Analysis of Large-Scale Metabolic Networks: Organization Theory, Phenotype Prediction and Elementary Flux Patterns

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

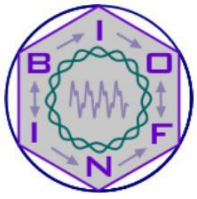
von Diplom-Bioinformatiker Christoph Kaleta
geboren am 16. Mai 1983 **in** Weimar, Deutschland

Gutachter:

1. Prof. Dr. Stefan Schuster (Universität Jena)
2. PD Dr. Peter Dittrich (Universität Jena)
3. Prof. Dr. Joachim Selbig (Universität Potsdam)

Tag der öffentlichen Verteidigung: 22.3.2010

Die vorliegende Arbeit wurde am Lehrstuhl für Bioinformatik der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena unter der Leitung von Prof. Dr. Stefan Schuster und in der Biosystemanalysegruppe unter der Leitung von PD Dr. Peter Dittrich angefertigt.



Abstract

Understanding the function and control of metabolic networks is one of the central topics in systems biology. The metabolic network allows an organism to recreate all of its components from molecules that can be found in its environment. Due to the inherent complexity of such networks different methods aimed at analysing their structure have been proposed. The methods used here can be broadly divided into two types. The first type comprises flux-centered approaches like elementary mode analysis, the closely related extreme pathway analysis and flux balance analysis. These approaches have in common that they concentrate on possible fluxes through reactions at steady state. The second type, represented by chemical organization theory, additionally explicitly takes into account metabolites. Thus, not only a specific set of reactions but all possible reactions that can be performed by a given set of metabolites is considered. All these methods differ in the size of the networks to which they can be applied and in the type of results that can be obtained. The major focus of this work is the integration of several of these concepts in order to allow a more comprehensive analysis of metabolic networks. Indeed, elementary mode analysis and chemical organization theory can complement each other in two ways. First, the set of chemical organizations of a reaction network allows for a meaningful clustering of elementary modes and helps to identify elementary modes that cannot be operative at steady state. Second, the set of elementary modes can be used as a base for an algorithm to compute chemical organizations. Furthermore, a comparison shows that taking into account the compounds present in a reaction network is of importance since

chemical organization theory allows for more accurate predictions concerning the viability of an organism after a perturbation of its metabolic network. In another direction, the combination of concepts from flux balance analysis and elementary mode analysis helps to overcome the problem that the entire set of elementary modes can only be computed in small or medium size metabolic networks. The framework behind this integration, elementary flux pattern analysis, allows one to identify all possible routes through a subsystem that are part of a pathway in a genome-scale metabolic network. A first important benefit of this concept is that many approaches building on elementary mode analysis can now be applied to genome-scale metabolic networks. Furthermore, using elementary flux patterns, we were able to identify several elementary modes in a subsystem of a metabolic model of *Escherichia coli* reported in a previous work that are not part of a steady-state flux if the entire network is considered. Additionally, we discovered several alternative routes to common metabolic pathways in the central metabolism of *E. coli*. Using elementary flux pattern analysis in a genome-scale metabolic model of humans we found several pathways that contradict a widely held assumption in biochemistry saying that the conversion of even-chain fatty acids into glucose is infeasible in humans.

Zusammenfassung

Die Untersuchung und Aufklärung von Zusammenhängen in metabolischen Netzwerken ist eines der zentralen Aufgabengebiete der Systembiologie. Ein metabolisches Netzwerk erlaubt es dem Organismus, alle seine Bestandteile aus Stoffen in seiner Umgebung herzustellen. Aufgrund der Komplexität dieser Netzwerke wurden verschiedene Methoden zur Untersuchung ihrer Struktur vorgeschlagen. Die Methoden, die in dieser Arbeit zur Anwendung kommen, können grob in zwei Richtungen unterteilt werden. Einerseits gibt es Methoden die sich auf die Flüsse in einem metabolischen Netzwerk konzentrieren. Beispiele für solche Methoden sind die Elementarmodenanalyse, die sehr ähnliche Analyse von Extreme Pathways und die Flussbilanzanalyse (Flux Balance Analysis). Andererseits gibt es Methoden, wie die Theorie der chemischen Organisationen, die zudem die Metabolite in einem Netzwerk in die Analyse mit einbeziehen. Dabei werden nicht nur Mengen an Reaktionen, sondern grundsätzlich alle Reaktionen die zwischen einer bestimmten Menge an Metaboliten möglich sind, betrachtet. Alle diese Methoden unterscheiden sich in der Größe der Netzwerke, auf die sie angewendet werden können und in den Ergebnissen, die sie liefern. Das zentrale Ziel dieser Arbeit ist es diese Methoden miteinander in Beziehung zu setzen und zu kombinieren, um eine umfassendere Analyse von Reaktionsnetzwerken zu ermöglichen. So können sich die Elementarmodenanalyse und die Theorie Chemischer Organisationen in zweierlei Hinsicht ergänzen. Einerseits erlauben es die chemischen Organisationen eines Reaktionsnetzwerks dessen Elementarmoden zu gruppieren und solche zu identifizieren, die nicht im stationären Zustand des Reaktionsnetzwerks auftreten können.

