OXFORD

Systems biology

Exact quantification of cellular robustness in genome-scale metabolic networks

Matthias P. Gerstl[1,2], Steffen Klamt[3], Christian Jungreuthmayer[1,2] and Jürgen Zanghellini[1,2,*]

[1]Austrian Centre of Industrial Biotechnology, Vienna, Austria, [2]Department of Biotechnology, University of Natural Resources and Life Sciences, Vienna, Austria, and [3]Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

*To whom correspondence should be addressed.
Associate Editor: Igor Jurisica

Abstract

Motivation: Robustness, the ability of biological networks to uphold their functionality in spite of perturbations, is a key characteristic of all living systems. Although several theoretical approaches have been developed to formalize robustness, it still eludes an exact quantification. Here, we present a rigorous and quantitative approach for the structural robustness of metabolic networks by measuring their ability to tolerate random reaction (or gene) knockouts.

Results: In analogy to reliability theory, based on an explicit consideration of all possible knockout sets, we exactly quantify the probability of failure for a given network function (e.g. growth). This measure can be computed if the network's minimal cut sets (MSCs) are known. We show that even in genome-scale metabolic networks the probability of (network) failure can be reliably estimated from MSCs with lowest cardinalities. We demonstrate the applicability of our theory by analyzing the structural robustness of multiple *Enterobacteriaceae* and *Blattibacteriaceae* and show a dramatically low structural robustness for the latter. We find that structural robustness develops from the ability to proliferate in multiple growth environments consistent with experimentally found knowledge.

Conclusion: The probability of (network) failure provides thus a reliable and easily computable measure of structural robustness and redundancy in (genome-scale) metabolic networks.

Availability and implementation: Source code is available under the GNU General Public License at https://github.com/mpgerstl/networkRobustnessToolbox.

Contact: juergen.zanghellini@boku.ac.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A major characteristic of biological systems is their 'robustness', i.e. the ability to perform normally under the presence of perturbations (Kitano, 2002). Despite its fundamental importance, robustness is hard to quantify and a comprehensive, quantitative understanding of robustness has yet to be developed (Kitano, 2007). The problem is augmented by the fact that robustness is an extremely general concept which obstructs efforts for an exact

definition (Stelling *et al.*, 2004). Clearly, robustness arises as a consequence of the interactions between the components in a system. It therefore requires a network-based approach for its analysis (Larhlimi *et al.*, 2011).

In the following, we study a key aspect of cellular robustness, namely structural robustness in metabolic networks (Wilhelm *et al.*, 2004), which is a measure of the apparent redundancy in metabolic networks. Redundancy is one reason why specific

(network) functions may persist in spite of changes in the networks' topologies. Metabolic networks are particularly suitable to study structural robustness as their topologies are either well known or can be reliably reconstructed by well-established protocols (Thiele and Palsson, 2010; Vlassis *et al.*, 2014).

An unbiased way to characterize all topologically feasible flux distributions in a metabolic network can be achieved in terms of elementary flux modes (EFMs) (Schuster *et al.*, 2000). EFMs are steady-state pathways that use a minimal set of reactions (enzymes) in the appropriate thermodynamic directions. Biologically, they can be interpreted as simple, indivisible functional building blocks, which represent all feasible metabolic capabilities of the organism. Thus, EFMs are likely candidates to be used in a definition of structural robustness. In fact, Wilhelm *et al.* (2004) were the first to define structural robustness by counting the number of EFMs in different situations.

An alternative approach to EFMs is minimal cut sets (MCSs) (Klamt and Gilles, 2004). These are minimal sets of reaction (enzyme) knockouts that completely disable a particular network functionality. MCSs are intimately connected to EFMs. Any network function can be represented by a set of EFMs. Thus, MCSs can equivalently be defined as the minimal set of reaction knockouts that prevent a set of target EFMs from carrying a steady-state flux. An MCS is therefore a so-called minimal hitting set of the target EFMs (Klamt, 2006). If the EFMs or the MCSs are known, the respective other can be calculated. Mathematically, this relationship is known as duality (Berge, 1989). While EFMs describe the possibilities to achieve a particular function, MCSs describe the effort it takes to remove that function from the network. Thus, the size distribution of the MCSs provides a measure for the structural fragility of a network (Klamt and Gilles, 2004).

One major disadvantage for both measures, the EFM-based structural robustness measure as well as the MCS-based structural fragility measure, is that they presuppose the availability of the complete set of EFMs and MCSs, respectively. This is a computationally intractable problem in genome-scale metabolic models (GSMMs).

Here, we present a novel method that overcomes the computational limitations and allows one for the first time to estimate the structural robustness even in GSMMs. We use MCSs to calculate the exact (structural) probability of failure (PoF) of a network. This puts the estimation of structural robustness on a theoretically sound fundament.

2 Theory

We consider a metabolic network made up of m (internal) metabolites and r reactions. The network contains n EFMs. We use the support representation $E = \text{supp}(e)$ to uniquely characterize an EFM, e. E is the set of all (reaction) indices i for which the EFM vector e carries any flux, i.e. $\text{supp}(e) = \{i | e_i \neq 0\}$, with $1 \leq i \leq r$. We collect all EFMs in the set $\mathscr{E} = \{E_1, \dots, E_n\}$.

