of the enzyme concentrations on the growth rate are set up and investigated. The analysis reveals that only those model variants that include a special motif, a feed-forward loop, can describe the data with high accuracy. This feature is named quantitative robustness, since the reproduction of experimental data - here a characteristic curve - is required. The minimal value is achieved with model variant M2b ($\sigma = 13.19\%$) and the maximum with model variant M2f ($\sigma = 15.2\%$), indicating that the model variants show nearly equal accuracy. For the models without feed-forward loop the difference is much larger (the values are between 14.9% and 23.6%, see also Figure 4).

The analysis of the circuit reveals that the feed-forward loop is a robust element. Small variations or disturbances will affect the function of the PEP/pyruvate ratio only marginally (Figure 3). We expected that with model variants M2 it is possible to describe also slightly different experimental data with higher accuracy than with model variants M1. Therefore, a bootstrap approach was performed and the analysis reveals that indeed model variants M2 give better results than model variants M1 (Figure 7). The 95% percent interval for model variant M1 $CR^{95} = 16\%$ while for model variant

M2 $CR^{95} = 10.28\%$. To assess the influence of the parameters, the four parameters that were estimated are randomly modified and σ values are calculated. For Model M1d the 95% percent interval $CR^{95} = 9.36\%$ while for model M2d $CR^{95} = 7.17\%$. From the results it could be concluded that model variants M2, including the feed-forward loop are more appropriate to describe the available data because they show better structural and quantitative characteristics than models without the loop.

Acknowledgment

AK and KB are funded by the FORSYS initiative from the German Federal Ministry of Education and Research (BMBF).

REFERENCES

N. Barkai and S. Leibler. Robustness in simple biochemical networks. *Nature*, 387, 1997.

K. Bettenbrock, S. Fischer, A. Kremling, K. Jahreis, T. Sauter, and E. D. Gilles. A quantitative approach to catabolite repression in *Escherichia coli*. *J. Biol.Chem.*, 281:2578–2584, 2006.

K. Bettenbrock, T. Sauter, K. Jahreis, A. Kremling, J. W. Lengeler, and E. D. Gilles. Analysis of the correlation between growth rate, EIIA^{Crr} (EIIA^{Glc}) phosphorylation levels and intracellular cAMP levels in *Escherichia coli* K-12. *J. Bacteriology*, 189:6891–6900, 2007.

C. Chassagnole, N. Noisommit-Rizzi, J. W. Schmid, K. Mauch, and M. Reuss. Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotech. Bioeng.*, 79(1):53–73, 2002.

T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3), 1996.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

B. M. Hogema, J. C. Arents, R. Bader, K. Eijkemanns, H. Yoshida, H. Takahashi, H. Aiba, and P. W. Postma. Inducer exclusion in *Escherichia coli* by non-PTS substrates: the role of the PEP to pyruvate ratio in determining the phosphorylation state of enzyme IIA^{Glc}. *Mol. Microbiol.*, 30:487–498, 1998.

H. Holms. Flux analysis and control of the central metabolic pathways in *Escherichia coli*. *FEMS Microbiol Rev.*, 19:85 – 116, 1996.

M. Joshi, A. Seidel-Morgenstern, and A. Kremling. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metab. Eng.*, 8(5):447–455, 2006.

A. Kremling, S. Fischer, T. Sauter, K. Bettenbrock, and E. D. Gilles. Time hierarchies in the *Escherichia coli* carbohydrate uptake and metabolism. *BioSystems*, 73(1): 57–71, 2004.

A. Kremling, K. Bettenbrock, and E. D. Gilles. Analysis of global control of *Escherichia coli* carbohydrate uptake. *BMC Systems Biology*, 2007. In press.

R. Mahadevan, J. S. Edwards, and F.J. Doyle. Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys. J.*, 83(3):1331–1340, 2002.

S. Mangan, A. Zaslaver, and U. Alon. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.*, 334:197–204, 2003.

D. C. Montgomery, G. C. Runger, and N. F. Hubele. *Engineering Statistics*. John Wiley and Sons, Inc., 2001.

J. Plumbridge. Expression of *ptsG*, the gene for the major glucose PTS transporter in *Escherichia coli*, is repressed by Mlc and induced by growth on glucose. *Mol. Microbiol.*, 29(4):1053–1063, 1998.

M. Santillan and M.C. Mackey. Influence of catabolite repression and inducer exclusion on the bistable behavior of the lac operon. *Biophys. J.*, 86(3):1282–1292, 2004.

S. Seeto, L. Notley-McRobb, and T. Ferenci. The multifactorial influences of RpoS, Mlc and cAMP on *ptsG* expression under glucose limited and anaerobic conditions. *Res. Microbiol.*, 155:211–215, 2004.

J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E.D. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420:190–193, 2002.

E. B. Waygood and B. D. Sanwal. The control of pyruvate kinase of *Escherichia coli*. *J. Biol. Chem.*, 249(1):265–274, 1974.

I&EC research
Industrial & Engineering Chemistry Research

Subscriber access provided by MPI FUR DYNAMIK

Article

Computer-Aided Modeling of Chemical and Biological Systems: Methods, Tools, and Applications

M. Mangold, O. Angeles-Palacios, M. Ginkel, A. Kremling, R. Waschler, A. Kienle, and E. D. Gilles

Ind. Eng. Chem. Res., **2005**, 44 (8), 2579-2591• DOI: 10.1021/ie0496434 • Publication Date (Web): 18 November 2004

Downloaded from http://pubs.acs.org on May 8, 2009

More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

ACS Publications
High quality. High impact.

Industrial & Engineering Chemistry Research is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Computer-Aided Modeling of Chemical and Biological Systems: Methods, Tools, and Applications

M. Mangold,*,† O. Angeles-Palacios,† M. Ginkel,† A. Kremling,† R. Waschler,† A. Kienle,†,‡ and E. D. Gilles†,§

Max-Planck-Institut für Dynamik komplexer technischer Systeme, Sandtorstrasse 1, 39106 Magdeburg, Germany, Lehrstuhl für Automatisierungstechnik und Modellbildung, Otto-von-Guericke Universität Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany, and Institut für Systemdynamik und Regelungstechnik, Universität Stuttgart, Pfaffenwaldring 9, 70550 Stuttgart, Germany

Computer tools can support and accelerate the development and implementation of first-principle models for chemical and biological processes significantly. Several application examples illustrate this in the contribution. Models of a biochemical reaction network, of a catalytic fixed bed reactor, and of two chemical production processes are considered. The models are implemented in a structured way in the process modeling tool PROMOT, whose key features are discussed. The structuring of the models is based on a uniform structuring methodology whose main ideas are presented as well.

1. Introduction

In chemical engineering, dynamic process models based on conservation laws have become an indispensable tool for the development of new processes and the improvement of existing ones. In biology, the rapidly increasing knowledge of cellular processes guides the way for a quantitative description of cellular systems. However, the development of realistic and predictive models is a challenging and time-consuming task in both sciences, for several reasons: To a large extent, modeling consists of choosing, validating, and revising physical model assumptions. It is an iterative process. Virtually every model of a complex process is inadequate at the beginning and requires a lot of refinements before it delivers satisfactory results. Often, it is necessary that experts from different fields share their knowledge during the model development process. In such a case, engineers used to differential equations, on the one hand, and chemists and biologists thinking in qualitative models, on the other hand, must find a common language to exchange their ideas. The resulting detailed process models typically contain a large amount of information. Usually, they are implemented in a monolithic way without much internal structuring. This makes the understanding and debugging of the model difficult. Existing models are not very transparent and are hardly reusable for another modeler. Furthermore, the implementation of complicated differential equations in a flow-sheet simulator is tedious and error prone. Finally, in most simulation tools, it is in the responsibility of the modeler to formulate his models in a manner suitable for numerical treatment, for example, to avoid a high differential index of a differential algebraic system.

In past years, efforts have been made to support the model development process by computer tools. The main objectives of a modeling tool are (i) to let a user concentrate on the physical modeling task and to relieve him from mechanical coding work, (ii) to increase the reusability and transparency of existing models, (iii) to simplify the debugging process during model development, and (iv) to provide libraries of predefined building blocks for standard modeling tasks such as reaction kinetics, physical properties, or transport phenomena.

In the field of chemical engineering, general structuring methodologies have been proposed by several authors.[1−6] On the basis of these theoretical concepts, modeling languages as well as modeling tools have been developed.[7−13] In the field of mathematical modeling of biological systems, and especially of cellular systems, computer tools[14−19] as well as language standards for model formulation[20] have been published.

The purpose of the present contribution is to give a review on recent results in the field of computer-aided modeling that have been obtained at the Max Planck Institute in Magdeburg. These results are based on basic research done within the joint research project SFB 412 at the University of Stuttgart.[6,12,21,22] In the next section, a general model structuring methodology will be presented that is applicable to biological as well as to chemical engineering processes.[6] This method provides the theoretical foundation for the process modeling tool PROMOT,[12,19] whose key features will be discussed in the following section. A number of different applications have been implemented in PROMOT, so far. In the area of chemical engineering, this includes reactive distillation processes,[21] integrated chemical production plants,[23] and model libraries for membrane reactors[24] and fuel cell systems.[25] In the field of biological cellular systems, a very comprehensive model for the growth of the small bacterium *Escherichia coli* on carbohydrates has been developed.[26,27]

Three examples selected from the various applications are presented in the last part of this contribution to illustrate the concepts.

---

* To whom correspondence should be addressed. Tel.: +49 391 6110 361. Fax: +49 391 6110 513. E-mail: mangold@mpi-magdeburg.mpg.de.

† Max-Planck-Institut für Dynamik komplexer technischer Systeme.

‡ Otto-von-Guericke Universität Magdeburg.

§ Universität Stuttgart.

10.1021/ie0496434 CCC: $30.25 © 2005 American Chemical Society
Published on Web 11/18/2004

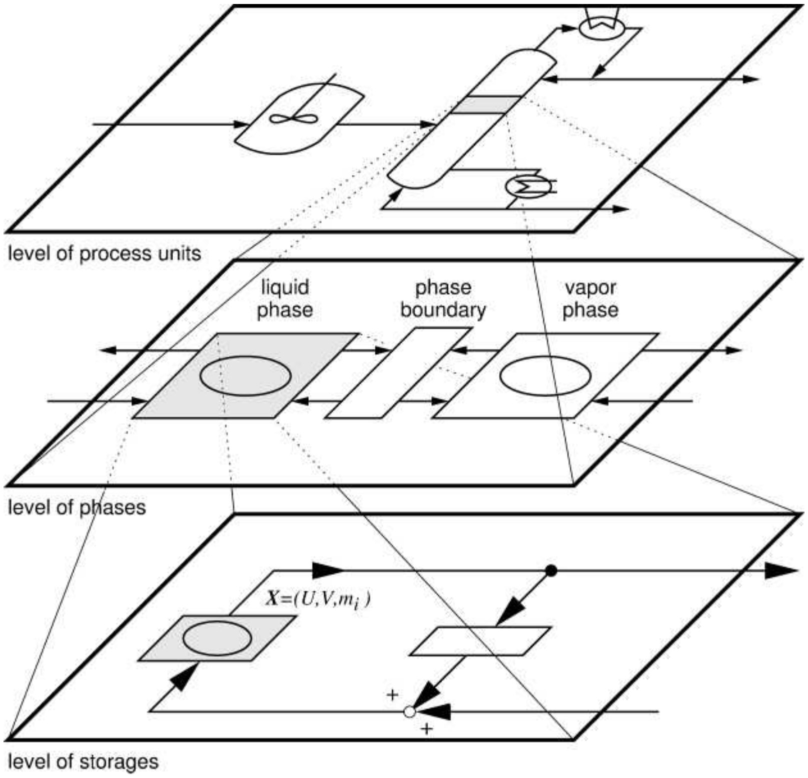


Figure 1. Levels of process structuring for the example of a simple chemical engineering process. Each level is structured into components and coupling elements. An elementary modeling entity on one level can be decomposed into systems of components and coupling elements on a lower level.

2. Model Structuring Concept

The Network Theory of Chemical and Biological Processes[6] proposes a way to decompose various processes into hierarchical units in a systematic manner. A model of a chemical plant, for example, can be decomposed into models of process units such as reactors, storage tanks, and separation units, the elementary modeling entities on the level of process units (see Figure 1). This is the level of modularization that standard flowsheet simulators are based on. Further, each process unit model consists of models of thermodynamic phases and therefore can be decomposed into phase models. Models of thermodynamic phases are elementary modeling entities on the level of phases. Finally, a thermodynamic phase consists of storages for mass, energy, and momentum and therefore can be decomposed further on the level of storages.

A completely analogous hierarchical decomposition can be made for cellular systems. However, in biology, the focus is on the storage level. Because the biological phase consists of hundreds of components that interact in a biochemical reaction network, a structuring of the storage level aims at grouping together elements with common physiological tasks. Therefore, the modeling concept for chemical processes is complemented in an ideal way by considering biochemical reaction networks.

The idea of the network theory is to describe each hierarchical level by two basic types of elements, components and coupling elements. Components possess a hold-up for physical quantities such as energy, mass, and momentum. Coupling elements describe the interactions and fluxes between components. Examples for components are reactors on the level of process units, thermodynamic phases on the level of phases, and mass storages on the level of storages. Examples for coupling elements are valves on the level of process units, phase boundaries or membranes on the level of phases, and reactive sinks and sources on the level of storages. Components are described by a thermodynamic state or state vector $\boldsymbol{X}$. The state of a component may be changed by fluxes $\boldsymbol{J}$, for example, fluxes of mass or energy. The general differential equation of a component therefore reads:

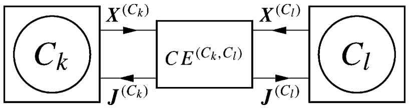


Figure 2. Connection of two components by a coupling element.

$$\frac{\partial \mathbf{X}}{\partial t} = \mathbf{J} \tag{1}$$

The task of the coupling elements is to determine the flux vectors $\mathbf{J}$. In accordance with the principles of irreversible thermodynamics, it is assumed that the flux vector is an algebraic function of potential differences or potential gradients. A simple example of a coupling element is the heat flux between two phases, which is driven by the temperature difference between the phases. The exchange between components can be visualized by a diagram as shown in Figure 2. The components C_k and C_l pass information on their states to the coupling element $\mathrm{CE}^{(C_k,C_l)}$. Depending on those states, the coupling element computes the flux vector and returns the result to the two components. This establishes a bidirectional signal transfer between components and coupling elements.

2.1. Example: Well-Mixed System. A simple example may illustrate this structuring concept on the level of storages. A CSTR with a single isothermal

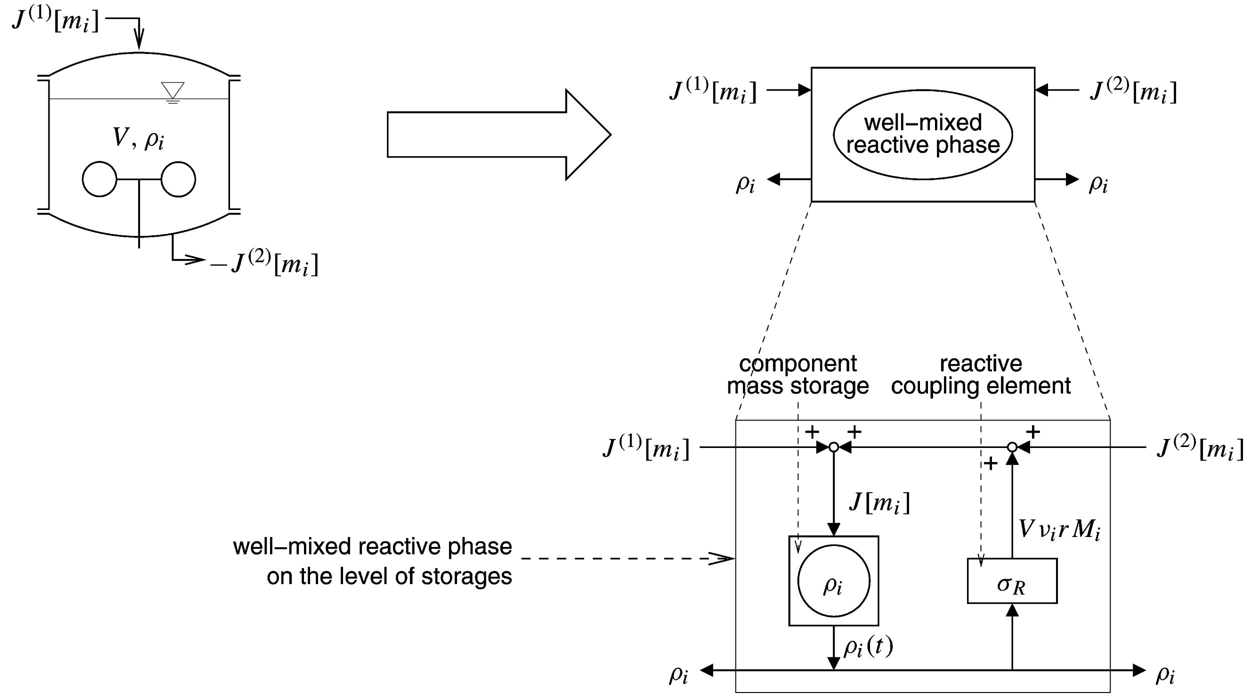


Figure 3. Structuring of a well-mixed reactive phase on the level of phases and on the level of storages.

reactive phase and constant volume is considered (see Figure 3). The component mass balances for this system read

$$V \frac{d\rho_i}{dt} = J[m_i] \qquad (i = 1, ..., \mathrm{NC}) \tag{2}$$

$$J[m_i] = \underset{\sigma_R}{V\nu_i r(\rho_i)\, M_i} + J^{(1)}[m_i] + J^{(2)}[m_i] \tag{3}$$

In eqs 2 and 3, V is the volume of the reactive phase, NC is the number of components, r is the reaction rate, ρ_i is the partial density of component i, M_i is its molar mass, and ν_i is a stoichiometric coefficient. $J^{(1)}[m_i]$ and $J^{(2)}[m_i]$ denote external component mass fluxes into the phase. The structured representation of the reactive phase on the level of storages is shown in the lower part of Figure 3. The phase consists of NC component mass storages. The component mass storages provide information on the partial densities ρ_i. They contain the mass balances in the general form (2). The mass fluxes $J[m_i]$ on the right-hand side of (2) are signal inputs for the component mass storages. They are determined by adding up the fluxes $J^{(1)}[m_i]$ and $J^{(2)}[m_i]$ across the phase boundary, on the one hand, and the mass fluxes caused by chemical reaction inside the phase, on the other hand. The mass fluxes caused by chemical reaction are the output of the reactive coupling element σ_R. This coupling element needs the partial densities ρ_i as signal inputs, because the reaction rate r depends on those values. Therefore, σ_R establishes a coupling between the different component mass storages.

On the level of phases, the reactive phase is an elementary component with the mass fluxes $J^{(1)}[m_i]$ and $J^{(2)}[m_i]$ as signal inputs and the partial densities ρ_i as signal outputs.

The structuring concept illustrated here for well-mixed systems can be extended in a straightforward manner to other classes of chemical processes. This was described in detail in previous publications for distributed and particulate systems[22] as well as for electrochemical processes.[25]

2.2. Extensions of the Concept to the Structuring of Biological Models. In the context of biological models, the overall aim is to provide a framework for modeling cellular systems that will serve as a basis for software tools. These tools should support model setup as well as model analysis. The focus is on the biochemical reaction network, which represents the storage level introduced above. The modeling concept is based on the analysis of the available knowledge on (i) metabolism, that is, the part of the biochemical reaction network that is responsible for the breakdown of the nutrients and the synthesis of the macromolecules, and (ii) signal-transduction systems, that is, the part of the biochemical reaction network that senses the environmental conditions and translates the extracellular stimulus into an intracellular signal. The procedure of decomposing the biochemical network thus has to be based on the molecular structure of the units that have to be defined in such a way that a cellular unit is represented by an equivalent mathematical submodel. This modular approach is a new feature in the modeling of cellular systems and guarantees a high transparency for biologists and engineers.

Analogous to the approach in chemical engineering, the basis of the framework is the definition of a complete but finite set of elementary modeling objects. They should be disjunct with respect to the biological knowledge they comprise to prevent overlapping. In the biological framework, components are represented by small metabolites such as sugars and amino acids, and macromolecules such as proteins or DNA. Coupling elements are represented by biochemical reactions. Here, two types have to be distinguished: enzymatic reactions and polymerization processes. Enzymatic reactions are responsible for uptake and breakdown of nutrients. Hundreds of enzymes are active inside a cell and show therefore a high substrate specificity. Polymerization processes are rather slow processes in com-

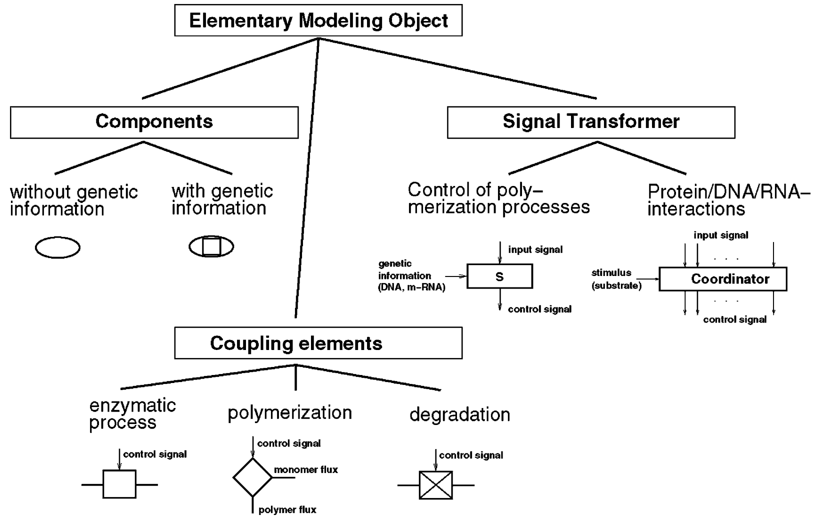


Figure 4. Hierarchy of elementary modeling objects for cellular systems. Substance exchange is marked by bold lines, whereas arrowheads are used to indicate signal connections.

parison to enzyme-catalyzed reactions. For instance, the transcription of the genetic information on the DNA into RNA and the translation of the RNA into proteins are typical polymerization processes.

Because the enzymatic and the polymerization processes are controlled by signal-transduction systems, a further class, the signal transformers, is introduced. Signal transformers process signals on different levels. As an example, gene expression, that is, transcription and translation, is considered. Here, the information from different transcription factors (specialized proteins) that are involved in the process has to be modeled in such a way that the expression efficiency, that is, the rate at which the protein is synthesized, can be calculated. The elementary modeling objects defined so far are summarized in Figure 4.

The increasing knowledge about the interconnection between different biological pathways allows the development of increasingly complex models. These models offer a highly detailed picture of metabolic, signal transduction, and regulatory networks, but the properties of these systems as a whole become difficult to grasp. The definition of functional units, that is, the aggregation of elementary modeling objects to higher structured units, might help to unravel this complexity. Once units are found, they are systematically analyzed and classified, creating a library of units that can be reused. This simplifies the setup of models because many parts of biological networks appear recurrently.

Two completely different approaches are considered to define the units. The first approach defines the units according to three biologically motivated criteria:[28] (i) A common physiological task. All elements of a functional unit contribute to the same physiological task. Easily recognizable examples are the specific catabolic pathways for individual carbohydrates, or the biosynthetic pathways for the different amino acids, nucleotides, and cofactors from their precursors. (ii) Common genetic units. The genes for all enzymes of a functional unit are organized in genetical units, that is, units that are expressed in a coordinated manner. (iii) A common signal-transduction network. All elements of a functional unit are interconnected within a common signal-transduction system. The signal flow over the unit border ("cross-talk") is small as compared to the information exchange within the unit, such that the coordinated response to a common stimulus helps to identify the members of a unit. The second approach delimits functional units from a theoretical point of view. An interesting criterion might be elements without retroactive effects, because they could be considered independently and analyzed by the means of system theory. Different models, starting from very simple models to models for complex signal-transduction systems, were analyzed, and functional units were defined.[29,30]

On the basis of the studies on the carbohydrate uptake in the bacterium *Escherichia coli*, a number of functional units could be defined and were implemented in the PROMOT environment.[19]

3. The Modeling Tool PROMOT

Models of chemical and biological systems that are structured and constructed according to the aforementioned modeling concept can be implemented in the Process Modeling Tool.[12] This tool allows the construction of structured models via a graphical user interface and with a modeling language. The final models are transformed into a differential-algebraic equation set, which can be analyzed in the simulation environment Diva[31] or in Matlab.[32] In the following section, the construction of models in this tool and some aspects of the model processing will be elaborated.

3.1. Modeling Elements. In the previous sections, a systematic approach for structuring models of chemical and biological systems was summarized. The components and coupling elements of this approach are implemented as modules in PROMOT. As an example, we use a model of a regulated metabolic pathway (see Figure 5), which is modeled on the level of storages. The whole pathway is represented by a module in PROMOT, an encapsulated entity containing a mathematical description of its behavior. This module is composed of submodules that represent storages (ellipses), reactions (squares with arrows), and signal-transformers (rectangles). This module is part of a larger model, describing the regulated carbohydrate uptake of *E. coli* (see section 4.1).

The submodules in the figure are instances of other module-classes; for example, the elements lac und allo are instances of the module-class storage-intra-x. The interface of the module is represented by terminals

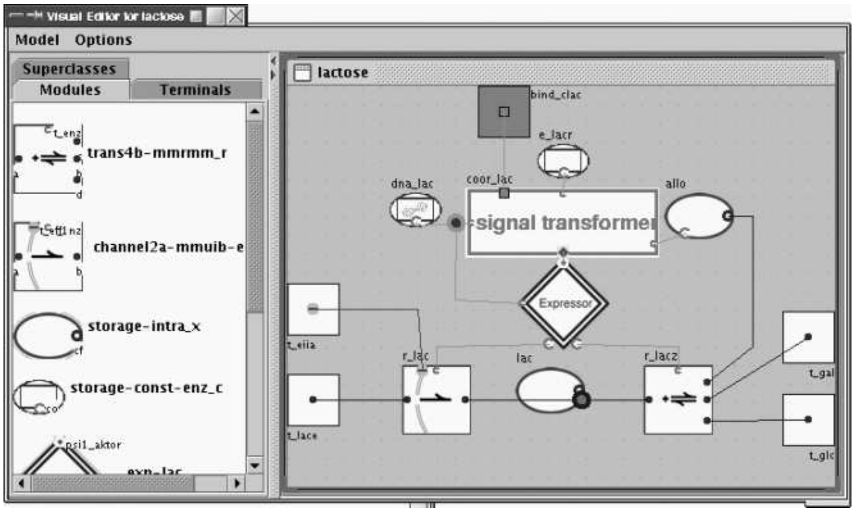


Figure 5. Module for lactose transport in *E. coli*.

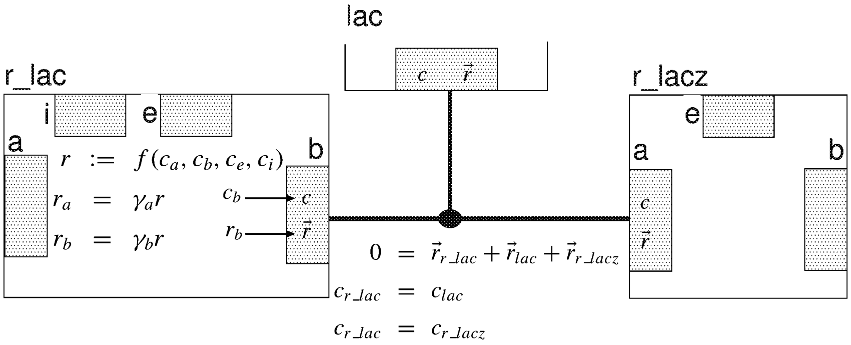


Figure 6. Generation of coupling equations for the link to lac in Figure 5.

(smaller squares on the outer edges of the module). Terminals are connection-points that contain a group of named variables and can be linked with other compatible terminals on the next higher level of the aggregation hierarchy. A terminal can represent flows of mass, energy, momentum, or signals. In biological systems, terminals represent mass flows (e.g., the terminal of lac) or concentrations of substances acting as signals. Terminals of submodules can be propagated as terminals of the containing module; for example, the left terminal a of r_lac is propagated as t_lace for the whole lactose module.

Users of PROMOT can build composed modules such as lactose by selecting module-classes on the left of the window shown in Figure 5 and placing them as submodules into the working-area on the right. Submodules can be parametrized with appropriate initial values and parameter settings and can be connected using their terminals. Two or more compatible terminals can be connected with links. Terminals are considered compatible, if each of them contains a set of variables with the same names. The generation of coupling equations for a link is shown in more detail in Figure 6.

PROMOT allows one to distinguish variables for potentials (e.g., *c* in Figure 6) and fluxes (e.g., $\vec{r}$) in the set of interface variables. When linking terminals, coupling relations are established between all variables in the set of connected terminals that have the same name. The flux variables sum up to zero in one link, whereas the potentials of all connected terminals are set equal. In the example, all *c*'s are potentials, whereas all $\vec{r}$'s are fluxes. This connection-method allows for a flexible and extensible setup of network-like modules, because additional modules can be easily added without changing the interface of the modules already present.

Elementary modules like r_lac contain local variables and model equations defined by the user with the Model Definition Language (MDL)[12] of PROMOT. The r_lac module in Figure 6 calculates a reaction rate *r* using the kinetic law *f*. To allow connections to the module interface, variables are assigned to the different terminals. In Figure 6, this is shown for c_b and r_b. Although PROMOT allows for the construction of models out of encapsulated modules, the modeling scheme is equation-based and the equations are fully transparent. Users can add their own modules with special equations or extend models from a library with their own equations and libraries. In general, modules can contain a linear-implicit differential−algebraic equation set which reads

$$\mathbf{B}(t,\mathbf{x},\mathbf{p})\,\dot{\mathbf{x}} = \mathbf{f}(t,\mathbf{x},\mathbf{p}) \tag{4}$$

In eq 4, **x**'s are the state variables, **p**'s are model parameters, **B** is a possibly singular and state-dependent descriptor matrix, and **f**'s are nonlinear functions for the calculation of the right-hand sides of the equations. Important for equations in the field of chemical engineering is the possibility to use arrays of variables, equations, and also modules for the efficient modeling of repeated elements in plant models (see also the column models in section 4.3.1).

When chemical engineering systems are modeled, purely continuous models are often not sufficient, because discontinuities in the model equations as well as discrete controllers have to be described. Therefore, PROMOT employs a concept of hybrid modeling which

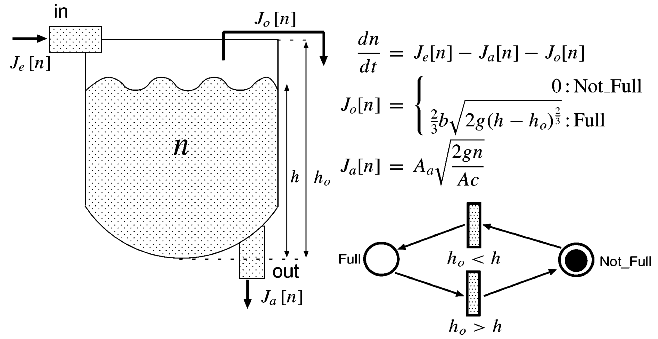

in
$J_e[n]$
$J_o[n]$
n
h
h_o
out
$J_a[n]$
$\frac{dn}{dt} = J_e[n] - J_a[n] - J_o[n]$
$J_o[n] = \begin{cases} 0 : \text{Not\_Full} \\ \frac{2}{3} b \sqrt{2g(h-h_o)^{\frac{3}{2}}} : \text{Full} \end{cases}$
$J_a[n] = A_a \sqrt{\frac{2gn}{Ac}}$
Full
$h_o < h$
Not_Full
$h_o > h$

Figure 7. Model of a tank with overflow, using a Petri net for switching the weir equation.

uses an additional Petri net[33] to describe the discrete part of the model. The places of the Petri net represent discrete model states, and the transitions describe changes of the discrete state that can be triggered by changes in the continuous variables. On the other hand, the current state of the Petri net (i.e., the marked places) changes the equation system locally through conditional equations or by changing characteristic parameter-values. The concept of Petri nets has been chosen in favor of the simpler state machines[34] because it allows one to describe parallel and synchronized discrete behavior, which is necessary for more complex control programs. As a very simple example, the modeling of a tank with an overflow is shown in Figure 7. The Petri net switches between Full and Not_Full depending on the current height h of the liquid in the tank. The weir equation for the overflow $J_o[n]$ is a conditional equation, because it is only active as long as Full is marked and the liquid reaches the overflow. More sophisticated examples of hybrid models with coupled discrete and continuous parts are presented in the reactor model of section 4.3.2.

3.2. Model Processing. All modeling entities in PROMOT are organized in a specialization hierarchy with multiple inheritance. The modeler can therefore use object-oriented implementation techniques such as abstraction and polymorphism to describe his modules. This becomes especially helpful when implementing general modeling libraries. In this case, general model elements such as balance equations and property correlations can be implemented in reusable superclasses. For a specific application, general modules can be extended and specialized by deriving subclasses that inherit all general parts and add application-specific equations and parameters. Examples for the use of object-oriented libraries are given in the next sections.

The structured and object-oriented view of a model is particularly effective for model formulation and manipulation. On the other hand, the numerical solution can be carried out more efficiently with the plain equations. Therefore, all further operations are carried out using a global DAE that is obtained from all local equations from the different modules and the coupling equations generated by links. Another issue is related to the numerical solution: Due to the structured way of modeling that generates coupling relations, and because the elements in model-libraries tend to be implemented in a general way, often using some extra equations, models typically contain a large number of simple algebraic equations. Before generating simulation code, PROMOT therefore analyzes and optimizes the system of equations. For this purpose, the incidence matrix of the equations is computed and an algorithm proposed by Tarjan[35] is used to transform this matrix to block-lower-triangular form. Explicit algebraic equations can be identified in this matrix and are symbolically transformed into a sequence of explicit assignments to intermediate variables. These assignments can be calculated in the simulation code without involving the numerical equation solver, which improves simulation performance. In this process of optimization, also repeated calculations of constant expressions and unused variables are removed from the model. The preprocessing step is not only useful for optimizing the efficiency of the numerical solution, but it also unveils structural inconsistencies (e.g., singularities) of the equation system. If such conditions occur, the user is provided with debugging information for detecting the error in the model structure quickly. Finally, the equation system is transformed into simulation code using the Code-Generator[36] for Diva.

4. Applications

In this section, application examples for the structuring methodology and the modeling tool PROMOT are presented. Guided by the hierarchical modeling concept, examples for the formulation on the storage level, on the level of phases, and on the level of process units will be discussed.

The example on the level of storages is chosen from biology. The carbohydrate uptake of bacteria is considered. The challenge here is to structure a huge reaction network into smaller functional units to understand the various interactions between the reaction steps. In this sense, the biological example has some similarity to chemical engineering systems with complex reaction kinetics such as combustion processes.

The example on the level of phases is an arrangement of thermally coupled fixed bed reactors. Spatially distributed models of thermodynamic phases have to be coupled in this case. The objective of the structuring on the level of phases is in this case to support the design and analysis of a novel integrated process.

On the level of process units, models of two chemical production processes are presented. It is shown that many of the modules defined for one process can be reused for modeling of the other process. Furthermore, the question of iterative refinement of process unit models is addressed. In many cases, it is reasonable to start with very simple process unit models, and to gradually increase the level of model detail of selected apparatuses during the model development process. This can be done conveniently by applying a top-down modeling strategy, that is, by building up a hierarchy of more and more detailed models of a certain process unit. In addition, the second production process serves as an example for the modeling of discrete events using PROMOT.