Andererseits können Elementarmoden verwendet werden, um chemische Organisationen zu berechnen. Weiterhin kann gezeigt werden, dass die explizite Beachtung von Metaboliten die Genauigkeit der Vorhersage bezüglich der Lebensfähigkeit eines Organismus nach der Störung seines metabolischen Netzwerks verbessert. Ein weiterer zentraler Punkt dieser Arbeit ist die Kombination von Flussbilanzanalyse mit Elementarmodenanalyse, die es erlaubt, die Problematik zu umgehen, dass Elementarmoden nur in metabolischen Netzwerken kleinerer und mittlerer Größe vollständig berechnet werden können. Das Konzept der Elementary Flux Patterns, das dieser Integration zu Grunde liegt, erlaubt es alle möglichen Wege von stationären Flüssen durch ein Teilnetwerk eines Ganzzellmodells (genome-scale model) zu bestimmen. Ein erster Vorteil dieses Konzepts liegt darin, dass viele Methoden, die auf Elementarmodenanalyse aufbauen, jetzt auch auf Ganzzellmodelle angewendet werden können. Weiterhin konnten wir in einem Teilsystem eines Ganzzellmodells des Metabolismus von *Escherichia coli* mehrere Elementarmoden identifizieren, die die Stationaritätsbedingung zwar in dem Teilmodell erfüllen, aber nicht Teil eines stationären Flusses durch das Gesamtmodell sein können. In demselben Modell fanden wir außerdem mehrere Reaktionswege, die als Alternativrouten für Stoffwechselwege im zentralen Metabolismus von *E. coli* verwendet werden können. In einem Modell des menschlichen Metabolismus identifizierten wir zudem mehrere Reaktionswege, die der weit verbreiteten Annahme widersprechen, dass die Umwandlung von geradzahligen Fettsäuren in Glucose im Menschen nicht möglich ist.

Danksagung

Nach nur einem Jahr neigt sich meine Arbeit als Doktorand schon dem Ende zu. Dass ich meine Doktorarbeit in dieser kurzen Zeit fertigstellen konnte, habe ich Peter Dittrich und Stefan Schuster zu verdanken. Durch meine Arbeit in der Biosystemanalysegruppe von Peter Dittrich in den Jahren 2005 – 2008 konnte ich erste wertvolle Erfahrungen in der wissenschaftlichen Praxis sammeln. Mein besonderer Dank gilt Stefan Schuster, da er es mir ermöglicht hat einen Teil dieser Arbeit und andere Arbeiten, die schon während meiner Zeit bei der Biosystemanalysegruppe, begonnen haben in seiner Gruppe fortzusetzen. Dem messe ich besondere Bedeutung bei, da ich so meine zuvor gewonnenen Erfahrungen weiter vertiefen konnte. Weiterhin gilt mein Dank auch Reinhard Guthke, bei dem ich im Rahmen meiner Diplomarbeit lernte, wie schwierig es ist Modelle von biologischen Systemen zu konstruieren. Es ist schwer einer Danksagung eine Reihenfolge aufzuerlegen. Gerade deshalb möchte ich genauso meiner Familie und ganz besonders meiner Freundin Claudia danken, die mich während dieser ganzen Zeit tatkräftig unterstützt und, wenn notwendig, auch mal abgelenkt haben. Hier möchte ich natürlich auch meinen Mitarbeitern danken, mit denen die Zusammenarbeit immer eine sehr große Freude war und sicherlich auch noch sein wird. Dabei möchte ich ganz besonders Luís, Jörn, Ines, Frank, Florian und Pietro danken.

Contents

Chapter 1

Introduction

Suddenly, Systems Biology is everywhere. Bernhard Ø. Palsson.

Systems biology is one of the most important and active fields of current research. Its primary focus is to deduce the interconnections between the various constituents of biological systems, integrating them into a larger picture in order to gain an understanding of properties that cannot be seen when considering the modules independently [Kitano, 2002, Westerhoff and Palsson, 2004, Cornish-Bowden and Cárdenas, 2005, Ventura et al., 2006]. However, due to the nature of living systems this leads to models that are growing more and more complex. Thus, it becomes difficult to analyze them by intuitive reasoning and methods that are able to deal with their inherent complexity are at need.

This problem is very apparent in the analysis of **metabolic networks**, which represent the biochemical factories of organisms (see Figure 1.1 for an example). Metabolic networks are made up by compounds and reactions transforming them into each other. Compounds can either be small molecules, like sugars and lipids, or macromolecules, like DNA and proteins. The principal task of the metabolic network is to convert a set of substrate molecules the organism can find in its environment to a set of products that are required for its survival such as the macromolecules of which it consists. The required substrates can vary from organism to organism. Autotroph organisms only do need some salts, carbon dioxide and

light to survive, while heterotrophic organisms like animals need to take up also more complex compounds, for instance, some amino acids. Parasites in contrast require a constant supply of many compounds from a host organism. Depending on these requirements also the size of the metabolic network varies. Thus, the smallest metabolic network model covering the metabolism of an entire organism, the parasite *Mycoplasma genitalium*, already contains 274 compounds and 262 reactions [Suthers et al., 2009]. In contrast the currently largest reconstructed model of human metabolism contains 2766 compounds and 3311 reactions [Duarte et al., 2007].

The metabolic network of an organism is controlled by a **regulatory network**. The task of the regulatory network is to fine tune the metabolic network according to the need of the organism if subjected to varying environmental conditions. Thus, the regulatory network acts as a control system that integrates signals from external and internal sensors and adapts the metabolic network accordingly. Such an adaption can be, for instance, the increase of the velocity of a reaction by increasing the expression of the catalyzing enzyme or the activation of an entire pathway that has previously been shut down. An important aspect of regulatory networks is that the boundaries to metabolic networks are sometimes blurred, for instance, if a compound is regulating its own production. Yet, most methods that have been developed in the analysis of metabolic networks do not take into account the interplay between the metabolic and the regulatory network. Instead, they assume that all of the reactions in a metabolic network are available at each timepoint. There are several reasons for this discrepancy. First, it is far more difficult to deduce the regulation of a metabolic network than the metabolic network itself since a regulatory effect might be subtle, overlapping with other effects or only encountered under very specific conditions. Many enzymes present in an organism in contrast can be deduced from its genome for proteins with homologues of known function in other organisms [Notebaart et al., 2006]. Second, many methods for the analysis of metabolic networks usually require a mass-flow in order to find a

relationship. Yet, regulatory effects do not necessarily entrain mass-flow and thus methods that aim at the integration of metabolic and regulatory networks either introduce an artificial mass-flow [Gianchandani et al., 2006, Behre and Schuster, 2009] to take into account regulatory events or model them by Boolean rules which govern the availability of reactions [Covert et al., 2001, Covert and Palsson, 2002, Shlomi et al., 2007]. Thus, one part of this work focuses on the development of a method that allows the analysis of an integrated regulatory and metabolic network.