If we delete d reactions, several EFMs will get disabled. We represent a set of d deleted reactions by a cut set (CS) C^d of cardinality d. C^d contains the set of (reaction) indices that get deleted. A CS C^d disables all EFMs involving any of the d deleted reactions. To study the impact of a deletion strategy, C^d, on the network, we collect all EFMs not affected by C^d in the set of the remaining EFMs,

$\mathscr{E}^{C^d} = \{E \in \mathscr{E} | C^d \cap E = \emptyset\}$. The complement of $\mathscr{E}^{C^d}$ is denoted by $\mathscr{H}^{C^d}$ and contains all EFMs that are disabled by C^d: $\mathscr{H}^{C^d} = \mathscr{E} \setminus \mathscr{E}^{C^d}$. A minimal (irreducible) set of reaction (enzyme) deletions that completely disables a given set of target EFMs, $\mathscr{T}$, is called an MCS, M (Klamt and Gilles, 2004). For any target set $\mathscr{T}$ we collect all μ MCSs in the set of MCSs, $\mathscr{M} = \{M_1, \dots, M_\mu\}$. Finally, we collect all $1 \leq i \leq b = \binom{r}{d}$ possible CSs C_i^d of cardinality d in the set of CSs, $\mathscr{C}^d = \{C_1^d, \dots, C_b^d\}$.

2.1 Structural robustness: $R(d)$

For a given number d of deletions, Behre *et al.* (2008) defined the network's structural robustness $R(d)$ as the average ratio of the number of remaining EFMs to the total number of EFMs over all possible combinations of exactly d knockouts, i.e.

$$R(d) := \frac{1}{|\mathscr{C}^d|} \sum_{i=1}^{|\mathscr{C}^d|} \frac{|\mathscr{E}^{C_i^d}|}{|\mathscr{E}|}. \tag{1}$$

We show in the Supplementary material that the network's structural robustness only depends on the cardinality of the network's EFMs and can easily be calculated for any number of deletions d by

$$R(d) = \frac{1}{n} \binom{r}{d}^{-1} \sum_{i=1}^{n} \binom{r - |E_i|}{d}. \tag{2}$$

Thus, network structural robustness is merely determined by the EFMs' length.

2.2 PoF: $F(d)$

For a given number of deletions, d, we define the network's PoF, $F(d)$, as the ratio of the number of CSs that disable a particular network objective, obj, to the total number of all possible combinations of exactly d knockouts, i.e.

$$F_{\text{obj}}(d) := \frac{|\mathscr{C}_{\text{obj}}^d|}{|\mathscr{C}^d|}. \tag{3}$$

CSs that disable a particular objective (i.e. the elements of $\mathscr{C}_{\text{obj}}^d$) are either MCSs or supersets of MCSs disabling 'obj'. We show in the Supplementary material that the PoF can be calculated by

$$F_{\text{obj}}(d) = \binom{r}{d}^{-1} \sum_{\emptyset \neq J \subseteq \{1, \dots, \mu\}} (-1)^{|J|-1} \binom{r - |M_J|}{d - |M_J|}, \tag{4}$$

where $M_J = \bigcup_{j \in J} M_j$ and J is a (multi-)index over all power sets of the indices $\{1, \dots, \mu\}$ of the set $\mathscr{M}$ of MCSs blocking 'obj'.

Note that the definition of the PoF requires specification of an objective, that can be the operation of the whole system or of a particular function. A typical cellular function is the production of biomass (BM). Thus, $F_{\text{BM}}(d)$ measures the probability that the deletion of d reactions is lethal. In what follows we will drop the index for the objective and implicitly assume that the production of BM is targeted, unless stated otherwise.

2.3 Overall structural robustness $\overline{R}$ and overall PoF (OPoF) $\overline{F}$

Both $R(d)$ and $F(d)$ are defined for a specific number of reaction deletions. Following Behre *et al.* (2008), we define the overall structural robustness, $\overline{R}$, if at most (all) r reactions are knocked out, as the weighted sum of all possible $R(d)$ (Behre *et al.*, 2008),

$$\overline{R} := \sum_{d=1}^{r} w_d^R R(d). \qquad (5)$$

Analogous to the structural robustness we define the OPoF, $\overline{F}$ as

$$\overline{F} := \sum_{d=1}^{r} w_d^F F(d). \qquad (6)$$

Behre *et al.* (2008) suggested to use the following weights

$$w_d^R := \binom{r}{d} p^d \text{ with } p = \sqrt[r]{2} - 1. \qquad (7)$$

However, their choice is difficult to interpret in terms of a probability distribution. That is why we opted for an appropriately selected probability distribution function for the calculation of $\overline{F}$. w_d^F may be naturally modeled by a binomial distribution,

$$w_d^F := \binom{r}{d} p^d (1-p)^{r-d}, \qquad (8)$$

which estimates the number of loss-of-function mutations with exactly d reaction deletions assuming a constant (reaction) mutation rate p. In typical metabolic networks, the number of reactions r is large and the mutation rate is small. We therefore approximate the binomial distribution with the simpler Poisson distribution,

$$w_d^F := \lambda^d \exp(-\lambda)/(d!), \qquad (9)$$

with $\lambda = rp$. $R(d)$, $\overline{R}$, $F(d)$ and $\overline{F}$ all range between 0 and 1. However, while $\overline{R} = 1$ indicates a totally robust network, $\overline{F} = 1$ denotes the reverse, a completely fragile network.

3 Implementation and computation

The computation of the structural robustness $R(d)$ according to Equation (2) is straight forward as long as the complete set of EFMs

is known. The computation of the PoF is computationally more demanding as the number of summands in the sum of Equation (4) explodes combinatorially. In the following, we will outline countermeasures to estimate the PoF even in GSMMs.