4.1. Application Example on the Level of Storages: Nutrient Uptake of a Bacterium. An interesting example for model setup of a biological system is concerned with the question of nutrient uptake. The control of the carbohydrate uptake in bacteria has been investigated for a long time. Starting with the pioneering work of Monod,[37] a number of components were detected that are responsible for the coordination of sugar uptake. It is widely accepted that the phosphotransferase system (PTS) is one of the important modules in the signal-transduction machinery of bac-

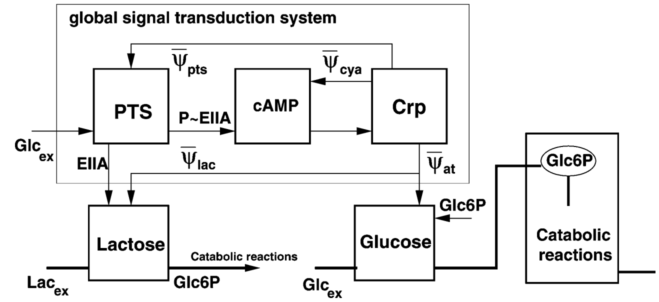


Figure 8. Survey of all submodules of the model.[27] The global signal-transduction unit comprises the PTS module, the synthesis of the second messenger cAMP, and the binding of the regulator protein Crp to the respective binding sites.

teria. The PTS represents a transport system (in microbiology, a protein that is membranstanding and transports components from the extracellular environment into the cell is named a transport system), and at the same time it is part of a signal-transduction system responsible for carbon catabolite repression.[38] Catabolite repression means the dominance of one carbohydrate uptake system over another one. If two sugars, for example, glucose and lactose, are present in the medium, glucose is taken up first while lactose is taken up only after the depletion of glucose. The connections of all submodules are shown in Figure 8. The lactose pathway is shown in detail in Figure 5. As can be seen, protein EIIA and its phosphorylated form P∼EIIA are the main output signals of the PTS. The output signal ψ from the Crp submodule describes the transcription efficiency of the genes and operons under control of Crp. The glucose and the lactose pathway are connected to the liquid phase, which is represented by Glc_{ex} and Lac_{ex}. Both pathways feed into the central catabolic pathways. The entire model comprises 30 state variables.[26,27]

The PTS controls via output EIIA the lactose pathway. EIIA is an inhibitor of the lactose transport protein. Providing both sugars at the very beginning of a batch experiment, glucose is taken up immediately, while lactose is taken up after glucose has run out (Figure 9). During the second growth phase, galactose is excreted in large amounts in the medium. The enzyme for splitting intracellular lactose, LacZ, is synthesized first in the second growth phase. Protein EIIA, the output of the PTS, shows an interesting dynamical behavior. After glucose has run out, EIIA switches very fast to the phosphorylated form and returns slowly back. This is based on the fact that the PTS is active during glucose uptake in the first growth phase and is active also during the second growth phase due to the splitting of intracellular lactose into intracellular galactose and glucose. The intracellular glucose is also phosphorylated by the PTS.

The model is available in the PROMOT/Diva environment.[19] Parameters are estimated using a number of experiments with different mixtures of carbon sources and mutant strains. The mutant strains differ only in one single gene. This strategy allows one to analyze the influence of different proteins in an isolated way and to estimate parameters from the time courses.

4.2. Application Example on the Level of Phases: Autothermal Fixed Bed Reactor. An ex-

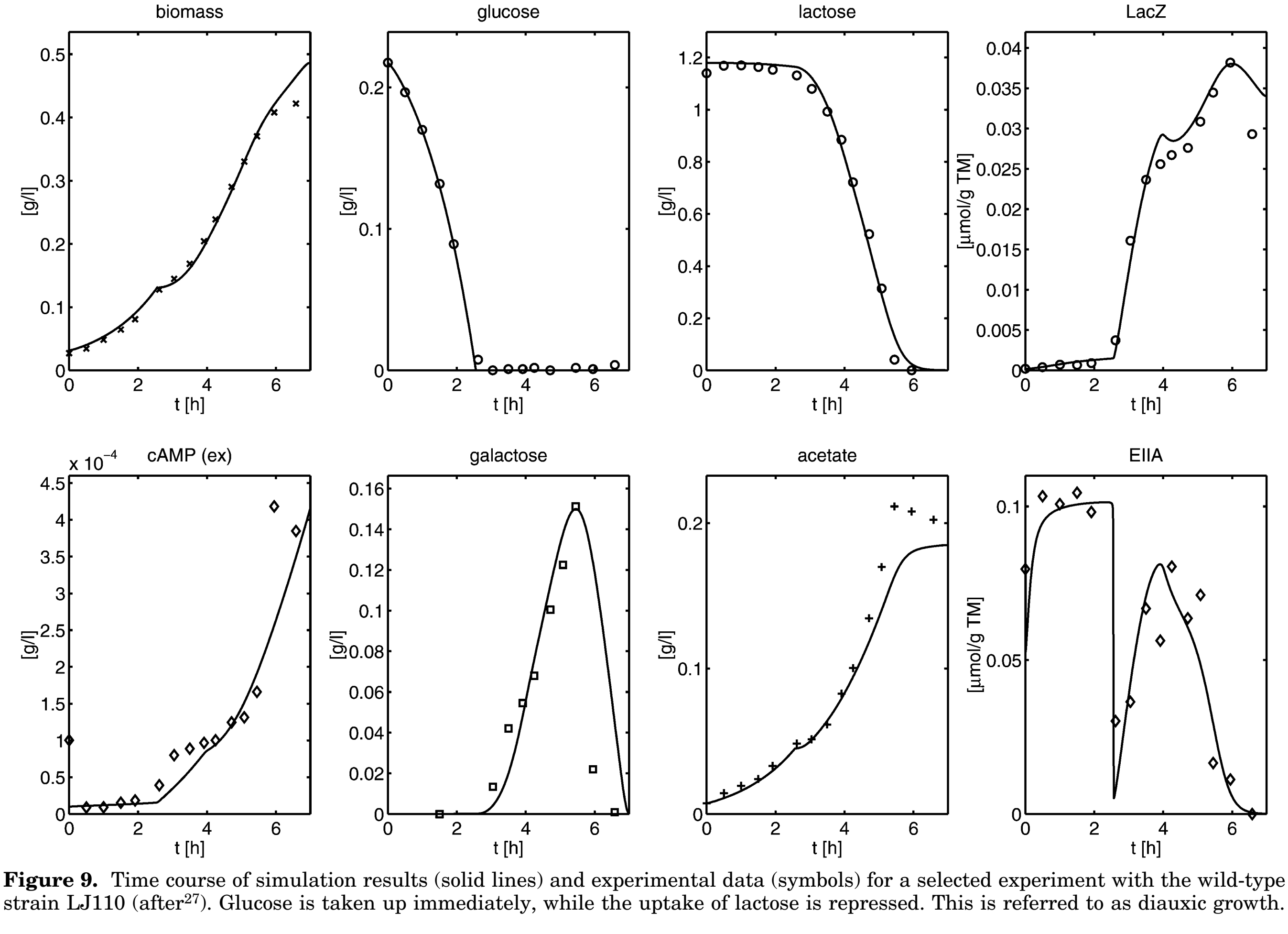


Figure 9. Time course of simulation results (solid lines) and experimental data (symbols) for a selected experiment with the wild-type strain LJ110 (after[27]). Glucose is taken up immediately, while the uptake of lactose is repressed. This is referred to as diauxic growth.

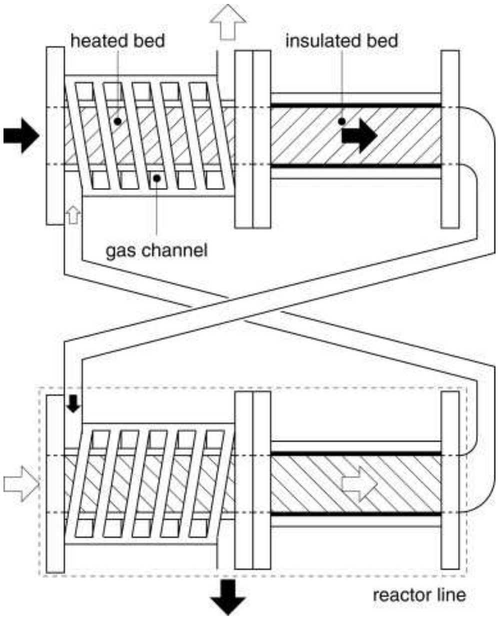


Figure 10. Autothermal reactor concept: Two catalytic fixed bed reactors coupled thermally by cocurrent heat exchange. White and black arrows indicate different mass fluxes.

ample from chemical engineering may serve to illustrate the structuring concept on the level of phases. An autothermal reactor concept is considered. The purpose of an autothermal reactor is to carry out weakly exothermic reactions without providing external heating energy. This can be achieved by integrating heat exchange and chemical reaction in one apparatus.[39] Dynamically operated autothermal reactors make use of the fact that a creeping reaction front in a catalytic fixed bed causes an overadiabatic temperature rise.[40] In the well-known reverse-flow reactor,[41] the creeping reaction front is generated by a periodic flow reversal, that is, in a forced periodic operation mode. Alternatively, it is also possible to design a reactor in such a way that the creeping reaction fronts and the resulting overadiabatic temperature rise are created by autonomous periodic oscillations without any external forcing.[42,43] A reactor concept of this kind will be modeled in the following.

The reactor is divided into two reactor lines with separate inlets and outlets (see Figure 10). Each reactor line consists of two catalytic fixed beds in series. The first bed is jacketed by a gas channel that establishes a cocurrent heat exchange. The second bed is insulated toward the environment. The reactants enter the heat exchanger section, flow through the insulated section, and leave the reactor via the gas channel of the other reactor line. In this way, a thermal feedback is established between the two reactor lines. This feedback can be used to generate circulating reaction fronts in the arrangement: A hot spot caused by a creeping reaction front in one of the reactor lines triggers a new reaction front in the other line when reaching the gas channel. For a simple oxidation reaction of first order, it can be shown that two types of autonomous periodic solutions coexist under certain operation conditions: a symmetric solution with a creeping reaction front in each of the reactor lines, and an asymmetric solution, where alternately one reactor line contains a reaction front and the other is in an extinguished state.[43]

A spatially distributed one-dimensional dynamic model for this process has been implemented in PROMOT. Figure 11 shows the top level structuring of the model. On this level, the two reactor lines are components in the nomenclature of the network theory. They are connected to reservoirs representing the inlet and outlet tanks of the system. The coupling between the two reactor lines is done by coupling elements describing the following internal boundary conditions between the two distributed systems: Continuity is assumed for the compositions and the temperatures on both sides of the boundary. Mass and energy conservation gives further

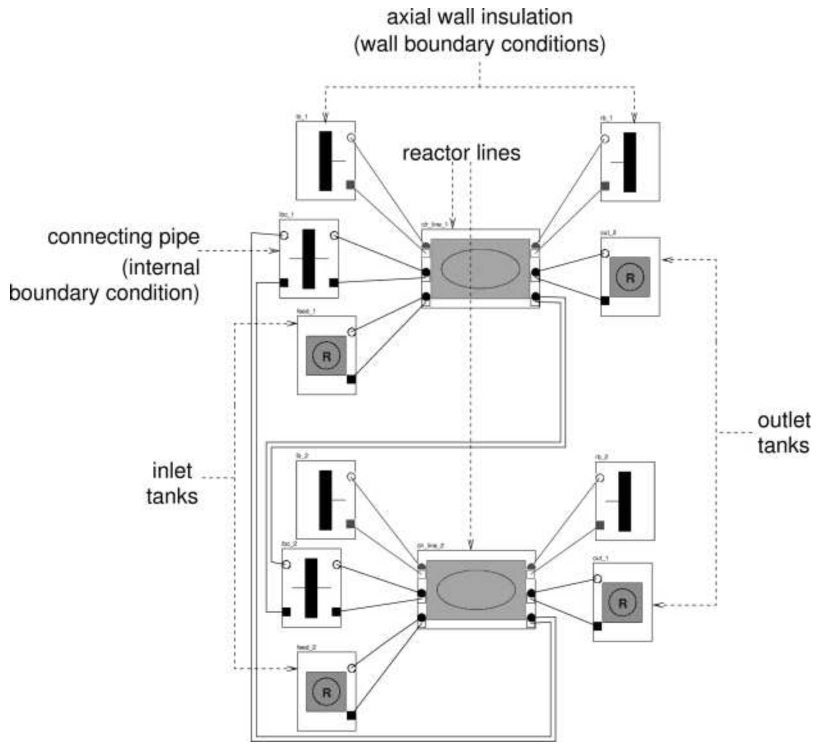


Figure 11. Structure of the fixed bed reactor model on the top level (from a screenshot of the PROMOT GUI).

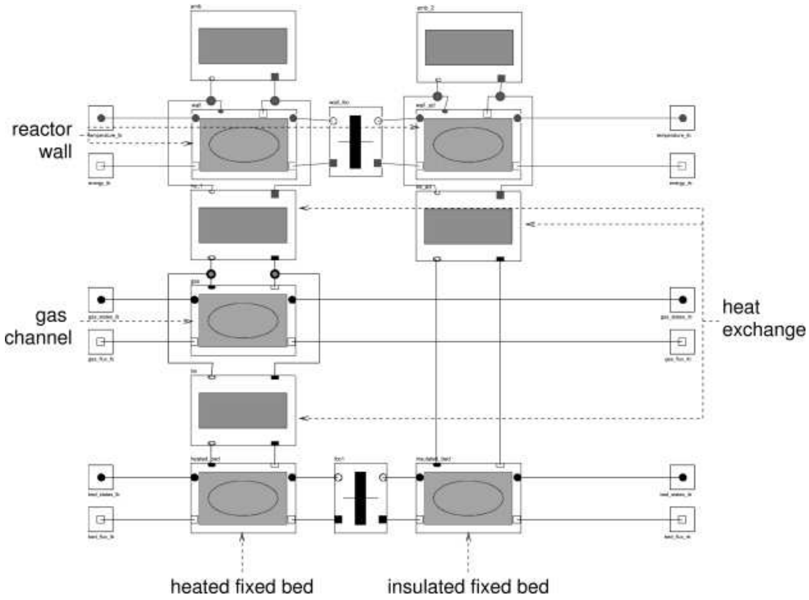


Figure 12. Structure of the fixed bed reactor model on the level of phases (from a screenshot of the PROMOT GUI).

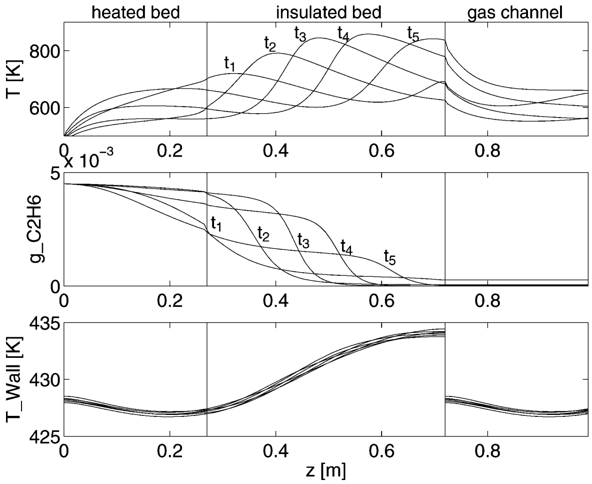


Figure 13. Spatial profiles of the gas temperature T, the ethane mass fraction $g_{C_2H_6}$, and the wall temperature T_{Wall} in autonomous periodic operation at time points t_i, $i = 1, ..., 5$.

conditions for the fluxes across the internal boundary. Further coupling elements are needed to define heat fluxes to the reactor walls that enter the boundary conditions for the energy balance of the walls. As can be seen in Figure 11, the connection between two modules is always bidirectional with one signal line passing the information on the state vector and one signal line passing the information on the fluxes.

The models of the two reactor lines can be decomposed into models of interacting thermodynamic phases, as is shown in Figure 12. The spatially distributed phases of the fixed beds, the gas channels, and the reactor walls are the components on this level. The coupling elements in Figure 12 define the internal boundary conditions between axially coupled phases, on one hand, and the radial heat exchange between fixed bed, gas channel, and reactor walls, on the other hand. The structuring makes changes of the model very easy. For example, one might want to add mass exchange to the heat exchange between the fixed bed and gas channel. This can be done by replacing the heat exchange coupling elements by membrane modules, as described in a previous publication.[24] The models of the distributed phases consist internally of components and coupling elements on the level of storages,[22,24] similar to the biological example in the previous section.

A simulation result obtained by the described model is shown in Figure 13. The total oxidation of ethane is considered. Under suitable inlet temperatures, inlet compositions, and flow rates, an autonomous periodic oscillation with creeping reaction fronts develops. At time t_1, the front stretches from the heated bed to the inlet of the insulated bed. At later times t_2, ..., t_4, the front moves into the insulated bed. At time t_5, the hot spot reaches the gas channel and ignites a new reaction front in the heated bed of the other reactor line. This marks the beginning of a new periodic cycle.

4.3. Application Examples on the Level of Process Units. In many chemical engineering processes, two-phase systems play an important role. In particular, vapor−liquid systems, as encountered for example in distillation or many reactive separation processes, are predominant in the area of thermal separation of fluid mixtures. While a rigorous description of the underlying heat and mass transfer processes in vapor−liquid systems would require detailed rate-based models, assuming thermodynamic equilibrium between the two phases is often appropriate and sufficient for many applications.

Therefore, a library of generic two-phase models assuming thermodynamic equilibrium between the two bulk phases has been implemented in PROMOT. This library contains generic, reusable models on the level of process units. One of the most general modeling entities in this context is a simple generic nonreactive two-phase model. It constitutes, for the given purposes, the simplest model formulation. As such, it serves as the superclass in a hierarchy of two-phase models. More specific two-phase models, as required, for example, for feed trays or reactive trays in distillation columns, are subsequently obtained by applying a top-down modeling strategy. This approach corresponds to an iterative refinement of existing models using the object-oriented concept of inheritance. A reactive tray, for instance, is straightforward implemented as a specialization of a nonreactive tray, that is, as a subclass of the more

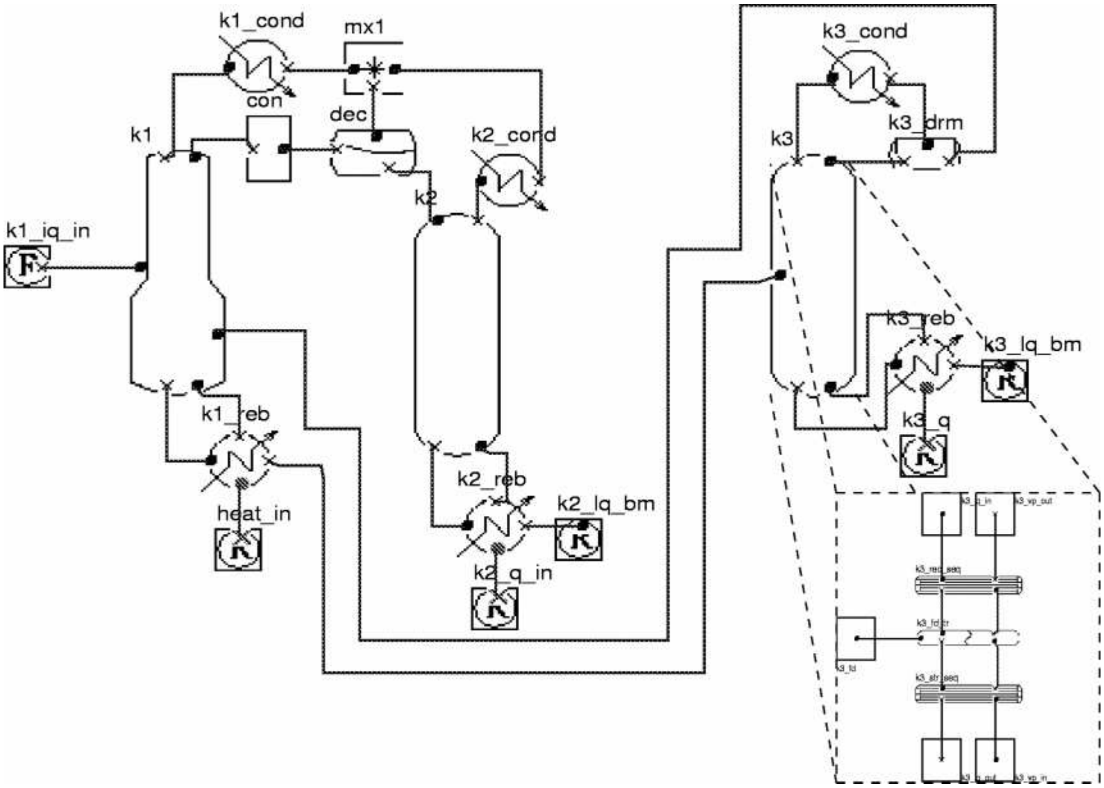


Figure 14. Flow sheet diagram of the plant for the production of butyl acetate in its PROMOT representation. The zoom illustrates column k3 as an aggregate from more basic modules.

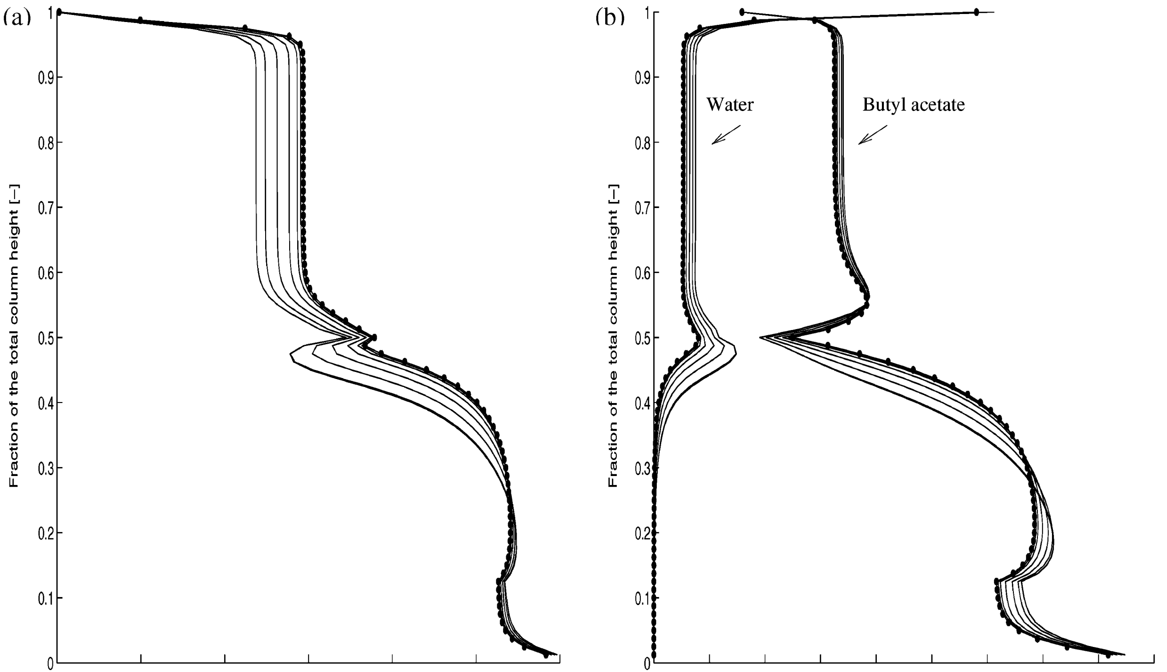


Figure 15. Transient behavior of the reactive-distillation column from an initial profile (denoted by dots) to a new steady state, following a 30% increase in main feed flow.

general tray, by just adding the reaction terms in the material and energy balances.

By virtue of a well-defined hierarchical structuring concept,[23] the major advantage of the library is its reusability. This is shown in what follows by means of two industrial application examples, which have both been modeled relying on exactly the same basic two-phase model library.

4.3.1. Butyl Acetate Production Process. The plant for the production of butyl acetate depicted in Figure 14 serves as a first example for the application of our model library.

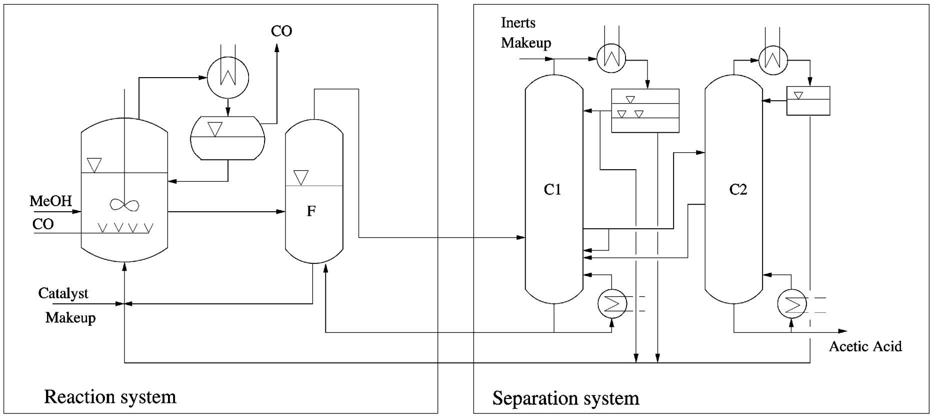


Figure 16. Simplified flowsheet of the process for the production of acetic acid via methanol carbonylation.

In this continuously operated process,[44] liquid *n*-butanol and acetic acetic are fed to column k1 and react in the stripping section of the column to the products butyl acetate and water according to the reversible esterification reaction

$$CH_3COOH + CH_3(CH_2)_3OH \rightleftharpoons CH_3COO(CH_2)_3CH_3 + H_2O$$

which occurs in the liquid phase and is homogeneously catalyzed by means of a strong acid. The more volatile water is boiled up and removed from the reactive section, thus allowing almost total conversion of reactants. At the top of the rectifying section, a vapor mixture rich in water is obtained, which is completely condensed in condenser k1_cond, and the condensate is collected in a decanter dec where a liquid−liquid phase split occurs. The organic phase rich in the ester is recycled to column k1, while the aqueous phase is fed to the stripping column k2 for further purification. Thanks to the formation of an azeotrope in column k2, almost pure water can be withdrawn as the bottoms product, while the reactants are recycled to column k1 via the decanter.

The bottoms stream of column k1 is fed to the distillation column k3 in which an almost pure bottom product is obtained. The second major recycle of the process from the top of column k2 back to column k1 serves for the recovery of unreacted educts.

As depicted in Figure 14, a complete column model is obtained by aggregating individual trays. Similarly, the overall flowsheet model of the complete plant is generated by aggregating and connecting the individual process units by means of the graphical editor in PROMOT.

Figure 15 shows the response of the reactive-distillation column k1 when a disturbance occurs in the main feed flow to the plant. The transition from an initial column profile (indicated by the dots in Figure 15) to a new a steady-state profile following an increase of 30% in feed flow is displayed. Figure 15a illustrates the temperature dynamics after the perturbation. While the rectifying section of the column cools, the temperature increases slightly in the trays near the bottom, as a consequence of the increased concentration of heavy boiling butyl acetate, as shown in Figure 15b.

4.3.2. Production of Acetic Acid. Based on the same library of process unit models used for modeling the butyl acetate process, the process of acetic acid production via methanol carbonylation[45] depicted in Figure 16 was modeled and implemented in PROMOT. The entire plant basically consists of two main parts: The reaction system contains a continuous synthesis reactor with evaporative cooling, in which the liquid reactant methanol and the gaseous reactant carbon monooxide form acetic acid according to the exothermic carbonylation reaction

$$CH_3OH + CO \rightarrow CH_3COOH, \quad \Delta H_R = -138.6\ \text{kJ/mol} \quad (5)$$

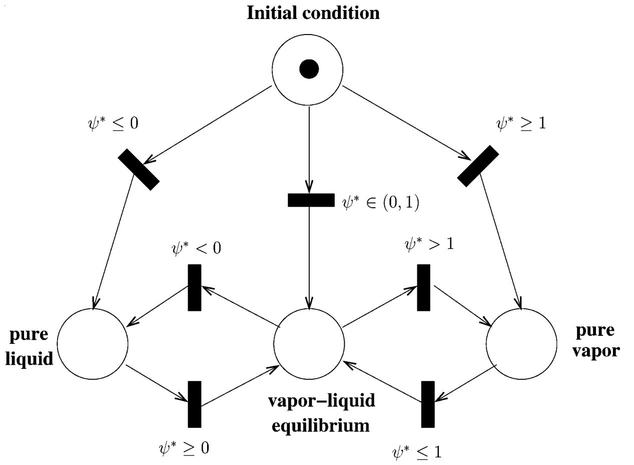


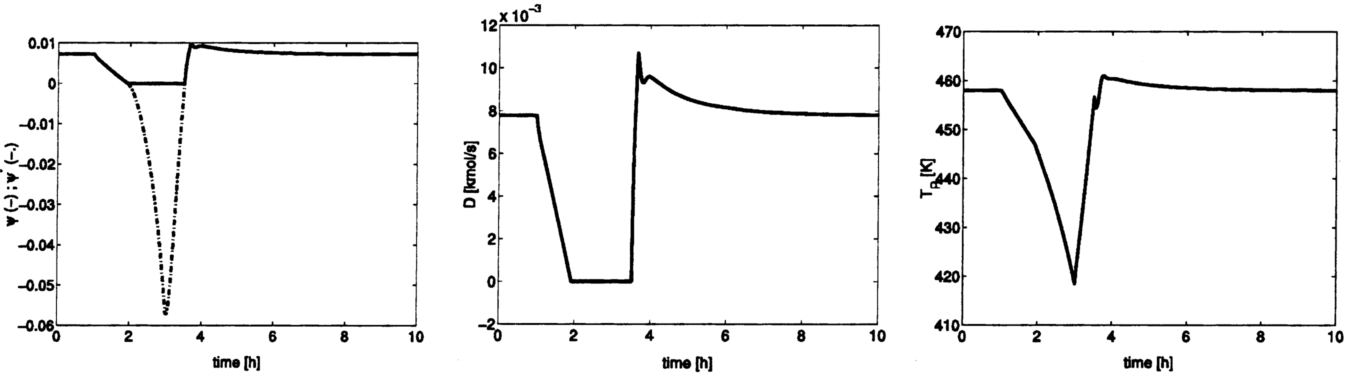
Figure 17. Petri net implementation of discrete events in the acetic acid reactor. Depending on the operating conditions, the reactor is either a pure liquid ($\psi^* < 0$), a vapor−liquid ($0 \leq \psi^* \leq 1$), or a pure vapor ($\psi^* > 1$) system.

In addition, an adiabatic flash separator (F) serves for the recovery of the rhodium-based catalyst, which, as being essentially nonvolatile, is completely recycled to the reactor.

The vapor product stream from the flash is fed to the liquid separation system, the second main part of the overall process. This part basically consists of two distillation columns that serve for the purification of the product acetic acid as well as for the recovery of the inert components methyl iodide (the promotor for the catalyst) and solvent water.

Both main parts of the plant are integrated via a couple of recycle streams from the separation section back to the reaction section.

A particular challenge in modeling the reactor stems from potential phase changes: Although the reactor is in two-phase vapor−liquid equilibrium under standard operating conditions, a disturbance accompanied by a



(a) Vapor fractions $\psi \in [0,1]$ and $\psi^* \in (-\infty, +\infty)$

(b) Vapor flowrate leaving the reactor

(c) Reactor temperature

Figure 18. Implicit discrete events in the reactor following disturbances.

decrease in temperature may cause the system to switch to the purely liquid regime. As explained in section 3, PROMOT offers the possibility to model and efficiently implement such implicit discrete events using Petri nets.

Figure 17 shows the Petri net used to perform the potential switches in the acetic acid synthesis reactor. Without going into modeling details, let us note that a, potentially fictitious, vapor fraction ψ^* determines the thermodynamic state of the system. That is, the real vapor fraction ψ, defined as the ratio of the molar holdup of the vapor phase n'' to the overall molar reactor holdup $n' + n''$, $\psi := n''/n' + n''$, coincides with ψ^* in the two-phase regime ($\psi = \psi^*$ for $\psi^* \in [0, 1]$); it is zero in the pure liquid regime ($\psi^* < 0$) and equal to one in the case of a pure vapor system ($\psi^* > 1$). Depending on the value of ψ^*, the Petri net performs the switches between the different regimes of the system.

Figure 18 illustrates the response of the acetic acid plant after a temporary failure of the feed preheater, which under normal operating conditions is used to control the reactor temperature. The failure leads to a drop in reactor temperature that is accompanied by a drop in reactor vapor fraction up to the point that the system is no longer at boiling conditions (i.e., the implicit discrete event $\psi^* < 0$ is triggered), and the vapor flow rate leaving the reactor becomes zero. Once the preheater works again, the temperature controller drives the system back to the desired steady-state operating conditions in the two-phase regime ($\psi^* \geq 0$).

Summarizing, we can say that, on the basis of our library of process units models and the modeling strategy explained in detail in a previous publication,[23] the modeling of large plants can be managed by a single modeler within a reasonable amount of time. Above all, employing a top-down modeling approach as opposed to a pure bottom-up strategy has been identified to be particularly effective whenever quick refinements of an existing model are required.

5. Conclusions

Computer assistance during the model development process can accelerate the modeling process and improve the quality of the resulting models. Three key components have been identified for the successful application of computer-aided modeling: (i) A structuring methodology is needed that permits a uniform, consistent, and systematic way to formulate different kinds of process models. (ii) A software tool must be available that is able to convert the structured model information into running program code suitable for numerical analysis in some flow-sheet simulator. (iii) A model library must exist that is based on the theoretical concepts of model structuring and that is implemented in the modeling tool. The model library must be comprehensive enough to enable a user to create his/her own model from predefined building blocks.

The work reported in this contribution addresses all of the three fields. The network theory provides a well-developed structuring methodology that is applicable to different types of process models, as the examples from biology and chemical engineering show. The process modeling tool PROMOT permits a direct realization of the theoretical structuring concept. It is able to handle models of high order and complexity. Using its object oriented modeling language, modeling experts can implement new models very efficiently. Due to its graphical user interface, PROMOT is also a tool that can be used conveniently by nonspecialists in the area of process modeling. The feasibility of the structuring approach and of the modeling concept could be demonstrated for quite different applications in the area of systems biology and chemical engineering. Currently, the database of models implemented in PROMOT is increasing, and the software is used more and more in the framework of ongoing research projects.

Future challenges will be extensions of the concepts and the tools to more complicated models, especially distributed systems with multiple dimensions. An example is coupled systems involving fluid dynamics and property coordinates, as they become more and more important for biological and chemical engineering applications.

Acknowledgment

The financial support of the Deutsche Forschungsgemeinschaft in the framework of joint research Projects FOR 447 and SFB 412 is gratefully acknowledged.

Literature Cited

(1) Stephanopoulos, G.; Henning, G.; Leone, H. MODEL.LA. A modeling language for process engineering. I. The formal framework. *Comput. Chem. Eng.* **1990**, *14*, 813.

(2) Ponton, J.; Gawthrop, P. Systematic construction of dynamic models for phase equilibrium processes. *Comput. Chem. Eng.* **1991**, *15*, 803.

(3) Perkins, J.; Sargent, R.; Vázquez-Román, R.; Cho, J. Computer generation of process models. *Comput. Chem. Eng.* **1996**, *20*, 635.

(4) Marquardt, W. Trends in computer aided modeling. *Comput. Chem. Eng.* **1996**, *20*, 591.

(5) Preisig, H. Computer-aided modelling: two paradigms on control. *Comput. Chem. Eng. (Suppl.)* **1996**, S981.

(6) Gilles, E. D. Network theory for chemical processes. *Chem. Eng. Technol.* **1998**, *21*, 121.

(7) Stephanopoulos, G.; Henning, G.; Leone, H. MODEL.LA. A modeling language for process engineering. II. Multifaceted modeling of processing systems. *Comput. Chem. Eng.* **1990**, *14*, 847.

(8) Andersson, M. Omola—An object-oriented language for model representation. Ph.D. Thesis, Lund Institute of Technology, 1990.

(9) Mattson, S.; Elmquist, H.; Otter, M. Physical system modeling with Modelica. *Control Eng. Pract.* **1998**, *6*, 501.

(10) Jensen, A.; Gani, R. A computer aided modeling system. *Comput. Chem. Eng. (Suppl.)* **1999**, *23*, S673.

(11) Westerweele, M.; Preisig, H.; Weiss, M. Concept and design of Modeller, a computer-aided modelling tool. *Comput. Chem. Eng. (Suppl.)* **1999**, S751.

(12) Tränkle, F.; Zeitz, M.; Ginkel, M.; Gilles, E. D. ProMot: A modeling tool for chemical processes. *Math. Comput. Modell. Dynam. Syst.* **2000**, *6*, 283.

(13) Bogusch, R.; Marquardt, W. Computer-aided process modeling with ModKit. *Comput. Chem. Eng.* **2001**, *25*, 963.

