Integration von Omics-Daten in der marinen Ökosystemforschung

Hanno Teeling, Frank Oliver Glöckner, Max-Planck-Institut für Marine Mikrobiologie und Jacobs Universität Bremen

Die Lebenswissenschaften erleben derzeit einen Paradigmenwechsel: Rasante Weiterentwicklungen der DNA-Sequenziertechnologien erlauben es, immer größere Sequenzmengen zu immer geringeren Kosten zu erzeugen. Dieser Trend wird sich auf absehbare Zeit fortsetzen und erzwingt einen Wechsel des Fokus von der Sequenz-Akquise hin zur bioinformatischen Sequenz-Prozessierung. In den vergangenen Jahren hat sich die marine Genomforschung zu einem der größten Lieferanten von Sequenzdaten entwickelt, und zwar maßgeblich durch Aktivitäten im Bereich der mikrobiellen Metagenomik. Damit steht die marine Genomforschung stellvertretend für die Herausforderungen einer kontextbezogenen Interpretation großer Sequenzvolumina. Diese umfasst die Integration genomischer Daten mit akzessorischen Daten, wie zum Beispiel begleitenden Expressionsdaten aus Metatranskriptom- oder Metaproteomstudien aber auch mit *in situ* erhobenen Daten zur Biodiversität sowie gemessenen und interpolierten physikochemischen Umgebungsparametern. Ein Problem für die Umweltgenomik war bislang, dass die Funktionsaufklärung neu gefundener Gene deutlich hinter deren Sequenzierung zurückblieb. Der jüngst vorgestellte Reaktom-Array stellt hier einen vielversprechenden Lösungsansatz dar. Zusammengefasst ist auf diese Weise eine neue Art von Umweltforschung möglich, die das Potential hat, entscheidend zum Verständnis global relevanter Stoffumsetzungen sowie der Folgeabschätzung voranschreitender Veränderungsprozesse in den Ozeanen beizutragen.