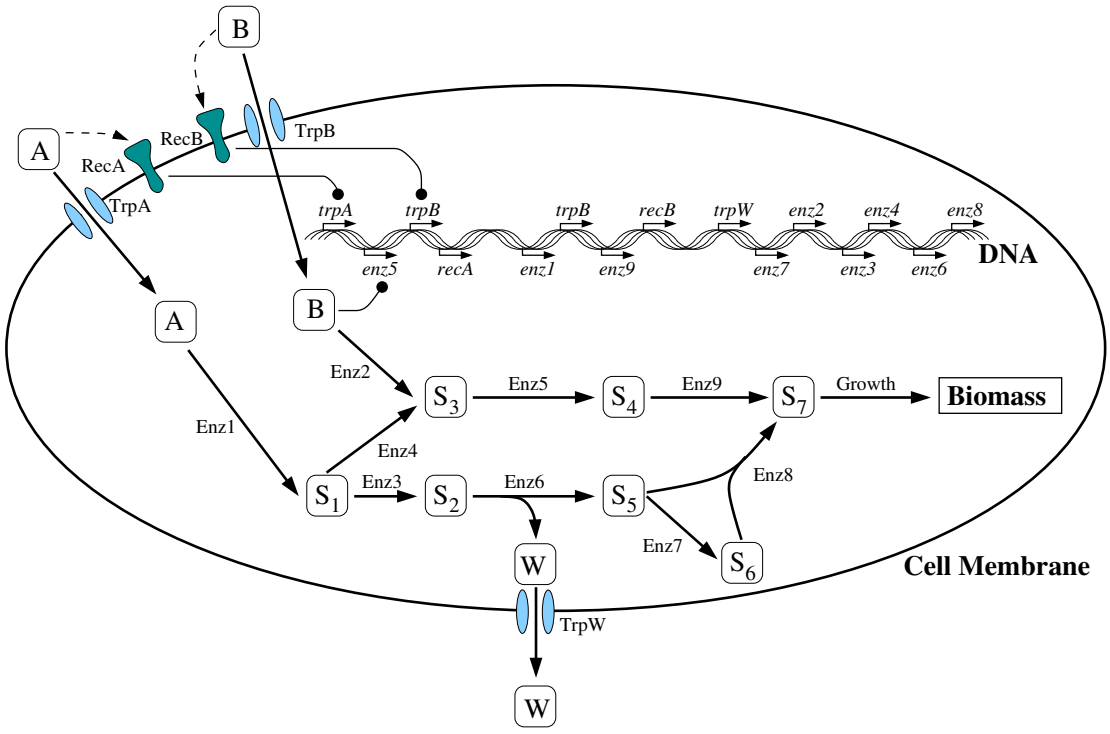


Figure 1.1: Simplified representation of a metabolic network integrated into a cellular context. The extra-cellular forms of A and B can be taken up through the transporters TrpA and TrpB. The expression of the transporters is regulated by RecA and RecB which sense the concentration of A and B, respectively. Both compounds can be used to produce S_7 which is converted into biomass, for instance, in form of cell-membrane constituents, amino acids or nucleotides. In one pathway a waste metabolite W that needs to be excreted through the transporter TrpW is produced. Furthermore, intracellular B activates the expression of Enz5, an enzyme necessary for the conversion of B into biomass. Lines ending in circles indicate activation. In cases where no regulatory event is indicated, the corresponding gene is assumed to be expressed constitutively.

An important point in the analysis of metabolic networks is the level of abstraction on which the analysis is based (Figure 1.2). With greater chemical or

physical detail the quality of predictions improves. However, this can lead to a too large computational burden for a comprehensive analysis like it occurs in molecular dynamics [Lindahl, 2008]. Furthermore, the parameters of the model might be difficult or impossible to obtain like it is often the case when ordinary differential equations are used (*cf.* Tsai and Wang [2005]). The level of abstraction used in this work is the stoichiometry of a metabolic network represented by the stoichiometric matrix (see Figure 1.2B for an example). The stoichiometry indicates for each reaction the number of compounds consumed and produced. This also includes compounds just participating in a reaction without being consumed or produced, for instance, catalyzing compounds. However, such compounds are not displayed in the stoichiometric matrix since their net production is zero. In contrast to reaction kinetics, stoichiometries can be more easily obtained and stoichiometric models for the metabolism of entire organisms, so-called genome-scale metabolic network models, are available (see Table 1.1 for a list of genome-scale metabolic networks of important model organisms). Predictions that can be made from the analysis of the stoichiometry of a reaction network are by principle less accurate than those obtained from more detailed methods [Dittrich and Speroni di Fenizio, 2007]. Yet, the dynamic behavior of a metabolic network is constrained by its stoichiometry. For instance, the non-producibility of a certain compound due to the knockout of a reaction in a stoichiometric analysis will also be seen if more detailed methods are used.

Organism	# Genes	# Metab.	# React.	Reference
Saccharomyces cerevisiae	832	1168	1857	Herrgård et al. [2008]
Homo sapiens	1496	2766	3311	Duarte et al. [2007]
Escherichia coli	1260	1039	2077	Feist et al. [2007]
Mycoplasma genitalium	189	274	262	Suthers et al. [2009]
Bacillus subtilis	844	988	1020	Oh et al. [2007]
Corynebacterium glutamicum	-	411	446	Kjeldsen and Nielsen [2009]

Table 1.1: List of selected genome-scale metabolic networks. The second column indicates the number of genes the corresponding model takes into account. The third and fourth column give the number of compounds and reactions, respectively.

