DISS. ETH NO. 17812

Integrating Thermodynamics-based Modeling and Quantitative Experimental Data for Studying Microbial Metabolism

A dissertation submitted to

ETH ZURICH

for the degree of

Doctor of Sciences

presented by

Anne Maria Rita Kümmel

Dipl.-Ing. RWTH Aachen, Germany

born July, 18th, 1977

German

accepted on the recommendation of

Prof. Dr. Sven Panke

Prof. Dr. Uwe Sauer

Dr. Matthias Heinemann

2008

Summary

Systems biology strives to understand how the behavior of living systems is emerging from the underlying diverse and complex molecular interactions. It builds (i) on large-scale data sets that harbor preferably the quantitative amounts of all interacting molecules, (ii) on computational tools for the analysis of such data sets, and (iii) on mathematical models to finally obtain system-level understanding. Nowadays, advances in high-throughput experimental omics techniques drives systems biology forward through providing a system-level perspective on biological networks. However, due to the large size of the generated data sets, they can hardly be understood or interpreted directly. In order to benefit from the information that is contained in these data sets, computational data analysis tools are essential. Here, models are valuable tools to enhance the computational data analysis next to their essential role in describing the system behavior to gain system-level understanding. Consequently, with the advances in experimental technologies the importance of computational tools to interpret these data and models as semantic tools to administer the gained knowledge increases.

Centered on microbial metabolism, the different chapters of this thesis cover on the one hand models and model-based data analyses and on the other hand the application of advanced experimental technologies. On the modeling side, computational tools (i) for assisting the reconstruction of genome-scale stoichiometric models of the metabolic network, (ii) for analyzing quantitative metabolomics data, and (iii) for inferring metabolic activities from measured metabolite concentrations were developed. On the experimental side, state-of-the art quantitative omics data sets were generated and exploited (iv) to compare the metabolic state of two *S. cerevisiae* strains in detail and (v) to monitor

the metabolic reprogramming of *S. cerevisiae* undergoing a substrate switch from glucose to ethanol.

Chapter 2 focuses on the reconstruction of large-scale models of metabolic networks. The genome-scale stoichiometric network models have particularly proven to be a valuable tool in systems biology. They are composed of the metabolic reactions' stoichiometry and assignments of the reactions' reversibility or irreversibility and thus define an organism's metabolic capacity to convert substrates into biomass, energy, and by-products. In the reconstruction process, a first version of the model typically comprises a list of metabolic enzymes that are identified from the annotated genome. Whereas significant efforts have been put into the development of computational tools for identifying and compiling such organism-specific lists of metabolic reactions, the definition of the thermodynamic constraints in terms of reaction directionalities is a tedious manual process. For an automated and systematic assignment of thermodynamically reasonable reaction direction restrictions, an algorithm based on thermodynamics, network topology, and heuristic rules was developed. It assigns reaction directions in metabolic models such that the reaction network is thermodynamically feasible with respect to the production of energy equivalents. The applicability of the algorithm was demonstrated on a genome-scale metabolic model of *Escherichia coli*. Although not being fully comprehensive, the presented algorithm could define a significant number of irreversible reactions automatically with low computational effort. It can be a valuable part of a computational framework that assists the automated reconstruction of genome-scale metabolic models.

In Chapter 3, next to such stoichiometric metabolic network models thermodynamic constraints were employed to analyze large-scale quantitative metabolomics data: Here, a framework for mechanistic and model-based analysis of these data - the network-embedded thermodynamic analysis (NET analysis) - is introduced. By exploiting the second law of thermodynamics and the metabolites' Gibbs energies of formation, NET analysis allows for inferring functional principles and identifies reactions that are most likely subject to active allosteric or genetic regulation as exemplified with quantitative

metabolome data from *E. coli* and *Saccharomyces cerevisiae*. Moreover, the optimization framework of NET analysis demonstrated to be a valuable tool to systematically investigate data sets for consistency, for the extension of sub-omic metabolome data sets, and for resolving intracompartmental concentrations from cell-averaged metabolome data. Without requiring any kind of kinetic modeling, NET analysis represents a scalable and unbiased approach to uncover insights from quantitative metabolome data.