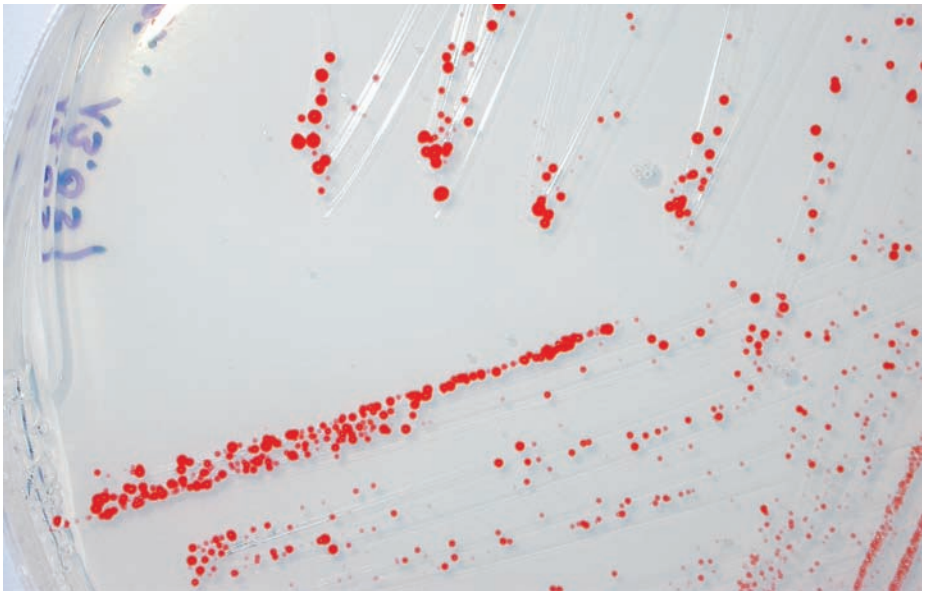

Abb. 1: Agarkultur von Rhodopirellula baltica SH1^T

Die Molekularbiologie wird heute zunehmend von Hochdurchsatztechnologien dominiert. Die damit verbundenen großen Datenmengen erfordern veränderte Arbeitsweisen. Speziell in der Genomik lassen sich seit der Einführung hochparalleler Next Generation Sequencing (NGS)-Technologien stetig steigende DNA-Sequenzmengen erzeugen. Das Tempo dieser Entwicklung ist derart hoch, dass mitunter von einer „genomischen Revolution" gesprochen wird (Abb. 2 A). Bemerkenswerterweise hat sich dabei die marine Genomik zu einem der größten Datenproduzenten entwickelt. Die Gründe dafür sind mannigfaltig. So spielen die Meere, die 71% der Erdoberfläche bedecken, eine entscheidende Rolle für die globalen Kreisläufe der Elemente. Die daran beteiligten Stoffumsetzungen werden überwiegend von Mikroorganismen durchgeführt, weshalb ein Studium ihres genetischen Potentials einen Schlüssel zum Verständnis globaler Stoffumsetzungsprozesse darstellt. Schätzungen zufolge gehen 99% aller Nährstoff- und Gasumsetzungen sowie 50% der Weltprimärproduktion auf marine Mikroorganismen zurück, die somit zum Beispiel die Basis aller nutzbaren Fischbestände bilden. Darüber hinaus stellen marine Habitate eine Quelle für biotechnologisch anwendbare oder prozesstechnisch optimierbare Enzyme dar, insbesondere für Synthesen bei hohen Ionenstärken. Prominente Beispiele für breit angelegte Studien in der marinen Genomik sind das von John Craig Venter durchgeführte Global Ocean Sampling mit Fokus auf mikrobielle Metagenome, die Marine Microbiology Initiative der Gordon & Betty Moore Foundation mit Fokus auf komplette Genome sowie das gerade angelaufene TARA Oceans Project mit Fokus auf Protisten. Auch in Deutschland hat das Bundesministerium für Bildung und Forschung (BMBF) bereits 2001 mit den REGX-Projekten (Real Environmental Genomics) und aktuell mit dem MIMAS-Projekt (Mikrobielle Interaktionen in Marinen Systemen, *www.mimas-projekt.de*) in die marine Umweltgenomik investiert.

Perspektivwechsel: Vom Organismus zur Systembiologie

Die Sequenzierung individueller Genome liefert wertvolle Erkenntnisse über das genomische Potential und die Überlebensstragien mariner Modellorganismen (Abb. 1). Allerdings ist die Gesamtgenomsequenzierung bislang auf kultivierbare Mikroorganismen beschränkt. Zwar gibt es Bemühungen, Genome einzelner Bakterienzellen zu sequenzieren, aber Erfolge bei solchen Einzelzell-Genomics-Ansätzen sind derzeit noch die Ausnahme[1]. Leider lassen sich mit etablierten Techniken – je nach Habitat – lediglich 1% bis 10% der vorhandenen Mikroorganismen in Kultur bringen. Um das genetische Repertoire der Mehrzahl der Umweltmikroorganismen dennoch studieren zu können, werden kultivierungsunabhängige Verfahren wie die Metagenomik genutzt. Dazu wird DNA aus Umweltproben isoliert, ohne vorangehende Isolierung einzelner Spezies fragmentiert und anschließend partiell sequenziert. Bis vor kurzem musste dabei die Umwelt-DNA vor der Sequenzierung mit Hilfe von Klonierungsvektoren vervielfältigt werden. Die Leistungsfähigkeit moderner Next-Generation Sequencing (NGS)-Geräte ermöglicht es jedoch, die arbeits-, zeit- und kostenaufwändige Klonierung zu umgehen und Umwelt-DNA direkt zu sequenzieren. Von den NGS-Verfahren am Markt ist das Pyrosequenzierungsverfahren von 454 Life Sciences (Roche) am besten für die Metagenomik geeignet, da es derzeit die längsten