Algorithm 1. Recursive CS count

Require: *mcs, deletions d, reactions r*
1: **for** $i = 1$ **to** LENGTH(*mcs*) **do**
2:     *csCount*[*i*] ← GET_CS_COUNT(*i*, *mcs*[*i*], 0, **false**)
3: **end for**
4:
5: **function** GET_CS_COUNT(*index, cs, csCount, inRecursion*)
6:     *foundSubset* ← **false**
7:     *combCs* ← *mcs*[*index*] **OR** *cs*
8:     *combCd* ← CARDINALITY(*combCs*)
9:     *combCsCount* ← CHOOSE(*r* − *combCd*, *d* − *combCd*)
10:    **for** $i = 1$ **to** (*index* − 1) **do**
11:        *testCs* ← *mcs*[*i*] **OR** *combCs*
12:        *testCd* ← CARDINALITY(*testCs*)
13:        **if** *testCd* <= *deletions* **then**
14:            **if** *testCd* > *combCd* **then**
15:                *combCsCount* ←
                        GET_CS_COUNT(*i*, *combCs*,
                        combCsCount, **true**)
16:            **else**
17:                *foundSubset* ← **true**
18:                **exit for** loop
19:            **end if**
20:        **end if**
21:    **end for**
22:    **if not** *foundSubset* **then**
23:        **if** *inRecursion* **then**
24:            *csCount* ← *csCount* − *combCsCount*
25:        **else**
26:            *csCount* ← *combCsCount*
27:        **end if**
28:    **end if**
29:    **return** *csCount*
30: **end function**

Table 1. Major topological properties of the MSMMs and GSMMs for the investigated *Enterobacteriaceae* and *Blattibacteriaceae*

Model ID	Organism	Medium	M	r	Rank(S)	n	n_{BM}
EColi_core + glc (Gerstl *et al.*, 2015)	*E. coli* K-12 MG1655	Glucose	70	90	65	169 916	121 753
EColi_core + gly (Gerstl *et al.*, 2015)	*E. coli* K-12 MG1655	Glycerol	71	91	66	60 495	48 944
EColi_core + ac (Gerstl *et al.*, 2015)	*E. coli* K-12 MG1655	Acetate	69	88	64	1299	736
iJO1366 + glc (Monk *et al.*, 2013)	*E. coli* K-12 MG1655	Glucose	1165	1726	1131	n/a	n/a
iJO1366 + mel (Monk *et al.*, 2013)	*E. coli* K-12 MG1655	Melibiose	1163	1718	1128	n/a	n/a
iECs_1301 + mel (Monk *et al.*, 2013)	*E. coli* O157:H7 Sakai	Melibiose	1098	1666	1057	n/a	n/a
iS_1188 + mel (Monk *et al.*, 2013)	*S. flexneri* 2a 2457T	Melibiose	1 026	1 517	982	n/a	n/a
iCG230 + full (González-Domenech *et al.*, 2012)	*B. cuenoti* Pam	Full	299	342	292	n/a	n/a
iCG238 + full (González-Domenech *et al.*, 2012)	*B. cuenoti* Bge	Full	306	350	299	n/a	n/a

Models were taken from Gerstl *et al.* (2015), Monk *et al.* (2013) and González-Domenech *et al.* (2012). Aerobic growth for *Enterobacteriaceae* was simulated on minimal medium with glucose or melibiose as sole carbon source and for *Blattibacteriaceae* with full medium (for media compositions see Supplementary Table S1). All reactions that could not carry a steady state flux under any circumstances for the given growth media were removed from the original models. *m* and *r* refer to the (remaining) number of internal metabolites and reactions, respectively, for fully consistent models (represented by the respective internal stoichiometric matrix, *S*). *n* and n_{BM} refer to the total number and the number of BM producing EFMs in the respective models. An EFM analysis was not applicable (n/a) for GSMMs.

Algorithm 1 implements Equation (4) by calculating the total number of CSs for given deletions d in a recursive way. The number of *reactions* in the model, as well as all *mcs* with a cardinality up to the current number of reaction deletions are needed for the calculation. The algorithm calculates the number of all possible CSs for a single MCS by the function CHOOSE, which computes the binomial coefficient for $\begin{pmatrix} r - combCd \\ d - combCd \end{pmatrix}$, where $combCd$ is the number of deletions resulting from the combination of MCSs. In order to avoid the consideration of a CS multiple times those CSs that have already been calculated by other MCS, indicated by their *index*, are subtracted. This is done by creating an intermediate CS by a logical **OR** operation and then following the inclusion–exclusion principal (Ryser, 1963) in a recursive procedure.

Algorithm 1 was implemented as multi-threaded C program which is available at https://github.com/mpgerstl/networkRobustnessToolbox.

4 Results

In the following analysis, we used $\lambda = 0.5$, unless stated otherwise.

4.1 $\overline{F}$ reliably identifies structural fragility

We used a medium-scale metabolic model (MSMM) that describes the core metabolism of *Escherichia coli* (see Table 1 for an overview of the topological properties of the models used) to study the influence of the carbon source on the structural robustness and the PoF

Table 2. Structural robustness and PoF in a core metabolic model of *E. coli* growing on minimal media (see Supplementary Table S1) and three different carbon sources (glc, glucose; gly, glycerol; ac, acetate)

model ID	$R(1)$	$R(2)$	$\overline{R}$	$F(1)$	$F(2)$	$\overline{F}$
EColi_core + glc	46.32	21.54	37.98	21.11	40.45	10.31
EColi_core + gly	46.19	21.69	37.95	31.87	54.63	14.84
EColi_core + ac	50.86	26.98	42.75	42.05	67.42	19.06