(14) Mendes, P. Biochemistry by numbers: simulation of biochemical pathways with GEPASI 3. *Trends Biochem. Sci.* **1997**, *22*, 361.

(15) Sauro, H. M. Jarnac: a system for interactive metabolic analysis. In *Animating the Cellular Map, 9th International BioThermoKinetics Meeting*; Hofmeyr, J.-H. S., Rohwer, J. M., Snoep, J. L., Eds.; Stellenbosch University Press: South Africa, 2000.

(16) Schaff, J.; Fink, C. C.; Slepchenko, B.; Carson, J. H.; Loew, L. M. A general computational framework for modeling cellular structure and function. *Biophys. J.* **1997**, *73*, 1135.

(17) Goryanin, I.; Hodgman, T.; Selkov, E. Mathematical simulation and analysis of cellular metabolism and regulation. *Bioinformatics* **1999**, *15*, 749.

(18) Tomita, M.; Hashimoto, K.; Takahashi, K.; Shimizu, T. S.; Matsuzaki, Y.; Miyoshi, F.; Saito, K.; Tanida, S.; Yugi, K.; Venter, J. G.; Hutchison, C. A. E-CELL: software environment for whole-cell simulation. *Bioinformatics* **1999**, *15*, 72.

(19) Ginkel, M.; Kremling, A.; Nutsch, T.; Rehner, R.; Gilles, E. D. Modular modeling of cellular systems with ProMoT/Diva. *Bioinformatics* **2003**, *19*, 1169.

(20) Hucka, M.; et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **2003**, *19*, 524.

(21) Tränkle, F.; Kienle, A.; Mohl, K.; Zeitz, M.; Gilles, E. D. Object-oriented modeling of distillation processes. *Comput. Chem. Eng. (Suppl.)* **1999**, *23*, S743.

(22) Mangold, M.; Motz, S.; Gilles, E. D. Network theory for the structured modelling of chemical processes. *Chem. Eng. Sci.* **2002**, *57*, 4099.

(23) Waschler, R.; Angeles-Palacios, O.; Ginkel, M.; Kienle, A. Application of the Process Modeling Tool ProMoT to large-scale chemical engineering processes. *Proceedings of the 4th MATHMOD, IMACS Symposium on Mathematical Modelling*, Feb 5−7, 2003; Vienna University of Technology: Vienna, Austria, 2003; Vol. 2.

(24) Mangold, M.; Ginkel, M.; Gilles, E. D. A model library for membrane reactors implemented in the process modelling tool ProMoT. *Comput. Chem. Eng.* **2004**, *28*, 319.

(25) Hanke, R.; Mangold, M.; Sundmacher, K. Application of hierarchical process modelling strategies to fuel cell systems − towards a virtual fuel cell laboratory. *Fuel Cells—From Fundamentals to Systems*, 2004, accepted.

(26) Kremling, A.; Gilles, E. D. The organization of metabolic reaction networks: II. Signal processing in hierarchical structured functional units. *Metab. Eng.* **2001**, *3*, 138.

(27) Kremling, A.; Bettenbrock, K.; Laube, B.; Jahreis, K.; Lengeler, J.; Gilles, E. D. The organization of metabolic reaction networks: III. Application for diauxic growth on glucose and lactose. *Metab. Eng.* **2001**, *3*, 362.

(28) Kremling, A.; Jahreis, K.; Lengeler, J.; Gilles, E. D. The organization of metabolic reaction networks: A signal-oriented approach to cellular models. *Metab. Eng.* **2000**, *2*, 190.

(29) Saez-Rodriguez, J.; Kremling, A.; Gilles, E. D. Dissecting the Puzzle of Life: Modularization of Signal Transduction Networks. *Comput. Chem. Eng.* **2004**, accepted.

(30) Saez-Rodriguez, J.; Kremling, A.; Conzelmann, H.; Bettenbrock, K.; Gilles, E. D. Modular Analysis of Signal Transduction Networks. *IEEE CSM special issue* **2004**, *24*, 35−52.

(31) Mohl, K. D.; Spieker, A.; Köhler, R.; Gilles, E. D.; Zeitz, M. DIVA − A simulation environment for chemical engineering applications. *ICCS Collect. Vol. Sci. Pap.*; Donetsk State Techn. University: Ukraine, 1997.

(32) Mathworks Inc., Matlab and Simulink, http://www.mathworks.com, 2004.

(33) Andreu, D.; Pascal, J.; Valette, R. Interaction of discrete and continuous parts of a batch process control system. Proceedings of the Workshop on Analysis and Design of Event-Driven Operations in Process Systems (ADEDOPS), London, U.K., 1995.

(34) Henzinger, T. The Theory of Hybrid Automata. Proceedings of the 11th Annual IEEE Symposium on Logic in Computer Science (LICS '96), New Brunswick, NJ, 1996.

(35) Tarjan, R. Depth first search and linear graph algorithms. *SIAM J. Comptg.* **1972**, *1*, 146.

(36) Köhler, R.; Räumschüssel, S.; Zeitz, M. Code Generator for Implementing Differential Algebraic Models Used in the Process Simulation Tool DIVA. Proceedings of the 15th IMACS World Congress, Berlin, 1997.

(37) Monod, J. *Recherches sur la croissance des cultures bacterienne*; Herrmann: Paris, 1942.

(38) Postma, P. W.; Lengeler, J. W.; Jacobson, G. R. Phosphoenolpyruvate: carbohydrate phosphotransferase systems of bacteria. *Microbiol. Rev.* **1993**, *57*, 543.

(39) Kolios, G.; Frauhammer, J.; Eigenberger, G. Autothermal fixed-bed reactor concepts. *Chem. Eng. Sci.* **2000**, *55*, 5945.

(40) Wicke, E.; Vortmeyer, D. Zündzonen heterogener Reaktionen in gasdurchströmten Körnerschichten. *Z. Elektrochem.* **1959**, *63*, 145.

(41) Matros, Y. *Catalytic Processes under Unsteady-State Conditions*; Elsevier: Amsterdam, 1989.

(42) Lauschke, G.; Gilles, E. D. Circulating reaction zones in a packed-bed loop reactor. *Chem. Eng. Sci.* **1994**, *49*, 5359.

(43) Mangold, M.; Klose, F.; Gilles, E. D. Dynamic behavior of a counter-current fixed-bed reactor with sustained oscillations. In *European Symposium on Computer Aided Process Engineering—ESCAPE-10*; Pierucci, S., Ed.; Elsevier: Amsterdam, 2000.

(44) Hartig, H.; Regner, H. Verfahrenstechnische Auslegung einer Veresterungskolone. *Chem.-Ing.-Tech.* **1971**, *18*, 1001.

(45) Waschler, R.; Kienle, A.; Anoprienko, A.; Osipova, T. Dynamic Plantwide Modelling, Flowsheet Simulation and Nonlinear Analysis of an Industrial Production Plant. In *European Symposium on Computer Aided Process Engineering—12—ESCAPE-12*, The Hague, The Netherlands, May 26−29, 2002; Grievink, J., van Schijndel, J., Eds.; Elsevier: Amsterdam, 2002.

Received for review April 30, 2004
Revised manuscript received September 29, 2004
Accepted September 30, 2004

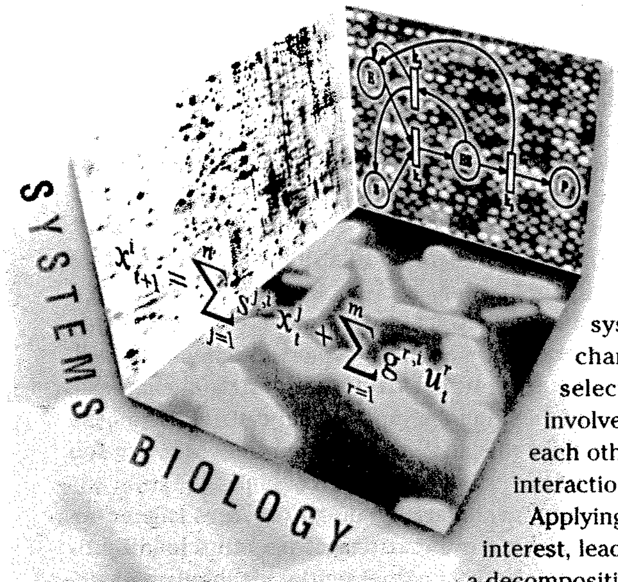
IE0496434

FEATURE

By Julio Saez-Rodriguez, Andreas Kremling,
Holger Conzelmann, Katja Bettenbrock,
and Ernst Dieter Gilles

Modular Analysis of Signal Transduction Networks

How engineering tools can be applied to the analysis of cellular machinery


SYSTEMS BIOLOGY
$x^i_{t+1} = \sum_{j=1}^{n} s^{i,j} x^j_t + \sum_{r=1}^{m} g^{r,i} u^r_t$

The study of detailed models for intracellular networks has become popular in recent years. New experimental techniques provide significant amounts of data that can be used to develop detailed models of intracellular networks, including signal transduction pathways. In the simplest case, the two-component system, the signaling system consists of two elements: a sensor that detects environmental changes and a regulatory element that influences the transcription of selected genes. In higher cells, however, signal transduction networks involve components that are embedded in feedback loops and interact with each other. The number of elements involved and their complicated nonlinear interactions give rise to a picture of significant complexity.

Applying system-theoretic tools to biological systems has attracted increasing interest, leading to the emergence of the field of systems biology [1]. In particular, a decomposition into smaller units or modules, as well as the subsequent analysis of the resulting elements, has been proposed as a useful tool for shedding light on the rationale of signaling networks [2]. How these modules should be defined, however, remains an open question [3]. Furthermore, thorough systems-theory-oriented analyses of signaling networks based on their modularity are still scarce.

This article is organized as follows. First, we describe the mechanisms that cells have developed to process information. Second, we discuss the decomposition of signaling networks into subsystems. We then present a novel criterion for defining modules based on the absence of retroactivity [4]. Some simple criteria for the analysis of the resulting units are introduced. Finally, we apply these tools to several examples.

Cellular Signal Transduction

Cells are autonomous entities. To function as single-celled organisms or as part of a higher multicellular organism, cells must sense their environment and must be able to react to it. As a result, cells are equipped with a wealth of sensor systems that allow

0272-1708/04/$20.00©2004IEEE

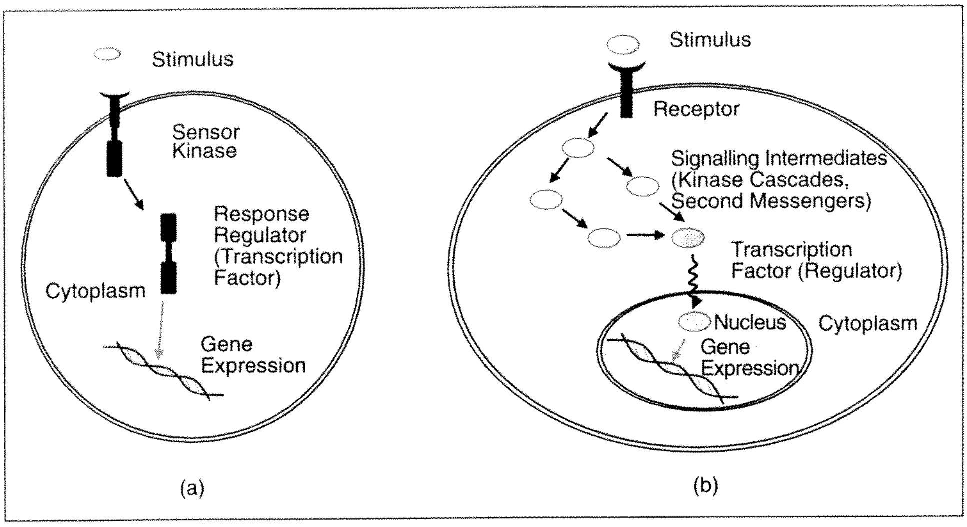

Stimulus
Sensor Kinase
Response Regulator (Transcription Factor)
Cytoplasm
Gene Expression
(a)
Stimulus
Receptor
Signalling Intermediates (Kinase Cascades, Second Messengers)
Transcription Factor (Regulator)
Nucleus
Cytoplasm
Gene Expression
(b)

Figure 1. *Simplified schemes for general signal transduction systems. (a) illustrates a two-component system in a prokaryotic cell, and (b) illustrates a receptor-coupled signal transduction system in a eukaryotic cell. Most of the signaling pathways can be described according to these schemes, despite the high variability among them.*

them to monitor external and internal states with sophisticated signal-processing units that secure the optimal reaction for these conditions.

Cells are typically surrounded by a plasma membrane, which is impermeable to most chemicals and represents a barrier necessary to maintain the autonomy of the cell and to protect it from external stresses. Since the membrane represents the barrier of the cell, it is the place where external conditions are sensed. The monitoring of external conditions is important for securing survival, especially in the case of single-celled organisms and for communicating with other cells, which is important for guaranteeing the functioning of an organism. The sensing of stimuli at the cell membrane demands the transfer of the signal to the place of action, a process called signal transduction. The target may be either enzymes within the cytoplasm (the part of the cell contained within the plasma membrane) or the DNA.

Enzymes, which are proteins that act as catalysts for biochemical reactions, are modified (for example, by phosphorylation as described below) so that their catalytic activities are increased or decreased in response to the extracellular signal. For example, the presence of a substrate can trigger a signaling pathway that leads to an increase in the activity of the enzyme, which in turn decomposes the substrate.

For DNA, the signal transduction process targets transcription factors, which are proteins that regulate gene expression; that is, the processes through which the information contained within the DNA is realized into new proteins. In simple organisms such as bacteria, the DNA lies freely in the cytoplasm; these organisms are called prokaryotes. On the other hand, cells of higher organisms, such as mammalian cells, have a compartment, the nucleus, where the DNA is located. These organisms are known as eukaryotes (see Figure 1). An important element is the operon, which is a group of genes that lie side by side on the DNA and are translated as one unit. The resulting variations in gene expression allow the cell to modify its enzymatic activities and metabolism according to the input signal. For example, the genes can express enzymes that protect the cell from stress, such as starvation or extreme temperature, upon receiving a stress signal.

In addition to the sensing and transduction of a signal, the term signal transduction traditionally includes the processing of signals. Correct processing is necessary to ensure the optimal response to the set of external and internal states the cell is facing. Branching and linkage of signal transduction pathways, a phenomenon known as crosstalk, are involved in signal processing. Signal processing is also performed at the level of gene expression, with multiple regulators modulating the expression of one gene.

The cytoplasm is structured into different areas. A signal might therefore be limited to a certain area of the cell, especially in eukaryotic cells. Due to crosstalk phenomena, as well as the high number and spatial distribution of the elements involved, signal transduction networks are too complex to be understood by intuitive thinking [5].

Information in cellular signaling processes is generally transferred by modifications of proteins leading to changes in their activity. An essential mechanism of signal transduction is the addition of a phosphate group to a protein, a process known as phosphorylation, which produces a conformational change in the protein that alters its activity. Proteins are chains of small molecules known as amino acids. Phosphorylation can occur in defined amino acids within a protein and is catalyzed by specific enzymes, called kinases, in response to a stimulus.

Phosphorylations can be reversed by specific enzymes called phosphatases. The activation and inactivation of phosphatases is also the result of signal transduction processes. This additional control mechanism increases the flexibility of signal transduction. There are often multiple phosphorylation sites within one protein, and the phosphorylation state determines the activity of the protein, again enhancing flexibility. Another typical mechanism used to transfer information in cells is physical interaction between proteins. Many proteins involved in signaling processes have sequences of amino acids, known as domains, which can bind to domains in other proteins, leading to the association of molecules. Combinations of these mechanisms are frequently present. For example, protein-protein interactions mediated by protein phosphorylation are a common element in signal transduction: a complex between two proteins will be only formed if the proteins are in the right conformation, such as when one of the proteins is in a phosphorylated state and the other in an unphosphorylated state.

Cells are equipped with a wealth of sensor systems that allow them to monitor external and internal states with sophisticated signal-processing units that secure the optimal reaction for these conditions.

Two-Component Signal Transduction and Phosphorelay Systems

The two-component system is regarded as the simplest signal transduction system [6], [7]. The system is widely used in prokaryotic signal transduction and can also be found in eukaryotes. In general, the system is composed of two proteins: the sensor kinase and the response regulator. The sensor kinase is often an integral membrane protein with the sensor domain located outside the cell. Sensing of a stimulus by the extracellular sensor domain provokes a change in the conformation of the sensor kinase. The conformational change leads to an increase in the kinase activity of the intracellular kinase domain, resulting in autophosphorylation. The phosphoryl group is subsequently transferred to the receiver domain of the response regulator. Phosphorylation results in a change of the activity of the regulator domain. The response regulator normally represents a DNA-binding protein that activates gene expression in response to phosphorylation, but it can also be an enzyme that changes its activity or regulates another enzyme (see Figure 1).

The two-component signaling pathway is linear, and this linearity yields a high specificity for signal transduction because the phosphotransfer normally occurs only between corresponding pairs (called cognate pairs) of sensor kinase and response regulator. Receiver dephosphorylation can be achieved by several routes. There is the inherent instability of the phosphate that limits the response time, but there can also be active dephosphorylation catalyzed by the sensor kinase. Typical examples of this kind of system are the PhoR/PhoB and the EnvZ/OmpR systems of the bacteria *Escherichia coli* (*E. coli*), as well as the KdpD/KdpE system.

In addition to these cases, researchers have identified two-component systems that possess the same principal organization but contain additional modules. Referred to as phosphorelay systems, these systems often contain additional phosphorylatable domains. These additional domains enhance the flexibility of signal transduction and allow the modulation of one system through phosphorylation by another system. Different organizations with respect to the arrangement of the different domains as independent proteins or as part of multidomain proteins are realized.

Not every signal transduction pathway is designed solely for its signal transduction purpose. The bacterial phosphoenolpyruvate-dependent phosphotransferase system (PTS) is an example of a signal transduction pathway that shares proteins with a metabolic pathway [8], which is a series of enzymatic reactions involved in the synthesis or breakdown of molecules. The metabolic property of the PTS is the uptake and phosphorylation of carbohydrates. Carbohydrate uptake in *E. coli* has been investigated for a long time. The most prominent observation is that a mixture of sugars, for example, glucose and lactose, provided in a batch experiment are taken up sequentially: glucose is taken up preferentially and after the depletion of glucose, lactose is taken up. This phenomenon is due to a complex signal transduction and control system where the PTS plays a central role.

Receptor-Coupled Signal Transduction

Eukaryotic signal transduction pathways are often more complex than prokaryotic ones due to a higher number of elements and the numerous interconnections between pathways. The activity of one pathway can be modulated by another pathway depending on its signaling state. This process, known as crosstalk, is mediated by proteins that are part of more than one signaling pathway. A common theme in eukaryotic signaling is receptor-coupled signal transduction. Receptors are membrane proteins with a signal recognition domain located at the outer surface of the cell. A transmembrane domain couples the signal recognition domain to the intracellular transmitter domain. Binding of the signaling molecule known as a ligand, which may be a hormone or growth factor, to the cell-surface receptor induces conformational changes in the receptor, which trigger a cascade of reactions leading to a particular cellular response [9] (see Figure 1).

An important family of receptors are the receptor tyrosine kinases (RTKs), which share many elements and mechanisms [10]. When a ligand binds an RTK, the ligand causes pairwise binding of the receptor proteins to create dimers, a process called dimerization, resulting in the activation of the intracellular tyrosine kinase [10]. The kinase can then phosphorylate either the receptor itself or a substrate protein. The phosphorylated residues are binding sites for several proteins.

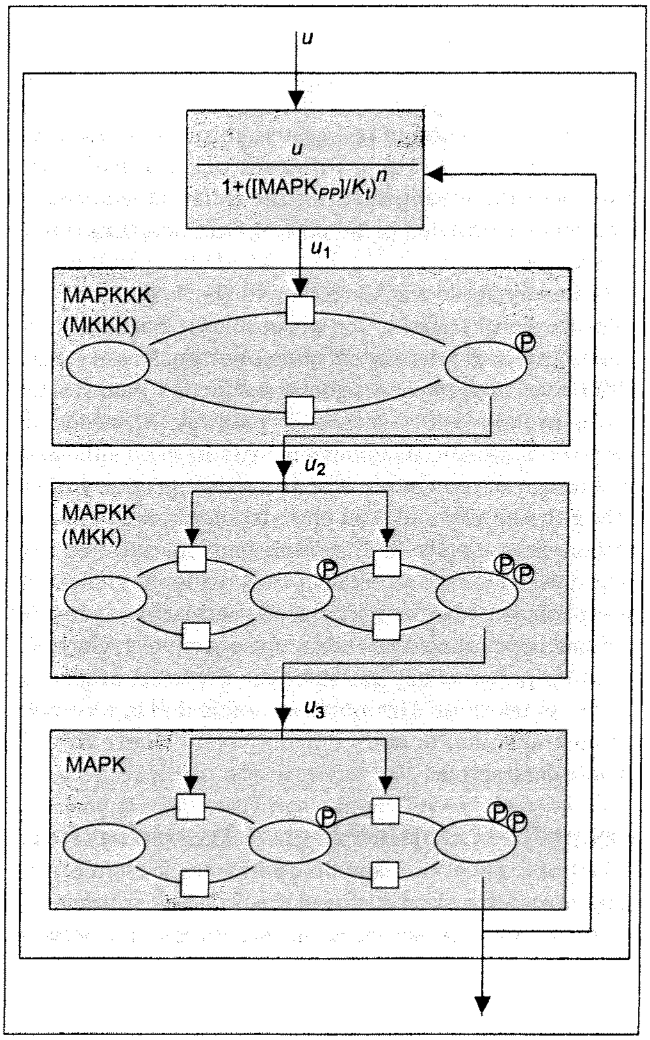


Figure 2. *Modular representation of the MAPK cascade. The cascade can be divided into three subunits corresponding to the three kinases involved (MAPKKK, MAPKK, and MAPK). A negative feedback is included as in [24]. If the reactions are assumed to follow Michaelis-Menten kinetics, as described in "Model Equations and Parameters," the connections between the modules are free of retroactivity [4].*

Key elements in signaling processes are the adaptor proteins, which have no enzymatic activity but have different domains that allow them to bind to other proteins [10]. The formation of these complexes brings together different proteins, allowing the interaction between certain elements and, hence, initiating signaling pathways. The activity of the complex can be reduced by dephosphorylation of the phosphotyrosine residues due to specific phosphatases. Adaptor proteins also play an important role in crosstalk phenomena among pathways. For example, an adaptor protein that has been phosphorylated by complex *A* might bind to complex *B*, inducing a conformation of the receptor complex *B*, thus changing the activity of *B*.

In these pathways the signal can be amplified: one or two molecules of ligand are needed for the formation of one receptor complex, and since the receptor complex is stable as long as the ligand molecule is bound, the corresponding enzymes are active for a long period of time and thus are able to produce high amounts of activated signaling molecules.

Downstream of the activated receptor complex there are two main possibilities for the intracellular transmission of the signal: the synthesis of second messengers or the activation of a cascade of successive protein kinases (see Figure 1).

Second messengers can be found in eukaryotes as well as in prokaryotes. These signaling molecules allow the transduction of the signal to different areas of the cell because they are easily diffusible. Especially in eukaryotic cells, gradients of second messengers in time and space might be present if a signal is specific to a certain area of the cell. Thanks to the high number of molecules and their structure, second-messenger molecules are able to interact with a diverse set of proteins.

A classic example of a cascade of kinases is the mitogen-activated protein kinase (MAPK) cascade. The MAPK cascade is a set of three kinases that activate each other sequentially, as depicted in Figure 2 and discussed below.

Modularization of Signaling Networks

As mentioned above, a decomposition into smaller elements might be an interesting approach to handle the complexity of cellular processes and signaling phenomena. First, functional modules, elements whose function is separable from those of other modules, have to be defined. Next, the modules should be thoroughly analyzed, regarding properties such as transfer functions to characterize the dynamical behavior, signal amplitude, or robustness with respect to kinetic parameters. A successful analysis can lead to a reduction of models if key elements can be identified and the main properties can be reproduced without having to model all the biological players. Finally, by regrouping the modules, the properties of the system as a whole become clearer.

Although the modularity of biological processes is generally accepted, a distinctive criterion for defining modules is still lacking. Different proposals, such as evolutionary conservation, robustness, and genetic coexpression, have been suggested [3]. We have recently introduced an alternative criterion for the delimitation of modules, based on the absence of retroactivity in the junction between different units [4] (see Figure 3). The so-defined modules possess the sense of independency inherent to the concept of functional units since the input/output behavior of a retroactivity-free unit does not depend on what it is connected to. Additionally, systems theory provides a battery of tools for analyzing systems free of retroactivity. The framework we use is based on network theory [11], which can be conveniently applied to biological systems [12]. Using network theory, the different cases of absence of retroactivity can be represented and analyzed [4].

In the analysis of signaling networks, special attention should be paid to their dynamic behavior since the biological response is often determined by the transient characteristics of the output signal, such as signaling time and signal duration, rather than steady-state properties [5]. Therefore, the classical steady-state analysis of such systems, though useful, might not provide sufficient insight into the properties of signaling pathways. Recently, Heinrich and colleagues [13] have comprehensively analyzed kinase-phosphatase cascades from a theoretical point of view. The analysis is based on the signal amplitude S, the signaling time τ, and the signal duration θ. For the output $y(t)$ of a module, these parameters can be calculated numerically according to [13] by means of

$$\tau = \frac{\int_0^\infty t\,y(t)\,dt}{\int_0^\infty y(t)\,dt}, \qquad \theta = \sqrt{\frac{\int_0^\infty t^2\,y(t)\,dt}{\int_0^\infty y(t)\,dt} - \tau^2},$$

$$S = \frac{\int_0^\infty y(t)\,dt}{2\,\theta}.$$

The definitions of τ and θ are analogous to the mean value and standard deviation of a statistical distribution, respectively. Therefore, τ and θ represent the average time to activate the output element and the average time during which this output component is activated, respectively. S gives the relationship between the total amount of output signal, the area under the curve, and the duration θ of the signal, hence providing a measurement of the average concentration of the output element [13]. These parameters are reasonable but if the output signal does not return to zero after a certain time (known in biology as adaptation), then τ and θ tend to infinity. Since the signal does not return to zero in some of the examples discussed here, another parameter, denoted by $\tau_{0.9}$, is used to measure how fast a system responds. The parameter $\tau_{0.9}$ is defined as the time at which 90% of the maximal output signal is reached, and, hence, can be more generally applied. The parameter $\tau_{0.9}$ was determined numerically based on the simulation data. Additionally, if the signal does not return to the basal level, the signal amplitude S depends on the simulation time. If the system reaches a steady state different from zero, it can easily be demonstrated that, as time tends to infinity, S converges to the steady-state value multiplied by $\sqrt{3}$. Therefore, when the output signal does not return to zero, we have computed the parameters for time tending to infinity.

Another useful parameter is the Hill coefficient. The Hill equation [14] describes enzyme kinetics and is defined as

$$r = \frac{r_{\max} x^h}{K_{0.5}^h + x^h}, \tag{1}$$

where r is the reaction rate, x is the substrate concentration, $r_{\max}$ is the maximal reaction rate, $K_{0.5}$ is the concentration of substrate for which $r = r_{\max}/2$ and can be considered the threshold value, and h is the Hill coefficient. If $h = 1$, then the curve is hyperbolic and is known as the Michaelis-Menten equation, as described in "Model Equations and Parameters." If $h > 1$, then the curve shows a sigmoidal form, a situation known as ultrasensitivity [15]. The higher the Hill coefficient, the more the curve tends to a step-form response. The Hill coefficient is thus a measurement of the ultrasensitivity and can be more generally applied to quantify the sigmoidity of the input/output behavior of a system, where in (1) x is the input and r is the output of the module or another variable of interest. The Hill coefficient can be estimated as $h = \log 81/(\log(S_{0.9}/S_{0.1}))$, where $S_{0.9}$ and $S_{0.1}$ are the substrate concentration values where r is 90% and 10% of the maximum ($r_{\max}$), respectively [15].

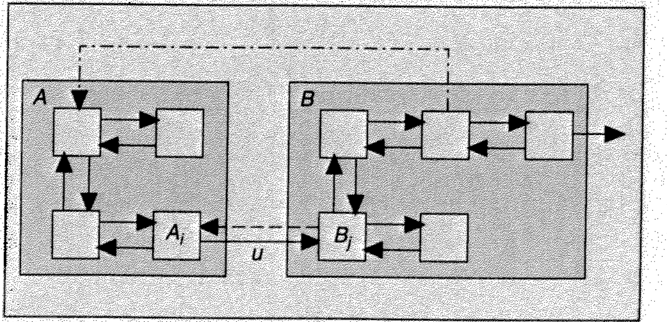


Figure 3. *Schematic representation of the concept of retroactivity. If the submodule A_i of the module A influences the submodule B_j of the module B (solid line) but the module B_j does not directly influence A_i (dashed line), then the connection between A and B is free of retroactivity. A feedback path from another part of B to A (dashed-dotted line) does not change the input/output behavior of module B but restricts the range of possible values for the input u.*

Model Equations and Parameters

Signal transduction processes are often modeled by ordinary differential equations, assigning one equation to the mass balance of each component. Generally, the reaction kinetics are modeled using the law of mass action. For example, a reversible reaction $A \rightleftharpoons B$ has a rate $r = kA - k_r B$, where k and k_r are kinetic constants and, with a slight abuse of notation, A and B denote the concentrations of the species A and B, respectively. Phosphorylation, or in general, reactions catalyzed by an enzyme, is usually described as

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} P + E,$$

where each reaction is modeled with mass action kinetics. In many cases, however, it is assumed that the dynamics of the complex ES can be neglected, which yields the simpler equation $S \rightarrow P$, where the reaction rate r is given by the Michaelis-Menten equation

$$r = \frac{v_{\max} \cdot S}{K_m + S} = \frac{k_2 \cdot E_0 \cdot S}{K_m + S},$$

where $K_m = (k_{-1} + k_2)/k_1$ and $E_0 = E + ES$ [14]. The Michaelis-Menten equation is widely used in the modeling of biological systems, even if it is not known whether the corresponding assumption holds because with Michaelis-Menten kinetics only two parameters are needed for each reaction instead of three.

The following equation system holds true for the two-component system:

$$\begin{aligned}
\frac{d\,S^p}{dt} &= r_1 - r_2 \\
\frac{d\,R^p}{dt} &= r_2 - r_{3a/b} - n\,r_4 \\
\frac{d\,\mathrm{DNA}_f}{dt} &= r_4 \\
\frac{d\mathrm{RNA}}{dt} &= k_{tr}\,\psi\,\mathrm{DNA}_0 - (k_z + \mu)\,\mathrm{RNA} \\
\frac{dS_0}{dt} &= k_{tl1}\,\mathrm{RNA}, - (k_d + \mu)S_0 \\
\frac{dR_0}{dt} &= k_{tl2}\,\mathrm{RNA}, - (k_d + \mu)R_0
\end{aligned}$$

and

$$\frac{dF}{dt} = k_{tl3}\,\mathrm{RNA}, - (k_d + \mu)F.$$

The rate equations read

$$\begin{aligned}
r_1 &= k_1\,S\,\mathrm{ATP} - k_{-1}\,S^p\,\mathrm{ADP} \\
r_2 &= k_2\,S^p\,R - k_{-2}\,S\,R^p \\
r_{3a} &= k_3\,R^p\,E_0 \\
r_{3b} &= k_3\,R^p\,S
\end{aligned}$$

and

$$r_4 = k_b\,R^{pn}\,\mathrm{DNA}_f - k_{-b}\,\mathrm{RDNA},$$

where E_0 represents the concentration of the phosphatase. States S, R, and RDNA can be calculated when the conservation equations $S_0 = S + S^P$, $R_0 = R + R^P + n\mathrm{RDNA}$, and $\mathrm{DNA}_0 = \mathrm{DNA}_f + \mathrm{RDNA}$ are considered. The parameter ψ describes the transcription efficiency according to [37]. The method takes into account that the RNA polymerase alone can begin the transcription but the formation of the complex between the transcription factor R^p and the polymerase increases the transcription efficiency. Note that different parameters k_{tl_i} are used for protein, sensor kinase, and response regulator to describe translation. For the open-loop model variant, the overall concentrations S_0 and R_0 are constant. All of the parameters are summarized in Table 1.

The MAPK cascade model is adapted from a model from Kholodenko [24] using Michaelis-Menten kinetics. The differential equations and parameters can be found in [24].

The model for the EGF-induced MAPK cascade was introduced in [30]. All of the reactions are modeled using mass action kinetics. The model can be downloaded from http://www.mpi-magdeburg.mpg.de/model/EGF/.

Two-Component Signal Transduction

Two-component systems are frequently used by bacteria to sense environmental conditions. As described above, these systems consist of a sensor kinase and a response regulator. Upon perception of a stimulus, the input domain of the sensor kinase modulates the signaling activity of its transmitter domain, resulting in autophosphorylation with the help of adenosine triphosphate (ATP). The phosphoryl group is then transferred to the response regulator, which activates the output to trigger the response.

According to the criterion of absence of retroactivity, the system can be subdivided into two modules, shown in Figure 4 [4]. Module 1 describes the activation of the sensor protein and the phosphoryl transfer to the regulator protein. Module 2 describes the binding of the activated regulator to the respective DNA binding site and the process of gene expression: transcription and translation. In the simplest case, sensor activation and phosphoryl transfer in module 1 can be described by a set of two reactions; in reaction R1 the input stimulus enhances the kinase activity, which causes autophosphorylation of the sensor kinase S by ATP. In reaction R2 the phosphoryl group is transferred to the response-regulator R. Here, the phosphorylated response-regulator R^p contains the active output domain. Although it is known that during enzymatic activities proteins form a number of temporary complexes, the kinetic reactions are kept simple to reduce the number of unknown and uncertain parameters. Hence, the model is defined by

$$\text{R1:} \qquad \text{ATP} + S \underset{k_{-1}}{\overset{k_1^*}{\rightleftharpoons}} \text{ADP} + S^p$$

and

$$\text{R2:} \qquad S^p + R \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} S + R^p.$$

The respective rates are summarized in "Model Equations and Parameters." In the model, the stimulus is considered by changes in the parameter k_1^* modeled by

$$k_1^* = k_1 u, \tag{2}$$

where u is a constant input, which is dimensionless. To turn off the system efficiently, the phosphoryl group is taken away in a dephosphorylation reaction R3. Here, two model variants are possible: a) the phosphoryl group is hydrolyzed by an additional enzyme possessing phosphatase activity, or b) the sensor kinase itself acts as the phosphatase. The latter is the case for a number of two-component systems in *E. coli*, for example, the KdpD/E system responsible for potassium uptake [16]. It is assumed that the dephosphorylation step is irreversible. The additional equations are given by

$$\text{R3a:} \qquad R^p \overset{r_{3a}}{\rightarrow} R + P$$

and

$$\text{R3b:} \qquad R^p + S \overset{r_{3b}}{\rightarrow} R + S + P.$$

The concentration of the phosphorylated response regulator R^p represents the output of the first module. In

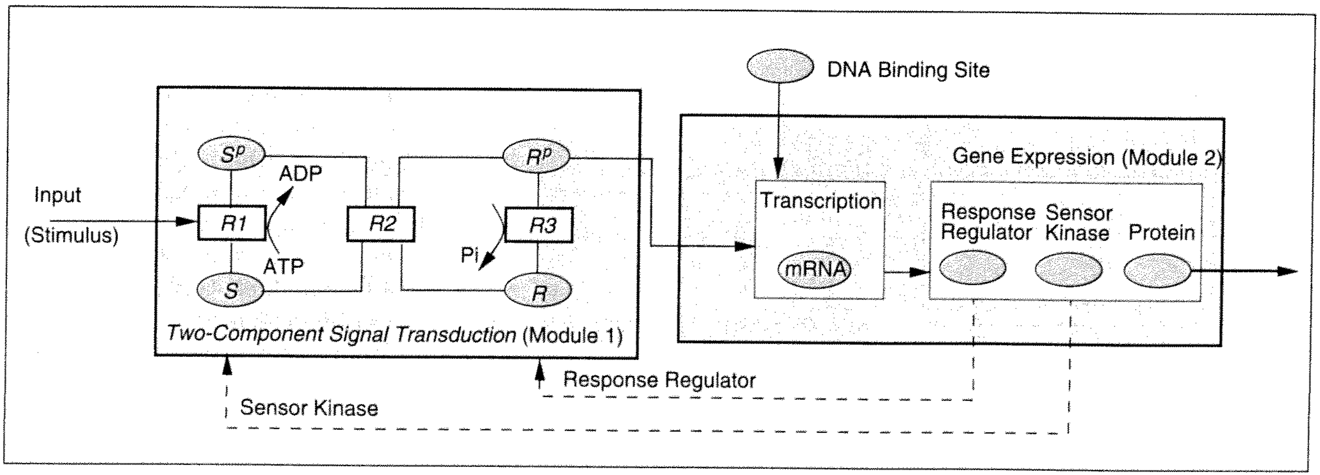


Figure 4. *Decomposition of the signal transduction pathway and gene expression of the two-component system as in [4]. The boxes represent reactions, and the ellipses represent compounds. The input of module 1 is an extracellular signal, such as the concentration of a nutrient or ion. The output of module 1, which is the input of module 2, is the phosphorylated response regulator R^p. The output of module 2 is the target protein. The dashed lines indicate possible positive feedback by the sensor kinase and response regulator.*

module 2 the binding of the phosphorylated regulator to the DNA binding site is described by

$$\text{R4:}\qquad n\,R^p + \text{DNA}_f \underset{k_{-b}}{\overset{k_b}{\rightleftharpoons}} \text{RDNA},$$

where n is the number of molecules that bind to the DNA.