Elucidating intracellular metabolic fluxes is crucial for gaining systems understanding of cellular processes. The standard technique to assess these fluxes experimentally, namely ^{13}C-based flux analysis, however, has certain experimental limitations. Having a broader experimental applicability, in Chapter 4, a novel metabolomics-based approach for inferring metabolic activities based on mass balances and thermodynamic constraints was developed. Here, by combining NET analysis with elementary flux mode (EFM) analysis, the feasible fluxes defined by both, mass balances and thermodynamic constraints, are such comprehensively analyzed: All EFMs obtained from a stoichiometric metabolic model are tested via NET analysis for their thermodynamic feasibility with the measured metabolite concentrations. To demonstrate the approach, reaction activities within the pyruvate/ethanol metabolism of either glucose- or ethanol-grown *S. cerevisiae* from measured metabolite concentrations were inferred.

In Chapter 5, the glucose repression state of *S. cerevisiae* CEN.PK 113-D7 and FY4 was compared by quantitative, large-scale proteomics and metabolomics data sets as well as metabolic fluxes. The wild-types display similar degree of glucose repression on the flux level since the biomass and ethanol production yields as well as the respiratory TCA cycle activities are comparable. Nevertheless, differences in protein and metabolite concentrations were observed which indicated that Hxk2-dependent signaling and regulation is stronger in CEN.PK. Differences in signaling and regulatory strengths became also evident as the glucose repression state on flux level - indicated by the respiratory TCA cycle activity - for a genetic perturbation differs significantly: While the glucose repression was maintained in FY4, CEN.PK switched to a respiratory metabolism upon a Δ*hxk2* deletion. It was demonstrated that a point mutation in the *CYR1* gene is

causing this distinct behavior.

Beyond these projects (as described in the appendix), omics techniques were applied to monitor metabolic changes in *S. cerevisiae* during a diauxic shift. All phases of this metabolic reprogramming were characterized by quantitative metabolomics and proteomics data next to the extracellular physiology. Some of the acquired data were already exploited for the development of the NET/EFM analysis (cf. Chapter 4). The complete dynamic data set represents a basis for future analysis to elucidate the temporal and causal series of regulatory events acting in response to the altered substrate availability.

Today, while the technology to generate quantitative omics data sets develops quickly, the development of generally applicable model-based analysis concepts for such data lags behind. Therefore, the here developed and also applied model-based approaches for analysis of large-scale data contribute to the field of computational systems biology. As a case study, *S. cerevisiae*'s metabolism was experimentally analyzed by quantitative and large-scale data sets.

Zusammenfassung

Ziel der Systembiologie ist es zu verstehen, wie aus komplexen Wechselwirkungen zwischen Molekülen das Verhalten lebender Systeme entsteht. Die Grundlage für dieses Verständnis sind (i) grosse Datensätze, die möglichst die vorherrschenden Mengen aller wechselwirkenden Moleküle quantifizieren, (ii) rechnergestütze Analyse dieser Datensätze und (iii) mathematische Modelle, die letztendlich Systemverständnis ermöglichen. Die heutzutage enormen Fortschritte im Bereich der Hochdurchsatz-Analytik, die mit sogennanten Omik-Daten eine systemweite Perspektive auf biologische Netzwerke bietet, treiben die Systembiologie voran. Solch grosse Datensätze sind jedoch unübersichtlich und schwer zu interpretieren. Daher ist rechnergestützte Datenanalyse notwendig, um die Information, die diese Datensätze beinhalten, aufzudecken. Neben ihrem unabdingbaren Einsatz zur Beschreibung des Systemverhaltens, welches ein Systemverständnis ermöglicht, können mathematische Modelle auch die Datenanalyse unterstützen und verbessern. Mit den Fortschritten in der experimentellen Analytik gewinnen daher sowohl rechnergestützte Werkzeuge zur Interpretation der erzeugten Datensätze als auch Modelle zur semantischen Beschreibung des erlangten Wissens an Bedeutung.