We used all EFMs for the calculation of $\overline{R}$ and all synthetic lethals of groups of up to $d_0 = 9$ reactions for the estimation of $\overline{F}$ with $\lambda = 0.5$. All values are listed in %
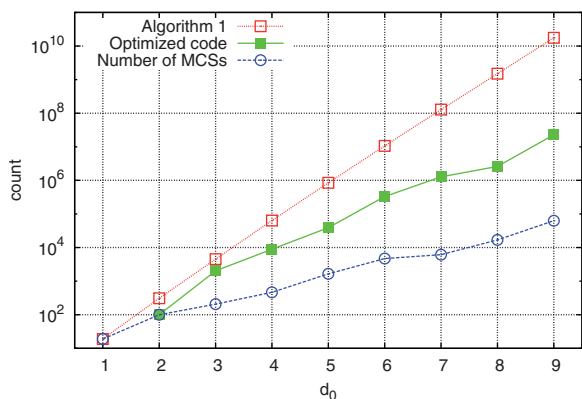
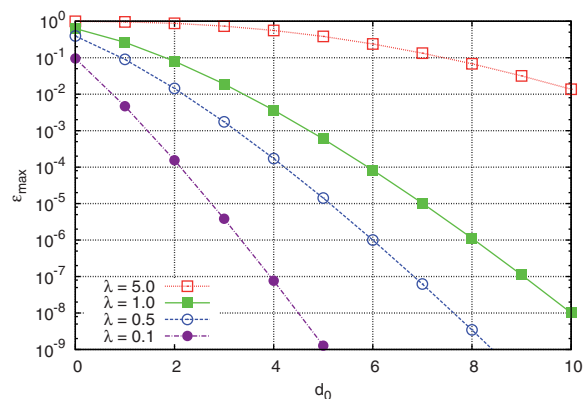(see Table 2). In all cases, a complete EFM and MCS analysis was feasible.

We found that both measures indicated a strong impact of the carbon sources although with opposing trends. The OPoF was lowest for growth on glucose and highest for growth on acetate. We observed that in all three growth environments the PoFs $F(d)$ could be well approximated (coefficient of determination, $R^2 > 0.999$) by $\alpha[1 - \beta\exp(-\gamma d)]$ with three free parameters α, β and γ (see Supplementary Fig. S1). $\overline{R}$ however ranked growth on acetate to be most robust and growth on glucose or glycerol similarly robust but far behind growth on acetate. According to Equation (2) this implied that for acetate EFMs were on average shorter then for glucose (see Supplementary Fig. S2). Finally, the simple ratio between the number of BM producing EFMs and the total number of EFMs revealed yet another ranking with the highest ratio for growth on glycerol and the lowest ratio for growth on acetate (see Supplementary Table S2). Comparing the different measures, we found that only the OPoF scored the (lack of) structural robustness in *E. coli* consistently to previous analysis (Klamt and Gilles, 2004; Stelling *et al.*, 2002).

4.2 $\overline{R}$ and $\overline{F}$ can be calculated exactly in MSMMs

Judged by the number of EFMs, the three growth models significantly differed in size (see Table 1). Yet all structural robustness-values were calculated within few seconds for any number of reaction deletions. As the number of summands in Equation (4) grew excessively for high number of reaction deletions, d, the calculation of $F(d)$ was more demanding. We found an exponential increase in the number of recursions for Algorithm 1 (see Fig. 1). Our optimized implementation scaled consistently better than that and allowed us to evaluate $F(d)$ in the largest model, EColi_core + glc, up to $d = 9$ in < 10 min using 10 threads on an Intel® Core™ i7-3930 K with two CPUs á six cores and 3.20 GHz operated with Ubuntu 12.04.

4.3 $\overline{R}$ and $\overline{F}$ are determined by the first few deletions

We analyzed the contribution of higher-order deletions to the overall structural robustness and OPoF, respectively. In the best case higher-order terms $R(d)$ and $F(d)$ vanish for $d > d_0$ causing no error in the sums of Equations (5) and (6). In the worst case higher-order

Fig. 1. Number of recursions in Algorithm 1 as function of the number of reaction deletions, d evaluated for the model EColi_core+glc in different computation scenarios. In addition, we plotted the number of MCS (dotted line) as function of their cardinality for the model used

Fig. 2. Maximum error, ε, as function of the expansion length, d_0, for various λ values. Note that according to Equation (10) the maximum error only depends on λ and is independent of the specific topology and size of the metabolic model

Table 3. PoFs in GSMMs of different *Enterobacteriaceae* growing on minimal media (see Supplementary Table S1) and glucose (glc) or melibiose (mel)

Model ID	F(1)	F(2)	F(3)	F(4)	$\overline{F}$	$\varepsilon(4)$
iJO1366 + glc	16.74	30.71	42.35	52.06	8.02	0.02
iJO1366 + mel	17.17	31.42	43.24	53.04	8.22	0.02
iECs_1301 + mel	17.59	32.11	44.10	53.98	8.41	0.02
iS_1188 + mel	19.97	35.99	48.83	59.12	9.51	0.00

The OPoF was estimated by considering all synthetic lethals of groups of up to $d_0 = 4$ reactions and $\lambda = 0.5$. All values are listed in %

terms maximally contribute to the sums. Thus, the maximum error ε is easily calculated by setting $R(d) = F(d) = 1$ for all terms $d > d_0$,

$$\varepsilon(d_0) = \sum_{d=d_0+1}^{r} w_d^{(i)} = 1 - \sum_{d=0}^{d_0} w_d^{(i)}, \qquad (10)$$

where the last sum represents the cumulative distribution function of $w_d^{(i)}$ with $i \in \{F, R\}$.

