26. HÜLSENBERGER GESPRÄCHE 2016

Die postgenomische Ära: Die Renaissance des Phänotyps

H. WILHELM SCHAUMANN STIFTUNG

Wolfram Weckwerth

Systemtheoretische Konzepte der genomweiten molekularen Analyse und Datenintegration in der Biologie

Systembiologie ist eine moderne Entwicklung in der Biologie, die genomweite molekulare Analysen, z.B. Metabolomics, Proteomics und/oder Transkriptomics, mit Computer-gestützten mathematischen und statistischen Modellen verknüpft, um einerseits kausale Mechanismen vom Molekül zum Organismus abzuleiten und andererseits Vorhersagemodelle für die Merkmalsausprägung bzw. die Genotyp-Phänotyp-Beziehung zu erhalten. Diesen Ansatz hat bereits Ludwig von Bertalanffy 1944 in seinem Buch „Vom Molekül zur Organismenwelt“ 1944 angedeutet (1). In einer seiner wichtigsten Publikationen „Der Organismus als physikalisches System betrachtet“ beschreibt Bertalanffy bereits 1940 die mathematische Modellierung eines sich selbst regulierenden Systems von biochemischen Pathways eines „offenen“ Organismus (2). In den folgenden Jahren hat Bertalanffy seine Theorie verallgemeinert und als auf alle komplexen nicht-linearen Systeme z.B. der Biologie, Ökologie oder auch Ökonomie anwendbare „Allgemeine Systemtheorie“ definiert (3).

Die technischen Beschränkungen Bertalanffy´s waren zu seiner Zeit im Hinblick auf unsere Möglichkeiten heutzutage schier unüberwindbar:

Das System, mit dem er sich 1940 auseinandersetzte, bestand aus 4 (5) Komponenten mit 4 Differentialgleichungen. Dieses System lässt sich analytisch beschreiben und man kann daraus grundlegende Prinzipien der Selbstregulation oder Selbstorganisation ableiten (2).

Wo stehen wir heutzutage mit einem „typischen“ Organismus, Pflanze oder Tier (Mensch)? Nach der Erkenntnis der molekularen Struktur und Prinzipien der Informationsspeicherung der DNA 1953 (4), dem zentralen Dogma der Molekularbiologie von Crick 1973 (5) und der rapiden Entwicklung von „next generation sequencing (NGS)“ (6) seit der Publikation des ersten Humangenoms und des ersten Pflanzengenoms 2000 sind zur Zeit ca. 80000 Genomprojekte und deren Datenbanken verfügbar. Eine genomische Rekonstruktion eines typischen tierischen oder pflanzlichen Stoffwechsels umfasst ca. 2500 Reaktionen und wesentlich mehr kaum abzuschätzende metabolischen Komponenten (7, 8). In anderen Worten, wir müssen mindestens 2500 vernetzte Reaktionen als Differentialgleichungen darstellen und modellieren, um eine kausale Verknüpfung des Systems zu beschreiben, bzw. Vorhersagen des dynamischen molekularen Phänotyps aus der Genomsequenz abzuleiten. Diese Art von mathematischer Beschreibung eines komplexen nicht-linearen Systems ist erst mit Hilfe der hochmodernen Computertechnologie lange nach Bertalanffy möglich geworden, und man ist heutzutage in der Lage, solche Systeme numerisch zu lösen bzw. zu approximieren (7).

Desweiteren haben sich Technologien für die genomweite molekulare Analyse entwickelt, von denen Bertalanffy keine Vorstellungen haben konnte: RNAseq, Proteomics und Metabolomics (9). Diese bioanalytischen Verfahren orientieren sich an dem molekularen Dogma von Crick und sind in der Lage

hochkomplexe Gemische aus Transkripten (RNA-seq), Proteinen (Proteomics) und Metaboliten (Metabolomics) aufzulösen, viele Komponenten zu identifizieren und zu quantifizieren. Schliesslich werden in diesen Daten Interaktionen der molekularen Komponenten gesucht, die molekulare oder auch andere phänotypische Merkmale erklären können. Hierzu werden hochkomplexe multivariate statistische Verfahren eingesetzt, die letztendlich eine Datenintegration und – interpretation ermöglichen (10). In einem letzten Schritt müssen diese statistischen Modelle mit den mathematischen Modellen verknüpft werden, um aussagekraeftige Genotyp-Phänotyp-Modelle zu generieren (9, 11).

Im folgenden werde ich Methoden der molekularen Hochdurchsatzanalyse vorstellen sowie einige Aspekte der mathematischen und statistischen Modellierung von molekularen Hochdurchsatzdaten und deren Verknüpfung mit genomweiten biochemischen Netzwerken erläutern.