The amount of the regulator DNS complex (RDNA) measures the rate of mRNA synthesis. In a further step, the mRNA is translated to protein Pr, the target protein. The two polymerization steps are connected in cascade, where the mRNA serves as an information input to the second polymerization step. The reaction equations have the form

$$\text{R5:}\qquad (\text{nucleotides}) \xrightarrow{r_{tr}} \text{RNA} \xrightarrow{r_{d1}} \text{degradation}$$

and

$$\text{R6:}\qquad (\text{amino acids}) \xrightarrow{r_{tl}} Pr \xrightarrow{r_{d2}} \text{degradation},$$

where the rate of transcription is r_{tr}, the rate of mRNA degradation is r_{d1}, the rate of translation is r_{tl}, and the rate of protein degradation is r_{d2}. Reaction rates, parameters describing experimental data [17], and the differential equations are summarized in "Model Equations and Parameters."

Information in cellular signaling processes is generally transferred by modifications of proteins leading to changes in their activity.

Dynamics and Steady-State Characteristics

In the first step, module 1 is under investigation. Two model variants are considered here; namely, model 1 describes the step of dephosphorylation with the rate law r_{3a}, while model 2 uses the rate law r_{3b}. The signaling time $\tau_{0.9}$ and the signal amplitude S_i are calculated as described above. For model 1, the parameter E_0 is chosen to be $E_0 = S_0$, the total concentration of the sensor kinase, to allow a comparison between the two model variants.

Figure 5 shows the signaling time and signal amplitude of both models as a function of the input. Model 2, which is always slower than model 1, traverses a maximum in the signaling time near the inflection point of the signal amplitude. Interestingly, this characteristic is not observed for model 1. These differences can be explained by the amount of enzyme that is available for dephosphorylation: since the enzyme concentration for model 1 is constant and always higher than in model 2 ($E_0 = S_0 \geq S$) the steady state is reached earlier but at the expense of high values of R^p, and, thus, the signaling amplitude is always lower for all input values. However, in model 2 the conversion of S into S^p not only increases the phosphorylation of R but also decreases the enzyme S available for dephosphorylation. Therefore, the rise of the signaling amplitude is steeper for model 2 than for model 1.

For both models the Hill coefficient is determined according to (1); model 1 and model 2 have Hill coefficients 1.0 and 1.98, respectively. In comparison with model 1, an advantage of the model 2 circuit is the switchlike behavior for higher input values, which allows the cell to respond with higher sensitivity in a certain input range.

In a second step, both module 1 and module 2 are analyzed (see Figure 4). Model 2 is used for module 1. As can be seen in Figure 6, module 2 reaches the maximal amplitude values even with low values of the input R^P. Based on this separate analysis of the submodels, it can be concluded that the chosen initial conditions for module 1 are sufficient to consistently reach the same target protein concentration. Therefore, the overall

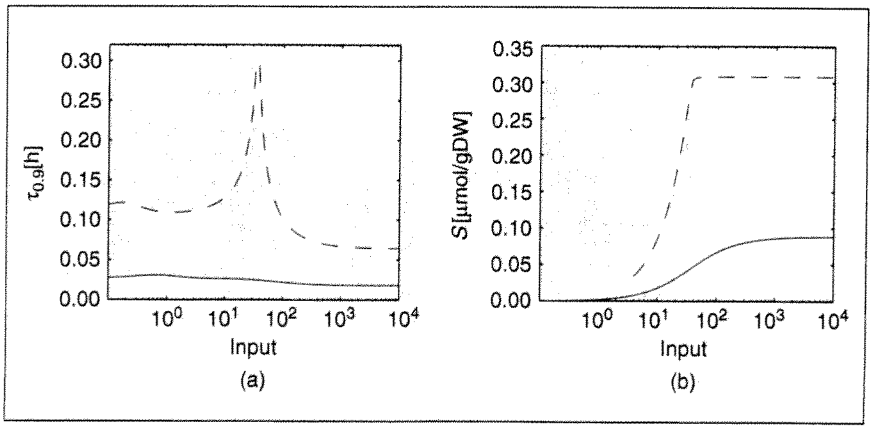


Figure 5. *Comparison of the model variants of the two-component system. The plots illustrate the signaling time $\tau_{0.9}$ and the signal amplitude S for the two-component system for both model variants (model 1 solid, model 2 dashed) for module 1. For model 2, both the signal amplitude and signaling time are higher. The signaling time shows a maximal value corresponding to the point of inflection of the signal amplitude.*

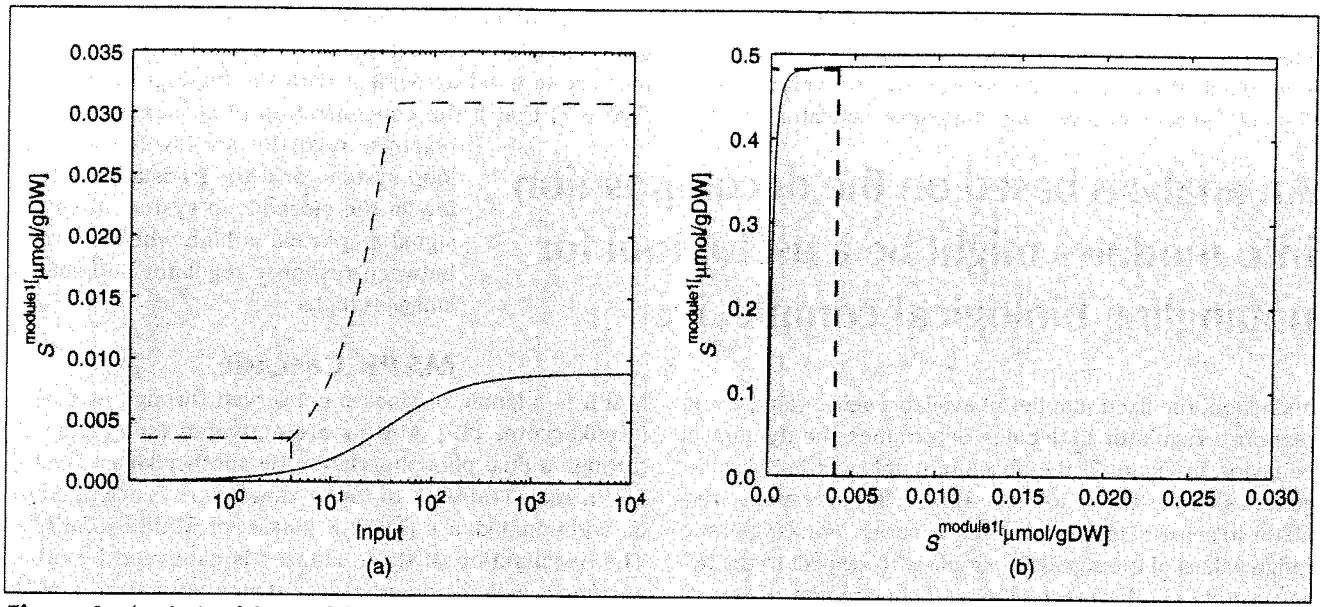


Figure 6. *Analysis of the module 2 of the two-component system. The plots show the steady-state values for module 1 and the signal amplitude S for module 2 of the two-component system. As illustrated with the dashed line, even for low input values in module 1 the maximal signaling amplitude in module 2 is achieved. The dashed line in (a) shows the signal amplitude of module 1 for a low input (u = 4),* $S \approx 0.004\ \mu\mathrm{mol/gDW}$*, much lower than the maximum. However, for this value, as depicted by the dashed line in (b), the signal amplitude of the second module reaches the maximum.*

output of the system shows the same amplitude over a wide range of input values.

Additionally, transfer functions G_i were obtained by fitting simulation data of the overall model against simple linear functions using the MATLAB system identification toolbox [18]. The system is linearized for a fixed stimulus $m = 10^{-6}$, where the system is switched off as in (2), and the initial conditions for the amount of sensor kinase and response regulator are as shown in Table 1. For these conditions, the modules can be described by

$$G_1(s) = \frac{R^p}{U} = \frac{0.0104}{s + 9.92}$$

and

$$G_2(s) = \frac{Pr}{R^p} = \frac{0.254\,s + 254.86}{s + 0.96}.$$

The denominator *den* of G_1 indicates the characteristic response time $\tau_1 \approx 0.1[1/h]$ of module 1. This value is in good agreement with the $\tau_{0.9}$ values determined in Figure 5. The numerator in G_2 reflects the dynamics at the beginning of the protein synthesis. Since the output of module 1 is low, protein synthesis needs time to accelerate from zero to the maximal rate of synthesis.

For the example above of the KdpD/E system in *E. coli*, it is known that the amount of sensor kinase and response regulator are not equal. Moreover, since the operon (prokaryotic operons and genes are written in italic, whereas the corresponding proteins are not, and the first letter is written with an initial capital letter) *kdpD/E* for both proteins lies adjacent to the operon for the target protein KdpFABC, the number of sensor kinases and response regulators can rise during the expression of the target protein based on the read-through effect of the RNA polymerase (expression of the sensor and regulator together with the target protein).

Therefore, the behavior of the open loop with no read-through effect is compared with the closed-loop behavior

Table 1. Parameter values for the two-component system. The parameter S_0 is used for the open-loop analysis.

$k_1 = 3.72$ 1/h	$DNA_0 = 1.32\ 10^{-5} \mu$mol/gDW
$k_{-1} = 1 \cdot 10^{-7}$ 1/h	$S_0 = 0.593\ 10^{-4}\ \mu$mol/gDW
$k_2 = 6.032 \cdot 10^4$ 1/h μmol/gDW	
$k_{-2} = 1 \cdot 10^{-5}$ 1/h μmol/gDW	ATP = 2 mmol/gDW
$k_b = 2.616 \cdot 10^{10}$ 1/h $(\mu$mol/gDW$)^2$	ADP = 0.082 mM
$k_{-b} = 360$ 1/h	
$k_3 = 1.73 \cdot 10^4$ 1/h μmol/gDW	
$k_{tr} = 2000$ 1/h	$k_z = 28.87$ 1/h
$k_{tl1} = 500$ 1/h	$k_d = 0.4$ 1/h
$k_{tl2} = 500$ 1/h	$\mu = 0.5$ 1/h
$k_{tl3} = 1200$ 1/h	

with the autocatalytic circuit based on the read-through effect. See Figure 4 to understand which of the circuits is more efficient and to clarify the effect of the initial conditions of the sensor kinase and response regulator. For the open loop, the fixed number of available sensor kinase and response regulator molecules determines the dynamical behavior. In Figure 7, the absolute number of initial molecules, as well as the ratio r ($r = 1/10, 1, 30$) of response regulator to sensor kinase molecules, is varied. For a high ratio r,high values of the signal amplitude with respect to the target protein can be reached even if the amount of sensor kinase and response regulator are low. Low values of the initial conditions result in large $\tau_{0.9}$ values independently of the ratio r. Remarkably, in the case of the KdpD/E two-component signal transduction system, experimental values for the initial number of molecules for sensor kinase and response regulator are in the range of $5 \cdot 10^{-5} \mu\text{mol}/gDW < S_0 < 5 \cdot 10^{-4} \mu\text{mol}/gDW$ and $r = 30$, where S is maximal and $\tau_{0.9}$ is minimal, indicating that these values are optimal even when no read-through effect occurs.

Finally, we analyze the robustness of the signaling time and signal amplitude with respect to parameter changes. Figure 8 shows the effect of the parameters k_{tl}, k_{tl2}, and k_3. Interestingly, the signal amplitude is not sensitive to the translation parameters k_{tl} and k_{tl2}, while the signaling time becomes larger for low parameter values. In contrast, the rate k_3 of dephosphorylation influences the signal amplitude, while the signaling time is barely affected. These findings are in good agreement with the findings above (see Figure 7) that if the concentration of sensor kinase and response regulator are low in the open-loop system, or if the translation rate is low in the closed-loop system, then the signal amplitude is high when the ratio between response regulator and sensor kinase is high.

An analysis based on the decomposition into modules might be a useful tool for untangling biological complexity.

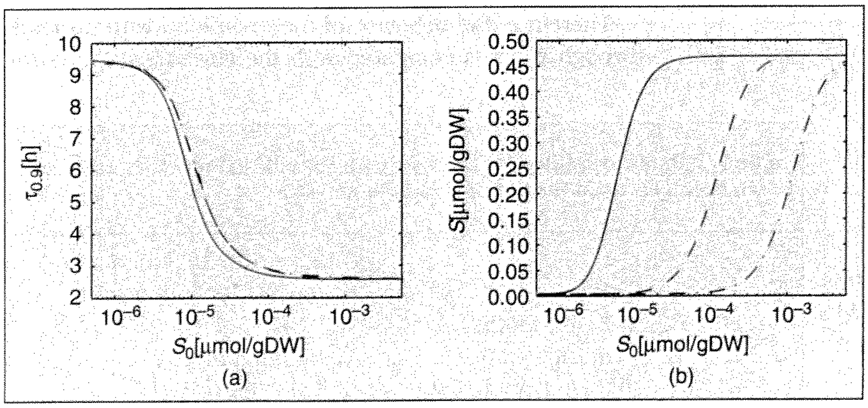


Figure 7. *Signaling time $\tau_{0.9}$ and signal amplitude S for the two-component system in open loop. The curves correspond to various initial conditions of the sensor kinase S_0 and the response regulator R_0, and the ratio r between sensor kinase and response regulator. The initial concentration of the response regulator R_0 is calculated as $R_0 = rS_0$, where r has values 0.1 (dash-dot line), 1.0 (dashed line), and 30 (solid line). While the effect of r on the signaling time is negligible, r has a strong influence on the signaling amplitude.*

MAPK Cascade

MAPK is a family of kinases conserved through evolution in eukaryotes [19]. MAPKs are activated (see Figure 2) through a dual phosphorylation by another kinase, called MAPK kinase (MAPKK or MKK), that in turn is activated by an additional kinase (MAPKK kinase, or MAPKKK/MKKK). The deactivation of these kinases is catalyzed by other enzymes, called phosphatases. Once activated, a MAP kinase can phosphorylate many proteins, including transcription factors, which in turn regulate gene expression. There are several families of MAPKs, with their corresponding MAPKKs and MAPKKKs, and the resulting cascades of reactions play a central role in signal transduction processes in eukaryotes. In mammals, MAPKs are typically downstream of signaling pathways that transmit the information delivered to the cell by stimuli such as growth factors or stress and trigger essential cellular responses such as proliferation and differentiation [19].

Dynamics and Steady-State Characteristics

The system-theoretical properties of the MAPK cascade have been studied extensively. The three-step structure shown in Figure 2 allows an amplification of the signal and, furthermore, provides ultrasensitive input/output behavior [20]. Such a sigmoidal characteristic curve arises due to the partial saturation of the enzymes as well as the dual phosphorylation mechanism of the kinases [20], [21]. Moreover, the different levels add their ultrasensitivity [22].

The addition of positive and negative feedback loops, which embed some MAPK cascades, enriches the versatility of the MAPK cascades. On the one hand, negative feedback can potentially drive the system to return to the basal state after a transient response to a constant input, a phenomenon known as adaptation [23], and introduce sustained oscillations [24]. On the other hand, the

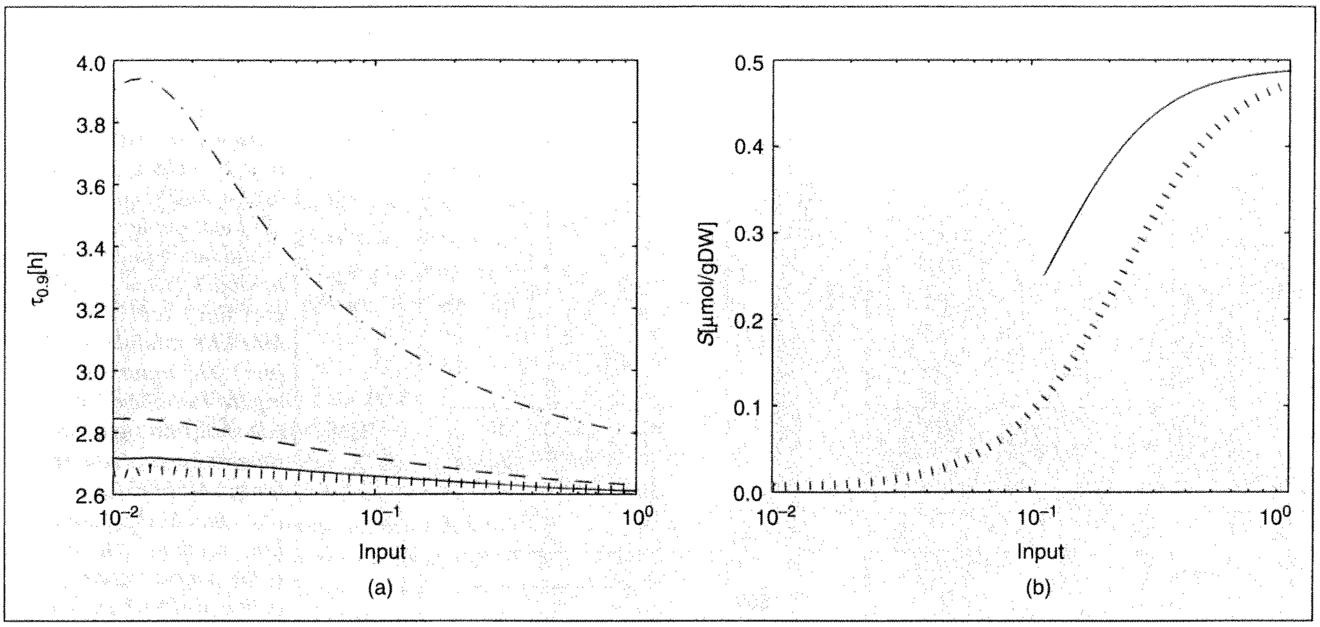


Figure 8. *Signaling time $\tau_{0.9}$ and signal amplitude S for the two-component system in closed loop. The plots show the influence of parameter variations on signaling time and signal amplitude. The signaling time and signal amplitude were first computed using the parameters from Table 1 (solid line), and then the parameters k_{tl} and k_{tl2} were reduced two-fold (dashed lines) and ten-fold (dash-dot lines). Finally, k_3 was increased two-fold, with the remaining parameters as in Table 1. The parameters k_{tl}, k_{tl2} show a large influence on the signaling time but show no effect on signal amplitude. The parameter k_3 shows an opposite effect, since a change in k_3 (dotted lines) strongly influences the signal amplitude but has only a slight effect on the signaling time.*

presence of positive feedback, in combination with the ultrasensitivity of the MAPK cascade, can potentially give rise to a bistable system, which can give an irreversible binary response to a continuous stimulus [25], [26].

From the variety of models available, we have chosen and analyzed the MAPK cascade model proposed by Kholodenko [24]. This model is set up simply, with all of the reactions described by Michaelis-Menten kinetics (see "Model Equations and Parameters"), which provides a connection between the modules free of retroactivity, as described in [4]. The model includes negative feedback, as shown in Figure 2. The parameter K_t determines the strength of the feedback, where larger K_t implies less important feedback. The feedback can thus be disabled by giving high values to K_t.

The parameters discussed in the introduction are computed for the three modules of the MAPK cascade (MAPKKK, MAPKK, and MAPK) and for the whole cascade operating in open and closed loop (see Figure 9) using steps of different magnitude as inputs. The three modules show ultrasensitivity, Hill coefficients of 4.0 for MAPKKK and 6.6 for MAPKK and MAPK, due to the saturation of the enzymes in the case of MAPKKK, and, in the case of MAPKK and MAPK, due to the dual-phosphorylation mechanism and to the saturation of the enzymes [20]. Interestingly, the threshold 0.085 of the cascade is close to, but slightly lower than, the threshold 0.103 of the first module, meaning that the system does not need the complete activation of the first module to reach full activation. The Hill coefficients of the subunits combine in a submultiplicative manner [22], producing high steepness in the curve of the total signal amplitude (Hill coefficient of 111). Additionally, the maximal signal amplitude of the whole cascade corresponds to the maximal possible signal amplitude of the last module (see Figure 9).

The three modules and the whole cascade show a sharp deceleration of the response around the threshold value, as shown by the parameter $\tau_{0.9}$. The peak is higher and narrower for the total cascade than for the first module parameter (see Figure 9). Far from this peak value, the whole cascade is, expectedly, slower than the single modules.

Additionally, the inclusion of a feedback loop decreases the response time, but only around the peak, and a decrease in the signal amplitude (see Figure 9).

Depending on the strength of the feedback, sustained and damped oscillations can be observed (see Figure 10). A bifurcation analysis shows that, over a wide range of feedback strengths determined by the value of K_t, there is a range of the values of the input V_1 for which the system shows sustained oscillations (see Figure 10). For strong feedback values, the oscillations disappear and the output signal decreases to almost zero.

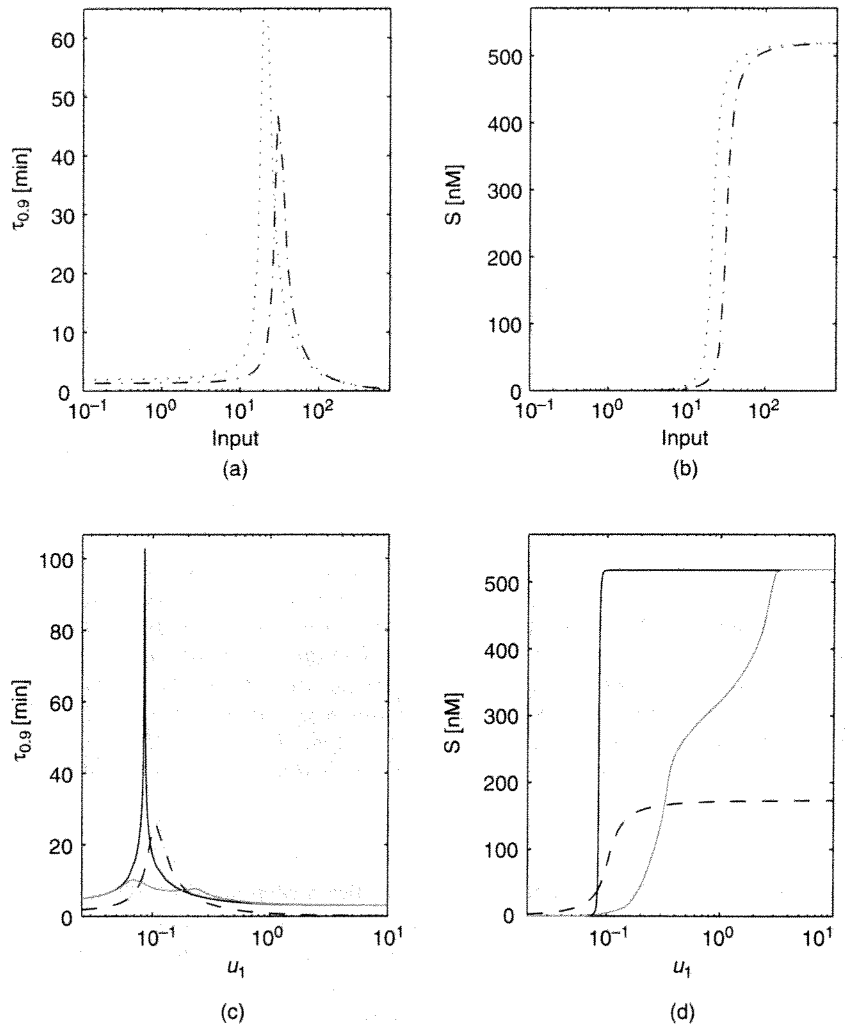


Figure 9. *Signaling time $\tau_{0.9}$ and signal amplitude S for the MAPK cascade and its subunits. (a) and (b) show the signaling time and signal amplitude for the MAPK cascade [24] without feedback (solid black line), with feedback ($K_t = 9$) (solid grey line), and of the MAPKKK module (dashed line). The input is u_1; see Figure 4 and text. (c) and (d) show the signaling time and signal amplitude for the MAPKK (dash-dot line) and MAPK (dotted line) modules. The input is the concentration of active MAPKKK (u_2) and MAPKK (u_3) for the modules MAPKK and MAPK, respectively; see Figure 4. All the subunits show an ultrasensitive behavior and a peak of the signaling time around the threshold value for the signal amplitude. The entire cascade combines the sigmoidity of the three subunits, and the negative feedback decreases the signal amplitude.*

Since the equations of the modules are simple, it is relatively easy to study the dynamic behavior of the MAPK cascade analytically for extreme input conditions. The output of the module MAPKKK (see Figure 2) is $[MKKK_P]$, which can be computed according to [24] by solving the differential equation

$$\frac{d[MKKK_P]}{dt} = \frac{k_1\,[MKKK]}{(K_{m,1}+[MKKK])\left(1+\left(\frac{MAPK_{PP}}{k_t}\right)^n\right)} - \frac{k_2\,[MKKK_P]}{K_{m,2}+[MKKK_P]}.$$

Defining $u = k_1/k_1^0$ and $u_1 = u/(1+(MAPK_{PP}/k_t)^n)$ we obtain

$$\frac{d[MKKK_P]}{dt} = \frac{k_1^0\,[MKKK]}{K_{m,1}+[MKKK]}u_1 - \frac{k_2\,[MKKK_P]}{K_{m,2}+[MKKK_P]}$$
$$= F_1([MKKK])\,u_1 - F_2([MKKK_P]). \qquad (3)$$

It should be noted that $[MKKK]$ and $[MKKK_P]$ are coupled, since $[MKKK]+[MKKK_P] = [MKKK^0]$. If the input is low, then the conversion of $MKKK$ into $MKKK_P$ will be low. Since the initial concentration of $MKKK$ is $[MKKK^0] = 100$ nM and $K_{m,1} = 10$ nM, then $[MKKK] >> K_{m,1}$ holds approximately, and, hence, F_1 can be roughly estimated by k_1^0. Additionally, since the value of $[MKKK_P]$ is low, $F_2 \approx (k_2/K_{m,2})[MKKK_P]$ and (3) implies

$$\frac{d[MKKK_P]}{dt} \approx k_1^0 u_1 - (k_2/K_{m,2})[MKKK_P],$$

which is a system with a first-order lag.

On the other hand, if the input value u is high, $F_1 u >> F_2$. The condition $[MKKK] >> K_{m,1}$ holds for a certain period of time, and, hence, $F_1 \approx k_1^0$, leading to the equation

$$\frac{d[MKKK_P]}{dt} \approx k_1^0 u_1,$$

which is an integrator. If $[MKKK] >> K_{m,1}$ is not satisfied, the system behaves as an integrator but with a variable gain F_1. Since the output is limited by the amount of MKKK ($[MKKK^0]$), the system saturates at a certain time.

Analogous considerations can be applied to the modules MAPKK and MAPK, which lead to the same conclusion; that is, each system is an integrator for high inputs and a proportional system with first-order lag for low inputs (see Figures 11 and 12).

Epidermal-Growth-Factor-Induced MAPK Cascade

The epidermal growth factor (EGF) signaling network is perhaps the best understood cellular signaling system in mammalian cells and part of what is now known is due to systems biology [27]. The EGF receptor (EGFR) is one of the four members of the EGFR family, which belongs to the RTK family. Activation of the EGFR can trigger responses that include growth and cell migration [28]. Due to the tight connection between EGFR and cancer, as evidenced by the fact that EGFR is overexpressed in many tumors, many novel cancer therapies target EGF signaling [29].

Upon ligand binding, EGFR dimerizes and crossphosphorylates. Once phosphorylated, the EGF receptor can bind several proteins leading to the formation of molecular complexes, which in turn trigger the MAPK pathway. The activation of the MAPK pathway requires the binding of the adapter molecules Gap, Grb2, Shc, and Sos to the EGF receptor, building a complex. The MAPK cascade can be activated by means of both an Shc-dependent pathway and an Shc-independent pathway. The recruitment of Sos to the membrane allows Sos to activate Ras, which in turn activates Raf, the first element of the Raf/MEK/ERK MAPK cascade (see Figure 13) [10], [28], [30].

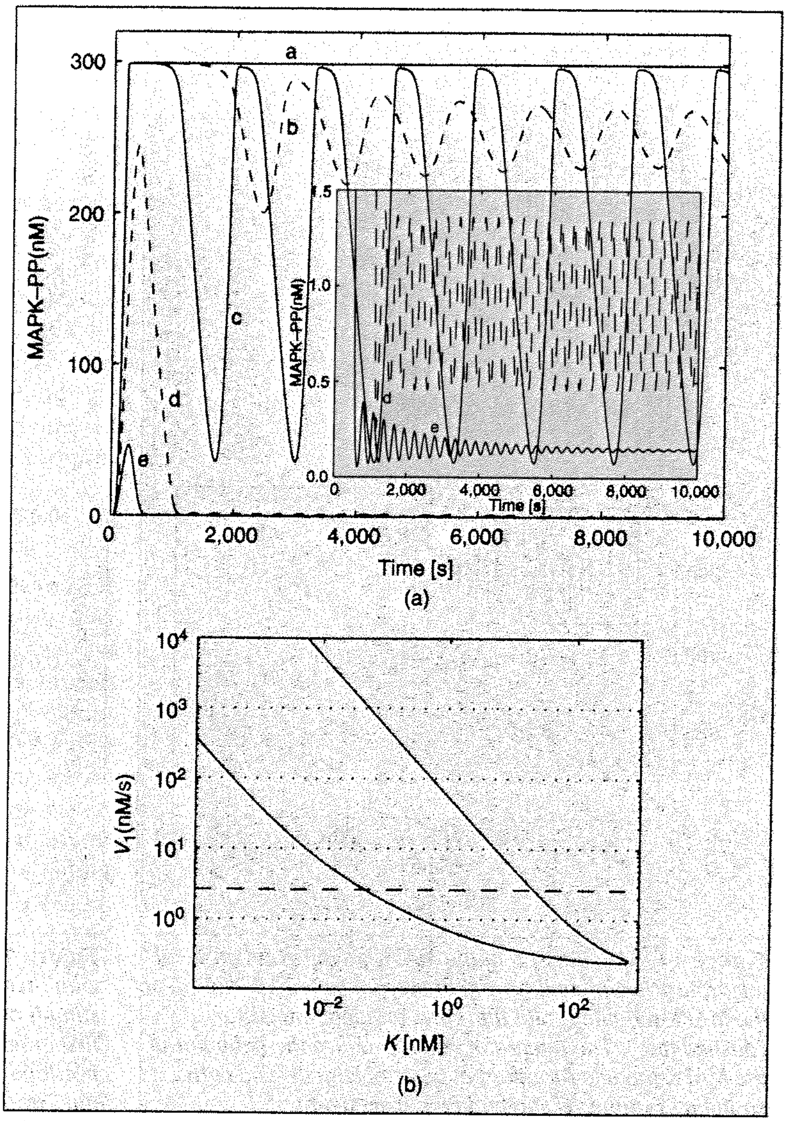


Figure 10. *Oscillations in the MAPK cascade. (a) illustrates simulation results for the model of Kholodenko [24] varying the feedback strength; a: $K_t = 1d^5$, b: $K_t = 23.5$, c: $K_t = 9$, d: $K_t = 6d^{-2}$, and e: $K_t = 1d^{-3}$. The input is $V_1 = 2.5$ nM/s for all cases. (b) illustrates bifurcation analysis of the MAPK cascade. The black line shows the Hopf bifurcation points as a function of the feedback strength (K_t value) and input (V_1 value). Sustained oscillations occur for values between both lines. The dashed line shows the input value V_1 used by Kholodenko [24], for which Hopf bifurcations occur at $K_t = 0.0532$ and $K_t = 23.6$.*

This pathway, as well as others influenced by EGF, has been modeled in [27] and [30]–[32]. The model in [30] includes 13 components from the EGF ligand to ERK. Since some of these elements can interact with each other to form various complexes; 94 states have to be included in the mathematical model. This model also includes internalization, a process in which the receptors are retrieved from the cell surface and moved into special compartments known as endosomes. The role of internalized receptors is still unclear. In the model, the EGF receptor, alone or bound to other proteins, can be internalized. Once internalized, the EGFR is still active and can bind to the same compounds as the receptor on the surface [30]. Therefore, the inclusion of the internalization doubles all the steps in the model.

We have decomposed the noninternalized part of the model according to the criterion of absence of retroactivity discussed above [4], obtaining a set of modules joined by connections free of retroactivity or having a weak retroactivity (see Figure 13). Some of the properties of the model are analyzed below.