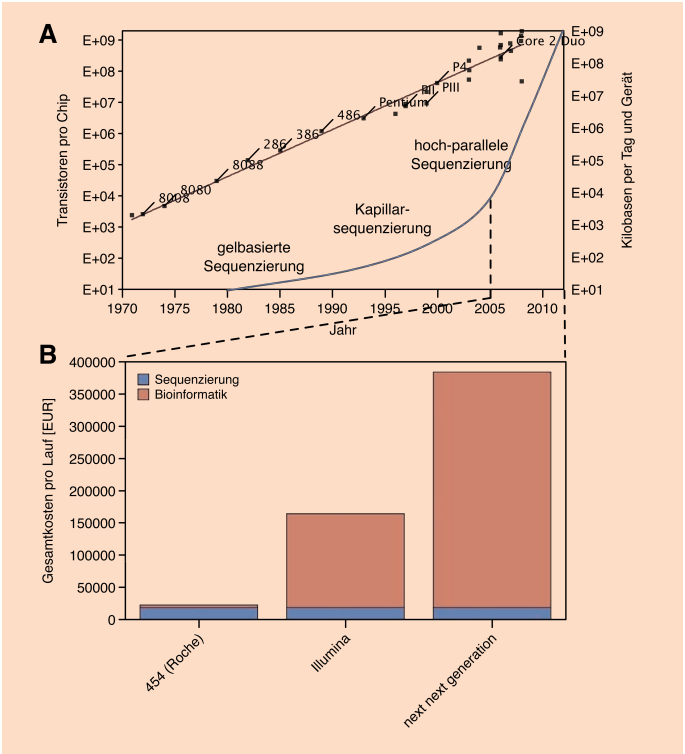

Abb. 2: Die genomische Revolution in den Lebenswissenschaften: A. Der Vergleich der Zunahme der Sequenzier- und Prozessorleistung zeigt, dass die Sequenzierleistung derzeit schneller als nach Moore´s law wächst (modifiziert nach Stratton *et al.*, Nature 458, 719–724). B. Die Folge ist, dass der Anteil der Bioinformatik an den Gesamtkosten bei vollständiger Aufarbeitung der Daten ständig steigt (modifiziert nach Folker Meyer, Argonne National Laboratory).

Leselängen liefert. Ein einziger Pyrosequenzierungslauf auf einem aktuellen 454 FLX Ti-Gerät kann mehr als 1 Millionen Sequenzen einer durchschnittlichen Leselänge von fast 400 Basenpaaren erzeugen. Die damit erzielte Sequenziertiefe ist groß genug, um für niedrig bis moderat diverse Habitate einen Teil der Sequenzen zu längeren Contigs zu assemblieren. Aus 300 bis 400 Megabasen Rohsequenzen werden so typischerweise 30 bis 60 Megabasen assemblierter Sequenz. Zwar ist der überwiegende Teil der Sequenzen relativ kurz, jedoch nicht viel kürzer als ein typisches bakterielles Gen.

Ein Milliliter Seewasser enthält in etwa eine Millionen Bakterienzellen und damit theoretisch mehrere Milliarden Gene. Da das Gros der Bakterienpopulation in der Regel jedoch von nur wenigen Dutzend Spezies dominiert wird, liegt die Anzahl der für die Hauptstoffumsetzungen verantwortlichen häufigen Gene lediglich in der Größenordnung von einigen 100.000. Dies ist eine Größenordnung, die sich bereits jetzt mit modernen NGS-Verfahren untersuchen lässt. Mit zukünftigen Sequenziertechniken der dritten Generation (Pacific Biosciences, Ion Torrent, u.v.a.m.) wird sich sogar der überwiegende Teil der relevanten Gene adressieren lassen.

Bioinformatik: Chancen und Grenzen

Um eine funktionierende sequenzbasierte, systembiologische Ökosystemforschung zu verwirklichen, ist eine leistungsfähige Bioinformatik unerlässlich. Tatsächlich führen die stetig fallenden Kosten bei der Sequenzierung zu stetig steigenden Kosten bei der bioinformatischen Prozessierung. Bei aktuellen Sequenzierverfahren liegen die Kosten für die Bioinformatik inzwischen deutlich höher als die Kosten der eigentlichen Sequenzierung. Es ist absehbar, dass sich dieser Trend