We found that the resulting maximum error dropped quickly with d_0 even in huge GSMMs (see Fig. 2). For instance, for $\lambda \leq 1$ the ε drops below 10^{-4} if all MCSs up to cardinality six are known. Thus, we conclude that for practical means the overall structural robustness $\overline{R}$ and the OPoF $\overline{F}$ can be approximated by the first few terms in their respective sums

$$\overline{R} \approx \sum_{d=1}^{d_0} w_d^R R(d), \text{ and } \overline{F} \approx \sum_{d=1}^{d_0} w_d^F F(d). \qquad (11)$$

For fixed λ the OPoF $\overline{F}$ depends only on the d_0 smallest MCS and consistently converges toward its true value with increasing MCS cardinality. In the *E. coli* example used above (see Table 2), all MCSs up to cardinality 4 were sufficient to estimate $\overline{F}$ within an error of 0.02 percentage points. However, a similar argument for the overall structural robustness $\overline{R}$ is not available. The knowledge of the d_0 shortest EFMs does not allow one to accurately estimate $\overline{R}$. Supplementary Figure S5 illustrates the error in the overall structural robustness as function of the d_0 shortest EFMs for the *E. coli* model used above. In that case, >55% of all EFMs were required to get within a 10% error margin.

4.4 For small λ, $\overline{F}$ is well approximated in GSMMs

According to Equation (11), the OPoF could be estimated if all low cardinality MCSs were known. Recently, von Kamp and Klamt (2014) showed that it is in fact possible to enumerate the smallest MCSs in GSMMs. In the following, we used their approach to calculate low cardinality MCSs.

We calculated all lethal MCSs up to a cardinality of 4 in GSMMs of three *E. coli* strains growing on minimal medium and glucose or melibiose as the sole carbon source and estimated the OPoF (see Table 3). The models represented *E. coli* K-12 MG1655 (iJO1366), the pathogenic enterohemorrhagic *E. coli* O157:H7 Sakai (iECs_1301) and the *Shigella flexneri* 2a 2457T (iS_1188).

Comparing the three organisms against each other revealed that growth on glucose is more failsafe than growth on the alternative carbon source melibiose. In fact, growth on glucose (and glucose 1-phosphate) was found to be more robust than on any other single, nitrogen-free carbon source (see Supplementary Fig. S4). The two *E. coli* were similarly robust (with a small advantage for *E. coli* K-12 MG1655), while the *S. flexneri* was most fragile. Although the

Table 4. PoFs in GSMMs of *B. cuenoti* Bge (iCG238 + full) and *B. cuenoti* Pam (iCG230 + full) growing on full media (see Supplementary Table S1)

Model ID	F(1)	F(2)	F(3)	F(4)	F(5)	F(6)	$\overline{F}$	$\varepsilon(6)$
iCG230 + full	64.62	87.68	95.81	98.61	99.55	99.86	27.63	10^{-4}
iCG238 + full	62.57	86.36	95.16	98.31	99.43	99.81	26.89	10^{-4}

The OPoF was estimated by considering all synthetic lethals of groups of up to $d_0 = 6$ reactions and $\lambda = 0.5$. All values are listed in %

Table 5. GC and RC PoF for various MSMMs and GSMMs

Model ID	F(1)		F(2)		$\overline{F}$	
	GC	RC	GC	RC	GC	RC
EColi_core + glc	5.00	21.11	10.01	40.45	2.50	10.31
EColi_core + gly	5.71	31.87	11.46	54.63	2.86	14.84
EColi_core + ac	15.00	42.05	29.10	67.42	7.38	19.06
iJO1366 + glc	16.94	16.74	31.03	30.71	8.11	8.02
iJO1366 + mel	17.55	17.17	32.05	31.42	8.39	8.22
iECs_1301 + mel	19.14	17.59	34.65	32.11	9.12	8.41
iS_1188 + mel	21.82	19.97	38.92	35.99	10.33	9.51
iCG230 + full	66.34	64.62	88.96	87.68	28.24	27.63
iCG238+full	62.26	62.57	86.33	86.36	26.79	26.89

For the GSMMs the OPoFs were estimated by considering all synthetic lethals of groups of up to $d_0 = 4$ reactions and $\lambda = 0.5$. For these settings the estimation error was < 0.02 percentage points. All values are listed in %

difference in the OPoF was small, it is indicative as the difference in $\overline{F}$ for any two models is larger than $\varepsilon(4)$. We observed the same trend not only for $\overline{F}$ but also for $F(d)$ for all tested cardinalities $d \leq 4$, as well. The GSMMs were also found to be more robust than the *E. coli* core model. Again, in all models, $F(d)$ could be nicely approximated by $\alpha[1 - \beta\exp(-\gamma d)]$ (see Supplementary Table S4).

Finally, we calculated all synthetic lethal reactions in groups of up to $d_0 = 6$ reaction deletions and evaluated the OPoF in the GSMMs of the endosymbiotic *Blattabacterium cuenoti* Bge (iCG238 + full) and *B. cuenoti* Pam (iCG230 + full), see Table 4, for full growth media containing all nutrients that were possibly taken up, see Supplementary Table S1 for detailed media compositions. Both strains were found to be extremely fragile with OPoFs > 26% and a maximal inaccuracy of 1×10^{-4} percentage points. Any three reaction deletions almost certainly killed these two strains (with a small chance of survival of <5%).