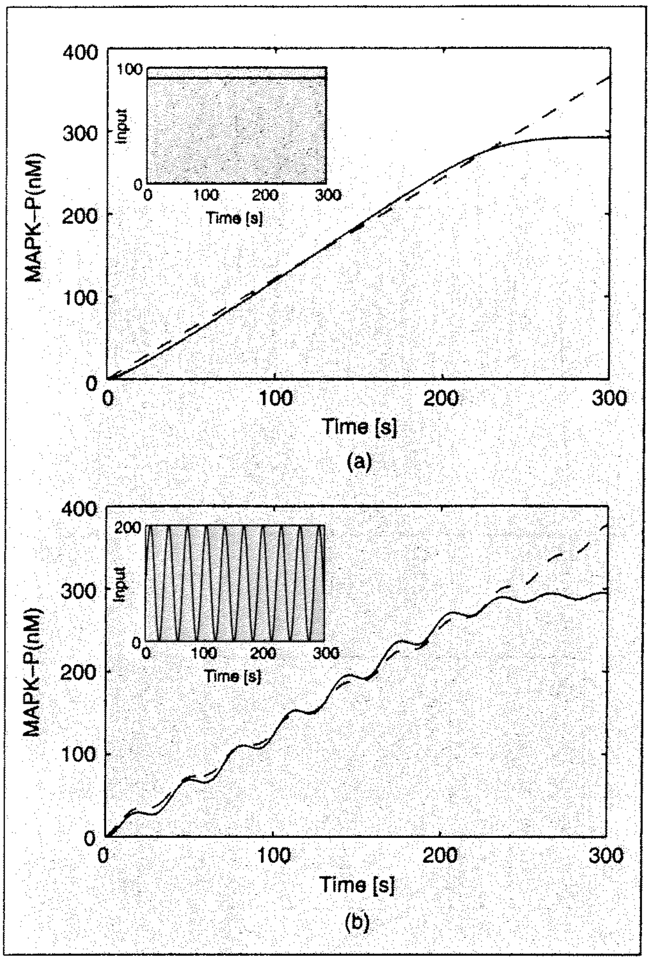


Figure 11. *Comparison of the MAPK module and an ideal integrator. The plots illustrate the response to high stimuli of the MAPK module (solid line) and an ideal integrator (dashed line). The integrator can reproduce the behavior of the MAPK module for different types of stimuli. The corresponding stimulus is shown in the inset figure.*

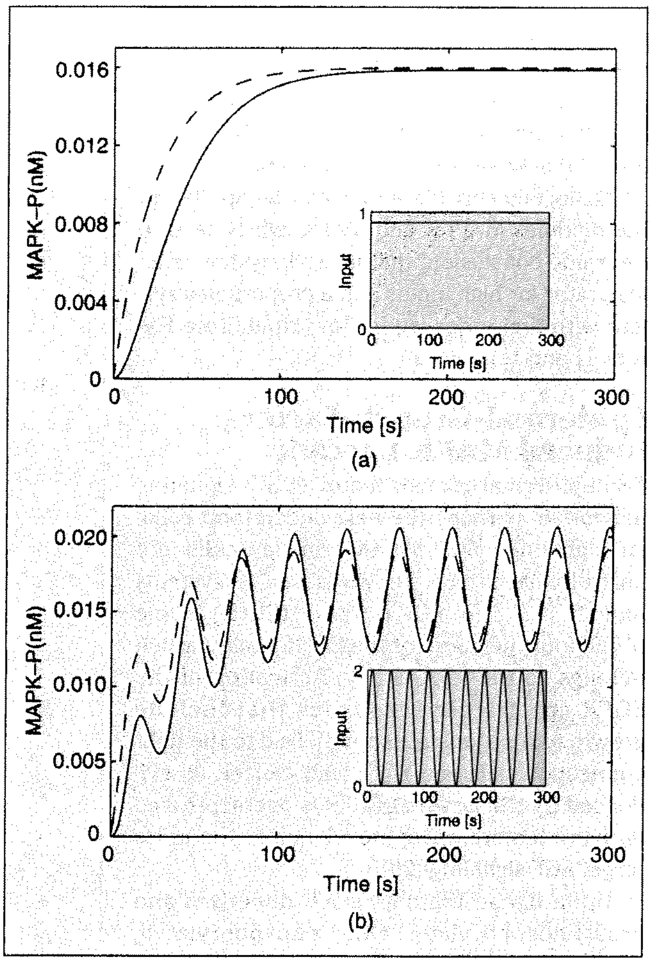


Figure 12. *Comparison of the MAPK module and a system with first-order lag. The plots illustrate the response to low stimuli of the MAPK module (solid line) and a system with first-order lag (dashed line). The MAPK module and the proportional system behave similarly for different types of stimuli. The corresponding stimulus is shown in the inset figure.*

Dynamics and Steady-State Characteristics

The signaling system shows a remarkable independence on the concentration of the ligand, as pointed out in [30]. The amount of EGF can vary over a wide range of biologically significant values without major effect on the output signal, the activated form of ERK. For any value higher than approximately 0.1 nM, neither the amplitude nor the signaling time are changed (see Figure 14). Interestingly, the output of the module MEK depends on the concentration of the ligand (see Figure 15). Therefore, it is the ERK module that produces independence from the ligand concentration. This phenomenon is due to the sigmoidal input/output relationship of the module ERK, the Hill coefficient of 2.44, and the threshold value of the ERK module ($K^h_{0.5} \approx 3100$ molecules/cell) reached for low stimuli, $\approx$0.007 nM. Hence, the output (ERK-PP) shows little variation for input values above 0.1 nM (see Figure 15).

Additionally, as shown in [30], the contribution of the internalized receptors is important only for input values below 0.1 nM, since for higher values the amplitude of the signal for the internalized pathway is negligible (see Figure 14).

Another point, still unclear, is the role of the adaptor molecule Shc, which provides a second mechanism for activating Ras. The Shc-dependent pathway is redundant and seems to be preferentially used [33]. When this pathway is disabled by setting the concentration of Shc to zero, simulating an Shc knockout, a genetic defect that disables Shc, the output signal is slightly lower for high EGF concentrations, higher than roughly 0.1 nM, and slightly higher for lower values, between approximately 0.01 nM and 0.1 nM. For EGF concentrations under 0.01 nM, the signal without Shc is again lower, but in this case the difference is relatively more important (see Figure 14). The system reacts faster if Shc is present, although the difference is small. Therefore, Shc seems to play an important role only at low EGF concentrations [34].

We have performed an analysis of the EGF signaling network, based on the modules defined in [4] (see Figure 13), where we aim to reduce the model [35]. Some of the modules, for example the module regarding the EGF reception, can be analytically linearized. For more complex modules, such as the complex formation, testing the response of the system to different inputs allows one to find the linear system that best reproduces its behavior. Linearization provides insight into the behavior of the different modules; for example, the reception module behaves as a differentiating system with a third-order lag behavior. Due to the strong role played by the nonlinearities, however, a linear system cannot completely reproduce the dynamics of those modules.

Conclusions

This article introduces compelling problems of biological systems and signaling networks. Our goal has been to show how engineering tools can be applied to the analysis of the cellular machinery. We reviewed three examples, which had been previously analyzed regarding their modularity [4]. We included two simple examples, two-component signaling and MAPK cascade, and a more complicated example, EGF signaling. An important point to discuss in model analysis for a cellular system is the availability of experimental data that is sufficient to verify the model structure and model parameters. Although model validation is not described here, the parameters of the models for the two-component system [17] and EGF pathway [30] can reproduce much experimental data.

An analysis based on the decomposition into modules might be a useful tool for untangling biological complexity. In some cases, a property of a large network can be assigned to a certain subunit. For example, the remarkable insensitivity of the EGF network to the input concentration turns out to be due to the last subunit. Consequently, this module can be thoroughly analyzed regarding this property. The absence of retroactivity ensures that the analysis performed on the isolated subsystem will not be distorted when translated into the whole system. In the case of ERK, there is a certain retroactivity to the module MEK, which means that the properties of the ERK unit might be slightly different when connected to the rest of the network. One possible approach to analyzing the modules is to determine the values of a relevant parameter at the output of a module as a function of the parameter at the input of this module. This dependence provides a kind of characteristic curve of the module, as in the case of the ERK module in the EGF model or the gene expression module in the two-component system.

A major problem in the analysis of signal transduction is the dynamic essence of signaling processes. The

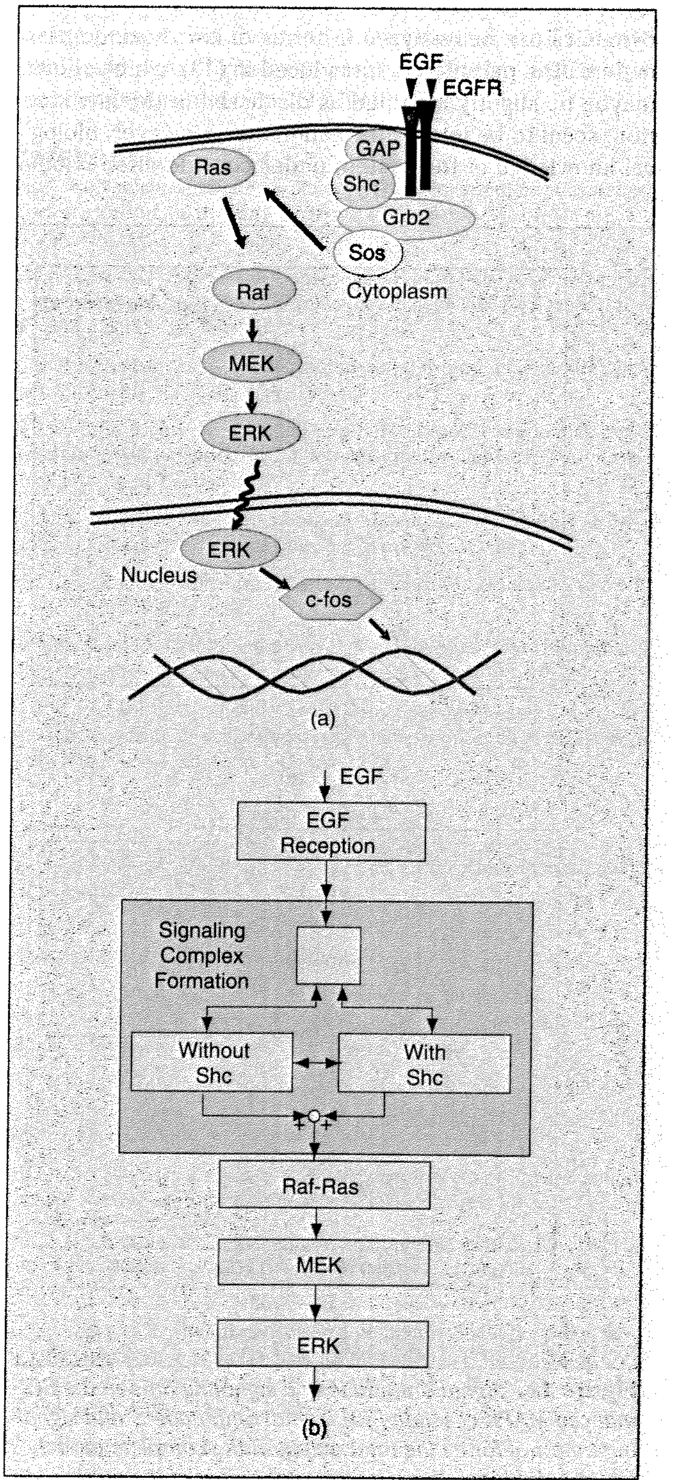


Figure 13. *Representation of the EGF-induced MAPK cascade. (a) is a schematic representation of the elements involved in the EGF-induced activation of the Raf/MEK/ERK MAPK cascade modeled in [30]. Note that the pathway follows the schema of Figure 1. (b) is a modular representation of the noninternalized part of the EGF signaling model, adapted from [4]. Unidirectional connections represent the absence of retroactivity or weak retroactivity, while bidirectional connections represent retroactive connections.*

dynamics can be analyzed in terms of key dynamic parameters. The parameters introduced in [13], which should maybe be slightly modified as discussed in the introduction, seem to be interesting candidates. However, biological knowledge of the system under consideration should be carefully studied to choose parameters that can be related to the biological outcome. For example, in the case of the activation of ERK, it has been proposed that the area under the curve of the activated ERK time course determines the response [36]. If this parameter is used, the results are similar to those obtained using the signal amplitude. The only distinguishable difference is that the area is always lower in absence of Shc, and that the differences are more important for high values of EGF than in the case of the signal amplitude. However, the influence is only of about 10%.

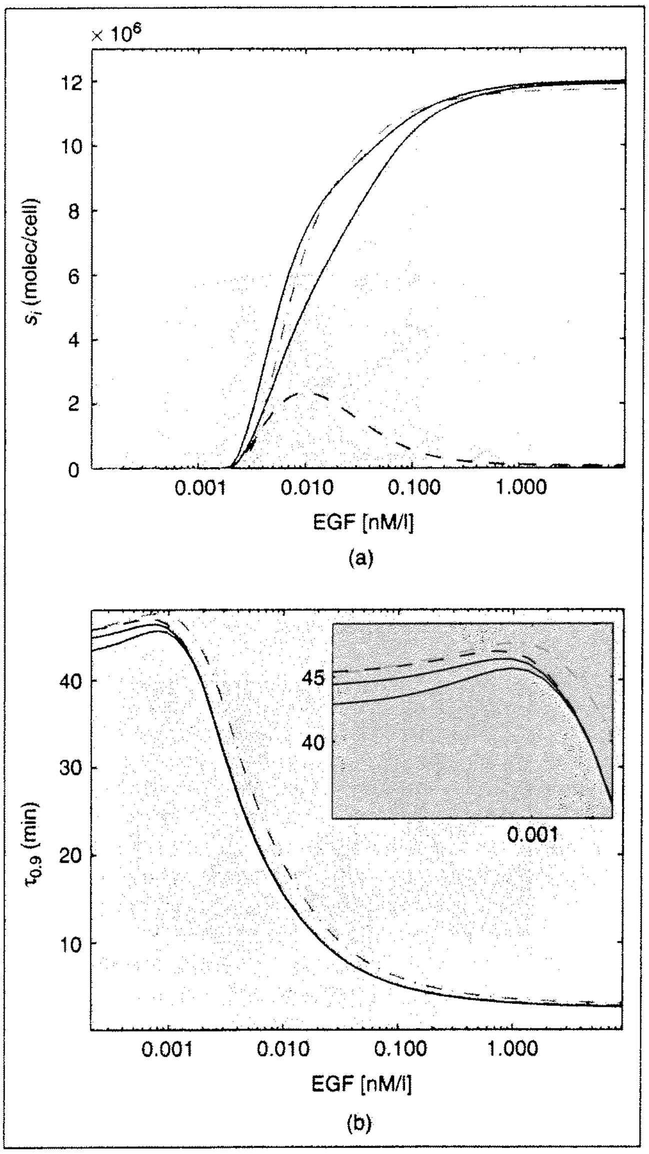


Figure 14. *Signal amplitude and signaling time in the EGF-induced MAPK cascade [30]. Signal amplitude S and signaling time $\tau_{0.9}$ for a) the total output of the complete model (thick solid black line), b) the total output of the model without Shc (thick, dashed-dot grey line), c) the output of the complete model due to the receptors on the surface (thin solid black line), and d) the output of the complete model due to the internalized receptors (thin dashed black line) from the EGF model [30]. The internalized receptors are important only for input values below 0.1 nM/l (nanomol/liter). The adaptor molecule Shc accelerates the response and increases the signal amplitude moderately for input values below 0.01 nM/l.*

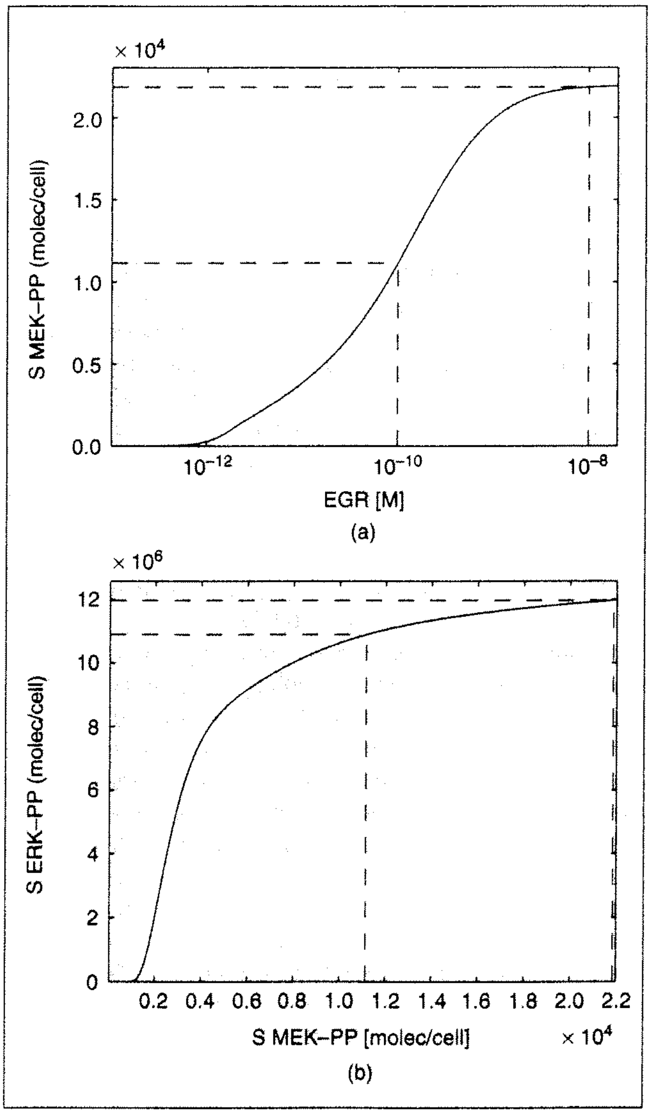


Figure 15. *Analysis of the ERK module of the EGF-induced MAPK cascade. The signal amplitude S for the output of the MEK module is represented as a function of the EGF concentration, and the signal amplitude of the output of the ERK module (ERK-PP) is represented as a function of the input (MEK-PP). The dashed lines show the values for two biologically relevant input values. Despite a 100-fold difference in the input, the difference in the output is only about 10%.*

The first module of the two-component system and the individual modules of the MAP kinase cascade are characterized by a transfer of phosphoryl groups. Interestingly, the time constants of the modules are comparable, and, for a large range of input signals, the systems reach steady-state within a few minutes. However, if the gene expression is considered, the values of the time constants clearly increase, due to the high number of individual steps involved in the protein synthesis. The signal amplitude cannot be easily compared quantitatively. The systems are from completely different cells, and even in a cell the number of components of different signaling systems is completely different. A more comparable parameter might be obtained by dividing the signal amplitude by the input, which could be interpreted as the gain of the system. Nevertheless, the systems can be qualitatively compared and the shape of the characteristic curves give some information about the physiological role of the subsystem. Both the two-component system and the MAPK cascade models show a sigmoidal dependence of the signal amplitude of the output with respect to the input, which can be used to convert a continuous signal into a digital one.

The two-component system has a Hill coefficient of two, in the case the kinase also acts as phosphatase. Otherwise the two-component system shows a Hill coefficient of one; that is, a Michaelis-Menten type curve without ultrasensitivity. In the case of the MAPK cascade, the subunits have Hill coefficients of 4, 6.6, and 6.6, respectively, showing the whole cascade a combined value of 111. Interestingly, there is a large peak in the signaling time of module 1 of the two-component system, as well as in the case of the MAPK for the whole system and for the three subunits. Remarkably, the corresponding input values are close to the respective threshold values of the modules (see Figures 5 and 9). This phenomenon, though interesting from a system-theoretic point of view, might be of little physiological relevance since the range of values for which it takes place represents a small region of the biologically relevant values.

Due to the increasing complexity and size of signaling models, model reduction has became an important point in the analysis of signaling networks [35]. In our examples, the behavior of the systems can be reproduced with simple linear system under certain conditions. A simplification under other conditions, where the nonlinearities play an essential role, is more difficult. The strong nonlinearities of signaling pathways make model reduction a complicated yet essential task for the future analysis of signaling networks.

Acknowledgments

The authors would like to thank B. Schoeberl, M. Ederer, O. Angeles-Palacios, and J. Lindquist for their critical reading of the manuscript. Work on the two-component system was supported by K. Jung and R. Heermann from TU Darmstadt and was financed in part by the Deutsche Forschungsgemeinschaft [JU 270/4-1 (K.J.)]

References

[1] O. Wolkenhauer, "Systems biology: The reincarnation of systems theory applied in biology?" *Brief. Bioinformat.*, vol. 2, no. 3, pp. 258–270, 2001.

[2] L. Hartwell, J. Hopfield, S. Leibler, and A. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, no. 6761 (Suppl.), pp. C47–C52, 1999.

[3] D. Wolf and A. Arkin, "Motifs, modules and games in bacteria," *Curr. Opin. Microbiol.*, vol. 6, no. 2, pp. 125–134, 2003.

[4] J. Saez-Rodriguez, A. Kremling, and E.D. Gilles, "Dissecting the puzzle of life: Modularization of signal transduction networks," *Comput. Chem. Eng.*, to be published.

[5] A. Asthagiri and D. Lauffenburger, "Bioengineering models of cell signaling," *Annu. Rev. Biomed. Eng.*, vol. 2, pp. 31–53, 2000.

[6] J.S. Parkinson, "Signal transduction schemes of bacteria," *Cell*, vol. 73, no. 5, pp. 857–871, 1993.

[7] A.M. Stock, V.L. Robinson, and P.N. Goudreau, "Two-component signal transduction," *Annu. Rev. Biochem.*, vol. 69, pp. 183–215, 2000.

[8] P.W. Postma, J.W. Lengeler, and G.R. Jacobson, "Phosphoenolpyruvate: Carbohydrate phosphotransferase systems of bacteria," *Microbiological Rev.*, vol. 57, no. 3, pp. 543–594, 1993.

[9] G. Krauss, *Biochemistry of Signal Transduction and Regulation.* Weinheim, Germany: Wiley-VCH, 2003.

[10] J. Schlessinger, "Cell signaling by receptor tyrosine kinases," *Cell*, vol. 103, no. 2, pp. 211–225, 2000.

[11] E.D. Gilles, "Network theory for chemical processes," *Chem. Eng. Technol.*, vol. 21, no. 2, pp. 121–132, 1998.

[12] A. Kremling, K. Kahreis, J. Lengeler, and E.D. Gilles, "The organization of metabolic networks: A signal-oriented approach to cellular models," *Metab. Eng.*, vol. 2, no. 3, pp. 190–200, 2000.

[13] R. Heinrich, B. Neel, and T. Rapoport, "Mathematical models of protein kinase signal transduction," *Molecular Cell*, vol. 9, no. 5, pp. 957–970, May 2002.

[14] I.H. Segel, *Enzyme Kinetics, Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems.* New York: Wiley, 1993.

[15] A. Goldbeter and D. Koshland, "An amplified sensitivity arising from covalent modification in biological systems," *Proc. Nat. Acad. Sci. USA*, vol. 78, no. 11, pp. 6840–6844, 1981.

[16] K. Jung and K. Altendorf, "Towards an understanding of the molecular mechanisms of stimulus perception and signal transduction by the KdpD/KdpE system of *Escherichia coli*," *J. Mol. Microbiol. Biotechnol.*, vol. 4, no. 3, pp. 223–228, 2002.

[17] K. Jung, personal communication.

[18] L. Ljung, *System Identification Toolbox: For Use with MATLAB.* Natick, MA: MathWorks Inc., 1995.

[19] H. Schaeffer and M. Weber, "Mitogen-activated protein kinases: Specific messages from ubiquitous messengers," *Mol. Cell. Biol.*, vol. 19, no. 4, pp. 2435–2444, Apr. 1999.

[20] J.E.J. Ferrell, "Tripping the switch fantastic: How a protein kinase cascade can convert graded inputs into switch-like outputs," *Trends Biochem. Sci.*, vol. 21, no. 12, pp. 460–466, 1996.

[21] C.F. Huang and J.E.J. Ferrell, "Ultrasensitivity in the mitogen-acti-

vated protein kinase cascade," *Proc. Nat. Acad. Sci., USA*, vol. 93, no. 19, pp. 10078–10083, 1996.

[22] J.E.J. Ferrell, "How responses get more switch-like as you move down a protein kinase cascade," *Trends. Biochem. Sci.*, vol. 22, no. 8, pp. 288–289, 1997.

[23] A. Asthagiri and D. Lauffenburger, "A computational study of feedback effects on signal dynamics in a mitogen-activated protein kinase (MAPK) pathway model," *Biotechnol. Prog.*, vol. 17, no. 2, pp. 227–239, 2001.

[24] B.N. Kholodenko, "Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades," *Eur. J. Biochem.*, vol. 267, no. 6, pp. 1583–1588, 2000.

[25] J.E.J. Ferrell and W. Xiong, "Bistability in cell signaling: How to make continuous processes discontinuous, and reversible processes irreversible," *Chaos*, vol. 11, no. 1, Mar. 2001.

[26] W. Xiong and J.E.J. Ferrell, "A positive-feedback-based bistable 'memory module' that governs a cell fate decision," *Nature*, vol. 426, no. 6965, pp. 460–465, 2003.

[27] H. Wiley, S. Shvartsman, and D. Lauffenburger, "Computational modeling of the EGF-receptor system: A paradigm for systems biology," *Trends Cell. Biol.*, vol. 13, no. 1, pp. 43–50, 2003.

[28] A. Wells, "EGF receptor," *Int. J. Biochem. Cell Biol.*, vol. 31, no. 6, pp. 637–643, 1999.

[29] F. Ciardiello and G. Tortora, "Anti-epidermal growth factor receptor drugs in cancer therapy," *Expert Opin. Investig. Drugs*, vol. 11, no. 6, pp. 755–768, 2002.

[30] B. Schoeberl, C. Eichler-Jonsson, E. Gilles, and G. Muller, "Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors," *Nat. Biotechnol.*, vol. 20, no. 4, pp. 370–375, 2002.

[31] G. Moehren, N. Markevich, O. Demin, I. Kiyatkin, A. Goryanin, J. Hoek, and B. Kholodenko, "Temperature dependence of the epidermal growth factor receptor signaling network can be accounted for by a kinetic model," *Biochemistry*, vol. 41, no. 1, pp. 306–320, 2002.

[32] U. Bhalla and R. Iyengar, "Emergent properties of networks of biological signaling pathways," *Science*, vol. 283, no. 5400, pp. 381–387, Jan. 1999.

[33] Y. Gong and X. Zhao, "Shc-dependent pathway is redundant but dominant in MAPK cascade activation by EGF receptors: A modeling inference," *FEBS Lett.*, vol. 554, no. 3, pp. 467–472, 2003.

[34] B. Schoeberl, "Mathematical modeling of signal transduction pathways in mammalian cells at the example of the EGF induced MAP kinase cascade and TNF receptor crosstalk," Ph.D. dissertation, Univ. of Stuttgart, Stuttgart, Germany, 2003.

[35] H. Conzelmann, J. Saez-Rodriguez, T. Sauter, E. Bullinger, F. Allgöwer, and E.D. Gilles, "Reduction of mathematical models of signal transduction networks: Simulation-based approach applied to EGF receptor signaling," submitted for publication.

[36] A. Asthagiri, C. Reinhart, A. Horwitz, and D. Lauffenburger, "The role of transient ERK2 signals in fibronectin- and insulin-mediated DNA synthesis," *J. Cell. Sci.*, vol. 113, no. 24, pp. 4499–4510, 2000.

[37] A. Kremling and E. Gilles, "The organization of metabolic reaction networks: Signal processing in hierarchical structured functional units," *Metab. Eng.*, vol. 3, no. 2, pp. 138–150, 2001.

Julio Saez-Rodriguez (saezr@mpi-magdeburg.mpg.de) received his chemical engineering degree from the University of Oviedo, Spain, in 2001. Since 2002, he has been a research assistant at the Systems Biology group at the Max-Planck-Institute for Dynamics of Complex Technical Systems in Magdeburg, Germany. His research interest focuses on the mathematical modeling and analysis of signal transduction networks in mammalian cells. He can be contacted at the Max-Planck-Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany.

Andreas Kremling obtained his degree in technical cybernetics from the University of Stuttgart, Germany, in 1992 and Ph.D. at the Institute for Systems Dynamics and Control Theory at the University of Stuttgart in 1997. Since March 1998, he has been a scientific assistant at the Max-Planck-Institute for Dynamics of Complex Technical Systems in Magdeburg, Germany. His scientific interests include mathematical modeling, analysis, and design of biochemical reaction networks; software tools for model setup and visualization of networks; and the application of methods from systems theory to cellular systems.

Holger Conzelmann received the Dipl.-Ing. degree in control engineering from the University of Stuttgart, Germany, in 2003. Since 2003, he has been a research assistant in the Systems Biology group at the Institute for System Dynamics and Control Theory at the University of Stuttgart. His research interests include mathematical modeling of signal transduction networks, model analysis, and model reduction.

Katja Bettenbrock received her M.Sc. in biology in 1993 and her Ph.D. in genetics in 1997, both from the University of Osnabrück, Germany. In 1998, she worked at the university hospital of Ulm in the field of medical microbiology. She has been a member of the Systems Biology group at the Max-Planck-Institute for Dynamics of Complex Technical Systems since 1998. Her research interests focus on signal transduction and metabolic regulation in bacteria.

Ernst Dieter Gilles received his M.S. in electrical engineering and his Ph.D. from the TH Darmstadt, Germany, in 1960 and 1964, respectively. After a postdoctoral period at the TH Darmstadt he became a professor at the Institute for System Dynamics and Control Engineering at the University of Stuttgart in 1968. He is the founding director of the Max Planck Institute for Dynamics of Complex Technical Systems (1997) and an honorary professor at the University of Magdeburg, Germany. His research interests include systems biology, network theory applied to chemical and biological processes, control engineering, and system dynamics. He is the author of some 350 scientific publications, including several books.

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Computers and Chemical Engineering 29 (2005) 619–629

Computers & Chemical Engineering

www.elsevier.com/locate/compchemeng

Dissecting the puzzle of life: modularization of signal transduction networks

J. Saez-Rodriguez, A. Kremling*, E.D. Gilles

Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany

Received 22 September 2003; received in revised form 12 January 2004
Available online 19 November 2004

Abstract

Cells have developed complex control networks which allow them to sense and response to changes in their environment. Although they have different underlying biochemical mechanisms, signal transduction units in prokaryotes and eukaryotes fulfill similar tasks, such as switching on or off a required process or amplifying a certain signal. The growing amount of data available allows the development of increasingly complex models which offer a detailed picture of signaling networks, but the properties of these systems as a whole become difficult to grasp. A sound strategy to untangle this complexity is a decomposition into smaller units or modules. How modules should be delimited, however, remains an unanswered question. We propose that units without retroactive effects might be an interesting criterion. In this contribution, this issue will be explored through several examples, starting with a simple two-component system in *Escherichia coli* up to the complex epidermal growth factor signaling pathway in human cells.
© 2004 Elsevier Ltd. All rights reserved.

Keywords: Mathematical modeling; Signal transduction; Systems biology; Modularity

1. Introduction

Cells are equipped with exquisite sensing systems which allow them to be continously aware of the conditions in their environment and react appropriately to these conditions. The basic elements of a cellular signaling system are a sensor protein, made of a receptor domain and a transmitter domain, and a response regulator, consisting of a receiver domain and a regulator domain (Lengeler, 2000). Stimulation of the sensor (normally bound to the cell membrane) leads to activation of the transmitter, which produces an intracellular signal. This signal is processed by a cascade of molecules and finally arrives at the receiver, which in turn activates the regulator (see Fig. 1). Regulators produce a response by modulating gene expression or enzyme activities.

The key components in this transfer of information are proteins, which form networks and are able to perform computational tasks. Proteins can change their state by interaction with other proteins or by biochemical modifications (such as phosphorylations) catalyzed by other proteins (Bray, 1995). Another common mechanism is the release of small molecules called second messengers, which diffuse in the cell and activate other proteins (Krauss, 2001). Interestingly, although eukaryotes systems are generally more complex, both prokaryotes and eukaryotes follow the same logic. Especially in eukaryotes, enhanced computation possibilities are achieved by inserting elements between the basic elements described above (Lengeler, 2000).

Bacteria, for example, have the capability to use a broad range of nutrient sources for life. Furthermore, they are also able to synthesize a number of monomers like amino acids if these are not provided in the medium. To sense their external environment, bacteria often use rather simple signal transduction systems. A paradigm of bacterial signal transduction is the two-component system that consists just of two elements, the sensor kinase and the response regulator (Hoch & Silhavy, 1995). Bacteria are also able to sense intracellular conditions. One representative is the phosphotransferase system (PTS) (Postma, Lengeler, & Jacobson, 1993). The PTS is an uptake system for several carbohydrates in *Escherichia coli*. In addition, it acts as a sensor and is

---

* Corresponding author. Tel.: +49 391 6110 466; fax: +49 391 6110 526.
E-mail address: kremling@mpi-magdeburg.mpg.de (A. Kremling).

0098-1354/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2004.08.035

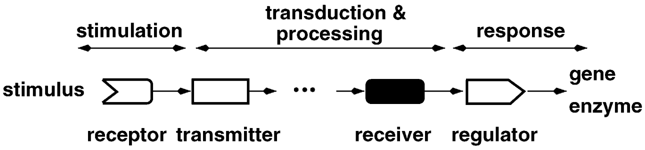


Fig. 1. General scheme of a signal transduction system. Adapted from Kremling et al. (2000) ©, with permission from Elsevier.

involved in the control of uptake of a number of carbohydrates.

Human cells also posses a complex signaling system which allows them to exchange information and thus coordinate themselves. In most cases, the signaling processes follow the general schema described above. The binding of extracellular signals such as hormones or growth factors to receptors results in changes in the intracellular part (*transmitter*) of the receptor. The thus activated receptor transmits the signal to intracellular signaling intermediates, triggering signaling cascades which finally activate transcription factors (*regulators*) which move into the nucleus, changing the gene expression of the cell (Downward, 2001). Essential processes like proliferation, cell development or even the suicide of the cell are controlled by cell signaling. Since it is related to such basic properties of the cells, defective signal processing can lead to important diseases such as cancer or diabetes, and thus signaling pathways are important targets for disease therapy (Levitzki, 1996).

The high number of components involved, the complex crosstalk phenomena among the different pathways and the biophysical regulation set up a picture difficult to grasp (Asthagiri & Lauffenburger, 2000). A useful tool to untangle this complexity might be mathematical modeling. The knowledge and amount of data available about signaling networks grows steadily, boosting the development of increasingly complex models. These models offer a highly detailed picture of signaling pathways, but the properties of these systems as a whole become difficult to understand. This holistic understanding is the target of the emerging discipline of systems biology (Kitano, 2002). Engineers usually face *synthesis* problems: design a system with certain characteristics using a set of well-characterized elements. A system-biologist has to face an *analysis* problem: understand the properties of a complex network. Therefore, the definition of functional units, i.e. entities whose function is separable from those of other units, might help to analyze biological systems (Hartwell, Hopfield, Leibler, & Murray, 1999) since, once modules are defined, they could be systematically analyzed regarding properties such as stability, robustness and dynamic behavior and classified, creating a library of reusable units. Once these *relatively* simple units are well understood, they can be re-assembled in order to analyze the emergent properties of the resulting systems, as engineers do. Furthermore, this set of reusable elements would simplify the set-up of models, since many parts of biological networks are found in several signal transduction pathways.

How biochemical modules should be delimited still remains an unanswered question, and is a topic under intense investigation. While several approaches are based on network-clustering methods applied to experimental data (e.g. Rives & Galitski, 2003), others try to develop a suitable theoretical framework for the analysis of modular networks (e.g. Bruggeman, Westerhoff, Hoek, & Kholodenko, 2002). We have previously introduced three *biologically motivated* criteria for defining functional units: (1) common physiological task (all the elements of a functional unit perform the same task, e.g. the specific catabolic pathways for individual carbohydrates), (2) common genetic units (the genes for all enzymes of a functional unit are organized in genetical units, e.g. operons and modulons in bacteria) and (3) common signal transduction network (all elements of a functional unit are interconnected within a common signal transduction system and the signal flow over the unit border is small compared to the information exchange within the unit (Kremling, Jahreis, Lengeler, & Gilles, 2000).

In this contribution, a novel criterion for the definition of modules, namely the absence of retroactivity in the connections between the modules, will be proposed. The different situations that can lead to a retroactivity-free connection will be first examined by means of the network theory, and later applied to two signaling systems in prokaryotes (the two-component system and the control of carbohydrate uptake) and two in eukaryotes (the MAPK Cascade and the EGF signaling network). Our approach does not intend to provide an algorithm to find modules from a set of experimental data (using e.g. network-clustering methods), but rather a theoretical framework to analyze signaling networks in a modular and systems-theoretical manner.

2. Modularization of signaling networks

A suitable frame for developing modular models is provided by the network theory introduced by Gilles (1998). Systems are described as a combination of two types of elementary units: *components*, which have storages of physical quantities and *coupling elements*, which describe the interactions between the components. These elements can be aggregated into a single elementary unit on a higher level, which can be again described by means of components and coupling elements, leading to a hierarchical structure (Mangold, Motz, & Gilles, 2002). Components and coupling elements are connected by two types of vectors: potential vectors, which are outputs of components and inputs of coupling elements, and current vectors, which are outputs of coupling elements and inputs of components. For example, in a chemical network the compounds would be the components, the reactions the coupling elements, potential vectors would carry information about the concentrations from the compounds to the reactions and current vectors would bring information about the rates back to the compounds, see Fig. 3(a).

The application of the network theory to biochemical systems leads to a modular modeling concept introduced elsewhere (Kremling et al., 2000). This concept is based on the definition of a complete but finite set of elementary objects at the highest level of resolution. Three types of elementary objects are defined: substance storages, substance transformers and signal transformers. By aggregating these objects, complex processes such as gene expression or signaling networks can be described (Kremling & Gilles, 2001; Kremling, Jahreis, Lengeler, & Gilles, 2001).

One argument supporting the application of the network theory to cellular pathways is the proposed hierarchy of biological systems (Kremling et al., 2000; Lengeler, 2000). Actually, this hierarchical structure can be represented similarly for biological systems (Fig. 2) and chemical processes. If we consider a human body, we can divide it into different systems which fulfill different tasks (e.g. the digestive system, the locomotive system, etc.), connected mainly (but not only) by blood vessels. Each of these systems can be described as a sum of organs connected also mainly by blood vessels. Organs are made up of several tissues, each of them made of one type of cells. The coupling between cells takes place by means of exchange of different substances as well as by direct contact between the cells. The machinery of a cell can be decomposed into functional units which perform different functions. These biochemical pathways are connected by common compounds. Finally, each of these modules can be decomposed into molecules which interact by means of molecular interactions or reactions (Fig. 2). This contribution focuses on the last two levels of detail.