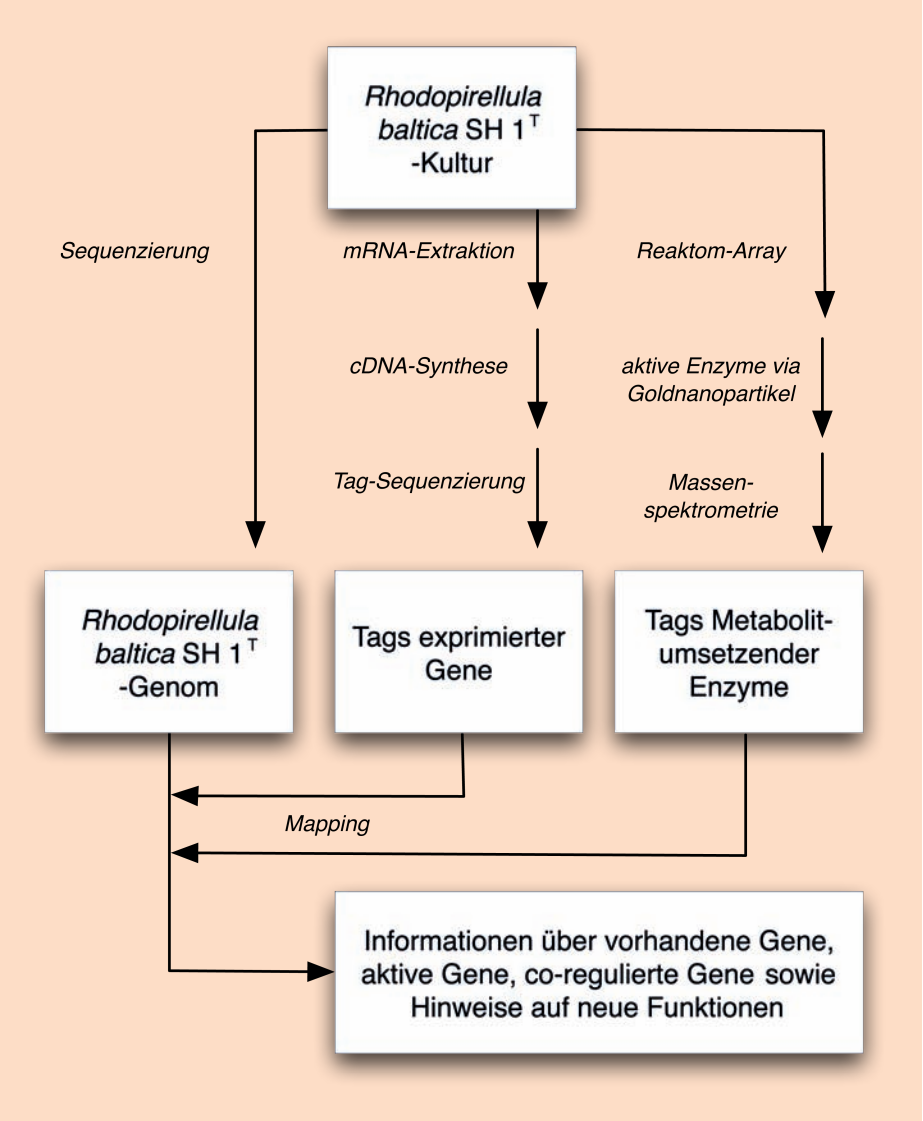

Abb. 3: Prinzip eines integrierten Ansatzes zur Aufklärung von Funktionen hypothetischer und konserviert-hypothetischer Gene unter Einbeziehung des Reaktom-Arrays

in Zukunft weiter verstärken wird (Abb. 2 B). Diese Kosten bestehen aus den Anschaffungs-, Betriebs- und Wartungskosten von Rechenclustern, die hinsichtlich ihrer Leistung mit den ständig steigenden Sequenzvolumina Schritt halten müssen. Für den Betrieb der Systeme sind jedoch vor allem gut ausgebildete und entsprechend bezahlte Bioinformatiker und Programmierer entscheidend. Tatsächlich herrscht derzeit ein Mangel an geeigneten bioinformatischen Werkzeugen zur Auswertung, Integration und Visualisierung von NGS-Metagenomdaten.