Der mikrobielle Stoffwechsel steht im Mittelpunkt dieser Arbeit, die sich auf der einen Seite mit Modellen und modellbasierter Datenanalyse und auf der anderen Seite mit dem Einsatz moderner experimenteller Analytik befasst. Im Bereich der Modellierung wurden rechnergestützte Methoden entwickelt, um (i) die Modellbildung genomweiter stöchiometrischer Modelle des metabolischen Netzwerks, um (ii) quantitative Metabolom-Datensätze zu analysieren, und um (iii) von gemessenen Metabolitkonzentrationen auf metabolische Aktivitäten zu schliessen. Im experimentellen Teil wurden quantitative

Omik-Datensätze erzeugt und dazu genutzt, den Metabolismus zweier *S. cerevisiae*-Stämme detailliert zu vergleichen, und die Umstellung des Metabolismus vom Substrat Glukose zu Ethanol zu verfolgen.

Kapitel 2 behandelt die Modellierung grosser metabolischer Netzwerke. Es hat sich gezeigt, dass insbesondere die genomweiten stöchiometrischen Netzwerkwerkmodelle wertvolle Werkzeuge in der Systembiologie sind. Sie bestehen aus der Stöchiometrie der metabolischen Reaktionen und der Festlegung der Reversibilität oder Irreversibilität dieser Reaktionen. Somit definieren sie die metabolische Kapazität eines Organismus, Substrate in Biomasse, Energie und Nebenprodukte umzuwandeln. Bei der Modellbildung wird üblicherweise eine erste Version des Modells aus der Liste der Enzyme, die im Genom annotiert wurden, erstellt. Während rechnergestützte Werkzeuge zur Identifizierung und Zusammenstellung solcher organismusspezifischer metabolischer Reaktionslisten entwickelt wurden, ist die Festlegung der thermodynamischen Restriktionen, d.h. der zulässigen Reaktionsrichtungen, ein mühsamer manueller Prozess. Zur automatisierten und systematischen Festlegung thermodynamisch konsistenter Reaktionsrichtungen wurde ein Algorithmus entwickelt, der auf Thermodynamik, Netzwerktopologie und heuristischen Regeln basiert. Die Reaktionsrichtungen in metabolischen Modellen werden so festgelegt, dass die Produktion von Energieäquivalenten im Reaktionsnetzwerk nur in thermodynamisch zulässiger Art und Weise möglich ist. Die Anwendung dieses Algorithmus wurde an einem genomweiten metabolischen Modell des Bakteriums *Escherichia coli* demonstriert. Der Algorithmus konnte zwar nicht die vollständige, jedoch eine wesentliche Anzahl der irreversiblen Reaktionen automatisch und mit geringem Rechenaufwand identifizieren. Damit stellt er ein wertvolles Softwaremodul zur automatisierten Modellbildung genomweiter metabolischer Modelle dar.

Neben diesen stöchiometrischen Modellen des metabolischen Netzwerks wurden in Kapitel 3 thermodynamische Restriktionen zur Analyse grosser, quantitativer Metabolomik-Datensätze benutzt. Dieses Kapitel beschreibt eine Methode zur mechanistischen, modellbasierten Analyse dieser Daten, die thermodynamische Netzwerkanalyse (*network-embedded thermodynamic analysis*, kurz NET-Analyse). Unter Berücksichti-

gung des zweiten Hauptsatzes der Thermodynamik und der Gibbs'schen Formationsenergien der Metabolite leitet die NET-Analyse funktionelle Zusammenhänge ab und identifiziert Reaktionen, auf die vermutlich aktive allosterische oder genetische Regulation wirkt. Dies wurde beispielhaft an zwei quantitativen Metabolomik-Datensätzen von *E. coli* und *Saccharomyces cerevisiae* gezeigt. Darüberhinaus ist die NET-Analyse ein wertvolles Werkzeug, um systematisch Datensätze auf Konsistenz zu überprüfen, um nicht messbare Metabolitkonzentrationen vorherzusagen, und um intrakompartimentelle Konzentrationen aus den gemessenen Konzentrationen (Zelldurchschnittswerte) zu berechnen. Die NET-Analyse ist eine skalierbare Methode, neues Wissen aus quantitativen Metabolomik-Datensätzen zu gewinnen, bei der nur wenige Annahmen und insbesondere kein kinetisches Modell notwendig sind.