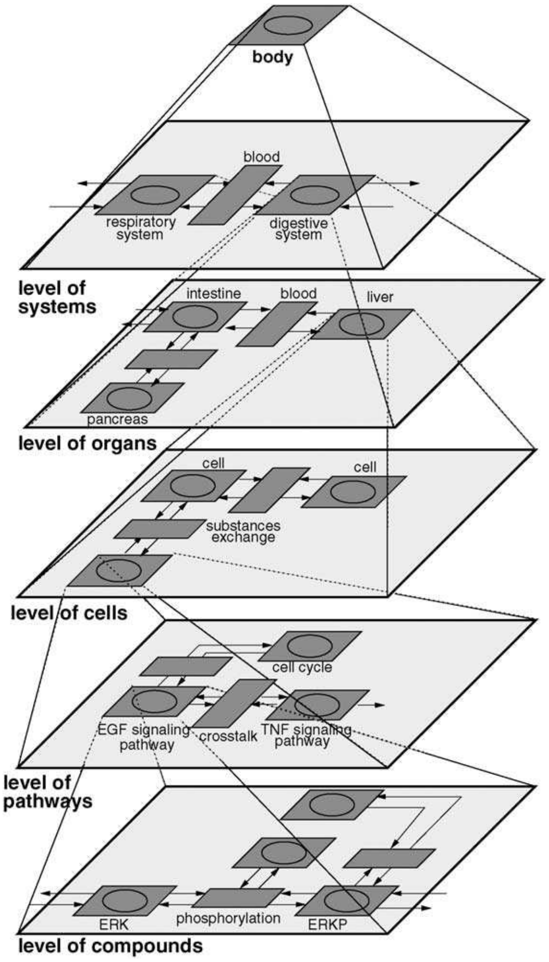


Fig. 2. Hierarchical structure of biological systems. An analogous figure for chemical processes can be found in Mangold et al. (2002).

2.1. *Absence of retroactivity as a criterion*

Since engineering sciences are used to working in a modular manner, it is tempting to approach the definition of biological modules from a technical perspective. From a *system-theoretical* point of view an interesting criterion might be the definition of elements *without retroactive effects* (i.e. where both the input and the output are unidirectional). Such units fulfill the requisite of independence of functional units: the properties of a retroactivity-free unit only depend on its input and are independent of what is downstream of it. Importantly, units without retroactive effects can be relatively straightforwardly analyzed by means of system theory's tools.

Consider the simple general schema depicted in Fig. 3(a), which represents one reaction (coupling element) r and three compounds (components), A, B and C, involved in the reaction r, according to the network theory. If one of the potential or current vectors can be neglected, the system shows a junction free of retroactive effects. But, under which conditions can a current vector (i.e., information about a rate) or a potential vector (information about a concentration) be neglected? In the following we discuss some typical simple cases.

2.1.1. *Neglect of a potential*

A potential can be neglected if the concentration of one of the compounds, say C, does not affect the reaction rate, which corresponds to neglect vector 1 in Fig. 3(a). An example is an irreversible reaction, where the product does not affect the reaction rate. Hence, an irreversible reaction of A and B to give C would be represented as in Fig. 3(b). There are some common irreversible reactions in biochemistry, like some types of phosphorylation reactions.

2.1.2. *Neglect of a current*

A retroactive-free connection by neglect of a current is possible if a compound influences a reaction rate, but the reaction rate does *not* influence this component (i.e., if the vector 2 in Fig. 3(a) can be neglected), leading to the system depicted in Fig. 3(c). One possibility would be if a compound is consumed or produced in a reaction, but the amount involved in the reaction is negligible compared to the total amount of the compound. For example, if the concentration of one of the substrates is much higher than the other, say $A \ll B$, then the amount of B consumed in the reaction will be negligible compared to the total amount of B, leading to a unidirectional connection.

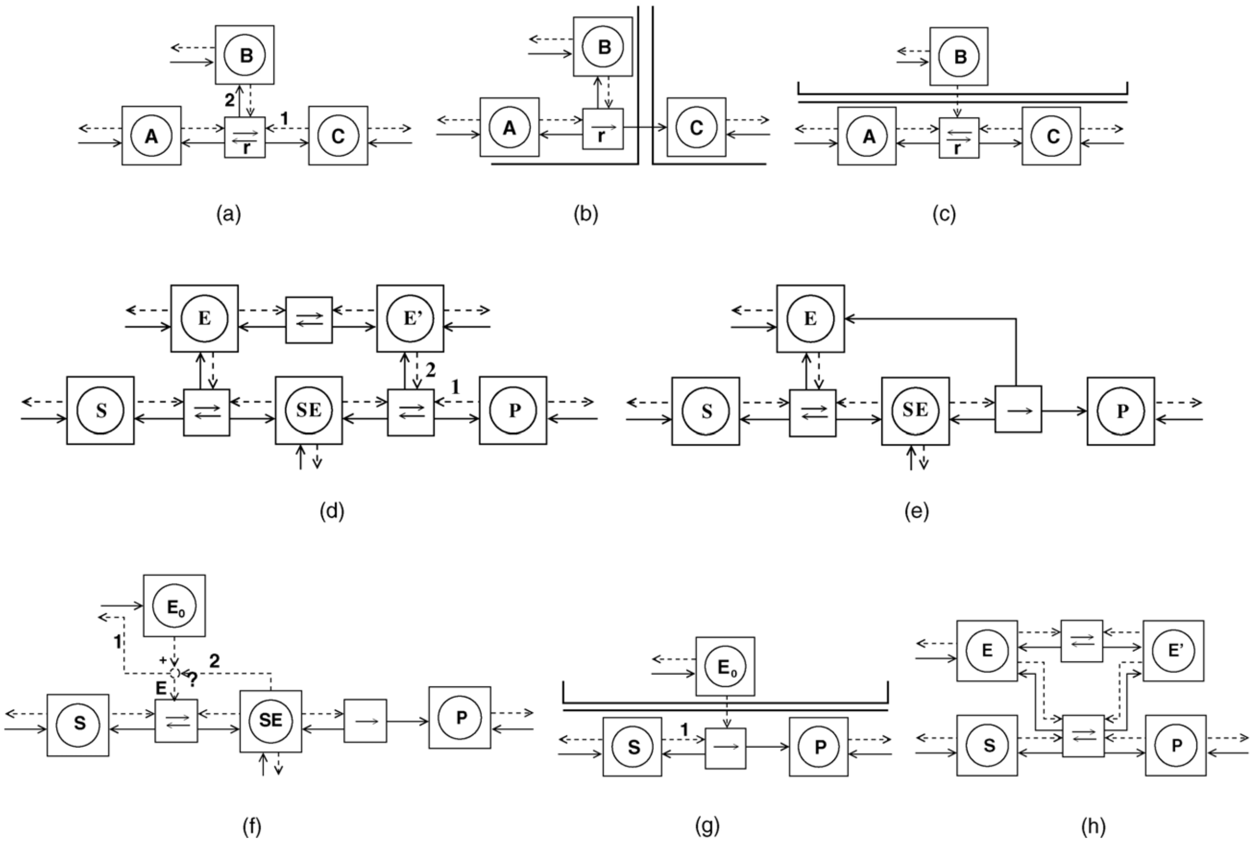


Fig. 3. Representation of different reactions schemes according to the network theory. Dashed lines represent potential (concentration) vectors, solid thin lines current (rates) vectors, and solid thick lines the borders of the modules. (a) General case; (b) neglect of a potential; (c) neglect of a current; (d) system defined by Eqs. (1) and (2); (e) system defined by Eq. (4); (f) same system as in (e) but with a change of variable $E_0 = E + SE$; (g) system defined by Eq. (5); (h) system defined by Eqs. (9) and (10).

If we consider the general case where a compound S is transformed into P, by reaction with another compound E, being E regenerated in an additional step, as defined by the equations:

$$E' \rightleftharpoons E \qquad (1)$$

and

$$S + E \rightleftharpoons SE \rightleftharpoons P + E', \qquad (2)$$

we arrive at the schema depicted in Fig. 3(d). The system is highly interconnected, without unidirectional connections. If the second step of the second reaction (Eq. (2)) is considered irreversible we obtain

$$S + E \rightleftharpoons SE \rightharpoonup P + E' \qquad (3)$$

instead of Eq. (2). The representation of the new system is obtained by deleting the vectors 1 and 2 in Fig. 3(d). In this system, there is a unidirectional connection defined by the irreversible step, but the connection between E/E' and S/P has still retroactivity (see Fig. 3(d)). If, additionally, $E = E'$, the system

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} SE \overset{k_2}{\rightharpoonup} P + E \qquad (4)$$

is obtained, which is shown in Fig. 3(e) and represents the irreversible conversion of S into P catalyzed by an enzyme E. The reactions are normally described according to the mass action law. Defining a new variable $E_0 = E + SE$ we obtain an alternative representation (Fig. 3(f)). Analyzing this schema we can see that a connection free of retroactivity from the enzyme to the reaction can be achieved if:

(i) The reactions that influence E_0 but are not represented in Fig. 3(f) are not influenced by E, which is equivalent to neglect the vector 1 in Fig. 3(f). This is actually the case introduced above of absence of retroactivity by an irreversible reaction.

(ii) The dynamics of the compound SE can be neglected (i.e., if $\mathrm{d}c_S E/\mathrm{d}t \approx 0$, which means that the vector 2 in Fig. 3(f) is negligible). This approximation is known as the quasi-steady-state assumption, and leads to the reduced system (see for example Segel, 1988)

$$S \to P, \qquad (5)$$

following the reaction rate r the classical Michaelis Menten equation

$$r = \frac{V_{\max} S}{K_m + S} = \frac{k_2 E_0 S}{K_m + S}, \qquad (6)$$

where $K_m = (k_{-1} + k_2)/k_1$. We obtain thus a connection free of retroactivity by absence of a current vector, as represented in Fig. 3(g). If, additionally, the enzyme is saturated by the substrate ($K_m \ll S$), then the reaction rate r becomes

$$r = k_2 E_0 \tag{7}$$

and the system can be represented as in Fig. 3(g) deleting the vector 1. We obtain hence an additional connection free of retroactivity between the reaction r and the substrate S.

The assumption $\mathrm{d}c_{SE}/\mathrm{d}t \approx 0$ is correct for the system defined in Eq. (4) if $\varepsilon \ll 1$, where $\varepsilon = E_0/(K_m + S_0)$ (Segel, 1988). This condition is fulfilled if $E_0 \ll S_0$ and if $E_0 \ll K_m$. $E_0 \ll S_0$ (much less enzyme than substrate, a usual situation in many in vitro experiments) is the usual assumption for the application of Michaelis Menten equation.

The condition $E_0 \ll K_m$ can be rewritten as $E_0 k_1 \ll k_2 + k_{-1}$. Since k_1 is the kinetic constant for the formation of the complex *SE*, and k_{-1} and k_2 the kinetic constants for the dissociation of the complex *SE* (see Eq. (4)), this condition can be interpreted as the decomposition of *SE* being much faster than the formation of *SE*. If it is assumed that the initial amount of *SE* is zero, the amount of *SE* is always around zero, holding the quasi-steady-state assumption. This situation is analogous to many electrical measuring devices, e.g. a thermocouple. In a thermocouple, a difference of temperature generates a voltage *V*, which in turn produces a current *I* through a conductor. A very high value is given to the resistance *R* and therefore the current is very low ($V = IR$). This current provides a measurement of the voltage that does not affect the source of the signal. In the case of an enzymatic reaction where $E_0 \ll K_m$, the reaction rate (or the amount of product) is a "measurement" of the concentration of the enzyme, but, since $E_0 \ll K_m$, there is a high resistance against the consumption of the enzyme, which is thus not affected by its "measuring device".

The Michaelis Menten expression (Eq. (6)) is widely used for enzymatic reactions without considering whether the assumptions described above are fulfilled or not.

If the Eq. (1) of the general case can be neglected, but the second term of the Eq. (2) can not be considered irreversible, we obtain a system defined by the equation

$$S + E \rightleftharpoons SE \rightleftharpoons P + E \tag{8}$$

which, under the quasi-steady-state assumption, can be transformed into a system with a unidirectional connection by neglect of a current as depicted in Fig. 3(c).

On the other hand, the neglect of the complex *SE* in the general case (Eqs. (1) and (2)) leads to the system

$$E' \rightleftharpoons E \tag{9}$$

$$S + E \rightleftharpoons P + E'. \tag{10}$$

As can be seen in Fig. 3(h) this system is still highly interconnected.

In the following, these criteria will be applied to several examples of mathematical models of signal transduction pathways. Two signaling systems in bacteria (the simple two-component system and the control of carbohydrate uptake system) and two in eukaryotes (the basic MAP Kinase Cascade and the complicated EGF Signaling Pathway in humans) will be considered.

3. Two-component signal transduction

Two-component systems are widespread in bacteria, archaea and plants. In *Escherichia coli*, 30 sensor kinases and 32 response regulators have been found. The two interacting components are a sensor kinase and a response regulator (Fig. 4). Upon perception of a stimulus, the input domain of the sensor kinase modulates the signaling activity of its transmitter domain, resulting in autophosphorylation with the γ-phosphoryl group of ATP. Then, the phosphoryl group is transferred to the response regulator receiver domain, resulting in an activation of the output domain(s) to trigger a response. In most cases the response is an alteration in the transcription level of an special gene or gene cluster (see e.g. Parkinson, 1993 or Stock, Robinson, & Goudreau, 2000 for reviews).

In the simplest case the system can be described by a set of two reactions (Kremling, Heermann, Centler, Jung, & Gilles, 2004)

$$S \rightleftharpoons S^p \tag{11}$$

$$S^p + R \rightleftharpoons S + R^p. \tag{12}$$

In the first reaction (Eq. (11)), the stimulus enhances the kinase activity that results in autophosphorylation of the sensor kinase (S, S^p). In the second reaction (Eq. (12)), the phosphoryl group is transferred to the response regulator (R, R^p). R^p contains the active output domain. To turn off the system efficiently, the phosphoryl group is taken away in a dephosphorylation reaction. Here, two model variants are possible: (a) the phosphoryl group is cut off by an additional enzyme possessing phosphatase activity or (b) the sensor kinase acts as the dephosphorylating enzyme, leading to the addition of

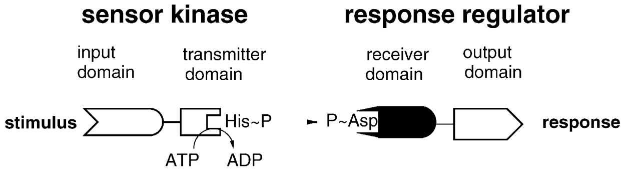


Fig. 4. General scheme of a two-component signal transduction system, as in Kremling et al. (2004) ©, with permission from Elsevier.

the equation

$$R^p \rightharpoonup R + P \tag{13}$$

or

$$R^p + S \rightharpoonup R + S + P \tag{14}$$

respectively. The latter is the case in a number of examples known from *E. coli*, e.g. the KdpD/KdpE system responsible for the regulation of protein KdpFABC, which is a potassium uptake system. It is assumed that the dephosphorylation is irreversible. In the model, the stimulus is a change in the reaction constant of the phosphorylation of S (Eq. (11)) and, as system output, the concentration of the phosphorylated response regulator is chosen. Experiments with purified enzymes of the *E. coli* KdpD/KdpE system revealed that the concentration of the phosphorylated response regulator is very low in absence of the DNA binding site (Kremling, Heerman, et al., 2004). Therefore, the DNA-binding step must be included into the model, leading to the addition of the equation

$$n\,R^p + \mathrm{DNA}_f \rightleftharpoons R - \mathrm{DNA}, \tag{15}$$

where n is the number of molecules which bind to the DNA binding site.

The amount of the regulator–DNA complex ($R - DNA$) is used as a measure for mRNA synthesis. In a further step, the mRNA is then translated to protein. The two polymerization steps are connected like a cascade:

$$\text{(nucleotides)} \rightharpoonup^{r_{\mathrm{tr}}} \text{RNA} \rightharpoonup \text{degradation} \tag{16}$$

$$\text{(amino acids)} \rightharpoonup^{r_{\mathrm{tl}}} Pr \rightharpoonup \text{degradation}, \tag{17}$$

where the rate of transcription r_{tr} is a function of $R - \mathrm{DNA}$ and the rate of translation r_{tl} a function of the available mRNA.

3.1. Modularization

The sensor kinase and the response regulator form a system (Eqs. (11) and (12)) which belongs to the strongly coupled type defined by Eqs. (9) and (10) (Fig. 3(h)). The additional dephosphorylation step (Eq. (13) or (14)) further increases the coupling of these subsystems which should be hence considered as a single module. The total number of DNA binding sites is much smaller than R^P. Hence, as discussed above, there is a unidirectional connection due to neglect of a current, as depicted in Fig. 3(c). $R -$ DNA and mRNA influence the synthesis of RNA (r_{tr}) and protein (r_{tl}), respectively, but are not consumed by them. This is similar to the manner in which the enzyme E influences the reaction rate in Eq. (5) (Fig. 3(g)). Therefore, there is a unidirectional connection due to neglect of a current. The system can thus be decomposed into different units free of retroactive effects, as depicted in Fig. 5.

4. Control of carbohydrate uptake

The control of the carbohydrate uptake in bacteria has been under investigation for a long time. Starting with the pioneering work of Monod, a number of components were detected which are responsible for the coordination of sugar uptake. It is widely accepted that the phosphotransferase system (PTS) is one of the important modules in the signal transduction machinery of bacteria. The PTS represents a transport system and at the same time is part of a signal transduction system responsible for carbon catabolite repression (Postma et al., 1993). Catabolite repression means the dominance of one carbohydrate uptake system over another one; if glucose and lactose are present in the medium, glucose is taken up first while lactose is taken up only after the depletion of glucose. The PTS covers a set of five reactions, where a phosphoryl group is transferred from phosphoenolpyruvate (PEP) through two common intermediates, enzyme I (EI, gene *ptsI*) and the phosphohistidine carrier protein (HPr, gene *ptsH*), to the substrate-specific EII and finally to the substrate (Fig. 6). If a sugar is taken up by the PTS and thereby converted to a phosphorylated form, e.g. glucose is converted to glucose 6-phosphate, the sugar is further metabolized; glycolysis is the link between the transport reactions and their energy supply. Metabolism of glucose 6-phosphate during glycolysis results

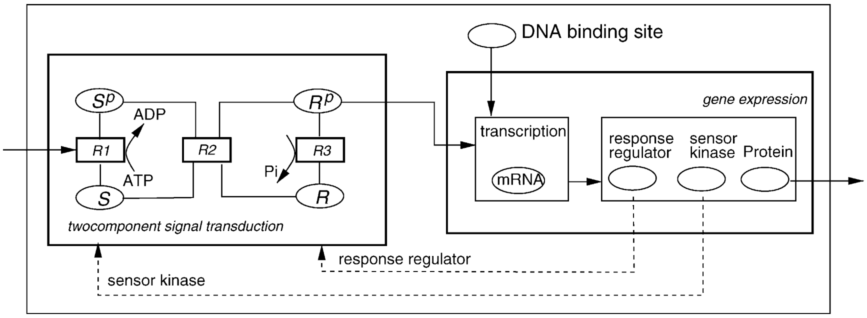


Fig. 5. Block diagram of the two-component system and gene expression, from Saez-Rodriguez et al. (2004)© 2004 IEEE. The entire system can be decomposed into units connected to each other in an unidirectional way. The output from the two-component unit is the phosphorylated form of the response regulator. Since the genes for the sensor and the response regulator are members of the same operon as the output protein, a positive feedback loop is established (dashed lines). However, the reactions inside the two-component do not influence *directly* the entire concentration of the proteins and the connection is, therefore, unidirectional.

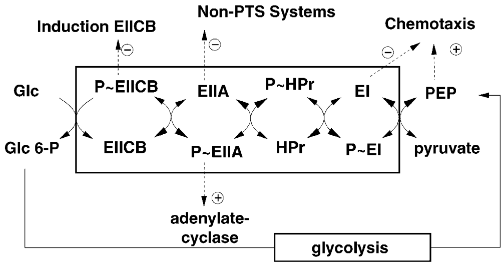


Fig. 6. Scheme of the five PTS reaction. Begin and end of the relay chain are connected by glycolytic reactions. Inputs are the concentration of the external glucose, PEP and pyruvate, from (Kremling, Fischer, Sauter, Bettenbrock, & Gilles, 2004)©, with permission from Elsevier.

in two moles of PEP. Beside glucose, a number of other carbohydrates are taken up by PTS in *E. coli*. In the case of glucose, EII^{Glc} consists of the soluble EIIA^{Crr} (gene *crr*) and the membrane-bound transporter EIICB^{Glc} (gene *ptsG*) for glucose uptake. All proteins involved in this phosphorylation cascade act as signaling molecules, e.g. EI in chemotaxis, HPr in glycogen metabolism, EIIA in inducer exclusion (inhibition of non-PTS transport systems), and the phosphorylated form of EIIA in activation of the adenylate cyclase. The adenylate cyclase produces the alarmone cAMP which is the activator of the transcription factor Crp. Currently, over seventy DNA binding sites for Crp are known.

4.1. Modularization

Fig. 7 summarizes the entire signal transduction unit starting from the PTS and ending with the binding of the transcription factor Crp complexed with cAMP to its corresponding binding sites, as modeled by Kremling et al. (2001). Considering the stoichiometry of the PTS, the reaction system extends the translocation of phosphoryl groups from two components, as shown above with the two-component system, to five components. The first reaction, the transfer of a phosphoryl group form PEP to pyruvate can be described as shown in Eq. (9), and the subsequent transfer of the phosphoryl groups can be described with a repeated sequence of the reaction given in Eq. (10). This system is, analogously to the two component system, strongly coupled and can not be decomposed into subunits. As can be seen in Fig. 7, PEP and pyruvate are inputs with retroactive effects, since they are involved in the metabolism; a flux through the PTS will affect the metabolism and, thus, alter the concentrations of both PEP and pyruvate. The connection through EIIA to other transport systems is also retroactive: EIIA binds to partner proteins (e.g. lactose permease or glycerol kinase), inhibiting non-PTS transport systems. The amount of EIIA complexed is not necessarily negligible, leading to a retroactive effect from the non-PTS to the PTS systems. The step to synthesize cAMP is assumed to be a function of the adenylate cyclase concentration. The phosphorylated form of EIIA (P$\sim$EIIA) acts as an activator binding to the adenylate cyclase and thus shifting it to an activated form. Since the concentration of P$\sim$EIIA is far larger than the cyclase concentration, the amount of P$\sim$EIIA bound to the enzyme can be neglected, leading to an unidirectional connection by neglect of a current. Since it is assumed that the concentration of the binding sites is smaller than the concentration of the cAMP·Crp complex, the output of the system, i.e., the binding of cAMP·Crp to its corresponding binding sites, is also regarded as unidirectional. Feedback loops (dashed lines in Fig. 7) from the Crp module occur because the PTS gene as well as the genes for the adenylate cyclase and Crp itself are under control of the cAMP·Crp complex.

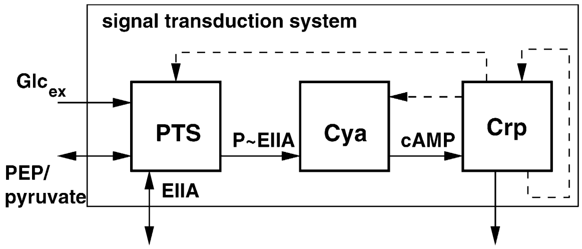


Fig. 7. Scheme of the signal transduction system responsible for carbohydrate uptake in *Escherichia coli*, adapted from Kremling et al. (2001)©, with permission from Elsevier. Inputs in the PTS are the external glucose concentration and the glycolytic intermediates PEP and pyruvate. EIIA in unphosphorylated form is responsible for inducer exclusion, while the phosphorylated form activates the adenylate cyclase to produce cAMP. cAMP is an alarmone which activates the transcription factor Crp. Dashed lines mark the feedback of Crp on the transcription of its own gene, the adenylate cyclase, and the PTS proteins.

5. MAP kinase cascade

The mitogen-activated protein kinases (MAPKs) are a family of highly conserved enzymes (a protein kinase is an enzyme which catalyzes the phosphorylation of a certain protein by ATP), which play a pivotal role in the transduction of signals in eukaryotes (Chang & Karin, 2001). There are several families of MAPKs, and at least four expressed in mammals: ERK-1/2, JNK, p38 and ERK5 (Chang & Karin, 2001). MAPKs have different names, but they share the same mechanism of activation: each MAPK (see Fig. 8), is phosphorylated at two points by another kinase—hence called MAPK kinase (MAPKK) or MAPK/ERK kinase (MEK)—which is also activated through a double phosphorylation by another kinase—called MAPKK kinase (MAPKKK) or MEKK. There also enzymes, called phosphatases, which reverse these phosphorylation steps (see Fig. 8).

In mammals, MAPK Cascades are involved in the response to a wide range of stimuli, ranging from growth factors to stress, which result in the regulation of essential cellular processes such as differentiation, cell proliferation and survival (Schaeffer & Weber, 1999). How MAPKs are able to produce specific responses to different stimuli is an issue not fully understood yet. Some of the mechanisms proposed are: (a) scaffold proteins, which bring together the elements of a

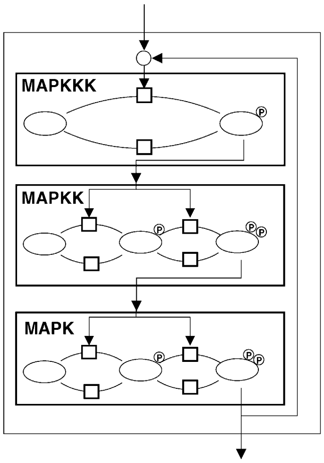


Fig. 8. Structure of the MAPK Cascade. The system can be decomposed into three modules. A positive or negative feedback from the last to the first module can be present.

MAPK Cascade, thus increasing its activity and specificity, (b) the spatial localization of signaling molecules and (c) the temporal organization (Schaeffer & Weber, 1999); for example, in PC12 cells, sustained ERK activation triggers cell differentiation, whereas a transient ERK activation leads to cell proliferation (Marshall, 1995).

5.1. Modularization

Considering its structure, the MAPK Cascade can be decomposed in three submodules corresponding to the three kinases, as depicted in Fig. 8. The connections between the three modules belong to the type discussed above (Eq. (4)). The assumption (i) introduced above does not hold, because the concentration information about E (e.g. MEK-PP) is needed in order to compute the dephosphorylation steps. However, the assumption (ii) might hold, depending on the values of the kinetic parameters and kinases concentrations. Some mathematical models (e.g. Brightman & Fell, 2000; Kholodenko, 2000) that include the MAPK cascade have been set up assuming (ii), i.e., the quasi-steady-state assumption—which implies the application of Michaelis Menten kinetics—while others have not (e.g. Schoeberl, Eichler-Jonsson, Gilles, & Müller, 2002). Even if the quasi-steady-state assumption does not hold, we think that a reaction catalyzed by an enzyme is still a suitable point for defining modules' borders, since the coupling is relatively weak (e.g. there is no net flux), thus providing pseudo-unidirectional connections.

The MAPK Cascade is a paradigm of modular system: through three subunits and eventually a feedback loop, the MAPK is able to perform several tasks. Probably, the most evident property of such a three-step structure is the amplification of the signal (Ferrell, 1996). However, the characteristic curve of the MAPK Cascade not only shows a high amplification, but also a sigmoidal form, a property termed by Goldbeter and Koshland (1981) ultrasensitivity. This switch-like behavior, which allows the cell to convert a gradual input into a binary response, is due to the double phosphorylation mechanism of activation (see Fig. 8) and the partial saturation of the kinases (Ferrell, 1996; Huang & Ferrell, 1996).

Several MAPK cascades have been found to be embedded in feedback loops, both positive and negative (see Fig. 8). A positive feedback can, together with the inherent ultrasensitivity of the MAPK Cascade, produce a bistable system and, if the feedback is strong enough, the system is able to give an *irreversible* on/off response to a transient continuous stimulus (Ferrell, 2002). Two MAPK Cascades, the JNK (Bagowski & Ferrell, 2001) and p42 (Ferrell, 2002) MAPK Cascades in *Xenopus* oocytes have been found to be bistable. On the other hand, a negative feedback could introduce oscillations (Kholodenko, 2000) and complete adaptation, which means that for a constant input the output signal goes back to the original value after a transient increase (Asthagiri & Lauffenburger, 2001).

6. EGF signaling pathway

The epidermal growth factor receptor (EGFR) is the prototype of the EGFR family, a group of receptors which belong to the tyrosine kinase receptors family (RTKs). RTKs are a large family of receptors for different ligands such as hepatocyte growth factor (HGF) and Insulin. The EGF receptor can bind to several growth factors including EGF and TGF-α (Yarden, 2001). Ligand binding promotes EGFR dimerization and autophosphorylation. This allows the formation of complexes formed by several signaling proteins, which activate many signaling pathways, including the MAPK Cascade. These steps are very similar in the case of other RTKs, and the molecules involved are to a wide range the same (Schlessinger, 2000).

EGFR signaling plays an essential role in mammalian development (Yarden, 2001). EGFR is overexpressed in a wide variety of human tumors (Wells, 1999). Therefore, the EGFR pathway has been intensively analyzed as a drug discovery target for cancer therapy and some of the resulting drugs are currently in clinical development (Ciardiello & Tortora, 2002).

EGFR is probably the best known receptor system, which has allowed the development of several models (Bhalla & Iyengar, 1999; Brightman & Fell, 2000; Kholodenko, Demin, Moehren, & Hoek, 1999; Schoeberl et al., 2002), recently reviewed by Wiley, Shvartsman, and Lauffenburger (2003). We have analyzed the EGF network model from Schoeberl et al. (2002), which describes the activation of the ERK MAPK Cascade by EGF. The model includes (a) the reception of EGF, (b) the formation of signaling complexes by interaction of several signaling proteins (namely Sos, Grb2 and Shc), (c) the activation of a signaling intermediate called Ras and (d) the activation of the Raf/MEK/ERK MAPK Cascade. Furthermore, the model also includes the internalization processes, hence duplicating all the steps de-

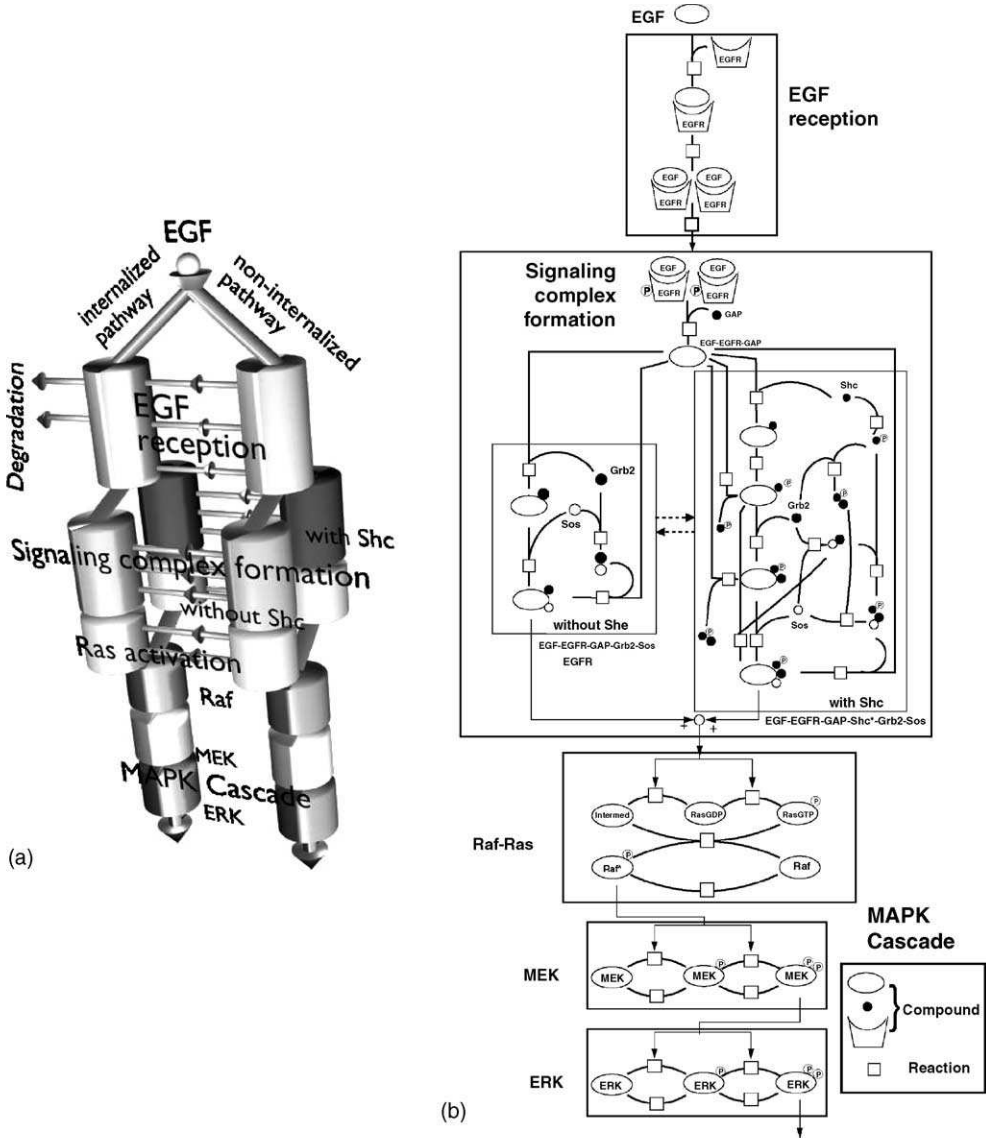


Fig. 9. Representation of the EGF signaling model of Schoeberl et al. (2002): (a) 3D representation of the whole model; (b) modular representation of the non-internalized part of the model. The module Raf–Ras corresponds to a modified version of the module MAPKKK of the MAPK Cascade, the module MEK to the module MAPKK and the module ERK to the module MAPK (see Fig. 8).

scribed above and increasing the complexity of the system (see Fig. 9(a)).

6.1. Modularization

If we focus on the non-internalized pathway of the model of Schoeberl et al. (2002), the system can be decomposed from a system-theoretical point of view as depicted in Fig. 9(b). The first module includes the EGF reception process up to the phosphorylation of the receptors. Although this phosphorylation ($v3$ in the model of Schoeberl et al. (2002)) is considered a reversible step in the model, it is still a suitable point to separate units, since the effect of the backward term is almost negligible, and the connection is hence almost free of retroactivity (data not shown). The next unit is the signaling complex formation. Here, we can distinguish two submodules, corresponding to the complex formation with and without the adaptor molecule Shc (see Fig. 9(b)). These two submodules are strongly coupled since they share several signaling molecules. However, the output signal of both units (compounds $c35$ and $c25$ in the model) is integrated into an enzymatic signal leading to a connection of the type discussed above (Eq. (8)). The next elements are the activation of Ras, Raf, MEK and ERK. The activations of Ras and Raf are strongly coupled, since Ras activates Raf through a reaction of the type of Eq. (3). Therefore, it is more reasonable to consider Ras and Raf as a unique module, obtaining three modules (Raf–Ras, MEK and ERK, see Fig. 9(b)) with enzymatic outputs ($c45$, $c51$ an $c59$, respectively). The connection between Raf–Ras and MEK modules, as well as the connection between MEK and ERK modules, belong to the type defined in Eq. (4) which, even if the quasi-steady-state assumption does not hold, are reasonable points for defining modules' borders, as discussed in the case of the MAPK Cascade.

7. Discussion

Thanks to new high-throughput techniques, the amount of experimental data about signaling networks is growing exponentially, leading to complex pictures of signaling pathways whose properties are not intuitively understandable. Dividing these networks in subunits might be a useful tool to tackle this complexity, but criteria for defining modules are still lacking. In this contribution it has been proposed that units without retroactive effects are reasonable modules since their properties show independence from their environment. Some criteria to find units free of retroactivity in signaling networks were introduced and applied to several examples. The analysis was performed by means of the network theory, which allows a clear identification of unidirectional connections due to neglect of current or potential vectors.