Important methods for the analysis of the stoichiometry of reaction networks are elementary mode analysis [Schuster et al., 1999, 2000], the very similar extreme

A

$$
\begin{aligned}
v_{A_{in}} &= k_1 \cdot A_{ext} \cdot a_1 \cdot A_{ext} && (1)\\
v_{B_{in}} &= k_2 \cdot B_{ext} \cdot a_2 \cdot B_{ext} && (2)\\
v_{Enz1} &= k_3 \cdot A && (3)\\
v_{Enz2} &= k_4 \cdot B && (4)\\
v_{Enz3} &= k_5 \cdot S_1 && (5)\\
v_{Enz4} &= k_6 \cdot S_1 && (6)\\
v_{Enz5} &= k_7 \cdot S_3 \cdot a_3 \cdot B && (7)\\
v_{Enz6} &= k_8 \cdot S_2 && (8)\\
v_{Enz7} &= k_9 \cdot S_5 && (9)\\
v_{Enz8} &= k_{10} \cdot S_5 \cdot S_6 && (10)\\
v_{Enz9} &= k_{11} \cdot S_4 && (11)\\
v_{W_{out}} &= k_{12} \cdot W && (12)\\
v_{Growth} &= k_{13} \cdot S_7 && (13)
\end{aligned}
$$

B

	A_{in}	B_{in}	Enz1	Enz2	Enz3	Enz4	Enz5	Enz6	Enz7	Enz8	Enz9	W_{out}	Growth
A	1	0	−1	0	0	0	0	0	0	0	0	0	0
B	0	1	0	−1	0	0	0	0	0	0	0	0	0
S_1	0	0	1	0	−1	−1	0	0	0	0	0	0	0
S_2	0	0	0	0	1	0	0	−1	0	0	0	0	0
S_3	0	0	0	1	0	1	−1	0	0	0	0	0	0
S_4	0	0	0	0	0	0	1	0	0	0	−1	0	0
S_5	0	0	0	0	0	0	0	1	−1	−1	0	0	0
S_6	0	0	0	0	0	0	0	0	1	−1	0	0	0
S_7	0	0	0	0	0	0	0	0	0	1	1	0	−1
W	0	0	0	0	0	0	0	1	0	0	0	−1	0

C

$$
\begin{aligned}
RecA &\begin{cases} \texttt{true}, & \text{if } [A_{ext}] > 0 \\ \texttt{false}, & \text{otherwise} \end{cases}\\
RecB &\begin{cases} \texttt{true}, & \text{if } [B_{ext}] > 0 \\ \texttt{false}, & \text{otherwise} \end{cases}\\
TrpA &\quad \texttt{if } RecA\\
TrpB &\quad \texttt{if } RecB\\
A_{in} &\quad \texttt{if } TrpA\\
B_{in} &\quad \texttt{if } TrpB\\
Enz5 &\quad \texttt{if } [B] > 0
\end{aligned}
$$

Figure 1.2: Examples for different types of representations of the metabolic network in Figure 1.1. **A** Rate laws governing the velocities of the reactions if simple mass-action kinetics are assumed. For simplicity, the activation of TrpA and TrpB is modeled to be instantaneously if the extra-cellular forms of A and B are present. **B** Stoichiometric matrix **N** of the metabolic network. The extra-cellular forms of A, B, and W as well as the pseudo-compound Biomass are assumed to be external, that is, they are considered to be buffered and are therefore not included in the stoichiometric matrix. **C** Boolean rules governing the expression of TrpA, TrpB and Enz5. In contrast to **A** the full detail of the activation is considered here. For instance, extracellular A binds to RecA which in turn activates the expression of TrpA. With TrpA present, A can be imported and thus a non-zero flux in $v_{A_{in}}$ is possible.

pathway analysis [Schilling et al., 2000], flux balance analysis [Fell and Small, 1986, Varma and Palsson, 1994b, Lee et al., 2006], and chemical organization theory [Dittrich and Speroni di Fenizio, 2007].

Elementary modes correspond to minimal sets of reactions that can operate at steady state with all reactions proceeding only in thermodynamical feasible directions ([Schuster et al., 1999], see Figure 1.3 for examples). Thus, reactions considered as irreversible since the backward direction is thermodynamically not feasible under physiological conditions need to have a positive flux. The set of reactions is minimal in the sense that setting the flux through one reaction of an elementary mode to zero implies that there is no steady-state flux through the

remaining reactions [Schuster et al., 1999]. An important property of elementary modes is that every steady-state flux of a reaction network can be written as a non-negative linear combination of elementary modes. Elementary modes have been used in many areas of biotechnology, such as assessing network flexibility [Stelling et al., 2002], finding pathways with optimal yields for certain metabolic compounds [Schuster et al., 2002, Krömer et al., 2006], finding possible targets for the engineering of metabolic networks [Klamt, 2006] and analyzing the effect of such an engineering [Carlson et al., 2002, Schwender et al., 2004]. Elementary mode analysis usually requires the enumeration of all elementary modes, which is not possible for larger systems [Klamt and Stelling, 2002]. Hence, only small systems focusing on a specific part of the entire known reaction network of an organism can be analyzed. The remaining systems is modeled by the introduction of exchange reactions and the definition of specific metabolites as external. Such metabolites are assumed to be buffered by the remaining system and thus do not need to obey the steady-state condition. However, these modifications heavily impact the results that are obtained. Thus, as will be shown in this work, some pathways detected in the subsystem might not belong to any pathway obeying the steady-state condition on the scale of the entire system. Furthermore, not all possible routes of steady-state fluxes in the entire network through the subsystem are found.

Flux balance analysis in contrast can be applied to larger systems. It consists of the search for a flux vector that satisfies the steady-state condition, a set of constraints, like maximal reaction fluxes, and optimizes a certain criterion. This criterion can be, for instance, the yield of a given metabolite or an overall biomass term [Fell and Small, 1986, Varma and Palsson, 1994b]. Flux balance analysis has been used to predict the rate of uptake and secretion of metabolites [Varma and Palsson, 1994a], to predict the lethality of gene knock-outs [Edwards and Palsson, 2000, Joyce and Palsson, 2008] and in the design of strains for improved metabolite production [Burgard and Maranas, 2001]. Furthermore, it has been