In Abbildung 1 ist eine komplette PANOMICS Plattform dargestellt. Die moderne Analyse von Organismen, Mikroorganismen, Pflanzen, Tiere, Mensch, startet heutzutage mit der Sequenzierung des Genoms mithilfe von „next generation sequencing (NGS)“ Technologien. Auf der Basis der vorhandenen Genomsequenz koennen dann weitere genomweite molekulare Analysen, Transcriptomics (RNAseq), Proteomics und Metabolomics, durchgeführt werden. In einem nächsten Schritt werden genomweite metabolische Netzwerke aus der vorhandenen Genomsequenz abgeleitet anhand des Vergleiches von bekannten orthologen Genen aus Datenbanken, wie z.B. Uniprot. Eine genaue Beschreibung des Vorganges dieser metabolischen Rekonstruktion habe ich in der Publikation „Unpredictability of Metabolism from Genome Sequences“ beschrieben (11).

Diese Rekonstruktion bildet allerdings nicht die phänotypische Plastizität ab, die jeden Organismus in seiner Wechselwirkung mit der Umwelt kennzeichnet. Somit kann aus der statischen Genotyp-Information nicht der plastische Phänotyp vorhergesagt werden, insbesondere nicht seine Wechselwirkungen mit der Umwelt (11). Auch genomweite Assoziierungstudien (Genomewide association studies GWAS) erlauben nur die korrelative Verknuepfung von genetischen Polymorphismen oder Mutationen mit Phänotypen. Polymorphismen sind in den meisten Fällen molekular-neutrale Mutationen im Genom, die keinen Effekt auf den Phänotyp haben. Es gibt allerdings „linkage disequilibrium“ Phänomene, d.h. Häufungen von „single nucleotide polymorphism (SNP)“, die darauf schliessen lassen, dass in diesen genomischen Regionen Mutationen zu Anpassungen an Umweltfaktoren

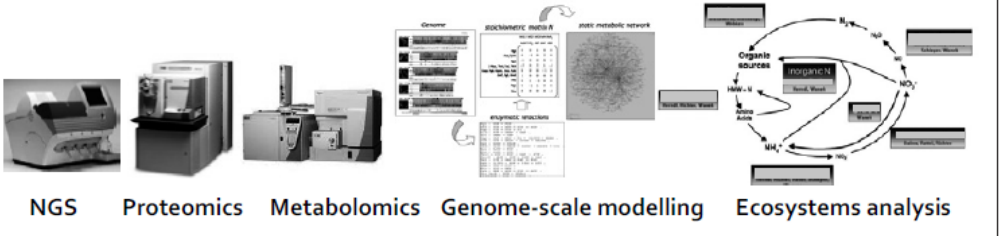

Abbildung 1: ANOMICS Plattform, welche NGS, Proteomics, Metabolomics mit metabolischer Modellierung des Modellorganismus bis hin zur Modellierung des zugehoerigen Ökosystems vereinigt (9).

führen. Allerdings müssen diese „kausalen“ Annahmen geprüft werden mithilfe klassischer biochemischer und bioanalytischer Methoden, wie Metabolomics oder Proteomics. Eine sehr elegante Studie, die demonstriert, wie in GWAS Daten ein kausaler Mechanismus für die Akkumulation von Ölsäure in Mais-Samen entdeckt worden ist, zeigt diesen Prozess auf (12).

Der entscheidende Punkt, um nun genomweite molekulare Daten und Genomsequenz-Informationen zu verknüpfen, ist die Modellbildung (11). Diese Modelle können entweder strukturbasiert oder kinetische Modelle sein (11). Es koennen auch Mischformen aus strukturkinetischen Modellen aufgebaut werden (13), die Aussagen über die Stabilität von metabolischen Netzwerken, also ihrer Plausibilität, zulassen.

In einem völlig neuen Ansatz haben wir Kovarianz-Modelle, also quasi Assoziierungs- oder Korrelationsnetzwerke von molekularen Komponenten in einem Organismus, die direkt aus den molekularen Daten abgeleitet werden koennen, mit einer Rekonstruktion der biochemischen Regulation verknüpft (14). Dieser Ansatz erlaubt die direkte Verknüpfung von genomweiten dynamischen molekularen Daten und der statischen metabolischen Rekonstruktion aus der Genomsequenz; er erweckt sozusagen das statische metabolische Netzwerk zum Leben. 2012 haben wir zum ersten Mal demonstriert, dass man in einem inversen Modellierungsansatz tatsächlich die biochemische Regulation berechnen und entscheidende biochemische Perturbationen damit vorhersagen kann (15). Dieser neue Datenintegrationsansatz erlaubt zum ersten Mal die direkte kausale Verknüpfung von Daten und genomischer Rekonstruktion in einem beliebig komplexen biochemischen Netzwerk und stellt damit eine fundamentale Genotyp-Phänotyp-Gleichung dar (9). In vielen folgenden Arbeiten haben wir gezeigt, dass die aus den molekularen Daten *berechneten* – nicht „spekulierten“ – Vorhersagen von Schlüsselpunkten biochemischer Perturbationen korrekt sind (7, 16, 17) und auch andere Forschungsgruppen haben diese Genotyp-Phänotyp-Gleichung inzwischen eingesetzt, um biochemische Perturbationen zu identifizieren (18, 19).