Two signaling networks of prokaryotes and two of eukaryotes have been analyzed. As a first example, the two-component system was chosen, since it is probably the simplest signaling system known. Next, the control of the carbohydrate uptake was considered. A part of it, the PTS system, is structurally an extension of the two-component system. The additional elements and, especially, the coupling to the metabolism, form a complex network with many retroactive effects. In eukaryotes, we first discussed the case of the MAPK Cascade, a central element of signal transduction in higher organisms. Finally, a part of the EGF signaling network, a complex system responsible of essential processes in human cells, was analyzed. The examples come from very distant organisms and seem to be very different. Remarkably, all the cases can, however, be analyzed using similar criteria. If other systems were analyzed, probably the same criteria would appear again, and maybe additional criteria for defining units without retroactive effects would arise.

Once these more manageable units are defined, they should be thoroughly analyzed, focusing especially on the dynamic properties (Saez-Rodriguez, Kremling, Conzelmann, Bettenbrock, & Gilles, 2004). Systems theory provides powerful tools to attempt this task. Due to the absence of retroactivity, the analysis performed on the isolated subsystem will not be altered when the module is embedded in the whole network.

A promising tool might be metabolic control analysis (MCA), which has proofed to be successful in the analysis of metabolic networks, and has been extended to signal transduction networks (Kahn & Westerhoff, 1991; Kholodenko et al., 1997). Furthermore, MCA provides a framework for modular analysis, which was firstly restricted to modules which do not share mass flows (Kahn & Westerhoff, 1991). This approach was later extended to modules which can share mass flows and have to fulfill conditions similar to the absence of retroactivity that was proposed within this paper (Schuster, Kahn, & Westerhoff, 1993). Recently, a new extention was presented which allows one to treat modules as black boxes, considering only the intermediates that mediate interactions between modules (Bruggeman et al., 2002). Importantly, MCA is based on a certain steady-state, which is a reasonable assumption for metabolic networks. However, signal transduction is prominently a dynamic process, where the transient behavior determines the response. Therefore, the application of MCA to signaling phenomena, yet useful, might be limited by the steady-state assumption. An additional limitation is that it provides information related to small changes in the signal, whereas signals often change drastically (Bruggeman et al., 2002).

Model reduction might also be a helpful tool for the analysis of the modules. A reduction of the complexity without losing the properties which are important for the function of the modules should provide more understandable units.

Feedback loops are ubiquitous in signaling networks and play an important role in the complex behavior of signaling processes. An appropriate modus operandi to unravel this complexity might be a stepwise analysis, investigating first the isolated modules, then the whole system without feedback loops and finally the complete network including the feedback effects (Saez-Rodriguez et al., 2004). While some emergent properties might be due to the interaction among the different modules, some might be determined by a certain subsystem. MCA can also be used to analyze feedback loops (Kahn & Westerhoff, 1991; Kholodenko et al., 1997). By understanding the *parts* and rejoining them, new insights into the properties of the *whole* system might be gained.

Acknowledgments

The authors would like to thank Prof. Lengeler for useful discussions. JSR would also like to thank B. Schoeberl for help with the EGF model and useful discussions. The authors are also thankful to anonymous reviewers for constructive comments. Work on the two-component system was supported by K. Jung and R. Heermann from TU Darmstadt and was financed in part by the Deutsche Forschungsgemeinschaft [JU 270/4-1 (K.J.)].

References

Asthagiri, A., & Lauffenburger, D. (2000). Bioengineering models of cell signaling. *Annual Review of Biomedical Engineering*, *2*, 31–53.

Asthagiri, A., & Lauffenburger, D. (2001). A computational study of feedback effects on signal dynamics in a mitogen-activated protein kinase (MAPK) pathway model. *Biotechnology Progress*, *17* (2), 227–239.

Bagowski, C. P., & Ferrell, J. E. J. (2001). Bistability in the JNK cascade. *Current Biology*, *11* (15), 1176–1182.

Bhalla, U., & Iyengar, R. (1999, January). Emergent properties of networks of biological signaling pathways. *Science*, *283* (5400), 381–387.

Bray, D. (1995). Protein molecules as computational elements in living cells. *Nature*, *376* (6538), 307–312.

Brightman, F. A., & Fell, D. A. (2000). Diferential feedback regulation of the MAPK cascade underlies the quantitative differences in EGF and NGF signalling in PC12 cell. *FEBS Letters*, *482* (3), 169–174.

Bruggeman, F. J., Westerhoff, H., Hoek, J., & Kholodenko, B. (2002). Modular response analysis of cellular regulatory networks. *Journal of Theoretical Biology*, *218*, 507–520.

Chang, L., & Karin, M. (2001). Mammalian MAP kinase signalling cascades. *Nature*, *410* (6824), 37–40.

Ciardiello, F., & Tortora, G. (2002). Anti-epidermal growth factor receptor drugs in cancer therapy. *Expert Opinion on Investigational Drugs*, *11* (6), 755–768.

Downward, J. (2001). The ins and outs of signalling. *Nature*, *411* (6839), 759–762.

Ferrell, J. E. J. (1996). Tripping the switch fantastic: how a protein kinase cascade can convert graded inputs into switch-like outputs. *Trends in Biochemical Science*, *21* (12), 460–466.

Ferrell, J. E. J. (2002). Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Current Opinion in Cell Biology*, *14* (2), 140–148.

Gilles, E. D. (1998). Network theory for chemical processes. *Chemical Engineering and Technology*, *21*, 121–132.

Goldbeter, A., & Koshland, D. (1981). An amplified sensitivity arising from covalent modification in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, *78* (11), 6840–6844.

Hartwell, L., Hopfield, J., Leibler, S., & Murray, A. (1999). From molecular to modular cell biology. *Nature*, *402* (Suppl. 6761), C47–C52.

Hoch, J. A., & Silhavy, T. J. (Eds.), 1995. *Two-component signal transduction*. Washington, DC: ASM Press.

Huang, C. F., & Ferrell, J. E. J. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proceedings of the National Academy of Sciences of the United States of America*, *93* (19), 10078–10083.

Kahn, D., & Westerhoff, H. (1991). Control theory of regulatory cascades. *Journal of Theoretical Biology*, *153*, 255–285.

Kholodenko, B. N., Hoek, J. B., Westerhoff, H., & Brown, G. C., 1997. Quantification of information transfer via cellular signal transduction pathways. *FEBS Letters*, *414*, 430–434.

Kholodenko, B. N. (2000). Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades. *European Journal of Biochemistry*, *267* (6), 1583–1588.

Kholodenko, B. N., Demin, O., Moehren, G., & Hoek, J. (1999). Quantification of short term signaling by the epidermal growth factor receptor. *Journal of Biological Chemistry*, *274* (42), 30169–30181.

Kitano, H. (2002). Systems biology: A brief overview. *Science*, *295* (5560), 1662–1664.

Krauss, G. (2001). Biochemie der Regulation und Signaltransduktion: das moderne Lehrbuch für Chemiker, Biochemiker. *Biologen und Mediziner*. Wiley-VCH.

Kremling, A., & Gilles, E. D. (2001). The organization of metabolic reaction networks: II. Signal processing in hierarchical structured functional units. *Metabolic Engineering*, *3* (2), 138–150.

Kremling, A., Heermann, R., Centler, F., Jung, K., & Gilles, E.D., 2004. Analysis of two-component signal transduction by mathematical modeling using the Kdpd/Kdpe system of *Escherichia coli*. Biosystems, in press.

Kremling, A., Jahreis, K., Lengeler, J., & Gilles, E. D. (2000). The organization of metabolic networks: a signal-oriented approach to cellular models. *Metabolic Engineering*, *2* (3), 190–200.

Kremling, A., Jahreis, K., Lengeler, J., & Gilles, E. D. (2001). The organization of metabolic networks III. Application for diauxic growth on glucose and lactose. *Metabolic Engineering*, *3* (4), 362–379.

Kremling, A., Fischer, S., Sauter, T., Bettenbrock, K., & Gilles, E. D. (2004). Time hierarchies in the Escherichia coli carbohydrate uptake and metabolism. *Biosystems*, *73* (1), 57–71.

Lengeler, J. W. (2000). Metabolic networks: a signal-oriented approach to cellular models. *Biological Chemistry*, *381* (9–10), 911–920.

Levitzki, A. (1996). Targeting signal transduction for disease therapy. *Current Opinion Cell Biology*, *8* (2), 239–244.

Mangold, M., Motz, S., & Gilles, E. D. (2002). A network theory for the structured modelling of chemical processes. *Chemical Engineering Sciences*, *57*, 4099–4116.

Marshall, C. J. (1995). Specificity of receptor tyrosine kinase signaling: Transient versus sustained extracellular signal-regulated kinase activation. *Cell*, *80* (2), 179–185.

Parkinson, J. S. (1993). Signal transduction schemes of bacteria. *Cell*, *73* (5), 857–871.

Postma, P. W., Lengeler, J. W., & Jacobson, G. R. (1993). Phosphoenolpyruvate: Carbohydrate phosphotransferase systems of bacteria. *Microbiology Reviews*, *57* (3), 543–594.

Rives, A., & Galitski, T. (2003). Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 1128–1133.

Saez-Rodriguez, J., Kremling, A., Conzelmann, H., Bettenbrock, K., & Gilles, E. D. (2004). Modular analysis of signal transduction networks. *IEEE Control Systems Magazine*, *24* (4), 35–52.

Schaeffer, H., & Weber, M. (1999, April). Mitogen-activated protein kinases: Specific messages from ubiquitous messengers. *Molecular and Cellular Biology*, *19* (4), 2435–2444.

Schlessinger, J. (2000). Cell signaling by receptor tyrosine kinases. *Cell*, *103* (2), 211–225.

Schoeberl, B., Eichler-Jonsson, C., Gilles, E., & Müller, G. (2002). Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature Biotechnology*, *20* (4), 370–375.

Schuster, S., Kahn, D., & Westerhoff, H. (1993). Modular analysis of the control of complex metabolic pathways. *Biophysical Chemistry*, *48* (1), 1–17.

Segel, L. A. (1988). On the validity of the steady state assumption of enzyme kinetics. *Bulletin of Mathematical Biology*, *50* (6), 579–593.

Stock, A. M., Robinson, V. L., & Goudreau, P. N. (2000). Two-component signal transduction. *Annual Review of Biochemistry*, *69*, 183–215.

Wells, A. (1999). EGF receptor. *The International Journal of Biochemistry and Cell Biology*, *31* (6), 637–643.

Wiley, H., Shvartsman, S., & Lauffenburger, D. (2003). Computational modeling of the EGF-receptor system: a paradigm for systems biology. *Trends in Cell Biology*, *13* (1), 43–50.

Yarden, Y. (2001). The EGFR family and its ligands in human cancer:signalling mechanisms and therapeutic opportunities. *European Journal of Cancer*, *37*, S3–S8.

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Metabolic Engineering 6 (2004) 364–377

METABOLIC ENGINEERING

www.elsevier.com/locate/ymben

Metabolic design based on a coupled gene expression—metabolic network model of tryptophan production in *Escherichia coli*

Joachim W. Schmid[a], Klaus Mauch[a], Matthias Reuss[a,*], Ernst D. Gilles[b], Andreas Kremling[b]

[a]*Institute of Biochemical Engineering, University of Stuttgart, Allmandring 31, D-70569 Stuttgart, Germany*
[b]*Max-Planck-Institut für Dynamik komplexer technischer Systeme, Sandtorstrasse 1, D-39106 Magdeburg, Germany*

Received 9 February 2004; accepted 24 June 2004
Available online 27 August 2004

Abstract

The presumably high potential of a holistic design approach for complex biochemical reaction networks is exemplified here for the network of tryptophan biosynthesis from glucose, a system whose components have been investigated thoroughly before. A dynamic model that combines the behavior of the *trp* operon gene expression with the metabolic network of central carbon metabolism and tryptophan biosynthesis is investigated. This model is analyzed in terms of metabolic fluxes, metabolic control, and nonlinear optimization. We compare two models for a wild-type strain and another model for a tryptophan producer.

An integrated optimization of the whole network leads to a significant increase in tryptophan production rate for all systems under study. This enhancement is well above the increase that can be achieved by an optimization of subsystems. A constant ratio of control coefficients on tryptophan synthesis rate has been identified for the models regarding or disregarding *trp* operon expression. Although we found some examples where flux control coefficients even contradict the trends of enzyme activity changes in an optimized profile, flux control can be used as an indication for enzymes that have to be taken into account in optimization.
© 2004 Elsevier Inc. All rights reserved.

Keywords: Tryptophan synthesis; Gene expression; Central carbon metabolism; Nonlinear optimization

1. Introduction

The most attractive processes to produce the aromatic amino acid tryptophan are enzymatic catalysis or fermentation from precursors like indole, serine, or anthranilic acid, and direct fermentation from carbohydrates (Leuchtenberger, 1996). Tryptophan is a very important amino acid that is widely used in medicine and also used as feed additive. In addition, other industrially relevant substances can be derived from tryptophan synthesis intermediates, like indigo (Ensley et al., 1983) or the anti-influenza drug Oseltamivir phosphate (Rasor and Voss, 2001). From an industrial point of view, high production rates are desirable, and by applying recombinant DNA technology, many attempts have been undertaken to improve product quality and quantity.

Early investigations aiming at the construction of tryptophan overproducing *Escherichia coli* strains (e.g. the systematic contribution by Tribe and Pittard (1979)) concentrate on genetic modifications in the anabolic pathways. The pathways examined are leading to tryptophan and the common intermediate for all aromatic amino acids, chorismate (e.g. Dell and Frost, 1993).

There have been several attempts in using a mathematical description of the pathways to optimize tryptophan production. The model of Xiu et al. (1997) considers repression of gene expression by the regulator Trp R as well as feedback inhibition of the enzymes in the tryptophan synthesis pathway. The effect of the growth rate and especially the demand of tryptophan for protein synthesis were taken into account by the authors. They conclude that the growth rate of the cells should be kept at

---

*Corresponding author.
E-mail address: reuss@ibvt.uni-stuttgart.de (M. Reuss).

1096-7176/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymben.2004.06.003

a minimal level. Marin-Sanguino and Torres (2000) refined the model and translated it into the S-system approach. Based on this description they used two optimization procedures to get a high tryptophan production rate. The parameters changed in their contribution are related to the efflux of tryptophan, the growth rate, the product inhibition, and the level of tryptophan repressor. The result of the optimization shows more than a four-fold increase of tryptophan production rate.

To summarize these and other activities (Sinha, 1988; Koh et al., 1997; Santillan and Mackey, 2001) for modeling and optimization of tryptophan production, it can be stated that the investigations so far have focused on the description of regulatory control structures inside the tryptophan pathway itself and the interplay with the product formation pathways of chorismate and tryptophan. This can be justified, since the tryptophan pathway can be regarded as a functional unit, showing a limited autonomy. This means that a precise stimulation of the unit by altering the environment leads only to changes in the unit under investigation, i.e. the response of the whole cellular network is restricted to a small part. However, it is well known from investigation of metabolic fluxes (e.g. Varma and Palsson, 1993a, b) that for optimization the whole network structure of the cell has to be taken into account.

As a first step in this direction, the supply of precursor metabolites by the central catabolic pathways comes to the fore. There have been several approaches to improve the supply of the tryptophan precursors, phosphoenolpyruvate and erythose-4-phosphate, as been reviewed by Bongaerts et al. (2001) and Nielsen (2001). Secondly, catabolic and anabolic parts of the network of precursor supply and product formation have to be explored as a whole rather than separated from each other. This necessity can be exemplified by the following fact: while the maximal yield of the first metabolite of tryptophan biosynthesis, 3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP), is 0.86 mole DAHP from 1 mole glucose (Liao et al., 1996), Schuster et al. (Schuster et al., 1999) found that only 0.45 mole tryptophan can be yielded from 1 mole glucose due to the additional consumption of phosphoenolpyruvate in the pathway of chorismate formation.

In this contribution, we combine a mathematical model describing the dynamics of important metabolites of the central carbon metabolism with a model describing in detail the regulation of *trp* operon expression, which codes the tryptophan pathway enzymes. The intention is to show that from a metabolic engineering point of view the design could be improved when the system of tryptophan biosynthesis from glucose is optimized considering carbohydrate uptake, precursor supply, the biosynthetic pathways, and gene expression regulation simultaneously rather than considering isolated parts of this network.

To describe transcription of the *trp* operon, two regulator proteins have to be described. Besides the description of the interaction between the repressor and the operator binding site, the RNA polymerase is necessary to start transcription. Since there is a large number of binding sites for the RNA polymerase and the number of binding sites for the repressor is restricted to the number of *trp* operon templates (normally between one and two, depending on the growth rate), a recently presented method to describe the interaction between two or more regulator proteins with the control sequence of the *trp* operon is applied (Kremling and Gilles, 2001). The method is based on a hierarchical view of the regulatory network where signals are transduced from the top level to the lower level, but not vice versa.

Consequently, we obtain a dynamic model that combines the behavior of the *trp* operon gene expression with the metabolic network of central carbon metabolism and tryptophan biosynthesis. This model is analyzed in terms of metabolic fluxes, metabolic control, and nonlinear optimization (Mauch et al., 2000).

2. Methods and model systems

Initial point of our investigation is the dynamic model of Chassagnole et al. (2002) that deals with the metabolic network of the central carbon metabolism of *E. coli* wild-type strain W3110. It comprises the phosphotransferase system (PTS) transporting glucose, the Embden–Meyerhoff–Parnas Pathway providing phosphoenolpyruvate, and the pentose phosphate pathway supplying erythrose-4-phosphate. The original model has been developed based on the measurement of metabolites in a continuous culture that has been perturbed by a glucose pulse. This model is expanded to describe the shikimate pathway, synthesizing chorismate, and the tryptophan production pathway, each pooled in a single reaction step. Enzyme activity in the tryptophan production pathway is considered as dynamic variable in order to take gene expression into account. Details of the differences to the model of Chassagnole et al. that are associated with this expansion are given below.

The model of gene expression accounts for both repression and attenuation of the *trp* operon genes and represents an application of a new method to describe the interaction between two or more regulator proteins with the control sequence of the *trp* operon (Kremling and Gilles, 2001). This method is explained in the following.

2.1. Transcription initiation

The method is based on the hierarchical structure of the regulatory network and calculates the transcription efficiency by neglecting minor important interactions betweens regulator proteins and DNA-binding sites. Since the RNA polymerase is essential for transcription, it represents the cellular or top level, while other regulator proteins have a special function (or are more specific) in metabolism, e.g. they are activators or inhibitors for the expression of specific genes. These regulator proteins are therefore assigned to further levels in the hierarchy. The hierarchical model structure allows signal transduction from the top to the lowest level but not vice versa. Therefore, some interactions of the proteins are neglected which leads to a simpler model structure in comparison to a complete model, including all interactions. In tryptophan enzyme synthesis, the *trp* aporepressor Trp R is activated by excess of tryptophan:

$$\begin{aligned} &Trp\,R + Trp \overset{K_1}{\rightleftharpoons} \mathrm{Trp\,R}^*, \\ &Trp\,R^* + Trp \overset{K_1}{\rightleftharpoons} \mathrm{Trp\,R}^{**}, \end{aligned} \tag{1}$$

where Trp R** represents the repressor occupied with two molecules of tryptophan. Arvidson et al. (1986) demonstrated that the two binding sites of the aporepressor are identical and independent, so the same constant K_1 can be used for both binding steps.

According to Fig. 6 in Kremling and Gilles (2001), cases A, B and B1 are valid for the interaction with the control sequence of the tryptophan operon D_{trp}. Case A denotes the formation of the initial complex by the RNA polymerase. In cases B and B1, the competition of the repressor and the RNA polymerase for the binding to the control sequence is described. The equations (equations according Table 3 in Kremling and Gilles (2001)) can be summarized as follows: the interaction of the repressor Trp R** with the operator site reads

$$\mathrm{Trp\,R}^{**} + D_{trp} \overset{K_{trp}}{\rightleftharpoons} TD_{trp}; \tag{2}$$

the binding of RNA polymerase to the promotor is described by

$$D_{trp} \overset{K(\psi)}{\rightleftharpoons} D^+_{trp}, \tag{3}$$

where D^+_{trp} represents the promoter occupied with RNA polymerase, and $K(\psi) = (1-\psi)/\psi$. If it is assumed that reactions (1)–(3) are very fast in comparison to the process of enzyme synthesis, a rapid equilibrium between the components is stated, resulting in the following algebraic equation system for the total amount of repressor c_{TrpR0} and the total amount of control sequence $c_{D0_{trp}}$:

$$\begin{aligned} &c_{D0_{trp}} = c_{D_{trp}} + c_{D^+_{trp}} + c_{TD_{trp}}, \\ &c_{TrpR0} = c_{TrpR} + c_{TD_{trp}} + c_{TrpR^*} + c_{TrpR^{**}} \end{aligned} \tag{4}$$

with

$$\begin{aligned} &c_{D^+_{trp}} = \frac{c_{D_{trp}}(1-\psi)}{\psi}; \; c_{TrpR^*} = \frac{c_{Trp}c_{TrpR}}{K_1}, \\ &c_{TrpR^{**}} = \frac{c^2_{Trp}c_{TrpR}}{K_1^2}; \; c_{TD_{trp}} = \frac{c^2_{Trp}c_{TrpR}}{K_1^2}\frac{c_{D_{trp}}}{K_{trp}}. \end{aligned} \tag{5}$$

For the model here, it is assumed that RNA polymerase and σ factor do not change and ψ is taken therefore as a constant parameter ($\psi = 0.91$ for normal promoters according to Kremling and Gilles (2001)). The rate of enzyme synthesis r is proportional to the fraction of occupied promoters $\bar{\psi}$ with

$$\bar{\psi} = \frac{c_{D^+_{trp}}}{c_{D0_{trp}}}. \tag{6}$$

To include attenuation also, a switch function is used. Since excess of tryptophan inhibits enzyme synthesis, the following equation is used to describe the rate of enzyme synthesis:

$$r_{\text{enzyme synthesis}} = k\bar{\psi}\frac{1}{1+k'(c^4_{Trp}/K + c^4_{Trp})}c_{D0_{trp}}, \tag{7}$$

where k indicates a rate constant of protein synthesis. It is assumed to be in the same range as estimated for the *lac* operon (Kremling et al., 2001). The term $c^4_{Trp}/(K + c^4_{Trp})$ represents attenuation and shows a threshold like behavior similar to a model proposed by Koh et al. (1997). k' is estimated from the assumption that attenuation can increase mRNA synthesis by a factor of 10 if tryptophan concentration is low (Neidhardt et al., 1990). The parameter K is chosen in such a way, that the transition from low mRNA synthesis to high mRNA synthesis takes place at a concentration of tryptophan of about 1 to 5 μM, as it is the case in the model of Koh et al. (1997). All parameter values of the gene expression model are summarized in Table 1.

2.2. Metabolism

In order to link this model of gene expression to the dynamic pathway model of Chassagnole et al. (2002),

Table 1
Parameter values of the gene expression model

Parameter	Value	Source
$c_{D^+_{trp}}$	5×10^{-3} μM	
c_{TrpR0}	4×10^{-2} μM	
K_1	30 μM	a
K_{trp}	2×10^{-3} μM	a
k	400 h^{-1}	
k_d	0.6 h^{-1}	
k'	9	See text
K	25 $\mu\mathrm{M}^4$	See text

Source a: Koh and Yap (1993).

the following expansions have been applied to the pathway model.

All enzymes of the tryptophan producing pathway are pooled into one state variable, the enzyme amount in the tryptophan pathway $c_{\mathrm{E},Trp}$. This state variable is balanced in consideration of gene expression, enzyme degradation, and dilution:

$$\frac{\mathrm{d}c_{\mathrm{E},Trp}}{\mathrm{d}t} = r_{\text{enzyme synthesis}} - r_{\text{Degradation}} - \mu c_{\mathrm{E},Trp} \tag{8}$$

Enzyme degradation is modeled as a first-order reaction with respect to $c_{\mathrm{E},Trp}$:

$$r_{\text{Degradation}} = k_{\mathrm{d}} c_{\mathrm{E},Trp} \tag{9}$$

The rate equations of chorismate synthesis, r_{ChoSynth}, and tryptophan synthesis, r_{TrpSynth}, are extended to mass action kinetics with respect to their substrates:

$$r_{\text{ChoSynth}} = r^{\max}_{\text{ChoSynth}}\, c_{\text{DAHP}}\, c_{\text{PEP}}\, c_{\text{NADPH}}, \tag{10}$$

$$r_{\text{TrpSynth}} = k^{\text{cat}}_{\text{TrpSynth}}\, c_{\mathrm{E},Trp}\, c_{\text{Cho}}\, c_{\text{PRPP}} c_{\text{Ser}}. \tag{11}$$

As a consequence, the balance equation for phosphoenolpyruvate now reads

$$\frac{\mathrm{d}c_{\text{PEP}}}{\mathrm{d}t} = r_{\text{ENO}} - r_{\text{PK}} - r_{\text{PTS}} - r_{\text{PEPCxylase}} - r_{\text{DAHPS}} - r_{\text{ChoSynth}} - r_{\text{MurSynth}} - \mu c_{\text{PEP}}. \tag{12}$$

r_{MurSynth} represents the consumption of phosphoenolpyruvate in mureine synthesis and is assumed to be constant.

Further balance equations have to be added for intermediate metabolites and the product, tryptophan:

$$\frac{\mathrm{d}c_{\text{DAHP}}}{\mathrm{d}t} = r_{\text{DAHPS}} - r_{\text{ChoSynth}} - \mu c_{\text{DAHP}}, \tag{13}$$

$$\frac{\mathrm{d}c_{\text{Cho}}}{\mathrm{d}t} = r_{\text{ChoSynth}} - r_{\text{TrpSynth}} - r_{\text{Synth3}} - \mu c_{\text{Cho}}, \tag{14}$$

$$\frac{\mathrm{d}c_{\text{Trp}}}{\mathrm{d}t} = r_{\text{TrpSynth}} - r_{\text{Trpremoval}} - \mu c_{Trp}, \tag{15}$$

$$\frac{\mathrm{d}c_{\text{PRPP}}}{\mathrm{d}t} = r_{\text{RPPK}} - r_{\text{Synth4}} - r_{\text{TrpSynth}} - \mu c_{\text{PRPP}}, \tag{16}$$

$$\frac{\mathrm{d}c_{\text{Ser}}}{\mathrm{d}t} = r_{\text{SerSynth}} - r_{\text{TrpSynth}} - r_{\text{Synth5}} - \mu c_{\text{Ser}}. \tag{17}$$

The rates r_{Synth3}, r_{Synth4}, and r_{Synth5} subsume the consumption of chorismate, phosphoribose pyrophosphate, and serine, respectively, in the production of biomass. They are assumed to be first order with respect to their substrates. The rate of tryptophan removal $r_{\text{Trpremoval}}$ is assumed to be first order with respect to tryptophan and represents the drain of tryptophan, but not exclusively in the production of biomass. For example, tryptophan might also be excreted by the gene products of *mtr*, *tnaB*, and *aroP* (Yanofsky et al., 1991).

Additionally, the inhibition of DAHPS by tryptophan has been taken into account in the wild-type strain models (models A and B, see below).

According to the approach of Rizzi et al. (1997) maximal rates are estimated from a stationary flux distribution, the concentration vector $\underline{c}$ and the vector of kinetic parameters $\underline{p}$:

$$r_i = r_i^{\max} f(\underline{c}, \underline{p}), \tag{18}$$

which results in

$$r_i^{\max} = \frac{r_{i,\text{stationary}}}{f(\underline{c}_{\text{stationary}}, \underline{p})}. \tag{19}$$

In r_{TrpSynth}, the activity $k^{\text{cat}}_{\text{TrpSynth}}$ is calculated in an analogous manner as follows:

$$k^{\text{cat}}_{\text{TrpSynth}} = \left[\frac{r_{\text{TrpSynth}}}{c_{\mathrm{E},Trp} c_{\text{Cho}} c_{\text{PRPP}} c_{\text{Ser}}}\right]_{\text{steady state}}. \tag{20}$$

The stationary concentration of serine is estimated to be 0.089 mM taking into account measurements from Piperno and Oxender (1968) and an assumed cell density of approximately 2.2 kg wet weight per liter cell volume. The stationary concentrations of tryptophan, chorismate and phosphoribose pyrophosphate are assumed to be 0.1 mM. The stationary concentration of tryptophan pathway enzymes is estimated from the parameters of the gene expression model using Eq. (8) and the steady-state condition $\mathrm{d}c_{\mathrm{E},Trp}/\mathrm{d}t = 0$.

2.3. Comparison of flux distributions

We studied the impact of different flux distributions comparing metabolic flux analysis of two wild-type models, differing in the assumption on the cofactor usage of isocitrate dehydrogenase, and a tryptophan overproducing strain from the contribution of Tribe and Pittard (1979). Model A is based on the flux distribution from Chassagnole et al. and describes the fluxes of a stationary, glucose-limited culture of *E. coli* wild-type strain W3110 at a growth rate of $0.1\,\mathrm{h}^{-1}$. It is assumed there that the cofactor of isocitrate dehydrogenase is NAD. Whereas, model B uses NADP as cofactor of isocitrate dehydrogenase, resulting in a NADP reduction flux parallel to the oxidative part of pentose phosphate pathway. Model C represents the strain NST 100 from the contribution of Tribe and Pittard (1979). The metabolic flux analysis is carried out based on the fluxes of glucose uptake, biomass production, and tryptophan excretion during the exponential growth phase. Model C also differs from the other models in the fact that there is no synthesis of phenylalanine and tyrosine. Since there is an unregulated copy of the *trp* operon in strain NST 100, gene expression regulation is removed in model C. In NST 100, feedback inhibition of DAHPS by tryptophan is removed as well.

Table 2 subsumes the differences between the models under investigation.

Table 2
Differences between the models under investigation

Model	Strain	Cofactor of isocitrate dehydrogenase	Gene expression regulated
A	W3110	NAD	Yes
A′	W3110	NAD	No
B	W3110	NADP	Yes
C	NST100	NADP	No

For model C, it is furthermore assumed that there is no phenylalanine and tyrosine production as well as no tryptophan inhibition of DAHPS.

2.4. Design of the dynamic system

The resulting dynamic system has been explored in terms of metabolic control analysis. The control on tryptophan synthesis rate in models A, B, and C has been compared.

Furthermore, tryptophan synthesis rate has been maximized according to the optimization problem

$$\max_{r_i^{\max}} r_{\text{TrpSynth}} \tag{21}$$

by variation of maximal reaction rates $r_i^{\max}$ for a set of reactions. Different choices of design parameter sets are explained together with the results of the corresponding optimization.

We performed the optimization considering the following constraints:

$$\frac{1}{m}\sum_{i=1}^{m}\frac{|c_{i,\text{Optimum}}^{\text{steady state}} - c_{i,\text{Referenz}}^{\text{steady state}}|}{c_{i,\text{Referenz}}^{\text{steady state}}} \leqslant \Theta, \tag{22}$$

$$\frac{1}{w}\sum_{i=1}^{w}\frac{r_i^{\max}}{r_{i,\text{Referenz}}^{\max}} \leqslant \Omega, \tag{23}$$

regarding homeostasis and total enzyme activity, respectively (Mauch et al., 2000). A substantial change in metabolite concentrations may be impedimental for vitally important cellular functions or lead to undesired flux diversion. Deviations of pool concentrations are constrained here to average 30% maximum ($\Theta = 0.3$). Total enzyme activity is constrained not to increase ($\Omega = 1$) to avoid a higher demand for total protein production that may result in a stress situation with unforeseeable regulatory impact. Furthermore, stability of models A, B, and C prior to perturbation of maximal rates has been demonstrated by an investigation of the eigenvalues of the corresponding Jacobian. Systems with perturbed maximal rates are simulated over a time span of at least 10 times the maximal time constant of the unperturbed system.

To test the uniqueness of optimization, three optimization strategies have been compared: (1) gradient method starting from the original values of all maximal rates, (2) gradient method starting from a tenth of the original enzyme concentrations, thus placing enzyme amount at the disposal from the beginning of optimization, and (3) simulated annealing (Kirkpatrick et al., 1983) starting from the original maximal rates.

The realization of an optimal enzyme activity distribution necessitates a manageable number of design parameters. Furthermore, the complexity of nonlinear optimization increases superproportional with the number of design parameters. Therefore, we try to reduce the number of optimized enzyme activities. Two approaches have been applied: (1) Only a part of the network is taken into consideration. For optimization, we considered the classical biochemical pathways glycolysis (including glucose transport), pentose phosphate pathway, and biosynthesis of tryptophan from DAHP. (2) Furthermore, we reduced the set of optimized enzyme activities based on the results of the analysis of flux control.

3. Results

We compare two wild-type models (A and B) and the model for the tryptophan producer NST100 from Tribe and Pittard (1979) (C). Our focus is especially on the interrelations between flux distribution, flux control, and optimization potential. In this contribution, optimization potential is referred to as the tryptophan production rate after optimization related to the same rate in the original steady state.

3.1. Flux distributions

The stationary flux distributions in the three models are shown in Fig. 1.

In model A, all NADPH has to be generated through the oxidative part of pentose phosphate pathway. Thus, fluxes through G6PDH and PGDH are well above the same fluxes in model B, where additional NADPH is generated in the isocitrate dehydrogenase reaction. Consequently, the fluxes in the nonoxidative part of pentose phosphate pathway are also smaller in model B. Flux through the lower part of glycolysis from GAP to Pyr is smaller in model A as compared to model B. This is also a result of the higher flux through PGDH, since PGDH activity is connected with the loss of one carbon atom per molecule ribulose 5-phosphate generated.

As compared to model B, there are only few fluxes in model C that differ remarkably. Most obvious is the 10-fold flux through tryptophan synthesis. In consequence, the serine and PRPP supplying fluxes also increase. Due to the higher usage of serine and PRPP in biomass synthesis, the percentage increase of these supplying fluxes is less pronounced. Surprisingly, the flux through

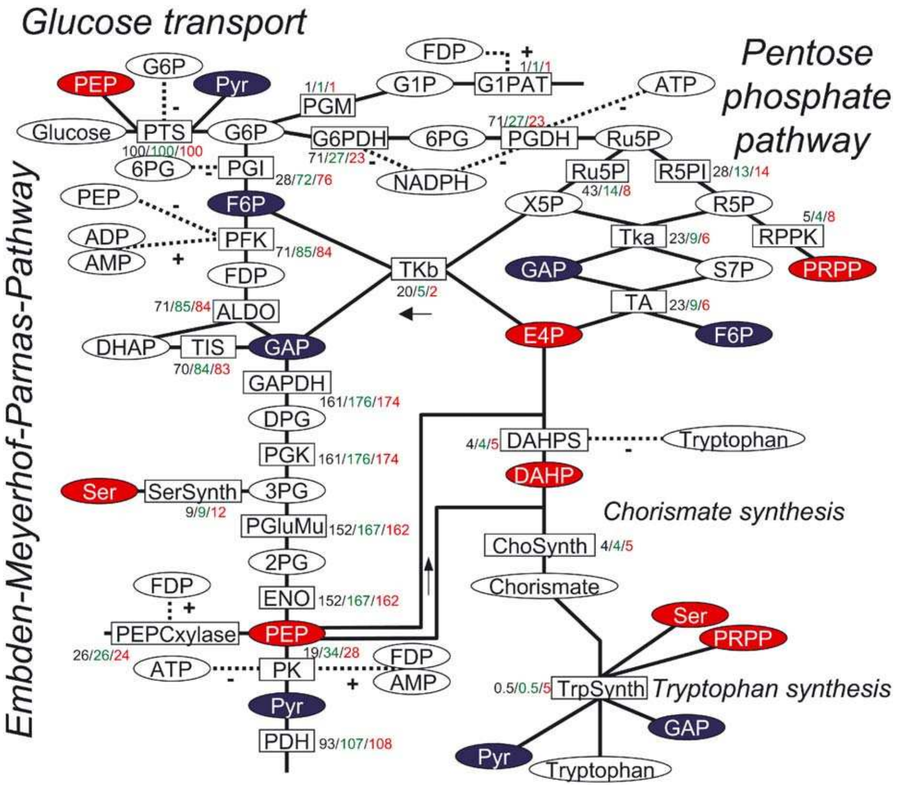


Fig. 1. Metabolic network of tryptophan biosynthesis from glucose as carbon source. Molar fluxes are given as numbers next to the enzyme symbols. They are normalized to glucose uptake flux (100). The first number results from the assumption that isocitrate dehydrogenase (ICD) uses NAD as cofactor (Model A), for the second number NADP is assumed to be the cofactor of ICD (Model B). The third number results from flux analysis for the strain NST100 (Tribe and Pittard, 1979) assuming NADP as cofactor of ICD (Model C). Colored metabolites represent important internal links of the reaction network. Except for E4P and DAHP, these metabolites are displayed more than once to allow for a clearer arrangement.

chorismate synthesis did not change significantly. Thus, according to our flux analysis, the main improvement in the tryptophan overproducing strain NST 100 is the successful channeling of all chorismate to tryptophan, as well as avoiding feedback phenomena due to product accumulation.