Zwei Bereiche sind dabei von besonderer Wichtigkeit. Zum einen braucht es Tools, die Metagenomsequenzen taxonomisch klassifizieren können, so dass sich die gefundenen Gene mit dedizierten Organismengruppen assoziieren lassen. Auf diese Weise werden Informationen zur Funktion und Taxonomie miteinander verknüpft und somit Subpopulationen sowie das Zusammenspiel verschiedener Organismengruppen untersuchbar. Zum anderen besteht ein Bedarf an Werkzeugen zur Datenintegration, denn nur durch eine Vernetzung genomischer Daten mit akzessorischen Daten wie Expressionsdaten (Metatranskriptomik, Metaproteomik), physikochemischen Parametern, geographischen Daten und Diversitätsdaten wird eine echte systembiologisch orientierte Ökosystemforschung möglich.

Trotz ihrer immensen Möglichkeiten unterliegt die Bioinformatik Grenzen, wenn es darum geht, für neue Gene aus der Umwelt Funktionen vorherzusagen. Eine bioinformatische Funktionsvorhersage beruht auf der Annahme, dass zwei Gene hinreichender Ähnlichkeit einen gemeinsamen evolutionären Ursprung aufweisen und damit dieselbe Funktion haben (Orthologie). Allerdings weisen orthologe Gene selbst auf Proteinebene selten mehr als 80% Sequenzidentität auf. Sogar bei Genen mit weniger als 50% identischen Aminosäuren wird oftmals noch eine Orthologiebeziehung angenommen. Erst unterhalb von 30% beginnt die sogenannte Twilight Zone der funktionellen Annotation. Spezifischer sind Methoden, die auf der Auswertung gemeinsamer Domänen von Proteinfamilien beruhen, wie etwa Pfam (protein families)-Profile. Allerdings erzeugen diese Methoden in der Regel deutlich weniger Treffer. Der Güte dieser, auf Sequenzhomologien fußenden Annotationen sind daher Grenzen gesetzt, und häufig lässt sich über eine allgemeine Funktionsklasse (Hydrolase, Oxidoreduktase, etc.) hinaus keine spezifischere Funktion vorhersagen.

Ein weiteres Problem der bioinformatischen Funktionsanalyse stellt die große Zahl potentieller Gene dar, denen durch Sequenzvergleich keine Funktion zugewiesen werden kann, weil es bislang keine Orthologen bekannter Funktion gibt. Eine aktuelle Analyse von insgesamt 305 vollständigen und partiellen marinen Genomen mit insgesamt 1,2 Millionen Genen zeigt, dass sich nur für rund 60% eine Funktion vorhersagen lässt. Dies sind vor allem Gene des Translations-, Transkriptoms- und Replikationsapparats sowie wohlbekannte Gene des Grund- und Energiestoffwechsels. Daher ist anzunehmen, dass Gene mit bislang unbekannter Funktion oftmals speziesspezifische Adaptionen und Fähigkeiten widerspiegeln. Mit Blick auf die Habitatvielfalt in den Ozeanen stellen diese Gene einen gewaltigen Pool an neuartigen und potentiell biotechnologisch nutzbaren Funktionen dar. Das Problem ist jedoch, ihre Funktionen aufzuklären.

Lösungsansätze

Um dieses Dilemma zu überwinden, sind verschiedene Ansätze denkbar. Kultivierbare Modellorganismen bieten etwa die Möglichkeit, die Auswirkungen definierter Stressoren oder Substrate auf das Expressionsmuster zu untersuchen. Co-regulierte Gene geben dabei Hinweise auf Beteiligungen an den untersuchten Stoffwechselvorgängen und erlauben somit, Funktionen bislang unbekannter Gene einzugrenzen.

Dieser organismenzentrische Ansatz lässt sich systembiologisch auf ganze Habitate erweitern, indem durch Integration von Umweltparametern mit Sequenz- und Expressionsinformationen charakteristische Genmuster identifiziert werden, die mit hoher Wahrscheinlichkeit an den für den Standort charakteristischen Prozessen beteiligt sind. Ein erster Ansatz, eine entsprechend großangelegte Datensammlung und Integration für das marine Ökosystem zu realisieren, wurde mit dem Megx.net (Marine Ecological Genomics) Projekt geschaffen[2].