4.5 Gene–centric PoFs showed same behavior as reaction–centric PoFs

For simplicity so far we analyzed structural robustness from a reaction–centric (RC) viewpoint. However, it is possible to change from an RC to a gene–centric (GC) viewpoint while retaining our formalism for the PoF calculation. PoF only requires the calculation of MCSs. These can be calculated in an RC model as well as in a GC model provided that a gene–reaction mapping is available as Boolean function of the genes (as it is typically the case). GC MCSs consider the effect of gene knock-outs on reactions based on the evaluation of the provided Boolean rules. If the complete set of EFMs is known, GC MCSs can be calculated using an established integer programming procedure (Jungreuthmayer and Zanghellini, 2012). In GSMMs, where the complete set of EFMs is unavailable, the dual systems approach (von Kamp and Klamt, 2014) can be

adopted to compute GC MCSs. The common dual approach calculates RC MCSs (ordered by ascending cardinality) by iteratively solving a mixed integer linear program (MILP) (De Figueiredo *et al.,* 2009). In order to obtain a GC MCS instead of an RC MCS the gene–reaction association has to be integrated into the MILP, which can be achieved by means of the so-called indicator constraints. These concepts are provided by modern MILP solvers, which allow for the Boolean coupling of reactions and their associated genes.

We evaluated GC PoFs for models listed in Table 1 and compared them against the corresponding RC PoFs (see Table 5). We found comparable (O)PoF values except for the MSMM of *E. coli*, where we observed larger differences. Yet the absolute ranking of the organisms with respect to structural robustness remained unchanged no matter if the GC or RC PoF was used.

5 Discussion

Here, we developed a new general measure to quantify the structural robustness (or fragility) of metabolic networks based on the (O)PoF. Both, PoF $F(d)$ as well as OPoF $\overline{F}$ are based on MCSs and share analogies with the previously defined structural robustness measures $R(d)$ and $\overline{R}$ which rely on EFMs (Behre *et al.,* 2008). In contrast to $\overline{R}$, the OPoF remains computationally feasible even in GSMMs for two reasons: (i) the OPoF can be well estimated from the shortest MCSs with up to d_0 reaction cuts; and (ii) these d_0 shortest MCSs can be calculated with a MILP even in GSMMs (von Kamp and Klamt, 2014). Previously, it was shown that MCSs in the (primal) network are EFMs in a network dual to the original network (Ballerstein *et al.,* 2012). Thus, finding the shortest EFMs in this dual network is equivalent to finding the low cardinality MCSs in the primal GSMM (von Kamp and Klamt, 2014). A high performance computing infrastructure is not required for this task. Thus, the OPoF is in principle an easily computable measure. However, due to the combinatorial explosion of the number of summands in $F(d)$, see Equation (4), the evaluation is practically limited to low cardinalities. This does not cause practical problems for the OPoF $\overline{F}$ as the impact of high-cardinality MCSs on $\overline{F}$ quickly decreases (see Fig. 2). The slope of this decrease depends on the selected probability density that models the number of deletions that occur per mutation. Here, we used a Poisson distribution which is typically used for the study of rare events, like mutations (Lee *et al.,* 2012, and references therein). We were able to show that up to $\lambda \leq 1$ our analysis is computationally feasible on standard computer infrastructure. This is sufficient for the study of naturally occurring mutations. For instance, in *E. coli* a mutation rate of about 10^{-3} per genome and generation was found (Lee *et al.,* 2012), which is far below our limit of $\lambda \leq 1$, as we can assume that not all mutations lead to a loss of function of metabolic enzymes. All our conclusions remained valid (data not shown) if we used other probability densities like a binomial distribution or the one used by Behre *et al.* (2008), see Equation (7).

Previously, MCSs were already used to compute a so-called fragility coefficient which characterized the sensitivity of a metabolic network to deletions (Klamt, 2006; Klamt and Gilles, 2004). Fragility was defined for each reaction as the inverse of the average cardinality of all MCSs that were supported by that reaction. Upon averaging the reaction-specific fragility over all reactions one obtained a measure for the overall network fragility. In contrast to the here introduced concept of the OPoF, network fragility lacks a rigorous probabilistic definition and more importantly requires the complete set of MCSs for its calculation. Network fragility is therefore not applicable to GSMMs.

A clear difference between the PoF and structural robustness is that the former is function oriented while the latter characterizes the network. $\overline{F}$ is calculated with respect to some particular function that fails. Typically, networks perform multiple functions at the same time, which may be differentially robust. Thus, the OPoF is not only a topological property but also a property of the associated function. For instance, it is interesting to ask (and scope of further analysis) whether two essential functions like cellular energy production and BM production are similarly robust. $\overline{R}$ on the other hand characterizes the network topology by essentially measuring the proportion of short EFMs in a network. This definition is independent of any functions performed by the network. $\overline{R}$ counterintuitively identified growth on acetate to be more robust than growth on glucose which indicates that the length of EFMs may not be an appropriate proxy for structural robustness. Even if we selected only those EFMs that support a particular target function (like BM production), and calculated $\overline{R}$ for the reduced set of EFMs, we had misidentified growth on acetate as the most robust growth environment as growth supporting EFMs on acetate were on average 'shorter' than on glucose (see Supplementary Fig. S2). For better understanding we illustrated this effect in a toy network in Supplementary Figure S3. Thus, we propose to use the complement of the OPoF: $\overline{R}_{new} = 1 - \overline{F}$ as a new measure for structural robustness in metabolic networks.

PoFs can be calculated equally well on the basis of reaction deletions as well as on the basis of gene deletions. Some reactions in an organism are not associated with a gene (e.g. non-catalyzed reactions, reactions with unknown annotations) and can therefore not be knocked out. Other reactions are catalyzed by several different enzymes and are therefore more difficult to knock-out. However, our analysis indicates that the question if an organism or a growth condition is more robust than another can be identically answered in both approaches.