3.2. *Effects of changes in flux distribution on control hierarchy*

The control of enzyme activities in the network on the tryptophan production flux has been quantified by means of flux control coefficients. These are compared in Fig. 2 for the different flux distributions from the previous section.

Flux control is distributed. PTS, PFK, PDH, DAHPS, Trp, and supply with serine and PRPP are carrying high positive control on tryptophan biosynthesis in the wild-type models. Among the enzymes that are exerting negative control, GAPDH, PK, G6PDH, PEPCxylase, and consumption of the intermediates serine, PRPP, and chorismate are noteworthy. Thus, high control can be associated with

1. *glucose transport*: Not only the transport system PTS itself comes into focus, but also the supply with the co-substrate phosphoenolpyruvate (negative control by PK and PEPCxylase), as well as removal of the products of glucose transport (PFK removing G6P via the very fast reaction of PGI, PDH removing pyruvate).
2. *tryptophan biosynthesis*: Especially branch point reactions (DAHPS, Trp) exert high control, whereas control by the reaction Cho is negligible in all models.
3. *precursor availability*: Both synthesis and consumption in side reactions have to be considered. However, side reactions that are leading to biomass synthesis should not be reduced in order to ensure sufficient growth.

Flux control also points at effects that are not as easily to be explained, or even surprising, such as the considerable negative control by GAPDH and G6PDH.

The main difference between models A and B in flux control can be traced back to the difference in flux distribution. Consistent with the low flux through the oxidative part of pentose phosphate pathway, the

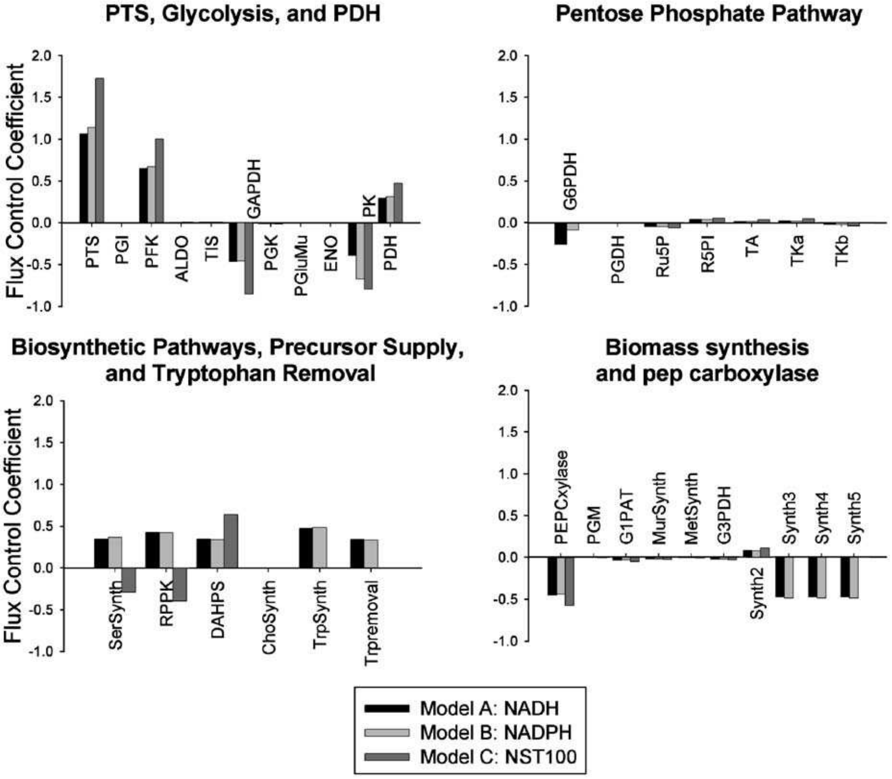


Fig. 2. Flux control coefficients on tryptophan production flux.

control of tryptophan synthesis by G6PDH has diminished. Instead, negative control by PK has increased distinctly.

In the overproducer model C, control shifted from tryptophan synthesis and supply of serine and PRPP to the central carbon metabolism. The increase in tryptophan synthesis flux as well as the enhanced supply with serine and PRPP release the system from restriction by the related enzymes. Serine and PRPP synthesis now even compete with supply of PEP and E4P and thus exert negative control. This competition is also indicated by the increased positive control by DAHPS, the enzyme that is channeling PEP and E4P to tryptophan synthesis. The release from restrictions in the biosynthetic pathways is accompanied by an increase in control by central carbon metabolism, especially by glucose transport (PTS, but also PFK) and GAPDH.

3.3. *Influence of gene expression regulation on flux control*

In the following, the impact of the gene expression model used here on flux control has been taken into focus. For comparison, a simplified model A′ has been derived from model A, where the concentration of *trp* operon enzymes is assumed to be constant, and thus gene expression regulation has been removed. In model A, the initial concentration of tryptophan c_{trp}^0 has been varied, whereas it has been kept at $c_{trp} = 0.1\,\mathrm{mM}$ in model A′. The initial tryptophan concentration is at the same time the stationary concentration, since the activity $k_{\mathrm{TrpSynth}}^{\mathrm{cat}}$ is calculated according to Eq. (20) after assuming the initial concentration.

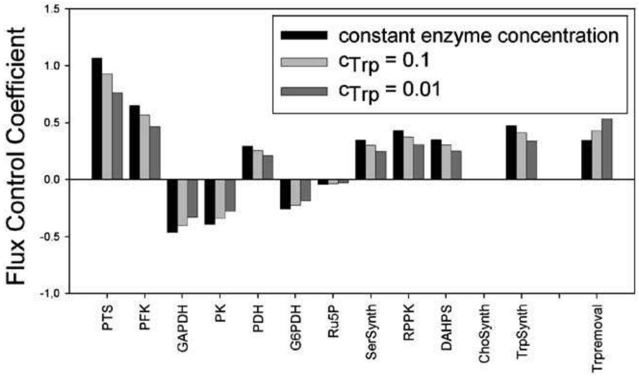


Fig. 3. Flux control coefficients on tryptophan production flux. Model A′ (constant enzyme concentration, tryptophan concentration 0.1 mM) compared to model A (gene expression model), where the tryptophan concentration was varied.

In Fig. 3 the flux control coefficients are compared for models A′ and A. All control coefficients, except for the control of tryptophan removal (Trpremoval), are smaller in model A. Thus, in this model gene expression has a damping influence on most control effects. Moreover, the ratio of flux control coefficients in models A and A′ has the same value for all reactions except Trpremoval. This finding stays valid if tryptophan concentration is varied in model A. The constant ratio between the flux control coefficients can be explained by the regulation of gene expression. In our model, expression of *trp* operon genes is regulated solely by the tryptophan concentration. The tryptophan concentration in turn depends on the fluxes through tryptophan biosynthesis and tryptophan removal (Trpremoval). All other reaction steps in the model have a rather indirect impact on tryptophan concentration. They determine the concentrations of precursors and intermediates and thus influence the rate of tryptophan biosynthesis. Consequently, their impact on *trp* operon expression is essentially coupled to the impact of the tryptophan biosynthesis reaction.

Fig. 4 shows the ratio between flux control coefficients on tryptophan biosynthesis in models A and A′, except for the coefficient for tryptophan removal. The tryptophan concentration has been varied in model A. At high tryptophan concentrations, the control in model A approximates that in model A′; gene expression influences flux control only a little. At low tryptophan concentrations, control approximates a constant level above the control in model A′. There are two effects that explain the pattern of the ratio in Fig. 4. Since flux control coefficients describe infinitesimal changes, both effects can be traced back to elasticities towards tryptophan concentration. In Fig. 4, the elasticities of DAHPS rate ($\partial r_{\mathrm{DAHPS}}/\partial c_{trp}$) and *trp* operon expression rate ($\partial r_{\text{enzyme synthesis}}/\partial c_{trp}$) on tryptophan concentration in model A are also depicted. DAHPS inhibition by tryptophan is included in both models A and A′. While kept constant in model A′, the concentration of tryptophan has been decreased in model A, leading to a release of DAHPS from inhibition. Accordingly, the flux control coefficients are larger in model A with low tryptophan concentration than in model A′ with unchanged tryptophan concentration.

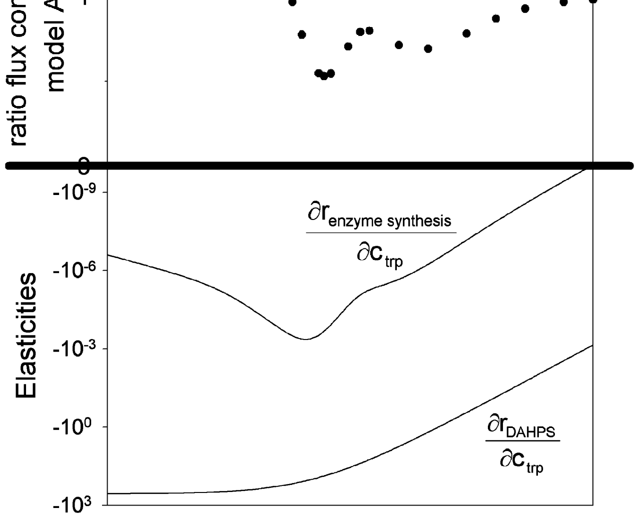


Fig. 4. Factor between control in model A′ (constant enzyme concentration, tryptophan concentration 0.1 mM) and in model A (gene expression model), where tryptophan concentration was varied.

The elasticity of *trp* operon expression is negative in the whole range of tryptophan concentrations. Thus, *trp* operon gene expression indeed has a damping effect on flux control of tryptophan synthesis. There is a pronounced minimum of expression elasticity at a tryptophan concentration of 0.0011 mM that can be brought in connection with attenuation. This minimum coincides with the global minimum of the flux control coefficient ratio. There is a local minimum of flux control coefficients at a tryptophan concentration of about 0.01 mM. This minimum coincides with a transitionally flattening trend of the gene expression elasticity due to repression.

3.4. Optimized enzyme amount distributions

So far we have investigated the impact of flux distribution and gene expression regulation on control of tryptophan flux. In the strict sense, flux control coefficients characterize infinitesimal changes in flux caused by infinitesimal changes in enzyme activities. In the following, we advance to a global optimization of tryptophan flux as described in Section 2.4.

The optimization of enzyme activities leads to the improved tryptophan production rates that are given in Fig. 5. First, we focus on the results where only a part of the network is taken into account as design variables for optimization. Fig. 6 shows the assignment of enzymes to subsystems of the network. For the following piecewise optimization, enzyme activities inside a subsystem are optimized, while all enzyme activities outside the particular subsystem remain unchanged.

3.4.1. Subsystem I: glycolysis

When enzyme activities in glycolysis are optimized, an improvement of tryptophan production rate of approximately 50% is achieved independent of the optimization strategy. The potential of rate enhancement is approximately the same for all models, including model C, which is the model of the overproducing strain NST 100. We obtain nearly the same improvement when only enzyme activities with flux control coefficients larger than 6% are optimized. In the latter case not only the

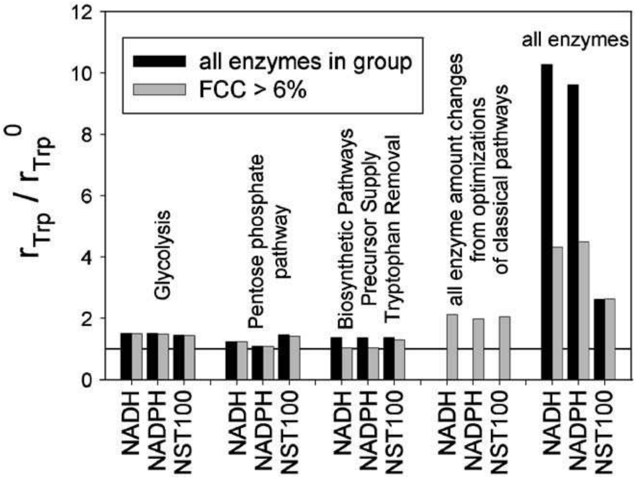


Fig. 5. Improvement of tryptophan production rate achieved by optimization of enzyme activities.

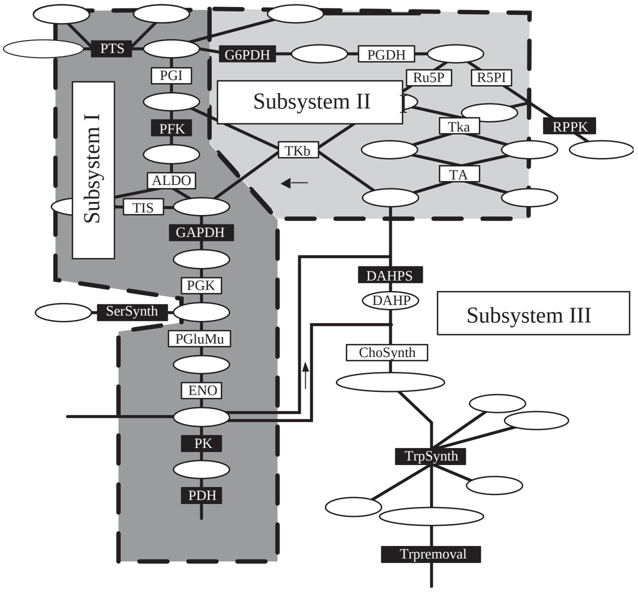


Fig. 6. Decomposition of the reaction network for choice of design parameters in optimization. Enzymes are assigned to the three subsystems (I) glycolysis, glucose transport, and pyruvate dehydrogenase, (II) pentose phosphate pathway, and (III) tryptophan biosynthesis including supply with intermediates. Reactions in black boxes carry a flux control coefficient larger than 6% (absolute value).

improvement, but also the optimized profile of enzyme amounts is independent of the optimization strategy. These findings indicate that there is one global optimum.

3.4.2. Subsystem II: pentose phosphate pathway

When enzyme activities in the pentose phosphate pathway are regarded, the optimized production rates differ for models A, B and C. While one could achieve an enhancement of 23% in model A, the potential reduces to 7% in model B. This decrease in optimization potential corresponds to the subordinate role of pentose phosphate pathway in model B, as being indicated by the lowered steady-state flux through G6PDH and the decreased control on tryptophan flux by this enzyme. In both models A and B the same improvement of tryptophan production rate could be achieved by deletion of G6PDH alone. In model C, however, an increase of tryptophan production rate by 46% is calculated. This result is also predicted when only Ru5P activity is reduced. The higher flux to tryptophan in the initial state leads to an accentuated role of pentose supply that causes the higher potential for improval in model C. Both deletion of G6PDH and reduction of Ru5P activity were found as local optima for all models, while the global optimum changes from deletion of G6PDH in models A and B to lowered Ru5P activity in model C.

3.4.3. Subsystem III: tryptophan biosynthesis

Variation of enzyme activities in the tryptophan biosynthesis, serine and PRPP supply, and in tryptophan removal yielded in a production rate that exceeds the starting value by approximately 35%. When the only reaction with a flux control coefficient below 6%, the shikimate pathway reaction ChoSynth, is not considered in the optimization, no significant increase in tryptophan production rate could be achieved. Since all other reactions exert pronounced positive control on tryptophan flux in models A and B, the activity of ChoSynth has to be decreased due to the constraint for total enzyme activity. In model C, three reactions carry a flux control coefficient above 6%: serine synthesis, PRPP synthesis, and DAHPS. An optimization of these three activities yields 28% improvement of tryptophan production rate. In contrast to the results for models A and B this enhancement is substantial, but still below that for all enzymes in the biosynthetic group of reactions. The existence of both reactions with negative and positive control allows for flux improvement without considering insensitive activities in the optimization. Yet there remains a lack of improvement due to the choice of varied activities based on flux analysis, which underestimates the role of the tryptophan synthesis.

3.4.4. Combination of piecewise optimization vs. integrated optimization

For all three models, we combined the changes in enzyme activities found by piecewise optimization of the three subsystems from Fig. 6. For subsystem I and II, we took the results, where only enzymes with control larger than 6% are regarded. The results for subsystem III show that in this subsystem all reactions have to be taken into account. When all these changes in enzyme activities are applied to the model, the tryptophan production rate is approximately doubled, as stated in

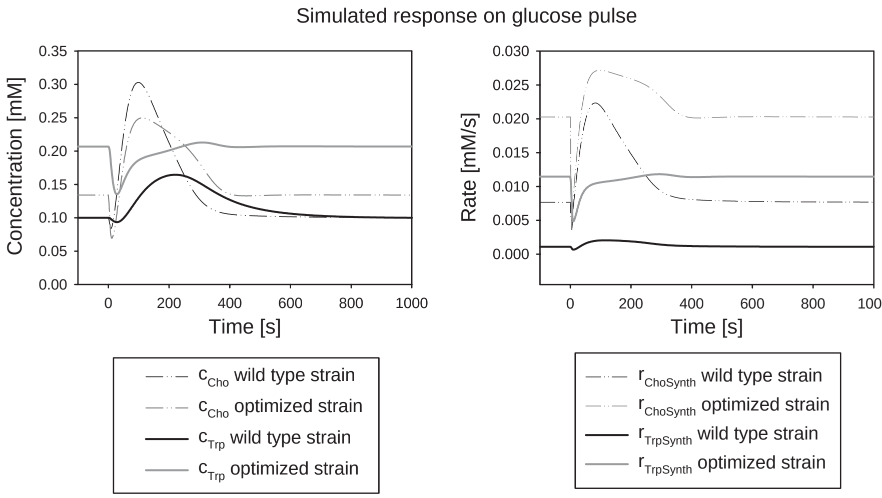


Fig. 7. Response on a glucose pulse according to Chassagnole et al. (2002). Comparison of model A and the respective optimized strain when all enzymes regardless of flux control are considered as design parameters.

Fig. 5. The combined changes, however, lead to a violation of the constraint for concentrations, and the suggested enzyme activity profile is therefore no solution of the optimization problem. Nevertheless, we compare this result to optimizations where enzyme activities of all three subsystems have been optimized at the same time. Considering all enzymes regardless of flux control, we found that tryptophan production rate could be increased approximately 10-fold in models A and B. Fig. 7 compares the response of model A and the respective optimized strain on a glucose pulse as described in Chassagnole et al. (2002). The investigated optimized strain turns out to be robust enough to return to its stationary state after this substantial perturbation. The pulse response of the optimized strain is qualitatively similar to that of the wild-type strain, but with noticeable quantitative differences.

When only reactions with a flux control coefficient above 6% are regarded, the achievable optimized rate reduces to 450% of that in the initial state. As it is the case for subsystem III, Eq. (23) represents a more demanding constraint on the amount of enzymes with large positive flux control if insensitive reactions are not considered, leading to a smaller flux enhancement potential compared to the case where all enzyme amounts are varied. In both cases, the combination of the separately optimized parts of the network lags behind the integrated optimization of the whole network.

This result indicates that this complex reaction network contains too many internal links to be optimized by a separated investigation of its parts. Nevertheless, a potential explanation of the difference between separated and integrated optimization could be provided by the optimized enzyme activities shown in Fig. 8. In all models, the activities of enzymes in biosynthesis, serine and PRPP supply, and tryptophan removal are increased on the cost of enzymes in the central carbon metabolism. This shift in activities to the production pathways is not possible if the pathways are optimized separately. In the central carbon metabolism, enzymes with high control on tryptophan production rate are found to be generally excluded from the activity shift. Surprisingly, this also holds for GAPDH, although it exerts a substantially negative control on tryptophan production. This enzyme may be of significant importance for homeostasis, keeping the concentrations at the initial level. Optimizations without constraint for concentrations showed, though, that this is not the only reason why GAPDH activity is kept at a high level in optimization. Thus, the role of GAPDH in the nonlinear model seems not to be identifiable by a linearizing analysis method alone, such as metabolic control analysis.

In contrast to the separated optimizations, not all optimization strategies lead to the maximal improvement in the integrated optimizations. The dependence of the achieved optimum on the starting point and on the optimization method points to the existence of local optima.

In model C, the maximized tryptophan production rate is 260% compared to that in the initial state, obtained by integrative optimization of all enzyme activities, regardless of flux control. Although the enhancement is well below that for the wild-type strain models A and B, this increase is remarkable for an overproducing strain. Considering only reactions with a flux control coefficient above 6%, the optimized tryptophan production rate is also reaching the maximal improvement. The activity of tryptophan synthesis, removal, and supply of serine and PRPP is increased to a lesser extent. This also results

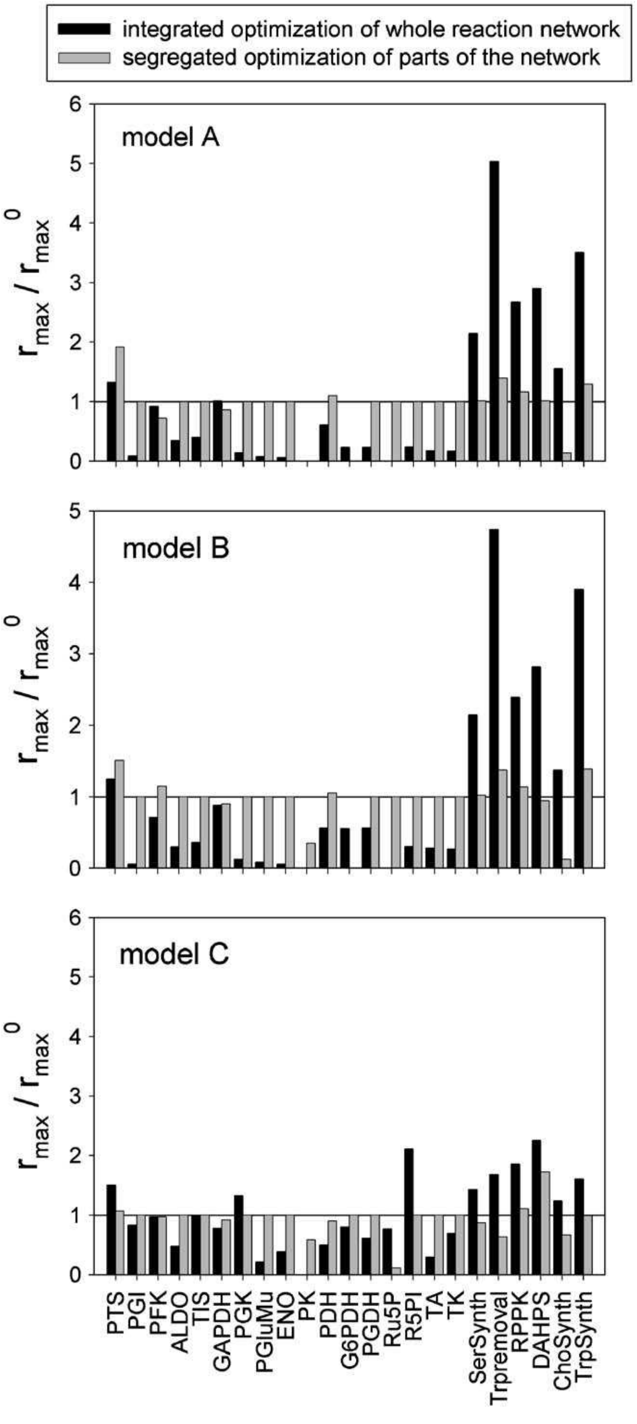


Fig. 8. Optimized enzyme activities for maximal tryptophan production rate.

in a smaller difference between the tryptophan production rate after separated and integrated optimization of network subsystems as compared to models A and B. Still, consideration of the network as a whole leads to a significantly higher optimized production rate.

4. Discussion

In the following, we discuss the improvements in tryptophan production rate obtained by optimization. Our focus will be on the interplay of flux distribution, flux control, and flux optimization, as well as on the role of gene expression regulation.

4.1. *Predicted improvement potential*

In all models we studied in this contribution, an integrated optimization of the whole network leads to a significantly higher increase in tryptophan production rate than the collection of segregated optimizations. Although this difference could partly be attributed to a lack in shift of activities towards anabolism in the segregated optimizations, this finding emphasizes the significance of network complexity.

It is noteworthy that even for the flux distribution of the overproducing strain NST 100 we found that the flux to tryptophan could be increased to 260% of that in the initial state. This enhancement has been reached in the model without changes in network topology, such as replacement of the PTS system for glucose transport. This exemplifies the potential of direct optimization in silico.

The existence of local optima points to the question how far a sequential improvement of strains by rational design might lead into local optima and thus suboptimal solutions.

4.2. *trp operon expression and flux control*

Regulation of gene expression leads to a constant ratio between the control of tryptophan synthesis rate in the model where *trp* operon expression was regarded and the model where it was not. Thus, while gene expression regulation has a strong impact on tryptophan flux itself, it does not change the hierarchy of flux control. This scaling holds for the whole tryptophan synthesis network and can be attributed to the fact that in this case regulating and regulated concentration are identical. While the situation will be different for regulation networks, this case is not unusual for biosynthesis pathways. The damping effect of inhibition and gene expression regulation on flux control leads to the obvious clue that both effects have to be eliminated for improvement of tryptophan production.

4.3. *Cross-links between flux distribution, control and optimization*

In our model, we found the following correlations between flux distribution, flux control, and the optimized enzyme activity distribution.

Only few fluxes differ remarkably between the two wild-type flux distributions and the flux distribution for the overproducer NST 100. These differences influence flux control, as well as optimization. In model B, the alternative source of NADPH in the tricarbonic acid cycle leads to a lowered flux through pentose phosphate pathway, as being compared to model A. This difference in flux distribution correlates with the lowered control by G6PDH, as well as with the reduced potential of rate enhancement by enzyme activity optimization in the pentose phosphate pathway in model B. Flux distribution in model C differs mainly in an increased flux through tryptophan synthesis, as well as the serine and PRPP supplying reactions. The change in flux distribution corresponds to a shift in control towards central carbon metabolism. The enhanced supply reactions, serine and PRPP synthesis, exert negative control due to competition for precursor molecules. In optimization, there is a reduced shift in enzyme activity towards biosynthetic and supplying reactions.

Furthermore, flux control might be an indication for enzymes that have to be taken into account in optimization. Optimization only of enzyme activities with high absolute values of flux control coefficients is promising in small, linear pathways which contain both enzymes with positive and negative control. Enzymes with low control only come into play, when the constraint of total enzyme activity gets dominant over the other constraints, especially for homeostasis. Such a situation emerges, for example, if all high absolute values of flux control coefficients are positive. For more complex systems, the indication works better if there is less shift between the parts that are optimized separately. Such a shift could be quantified and integrated into optimization using a preceding analysis, like e.g. the determination of group flux control coefficients (Stephanopoulos and Simpson, 1997), and a successive weighting of reaction group activities.

Still, for predicting promising changes in enzyme activity distribution, metabolic control analysis is no substitute for optimization of a detailed model. There are some examples in our optimization where flux control coefficients do not correlate with the trends of enzyme activity changes in an optimized profile. Especially in the wild-type models, the outstanding shift of enzyme activities towards anabolic reactions could not be foreseen by control analysis. In contradiction to this shift, some of the flux control coefficients in the central carbon metabolism are of the same magnitude or even outrange that of the anabolic reactions. Second, GAPDH would have been identified as a main target for decrease of enzyme activity by control analysis, while GAPDH is still present in considerable activity in all optimized profiles. In model C, control analysis would even predict a smaller activity of serine and PRPP supplying reactions to be advantageous, while both activities are significantly increased in the integrated optimization of the whole network.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 412.

Appendix A. List of symbols and indices

$c_{\mathrm{E},Trp}$	concentration of enzymes of the *trp* operon
c_{TrpR0}	total concentration of *trp* aporepressor
D_{trp}	free operator sites of *trp* operon
D^{+}_{trp}	promoter occupied with RNA polymerase
$D0_{trp}$	total *trp* operon
K_1	association constant of tryptophan to aporepressor
K_{trp}	association constant of corepressor–aporepressor complex to DNA
k	protein synthesis rate constant
k_d	protein degradation rate constant
k', K	parameters of threshold function for attenuation
TD_{trp}	operator site of *trp* operon with associated corepressor–aporepressor complex
$TrpR$	unoccupied *trp* aporepressor
$TrpR^{*}$	*trp* aporepressor occupied with one molecule of tryptophan
$TrpR^{**}$	*trp* aporepressor occupied with two molecules of tryptophan
μ	specific growth rate
ψ	fraction of occupied promoters taking only RNA polymerase into account
$\bar{\psi}$	fraction of occupied promoters taking RNA polymerase and Trp repressor into account

Appendix B. List of reactions and important metabolites

ALDO	aldolase
Cho	chorismate
ChoSynth	chorismate synthesis pathway
DAHP	3-deoxy-D-arabino-heptulosonate 7-phosphate
DAHPS	DAHP synthase
E4P	erythrose-4-phosphate
ENO	enolase
G1PAT	glucose-1-phosphate adenyltransferase
G3PDH	glycerol-3-phosphate dehydrogenase
G6P	glucose-6-phosphate
G6PDH	glucose-6-phosphate dehydrogenase
GAP	glyceraldehyde-3-phosphate

GAPDH	glyceraldehyde-3-phosphate dehydrogenase
MetSynth	methionine synthesis
MurSynth	mureine synthesis
NAD	diphosphopyridindinucleotide, oxidized
NADH	diphosphopyridindinucleotide, reduced
NADP	diphosphopyridindinucleotide-phosphate, oxidized
NADPH	diphosphopyridindinucleotide-phosphate, reduced
PDH	pyruvate dehydrogenase
PEP	phosphoenolpyruvate
PEPCxylase	PEP carboxylase
PFK	phosphofructokinase
PGDH	6-phosphogluconate dehydrogenase
PGI	glucose-6-phosphate isomerase
PGK	phosphoglycerate kinase
PGluMu	phosphoglycerate mutase
PGM	phosphoglucomutase
PK	pyruvate kinase
PRPP	phosphoribosylpyrophosphate
PTS	phosphotransferase system
Pyr	pyruvate
R5PI	ribosephosphate isomerase
RPPK	ribosephosphate pyrophosphokinase
Ru5P	ribulosephosphate epimerase
Ser	serine
SerSynth	serine synthesis pathway
Synth2	consumption of pyruvate
Synth3	chorismate
Synth4	PRPP
Synth5	and serine in biomass synthesis
TA	transaldolase
TIS	triosephosphate isomerase
TK	transketolase
TKa	transketolase, reaction a
TKb	transketolase, reaction b
Trp, trp	tryptophan
Trpremoval	drain of tryptophan from the system
TrpSynth	tryptophan synthesis pathway

References

Arvidson, D.N., Bruce, C., Gunsalus, R.P., 1986. Interaction of the Escherichia coli trp aporeressor with its ligand, L-tryptophan. Journal of Biological Chemistry 261, 238–243.

Bongaerts, J., Krämer, M., Müller, U., Raeven, L., Wubbolts, M., 2001. Metabolic engineering for microbial production of aromatic amino acids and derived compounds. Metabolic Engineering 3, 289–300.

Chassagnole, C., Noisommit-Rizzi, N., Schmid, J.W., Mauch, K., Reuss, M., 2002. Dynamic modeling of the central carbon metabolism of *Escherichia coli*. Biotechnology Bioengineering 79, 53–73.

Dell, K.A., Frost, J.W., 1993. Identification and removal of impediments to biocatalytic synthesis of aromatics from D-glucose: rate-limiting enzymes in the common pathway of aromatic amino acid biosynthesis. Journal of American Chemical Society 115, 11581–11589.

Ensley, B.D., Ratzkin, B.J., Osslund, T.D., Simon, M.J., Wackett, L.P., Gibson, D.T., 1983. Expression of naphthalene oxidation genes in *Escherichia coli* results in the biosynthesis of indigo. Science 222, 167–169.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220, 671–680.

Koh, B.T., Yap, M.G.S., 1993. A simple genetically structured model of *trp* repressor–operator interactions. Biotechnology and Bioengineering 41, 707–714.

Koh, B.T., Tan, R.B.H., Yap, M.G.S., 1997. Genetically structured mathematical modeling of *trp* attenuator mechanism. Biotechnology and Bioengineering 58, 502–509.

Kremling, A., Gilles, E.D., 2001. The organization of metabolic reaction networks: II. Signal processing in hierarchical structured functional units. Metabolic Engineering 3 (2), 138–150.

Kremling, A., Bettenbrock, K., Laube, B., Jahreis, K., Lengeler, J.W., Gilles, E.D., 2001. The organization of metabolic reaction networks: III. Application for diauxic growth on glucose and lactose. Metabolic Engineering 3 (4), 362–379.

Leuchtenberger, W., 1996. Amino acids—technical production and use. In: Rehm, H.-J., Reed, G. (Eds.), Biotechnology, vol. 6. VCH Verlagsgesellschaft mbH, Weinheim, Germany, pp. 465–502.

Liao, J.C., Hou, S.-Y., Chao, Y.-P., 1996. Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. Biotechnology and Bioengineering 52, 129–140.

Marin-Sanguino, A., Torres, N.V., 2000. Optimization of tryptophan production in bacteria. Design of a strategy for genetic manipulation of the tryptophan operon for tryptophan flux maximation. Biotechnology Progress 16, 133–145.

Mauch, K., Buziol, S., Schmid, J., Reuss, M., 2002. Computer aided design of metabolic networks. In: Rawlings, J.,Ogunnaike, T., Eaton, J. (Eds.), AIChE Symposium Series, vol. 98. pp. 82–91.

Neidhardt, F.C., Ingraham, J.L., Schaechter, M., 1990. Physiology of the Bacterial Cell: A Molecular Approach. Sinauer Associates, Sunderland, MA.

Nielsen, J., 2001. Metabolic engineering. Applied Microbiology and Biotechnology 55, 263–283.

Piperno, J.R., Oxender, D.L., 1968. Amino acid transport systems in *Escherichia coli* K12. Journal of Biological Chemistry 243 (22), 5914–5920.

Rasor, J.P., Voss, E., 2001. Enzyme-catalyzed processes in pharmaceutical industry. Applied Catalysis A: General 221, 145–158.

Rizzi, M., Baltes, M., Theobald, U., Reuss, M., 1997. In vivo analysis of metabolic dynamics in *Saccharomyces cerevisiae*: II. Mathematical model. Biotechnology and Bioengineering 55, 592–608.

Santillan, M., Mackey, M.C., 2001. Dynamic regulation of the tryptophan operon: a modeling study and comparison with experimental data. Proceedings of the National Academy of Science, 98, 1364–1369.

Schuster, S., Dandekar, T., Fell, D.A., 1999. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. TIBTECH 17, 53–60.

Sinha, S., 1988. Theoretical study of tryptophan operon: application in microbial technology. Biotechnology and Bioengineering 31, 117–124.

Stephanopoulos, G., Simpson, T.W., 1997. Flux amplification in complex metabolic networks. Chemical Engineering Science 52, 2607–2627.

Tribe, D.E., Pittard, J., 1979. Hyperproduction of tryptophan by *Escherichia coli*: genetic manipulation of the pathways leading to tryptophan formation. Applied Environmental Microbiology 38 (2), 181–190.

Varma, A., Palsson, B.O., 1993a. Metabolic capabilities of *Escherichia coli*: I. Synthesis of biosynthetic precursors and cofactors. Journal of Theoretical Biology 165, 477–502.

Varma, A., Palsson, B.O., 1993b. Metabolic capabilities of *Escherichia coli*: II. Optimal growth pattern. Journal of Theoretical Biology 165, 503–522.

Xiu, Z., Zeng, A., Deckwer, W.-D., 1997. Model analysis concerning the effects of growth rate and intracellular tryptophan level on the stability and dynamics of tryptophan biosynthesis in bacteria. Journal of Biotechnology 58, 125–140.

Yanofsky, C., Horn, V., Gollnick, P., 1991. Physiological studies of tryptophan transport and tryptophanase operon induction in *Escherichia coli*. Journal of Bacteriology 173, 6009–6017.