Die Berechnung von biochemischen Perturbationen aus molekularen, insbesondere Metabolomics, Daten ist implementiert in der Metabolomics Toolbox COVAIN (COvariance Inverse) (15) und kann von jedem Labor, welches metabolomische Daten zur Verfügung hat implementiert werden. COVAIN bietet ausserdem Algorithmen fuer die Datenintegration, Granger Causality Netzwerkanalyse, Strukturaufklärung und Pathway Vorhersage von unbekannten Metaboliten, multivariate Statistik und vieles mehr (15, 16, 20). Wir haben COVAIN in einem nächsten Schritt verwendet, um quantitative phänotypische Merkmalsbeschreibungen wie z.B. die mittels micro-Computertomographie (micro-CT) gemessene Entwicklungsmorphometrie eines Organismus mit diesen metabolischen Daten zu verknüpfen (21). Dieser Ansatz erlaubt es in Zukunft, kausale metabolische Modelle mit morphometrischen Daten zu integrieren, um phänotypische Merkmale in kausalen molekularen Zusammenhängen zu interpretieren.

Literaturverzeichnis

1. Bertalanffy Lv (1944) Vom Molekül zur Organismenwelt. *Akademische Verlagsgesellschaft Athenaion, Potsdam.*

2. Bertalanffy Lv (1940) Der Organismus als physikalisches System betrachtet. *Naturwissenschaften* 33:522-531.

3. Bertalanffy Lv (1969) General System Theory. *George Braziller, New York.*

4. Watson JD & Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171(4356):737-738.

5. Crick F (1970) Central dogma of molecular biology. *Nature* 227(5258): 561-563.

6. Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11(1):31-46.

7. Nagele T, *et al.* (2014) Solving the differential biochemical Jacobian from metabolomics covariance data. *PloS one* 9(4):e92299.

8. Nagele T & Weckwerth W (2012) Mathematical modeling of plant metabolism-from reconstruction to prediction. *Metabolites* 2(3):553-566.

9. Weckwerth W (2011) Green systems biology - From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *Journal of proteomics* 75(1):284-305.

10. Weckwerth W & Morgenthal K (2005) Metabolomics: from pattern recognition to biological interpretation. *Drug discovery today* 10(22):1551-1558.

11. Weckwerth W (2011) Unpredictability of metabolism--the key role of metabolomics science in combination with next-generation genome sequencing. *Analytical and bioanalytical chemistry* 400(7):1967-1978.

12. Beló A, *et al.* (2008) Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. *Molecular Genetics and Genomics* 279(1):1-10.

13. Furtauer L & Nagele T (2016) Approximating the stabilization of cellular metabolism by compartmentalization. *Theory Biosci* 135(1-2):73-87.

14. Weckwerth W (2003) Metabolomics in systems biology. *Annu Rev Plant Biol* 54:669-689.

15. Sun X & Weckwerth W (2012) COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics* 8:81-93.

16. Doerfler H, *et al.* (2013) Granger causality in integrated GC-MS and LC-MS metabolomics data reveals the interface of primary and secondary metabolism. *Metabolomics* 9(3):564-574.

17. Wang L, *et al.* (2016) System level analysis of cacao seed ripening reveals a sequential interplay of primary and secondary metabolism leading to polyphenol accumulation and preparation of stress resistance. *Plant J.*

18. Kugler P & Yang W (2014) Identification of alterations in the Jacobian of biochemical reaction networks from steady state covariance data at two conditions. *J Math Biol* 68(7):1757-1783.

19. Oksuz M, Sadikoglu H, & Cakir T (2013) Sparsity as cellular objective to infer directed metabolic networks from steady-state metabolome data: a theoretical analysis. *PloS one* 8(12):e84505.

20. Doerfler H, *et al.* (2014) mzGroupAnalyzer--predicting pathways and novel chemical structures from untargeted high-throughput metabolomics data. *PloS one* 9(5):e96188.

21. Bellaire A, *et al.* (2014) Metabolism and development - integration of micro computed tomography data and metabolite profiling reveals metabolic reprogramming from floral initiation to silique development. *New Phytol* 202(1):322-335.