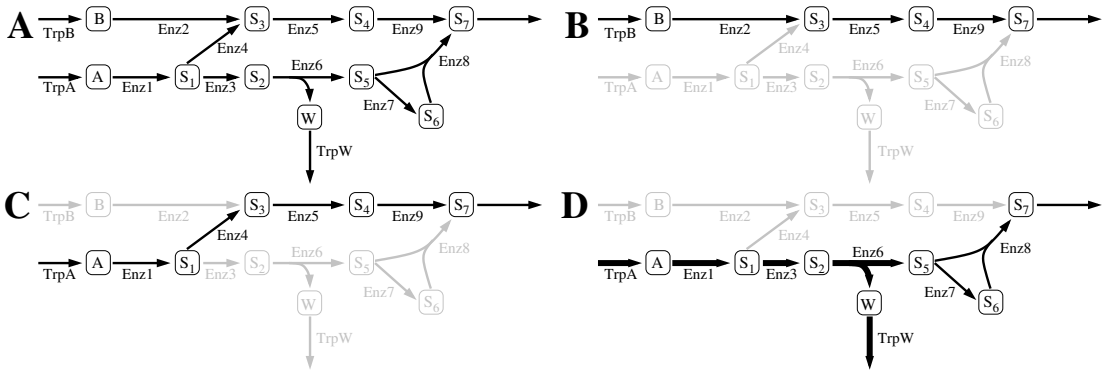


Figure 1.3: Elementary modes of the metabolic network in Figure 1.1. **A** Network representation of the stoichiometric matrix in Figure 1.2B considered for the computation of elementary modes. **B-D** Three elementary modes of the system. Reactions belonging to an elementary mode are drawn in black. The thickness of the arrows corresponds to the relative flux each reaction carries. Two aspects of elementary mode analysis can be seen from EM C and EM D. First, EM C has a higher molar yield in S_7 produced from A than EM D, since the former consumes one mole of A while the latter consumes two moles of A to produce one mole of S_7. The pathway producing the highest molar yield of a certain compound always corresponds to an elementary mode [Schuster et al., 1999]. Second, EM C contradicts a regulatory rule since it uses Enz5, which, as can be seen from Figures 1.1 and 1.2C, is only expressed if B is present in the cell.

extended to allow the inclusion of regulatory rules [Covert et al., 2001, Covert and Palsson, 2002, Shlomi et al., 2007] and reaction kinetics [Covert et al., 2008] into the analysis. One important caveat of flux balance analysis is that an appropriate objective function that is to be optimized has to be defined. Usually, is is assumed that the growth rate of the organism is maximized, a requirement which is not met under all circumstances [Schuetz et al., 2007, Schuster et al., 2008, Papp et al., 2009]. Furthermore, flux balance analysis only returns a single optimal flux. Thus, it is not possible to obtain alternative pathways [Mahadevan and Schilling, 2003].

The concept of **elementary flux patterns** introduced in Kaleta et al. [2009b] combines elementary mode analysis and flux balance analysis in order to overcome the shortcomings of both approaches. By definition, elementary flux patterns represent the basic routes of each steady-state flux of a large-scale metabolic network through a subsystem. Thus, an analysis using elementary flux patterns explicitly considers possible fluxes through the entire network when considering the fluxes through a subnetwork. In consequence a more thorough analysis of the integration of the subsystem into the entire network is achieved. Furthermore, through a

strong connection to elementary mode analysis, elementary flux patterns now allow the application of tools building on elementary mode analysis to genome-scale metabolic networks Kaleta et al. [2009b].

Finally, **chemical organization theory** takes a somewhat different approach since it does not only consider the reactions within a system but also the specific compounds that participate in them. Hence the information that is used is not constrained to compounds that are consumed or produced by a reaction. Based on the ideas of [Fontana and Buss, 1994] an chemical organization is defined as a set of compounds that is closed and self-maintaining [Dittrich and Speroni di Fenizio, 2007]. Closure indicates that there is no reaction within a metabolic network that produces a compound not within the organization from compounds of the organization. Self-maintenance requires that there exists a flux through the reactions among the compounds of an organization such that the concentration of no compound decreases. Since organizations may share the same species, the set of organizations together with the set inclusion $\subseteq$ form a partially ordered set that can be visualized in a Hasse diagram, providing a hierarchical view of the network under consideration: Organizations are vertically arranged by size, with small organizations at the bottom. Two organizations are connected by a line if the upper contains the lower organization and no other organization exists between them. The Hasse diagram of organizations of the network displayed in Figure 1.1 is depicted in Figure 1.4. An interesting property of chemical organizations is that every steady state and growth state of a reaction network[1], corresponds to a chemical organization if the reactions obey a certain condition [Dittrich and Speroni di Fenizio, 2007, Kaleta et al., 2009c]. Using this property, chemical organization theory has already been applied to the prediction of growth phenotypes [Centler et al., 2007] and the outcome of knockout experiments [Kaleta et al., 2008], as well as in the design of chemical programs to solve NP-complete problems [Matsumaru et al., 2007]. Furthermore, a concept identical to closure by the name of scopes

[1] a growth state is a situation in which the concentration of some compounds increases, while the concentration of no compound decreases

has been used to study the biosynthetic potentials of certain metabolites [Handorf et al., 2005, Matthäus et al., 2008]

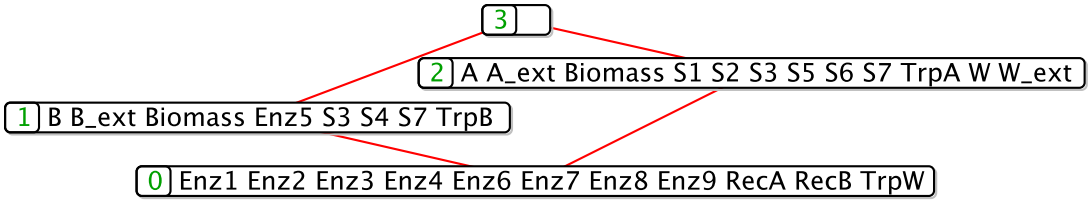


Figure 1.4: Hasse diagram of organizations of the model depicted in Figure 1.1. For simplicity, only species appearing for the first time, that is, which are not element of a lower organization, are displayed. The chemical organizations have been computed assuming that there is an outflow for every compound. The inputs of A and B have been modeled using a replicator reaction of the form $1A \rightarrow 2A$ for A and similarly for B. Thus, depending on the replicator reactions used, all four possible input scenarios for A and B are modeled. Hence, the following input-scenarios are used: Organization 0, no input; Organization 1, input of B; Organization 2, input of A; Organization 3, input of A and B. The four organizations correspond to the cellular phenotypes that can be seen under the four modeled conditions. Furthermore, in contrast to the elementary mode displayed in Figure 1.3D, $Enz5$ is correctly predicted to be present only if B, which activates the expression of this protein, is present.