It was criticized that structural robustness (as defined by Behre *et al.,* 2008) lacks the ability to identify critical enzymes, such as knot enzymes (Min *et al.,* 2011). This is true for the OPoF as well. For instance, according to our measure, network A in Figure 3 was correctly identified to be more robust than network B. The cause for the lack of structural robustness is not apparent from the numerical value of the OPoF. Min *et al.* (2011) proposed to solve this dilemma by considering the number of independent (i.e. non-overlapping) EFMs, which are not available in GSMMs. However, by measuring the specific PoF $f_R(d)$ for each individual reaction (enzyme) R we are at least able to identify critical enzymes. For this purpose, we define the (reaction) specific PoF $f_R(d)$

$$f_R(d) := \frac{|\boldsymbol{c}_R^d|}{|\mathscr{C}^d|} \tag{12}$$

to be the ratio between all specific CSs, $\boldsymbol{c}_R^d$, that are supported by reaction R and the number of all possible CSs, $\mathscr{C}^d$ (and similarly for $\overline{f}_R$). By ranking reactions according to their impact on the network we are able to identify critical reactions which may be useful drug targets.

We calculated OPoFs in MSMMs and GSMMs of multiple *Enterobacteriaceae* growing on minimal media to illustrate our concept. Among these, the ordinary lab strain *E. coli* K-12 MG1655 growing on glucose was found to be most robust. This was consistently observed in MSMMs and GSMMs in agreement with expectations (Stelling *et al.,* 2002). Moreover, catabolite repression mechanisms support this observation of glucose as a preferential growth source.

Escherichia coli K-12 MG1655 was also found to be more robust than the pathogenic *E. coli* O157:H7 Sakai, and the least

A

B

C

model ID	$\bar{f}_{R1}$	$\bar{f}_{R2}$	$\bar{f}_{R3}$	$\bar{f}_{R4}$
A	3.63	3.63	3.63	3.63
B	12.48	3.63	2.05	2.05
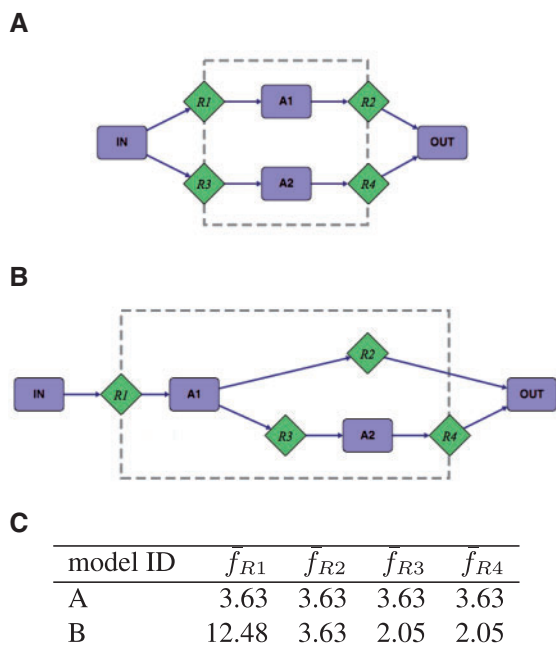
Fig. 3. Two differently robust networks (panel A, $\bar{F} = 6.48\%$; panel B, $\bar{F} = 15.32\%$) and their overall specific PoF, $\bar{f}$ with $\lambda = 0.5$ (panel C)

robust *S. flexneri* 2a 2457T. While the difference between K-12 MG1655 and O157:H7 Sakai was small (but larger than the possible error), *S. flexneri* was clearly set apart. Due to its adaption to a specific growth niche, *S. flexneri* has lost many catabolic pathways for various nutrient sources (Bliven and Maurelli, 2012). It was demonstrated computationally as well as experimentally that the number of growth-supporting conditions is larger for *E. coli* K-12 MG1655 than for *E. coli* O157:H7 Sakai than for *S. flexneri* 2a 2457T (Monk *et al.*, 2013). In contrast we here investigated the OPoF as a measure of structural robustness for these organisms in a single growth condition. Yet we found the same ranking as obtained from counting the number of growth-supporting conditions. Structural robustness is thought to come about from the need to survive in multiple environments. Many of these adaptations occur through rewiring of existing enzymes (Wagner, 2013), thus increasing the connectivity in the network, and—as a byproduct—also introducing many alternative routes. These topological alternatives can be exploited for other growth sources as well and were mirrored in our analysis in low OPoFs.

The two strains of the highly specialized endosymbiont *B. cuenoti*, which for millions of years have lived in the relatively constant environment provided by specialized cells of cockroaches (Lo *et al.*, 2007), were found to be extremely sensitive to loss-of-function mutations (González-Domenech *et al.*, 2012). This was confirmed by an extremely high OPoF of >26%. *B. cuenoti* was more than three times less robust than *E. coli*. It is known that *B. cuenoti* grows only on an extremely limited palette of metabolic substrates (Sabree *et al.*, 2009), which results in their low structural robustness as documented here. Consistent with our findings above on *S. flexneri* this lack of structural robustness is thought to be shaped by massive gene loss (Sabree *et al.*, 2010). A characteristic difference between *B. cuenoti* Pam and Bge is a defunct TCA cycle in the former due to the absence of the first three enzymatic steps (González-Domenech *et al.*, 2012). This fact was mirrored in a lower structural robustness of *B. cuenoti* Pam compared with Bge. Thus, we conclude that the OPoF is indeed an appropriate measure to correctly assess the structural robustness of cells.

Here, we analyzed the structural robustness of multiple organisms in several growth environments. Consistent with previous findings our results confirm that structural robustness is correlated with the ability to utilize multiple carbon sources (Papp *et al.*, 2004). They concluded that structural robustness is not a trait selected by evolution but a byproduct of environmental flexibility. However, very recently Yang *et al.* (2015) showed that structural redundancy also impacts a cell factory's ability to produce a product of interest in the presence of perturbations. Based on predictions from GSMMs, they argue that pathway diversification leads to robust production capabilities under large perturbations. For instance, we show in Supplementary Table S5 two strain designs with different OPoFs but identical production/growth characteristics. Thus, calculating the OPoF will identify promising designs for cell factories.