Um ein Systemverständnis von zellulären Abläufen zu gewinnen, ist es unabdingbar, die intrazellulären metaboischen Flüsse zu kennen, die sehr nah mit dem Phänotyp verbunden sind. Die heutige Standardmethode zur experimentellen Bestimmung dieser Flüsse, die ^{13}C-basierte Flussanalyse, ist in einigen Punkten experimentell limitiert. In Kapitel 4 wird ein neuartiger Metabolomik-basierter Ansatz vorgestellt, der metabolische Aktivitäten unter Berücksichtigung von Massenbilanzen und thermodynamischer Restriktionen identifiziert und einen grösseren experimentellen Spielraum besitzt. Indem die NET-Analyse mit Elementarmodenanalyse (EFM-Analyse) kombiniert wird, kann die Gesamtheit der aufgrund von Massenbilanzen und thermodynamischer Restriktionen zulässigen Flüsse umfassend analysiert werden. Alle EFMs eines stöichiometrischen metabolischen Netzwerkmodells werden mit der NET-Analyse auf thermodynamische Zulässigkeit mit gemessenen Metabolitkonzentrationen getestet. Um diese Methode zu demonstrieren, wurde die Aktivität von Reaktionen innerhalb des Pyruvat-Ethanol-Metabolismus in *S. cerevisiae* für Wachstum sowohl auf Glukose als auch auf Ethanol aufgrund von gemessenen Metabolitkonzentrationen bestimmt.

In Kapitel 5 wurde die Glukoserepression in den *S. cerevisiae*-Stämmen CEN.PK 113-D7 und FY4 anhand von quantitativen und grossen Proteomik- und Metabolomik-Datensätzen sowie anhand von metabolischen Flüsse verglichen. Die Wildtypstämme

zeigen auf der Ebene der metabolischen Flüsse einen ähnlichen Grad an Glukoserepression, da sowohl Biomasse- und Ethanolausbeuten sowie die respiratorische Zitratzyklusaktivität ähnlich sind. Trotzdem wurden unterschiedliche Protein- und Metabolitkonzentrationen beobachtet, die darauf hinweisen, dass die Hxk2-abhängige Signalübertragung und Regulation in CEN.PK stärker ausgeprägt ist. Unterschiede in den Signal- und Regulationsstärken traten auch dadurch zutage, dass sich die Glukoserepression auf Flussebene, die durch die respiratorische Zitratzyklusaktivität bestimmt wurde, infolge einer genetischen Mutation wesentlich verändert hat: Während in FY4 die Glukoserepression aufrecht erhalten wird, wenn *HXK2* eliminiert wird, geht CEN.PK zu einem respiratorischen Metabolismus über. Es konnte gezeigt werden, dass eine Punktmutation im *CYR1*-Gen für das unterschiedliche Verhalten verantwortlich ist.

Über diese Arbeiten hinaus wurden metabolische Veränderungen im Verlauf des diauxischen Wachstums von *S. cerevisiae*s mit Hilfe von Omik-Technologien beobachtet (siehe Appendix). Alle Phasen dieser metabolischen Umstellung wurden mit quantitativen Metabolomik- und Proteomik-Daten sowie der extrazellulären Physiologie charakterisiert. Ein Teil der aufgenommenen Daten wurden verwendet, um die NET/EFM-Analyse zu entwickeln (vgl. Kapitel 4). Der komplette dynamische Datensatz ist die Basis für zukünftige Analysen, um die zeitliche und kausale Folge von regulatorischen Ereignissen zu bestimmen, mit denen die Zelle auf die veränderte Substratzugänglichkeit reagiert.

Heutzutage entwickelt sich die Technologie, grosse quantitative Omik-Datensätze zu erzeugen, sehr schnell. Insbesondere die Entwicklung allgemein anwendbarer modellbasierter Analysemethoden für solche Daten bleibt dahinter zurück. Die in dieser Arbeit entwickelten und angewendeten modellbasierten Methoden, grosse Datensätze auszuwerten, tragen zur Datenanalyse in der rechnergestützten Systembiologie bei. Beispielhaft konnten globale quantitative Datensätze zum *S. cerevisiae*-Stoffwechsel generiert und interpretiert werden.