The aim of this work is to demonstrate how the integration of results obtained from different stoichiometric analysis methods can help to gain a better understanding of metabolic networks. Chapter 2 is dedicated to the comparison of elementary modes and chemical organizations in order to understand how both concepts can benefit from each other: on the one hand elementary modes can be grouped according to chemical organizations and, on the other hand, the chemical organizations of a system can be computed from its elementary modes.

The last part of Chapter 2 and Chapter 3 are aimed at the analysis of the phenotypes of metabolic networks. A phenotype is here considered as a set of compounds that can be present over long periods of (simulation) time which directly translates into the concept of chemical organizations. This is similar to the concept of metabolic phenotypes introduced in Varma and Palsson [1994a] which correspond to a particular optimal flux distribution in a metabolic network. In a first setting we relate the phenotypes arising by the knockout of certain genes to the viability of an organism by checking whether we still find an organization that contains all the compounds required for growth. In a second setting,

by identifying compounds that do not appear in any chemical organization, we can infer those parts of a metabolic network which might contain inconsistencies. Such inconsistencies can be, for example, due to missing knowledge, an insufficient growth media or errors in the model. In the course of this analysis we also demonstrate how the concept of chemical organizations can be extended in two directions. First, regulatory and metabolic networks are often analyzed independently from each other. Chemical organization theory allows for the translation of Boolean rules into reactions. Using this property we demonstrate how an integrative model of metabolism and the regulatory network governing it can be analyzed. Second, the correspondence of steady states and growth states requires that the reactions of a network fulfill a certain property. Given that further information on the reactions' kinetics are available, we show how it can be verified whether the reactions fulfill this property and how they can be adapted if this is not the case.

Chapter 4 is focused on the above described shortcomings of elementary mode analysis and flux balance analysis which we overcome by the introduction of the concept of elementary flux patterns. In the first part of Chapter 4, elementary flux patterns are introduced and several applications of which one is studied in detail are outlined. In the second part of Chapter 4, we use elementary flux patterns to infer pathways which contradict the widely held believe in biochemistry that a conversion of fatty acids into carbohydrates is not possible in humans [Stryer et al., 2002, Lehninger et al., 2008, Voet et al., 2005]. Analyzing the inferred pathways in detail we found that indeed there have been previous works which suggested some of the pathways we identified as probable gluconeogenic routes from fatty acids in humans [Weinman et al., 1957, Argilés, 1986, Landau and Brunengraber, 1987]. However, these studies seem to have been largely ignored.

Chapter 2

From Metabolic Pathways to Chemical Organizations

In Kaleta et al. [2006] we give an outline of the connection between elementary modes and chemical organizations. Thus, we demonstrate how the analysis of elementary modes can be facilitated using chemical organizations and how chemical organizations can be used to more faithfully asses the metabolic capabilities of organisms. In the other direction we show how the set of chemical organizations can be computed from the set of elementary modes in Centler et al. [2008]. Furthermore, we present another algorithm for the computation of chemical organizations based on the Hasse diagram of organizations and a heuristic algorithm for the computation of chemical organizations in genome-scale metabolic networks. Using this heuristic in a whole-cell model of *E. coli* we demonstrate how chemical organization theory can be used to predict parts of a reaction network whose maintenance can already be explained given a sufficient growth media and parts that need refinement due to missing knowledge.

The Supplemental Material of Centler et al. [2008] can be found in Chapter 6 on pages 108 ff.

C. Kaleta, F. Centler, and P. Dittrich. *Mol. Biotechnol.*, 34(2):117-123, Oct 2006.

RESEARCH

Analyzing Molecular Reaction Networks

From Pathways to Chemical Organizations

Christoph Kaleta,[1] Florian Centler,[1] and Peter Dittrich[1,*]

Abstract

Pathways are typically the central concept in the analysis of biochemical reaction networks. A pathway can be interpreted as a chain of enzymatical reactions performing a specific biological function. A common way to study metabolic networks are minimal pathways that can operate at steady state called elementary modes. The theory of chemical organizations has recently been used to decompose biochemical networks into algebraically closed and self-maintaining subnetworks termed organizations. The aim of this paper is to elucidate the relation between these two concepts. Whereas elementary modes represent the boundaries of the potential behavior of the network, organizations define metabolite compositions that are likely to be present in biological feasible situations. Hence, steady state organizations consist of combinations of elementary modes. On the other hand, it is possible to assign a unique (and possibly empty) set of organizations to each elementary mode, indicating the metabolites accompanying the active pathway in a feasible steady state.

Index Entries: Chemical organization; elementary mode; metabolic network; steady state flux distribution; stoichiometric network analysis.

1. Introduction

Elementary modes have proven to be a powerful means in the analysis of metabolic networks and their underlying properties. For example, they have been used to assess network flexibility ***(1)***, to find pathways with optimal yields for certain metabolites ***(2)*** and to study enzyme deficiencies ***(3)***. Because the number of elementary modes grows exponentially with the size of the network, the study of elementary modes in larger systems is difficult ***(4)***.