Interestingly, we found that the raise in $F(d)$ with increasing d was well approximated by an exponential function in all investigated models and independent of the investigated organism. Whether this is an expression of some general topological properties of metabolic networks or a coincidence will be scope of further work.

6 Conclusion

We developed a consistent theory of cellular redundancy based on a rigorous probabilistic definition of failure in metabolic networks. The new measure, called OPoF, allows quantification of the structural robustness of biochemical reaction networks with respect to certain functionalities. OPoF can be reliably estimated—even in GSMMs—from low cardinality MSCs. In contrast to other measures, the OPoF quantifies structural robustness and fragility of cellular metabolism concurrent with current biological paradigms and experimental findings. More specifically we showed that the number of growth-supporting nutrients indirectly correlates with the organisms' OPoF. As the OPoF is easily computable, we expect that it will be useful in increasing our understanding of key properties of metabolic networks.

References

Ballerstein,K. *et al.* (2012) Minimal cut sets in a metabolic network are elementary modes in a dual network. *Bioinformatics*, **28**, 381–387.

Behre,J. *et al.* (2008) Structural robustness of metabolic networks with respect to multiple knockouts. *J. Theor. Biol.*, **252**, 433–441.

Berge,C. (1989) *Hypergraphs, Volume 45: Combinatorics of Finite Sets*, 1st edn. North Holland, Amsterdam.

Bliven,K.A. and Maurelli,A.T. (2012) Antivirulence genes: insights into pathogen evolution through gene loss. *Infect. Immun.*, **80**, 4061–4070.

De Figueiredo,L.F. *et al.* (2009) Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, **25**, 3158–3165.

Gerstl,M.P. *et al.* (2015) Metabolomics integrated elementary flux mode analysis in large metabolic networks. *Sci. Rep.*, **5**, 8930.

González-Domenech,C.M. *et al.* (2012) Metabolic stasis in an ancient symbiosis: genome-scale metabolic networks from two *Blattabacterium cuenoti* strains, primary endosymbionts of cockroaches. *BMC Microbiology*, **12**, S5.

Jungreuthmayer,C. and Zanghellini,J. (2012) Designing optimal cell factories: integer programming couples elementary mode analysis with regulation. *BMC Syst. Biol.*, **6**, 103.

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

Kitano,H. (2007) Towards a theory of biological robustness. *Mol. Syst. Biol.*, **3**, 137.

Klamt,S. (2006) Generalized concept of minimal cut sets in biochemical networks. *Biosystems*, **83**, 233–247.

Klamt,S. and Gilles,E.D. (2004) Minimal cut sets in biochemical reaction networks. *Bioinformatics*, **20**, 226–234.

Larhlimi,A. *et al.* (2011) Robustness of metabolic networks: a review of existing definitions. *Biosystems*, **106**, 1–8.

Lee,H. *et al.* (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc. Natl Acad. Sci. USA*, **109**, E2774–E2783.

Lo,N. *et al.* (2007) Cockroaches that lack *Blattabacterium* endosymbionts: the phylogenetically divergent genus *Nocticola*. *Biol. Lett.*, **3**, 327–330.

Min,Y. *et al.* (2011) Pathway knockout and redundancy in metabolic networks. *J. Theor. Biol.*, **270**, 63–69.

Monk,J.M. *et al.* (2013) Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl Acad. Sci. USA*, **110**, 20338–20343.

Papp,B. *et al.* (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, **429**, 661–664.

Ryser,H. (1963) *Combinatorial Mathematics*. Mathematical Association of America, Washington.

Sabree,Z.L. *et al.* (2009) Nitrogen recycling and nutritional provisioning by *Blattabacterium*, the cockroach endosymbiont. *Proc. Natl Acad. Sci. USA*, **106**, 19521–19526.

Sabree,Z.L. *et al.* (2010) Chromosome stability and gene loss in cockroach endosymbionts. *Appl. Environ. Microbiol.*, **76**, 4076–4079.

Schuster,S. *et al.* (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotech.*, **18**, 326–332.

Stelling,J. *et al.* (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190–193.

Stelling,J. *et al.* (2004) Robustness of cellular functions. *Cell*, **118**, 675–685.

Thiele,I. and Palsson,B.Ø. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.

Trinh,C.T. *et al.* (2008) Minimal *Escherichia coli* cell for the most efficient production of ethanol from hexoses and pentoses. *Appl. Environ. Microbiol.*, **74**, 3634–3643.

Vlassis,N. *et al.* (2014) Fast reconstruction of compact context-specific metabolic network models. *PLoS Comput. Biol.*, **10**, e1003424.

von Kamp,A. and Klamt,S. (2014) Enumeration of smallest intervention strategies in genome-scale metabolic networks. *PLoS Comput. Biol.*, **10**, e1003378.

Wagner,A. (2013) Chapter 13—genotype networks and evolutionary innovations in biological systems. In: Walhout, A. J. M. *et al.* (eds.) *Handbook of Systems Biology*. Academic Press, San Diego, pp. 251–264.

Wilhelm,T. *et al.* (2004) Analysis of structural robustness of metabolic networks. *IEE Proc. Syst. Biol.*, **1**, 114–120.

Yang,L. *et al.* (2015) Characterizing metabolic pathway diversification in the context of perturbation size. *Metab. Eng.*, **28**, 114–122.