Eine weitere Möglichkeit stellt der unlängst veröffentlichte Reaktom-Array dar[3]. Dabei handelt es sich um einen Microarray, auf dessen Oberfläche eine Vielzahl unterschiedlicher Metabolite zusammen mit einem gequenchten und daher inaktiven Fluoreszenzfarbstoff gebunden sind. Wird ein Proteinextrakt appliziert, wird der Farbstoff dort aktiv, wo Metabolite umgesetzt werden und liefert ein detektierbares Fluoreszenzsignal. Mit dieser Technologie ist es möglich, eine Momentaufnahme der aktiven enzymatischen Funktionen eines Organismus' oder einer Umweltprobe zu erhalten und Veränderungen zu verfolgen. In einem weiteren Schritt können diejenigen Metabolite, die ein Signal erzeugt haben, an Goldnanopartikel gebunden und als Enzymfänger eingesetzt werden. Mittels Massenspektrometrie können anschließend Massenfingerprints dieser Enzyme erzeugt werden, mit deren Hilfe sich die zugehörigen Gene in den genomischen oder metagenomischen Sequenzdaten identifizieren lassen. Vorbehaltlich der noch ausstehenden positiven Evaluierung des Reaktom-Arrays wäre es damit möglich, bioinformatische Vorhersagen zu validieren und darüber hinaus zumindest einigen der hypothetischen Gene eine Funktion zuzuordnen. So konnten jüngst rund 300 Genen unbekannter Funktion aus dem marinen Modellbakterium *Rhodopirellula baltica* SH 1^T mittels des Reaktom-Arrays und Metabolit-Nanopartikel eine mögliche Funktion zugeordnet werden (Abb. 3). Sollte sich die Zuverlässigkeit des Ansatzes bestätigen, dann wäre dies ein entscheidender Schritt um ein besseres Verständnis der Funktion und Anpassungsfähigkeit von Umweltorganismen zu erlangen.

Perspektiven

Aktuelle technologische Fortschritte wie der Reaktom-Array bestärken unsere Einschätzung, dass wir bald in der Lage sein werden, grundlegende Funktionen der Meere sowie die Folgen anthropogener und klimatischer Einflüsse sehr viel besser zu verstehen als bisher. Rund 40% der Erdbevölkerung leben in weniger als 50 Kilometer Entfernung einer Küste. Damit stellen unsere Meere für Milliarden von Menschen Arbeitsplätze und Nahrung bereit. In Anbetracht der andauernden Belastung und Verschmutzung der Meere ist ein besseres Verständnis ihrer Funktionsweisen für unser aller Wohlergehen nicht nur wünschenswert sondern überlebenswichtig[4]. Auch in wirtschaftlicher Hinsicht hält das Meer eine Fülle an neuen Produkten für uns bereit. Mit interdisziplinären und integrativen Ansätzen aus der Bioinformatik, Ökosystemforschung und Biotechnologie erscheint es möglich, die noch in der Flut der Daten verborgenen Schätze des Wissens zu heben.

Literatur

[1] Woyke, T., G. Xie, A. Copeland, J. M. Gonzalez, C. Han, H. Kiss, J. H. Saw, P. Senin, C. Yang, S. Chatterji, et al. 2009. Assembling the marine metagenome, one cell at a time. PLoS One 4:Article No.: e5299.

[2] Kottmann, R., I. Kostadinov, M. B. Duhaime, P. L. Buttigieg, P. Yilmaz, W. Hankeln, J. Waldmann, and F. O. Glöckner. 2010. Megx.net: integrated database resource for marine ecological genomics. Nucleic Acids Res 38:D391-D395.

[3] Beloqui, A., M. E. Guazzaroni, F. Pazos, J. M. Vieites, M. Godoy, O. V. Golyshina, T. N. Chernikova, A. Waliczek, R. Silva-Rocha, Y. Al-ramahi, et al. 2009. Reactome array: forging a link between metabolome and genome. Science 326:252-257.

[4] Glöckner, F. O., and I. Joint. 2010. Marine microbial genomics in Europe: current status and perspectives Microbial Biotechnology:doi:10.1111/j.1751-7915.2010.00169.x.

Korrespondenzadresse

Prof. Dr. Frank Oliver Glöckner
Max-Planck-Institut für
Marine Mikrobiologie
Jacobs Universität Bremen
Celsiusstraße 1, 28359 Bremen
www.microbial-genomics.de
fog@mpi-bremen.de