Chemical organization theory has recently been proposed as another tool to examine capabilities and structure of metabolic networks. It allows to decompose a metabolic network into algebraically closed and self-maintaining subnetworks that form a hierarchy. One important conclusion drawn from organization theory is, that every steady state of a metabolic network can be mapped to an organization ***(5)***. Because steady states are broadly regarded as an important condition of metabolic networks, organization theory can serve as a base to study the capabilities of such networks. From elementary mode analysis it is known that each steady state of the system can be interpreted as a combination of elementary modes. This highlights the link between the two concepts: each steady state is on one hand a combination of elementary modes, and on the other hand an organization. The remaining part of this paper analyzes the relation between elementary modes and organizations in more detail. The article is organized as follows. In **Subheadings 2.** and **3.**, the concepts of elementary modes and chemical organizations

*Author to whom all correspondence and reprint requests should be addressed. [1]Bio Systems Analysis Group, Jena Centre for Bioinformatics (JCB), and Department of Mathematics and Computer Science, Friedrich-Schiller-University Jena, D-07743 Jena, Germany. Email: dittrich@minet.uni-jena.de

 ISSN: 1073–6085/Online ISSN: 1559–0305/2006/34:2/117–124/$30.00

are introduced. **Subheading 4.** examines the relatedness of both in detail. Some simple networks and a more elaborated network model of the central sugar metabolism of *Escherichia coli* are used to demonstrate the results in **Subheading 5.** Finally, we conclude in **Subheading 6.**

2. Elementary Modes

Elementary modes represent minimal sets of reactions that can operate at steady state with all reactions proceeding in their appropriate direction ***(6)***. The reaction set is minimal in the sense that there is no subset of reactions that could also operate at steady state. Metabolites are classified as either internal or external. Whereas internal metabolites are required to be in steady state, external metabolites are buffered by reactions not contained in the model. They are the potential subtrates and products of the pathway. Elementary modes are closely related to T-invariants in Petri nets ***(7)***.

Generally, steady state solutions for a metabolic network containing n reactions can be determined in the n-dimensional flux space of the system. Each flux vector $\vec{v} \in \mathbb{R}^n_+$ in the flux space assigns to each reaction a nonnegative value that represents the reaction's turnover rate. The steady state condition imposes constraints in the flux space so that the solution space containing all possible steady state flux distributions forms a convex cone ***(4)***. The edges of the cone are exactly the elementary modes. Hence, they form a basis for the solution space and accordingly, every feasible steady state flux distribution is a linear combination of elementary modes. In this sense, elementary modes represent the extreme boundaries of the network's potential behavior.

3. Chemical Organizations

The theory of chemical organizations ***(5)*** provides a new method to analyze complex reaction networks. Extending ideas by Fontana and Buss ***(8)***, the main objective is to determine combinations of network species that are more likely to be present over a long period of (simulation-) time than others. Such sets of species are called organizations. To be an organization, a species set has to fulfill two properties: algebraic closure and self-maintenance. The first property—closure—ensures that given the molecular species of an organization, there is no reaction within the reaction network that could create a species not yet present in the organization. The second property—self-maintenance—guarantees that every molecular species that is used-up within the organization can be recreated from organization species at a sufficient rate for its maintenance. Formal definitions of these concepts are given in **Subheading 3.1.** The method delivers a set of organizations, representing all closed and self-maintaining subnetworks of the system. It is shown by Dittrich and Speroni di Fenizio ***(5)*** that assuming that the dynamics is modelled using ordinary differential equations, all steady states of the system are instances of organizations, i.e., the species with concentrations greater zero in a particular steady state are exactly the species contained in a corresponding organization. Because organizations may share the same species, the set of organizations together with the set inclusion $\subseteq$ form a partially ordered set that can be visualized in a Hasse diagram providing a hierarchical view on the network under consideration (*see* **Fig. 2B** for an example). Organizations are vertically arranged according to their size, with small organizations at the bottom. Two organizations are connected by a line if the upper contains the lower organization and no other organization exists between them.

3.1. Formal Definition of Central Concepts

3.1.1. Algebraic Chemistry ***(5)***

Let $\mathcal{M}$ be a set of elements (called species, molecular species, or just molecules).

$\mathcal{P}_{\mathcal{M}}$ ($\mathcal{M}$) denotes the set of all multisets with elements from $\mathcal{M}$. A multiset differs from a set in the fact that it can contain the same element more than once. Reactions occuring among the species $\mathcal{M}$ can then be defined by a relation $\mathcal{R}$: $\mathcal{P}_{\mathcal{M}}(\mathcal{M}) \times \mathcal{P}_{\mathcal{M}}(\mathcal{M})$. We call the pair $\langle \mathcal{M}, \mathcal{R} \rangle$ an *algebraic chemistry*.

3.1.2. Closed Set ***(8)***

A set of species $S \subseteq \mathcal{M}$ is *closed*, if for all reactions $(A \to B) \in \mathcal{R}$, with $A \in \mathcal{P}_{\mathcal{M}}(S) \Rightarrow B \in \mathcal{P}_{\mathcal{M}}(S)$.

In other words: if the educts of a reaction are contained in S, then also its products must be in S. There is no reaction that could create any new species not yet in S from species contained in S.

3.1.3. Self-Maintaining Set **(5)**

Given an algebraic chemistry $\langle \mathcal{M}, \mathcal{R} \rangle$ with $m = |\mathcal{M}|$ species and $n = |\mathcal{R}|$ reactions, its dynamics can be described by $\dot{\vec{c}} = \mathbf{M}\vec{v}$ with concentration vector $\vec{c} \in \mathbb{R}_+^m$, stoichiometric matrix $\mathbf{M}$, and flux vector $\vec{v} \in \mathbb{R}_+^n$. A set of species $S \subseteq \mathcal{M}$ is called *self-maintaining* if a flux vector $\vec{v}$ exists, so that the following three conditions are fulfilled:

1. For every reaction $(A \rightarrow B) \in \mathcal{R}$ with $A \in \mathcal{P}_{\mathcal{M}}(S)$, its corresponding flux is $v_{A \rightarrow B} > 0$.
2. For every reaction $(A \rightarrow B) \in \mathcal{R}$ with $A \notin \mathcal{P}_{\mathcal{M}}(S)$, its corresponding flux is $v_{A \rightarrow B} = 0$.
3. For every species $i \in S$, its concentration change is nonnegative: $\dot{c}_i \geq 0$.

In other words: if we consider only the subnetwork made up by the species of S and additionally the species that can be created from S (but are not in S) (**Conditions 1** and **2**), we can find a positive flux vector, such that no species of S decays (**Condition 3**).

3.1.4. Organization **(5,8)**

A set of species $S \subseteq \mathcal{M}$ that is closed and self-maintaining is called an *organization*.

4. Linking Elementary Modes and Organizations

In contrast to organizations, elementary modes are defined as sets of reactions, not species. However, the concept of closure can be expanded to reactions easily. Because we are interested in pathways, we do not consider organizations that contain isolated species not participating in any reaction of the organization. The concept connecting elementary modes and chemical organizations is the self-maintenance property. To elucidate this connection we have to inspect the definition of self-maintenance more closely. Self-maintenance is defined with respect to a set of species. To show the self-maintenance of such a set, a flux vector must exist fulfilling certain conditions. If the solution space of these conditions is empty, the set is not self-maintaining. The union of the solution spaces of all species sets lies within a convex cone in flux space as will be shown in the following. Taking the set $\mathcal{M}$ of all m species of the network and its stoichiometric matrix $\mathbf{M}$ defining the n reactions among these species, the self-maintenance condition $\mathbf{M}\vec{v} \geq \vec{0}$ defines a set of m linear inequalities for the complete network. The restriction to non-negative fluxes $\vec{v} \geq \vec{0}$ defines another set of n inequalities. The solution space of these $m+n$ linear inequalities is a convex cone in the n-dimensional flux space. This cone, encompassing all flux distributions fulfilling the self-maintenance property, can serve as input to an algorithm that computes all organizations. As mentioned in **Subheading 2.**, elementary modes represent the edges of another convex cone: the solution space of the equalities $\mathbf{M}\vec{v} = \vec{0}$ and $\vec{v} \geq \vec{0}$, representing all steady state flux distributions of the system. Apart from flux vectors with zero components, it is obvious that any flux vector $\vec{v}$ fulfilling the steady state condition also fulfills the self-maintenance condition. Hence, the steady state cone lies within the self-maintenance cone except for zero component vectors that are only allowed by the steady state condition.

With the steady state condition being the stricter constraint, in some cases we might find flux vectors fulfilling the self-maintenance property but no flux vector fulfilling the steady state condition. In such a case mass is produced and accumulates in the network. But if there exists a decay reaction of the form $k \rightarrow \emptyset$ for every metabolite $k \in \mathcal{M}$, the overproduction of species can be compensated for by the decay reactions. In such a setting, we find that if a flux vector exists fulfilling the self-maintenance constraint, also a flux vector fulfilling the steady state condition exists as a linear combination of elementary modes.

This leads to the conclusion, that:

1. If all metabolites decay spontaneously, we can find all organizations by using the convex cone that is spanned by the elementary modes.
2. Elementary modes can be used in general reaction systems to search for organizations that

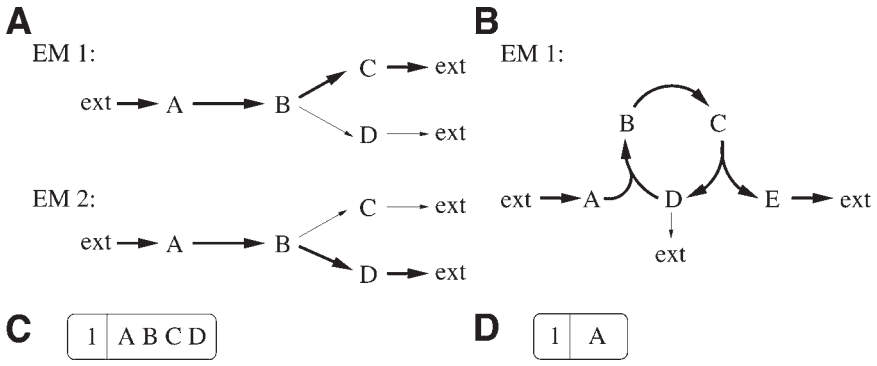


Fig. 1. Elementary modes and organizations in simple linear branching and looping networks. External substrate and product metabolites are denominated with "ext." **(A)** The linear branching pathway contains two elementary modes. The only organization consists of the whole pathway **(C)**. **(B)** The network contains only one elementary mode consisting of all metabolites and all reactions except the decay of **D**. The only organization of this network solely contains **A (D)**.

fulfill the steady state condition. Such steady state organizations are composed of a combination of elementary modes.

3. An elementary mode implies a unique set of organizations. The smallest organizations containing the mode constitute this set. If it is empty, the elementary mode cannot be present in any steady state of the system.
4. Organizations need not to contain elementary modes because they also account for positive productions of metabolites.
5. The set of metabolites taking part in an elementary mode is not necessarily self-maintaining.

5. Examples

5.1. Branching and Cycling Pathways

A simple linear metabolic network with one branching point is depicted in **Fig. 1A**. An external substrate metabolite is transformed into internal metabolite A, which is in turn transformed into B. From B, one path leads over C and another over D to external product metabolites. This networks contains two elementary modes as depicted in **Fig. 1A**. The first elementary mode uses the pathway including metabolite C whereas the second uses metabolite D. Up to the branching point B both modes are identical. Because the sets of metabolites making up the two pathways are not closed (from B it is possible to create C *and* D) they do not form organizations on their own. Indeed, the only organization of this network contains the whole network as seen in **Fig. 1C**. This is an example for a steady state organization composed of two elementary modes (**Subheading 4.**). The smallest organization containing the first elementary mode is the same as for the second; it is the only organization containing the whole network.

The second simple metabolic network shown in **Fig. 1B** features a loop consisting of metabolites B, C, and D. Metabolite A is created from an external substrate and reacts with D to form B, which is transformed to C. Finally, C is transformed into D and E. Both metabolites D and E are transformed into external product metabolites. Only one elementary mode exists in this network. It contains all metabolites and all reactions except the transformation of D into an external metabolite. Although the set of all metabolites is closed, it is not an organization. The set is not self-maintaining because within the loop, D is transformed into an external metabolite, leaving the network (**Subheading 4.**). Metabolite D is required to keep the loop running, but no reaction compensates for the outflow of D. We find that the set only containing A is the only organization in this network as seen in **Fig. 1D**. Because A is accumulating in the organizational reaction network

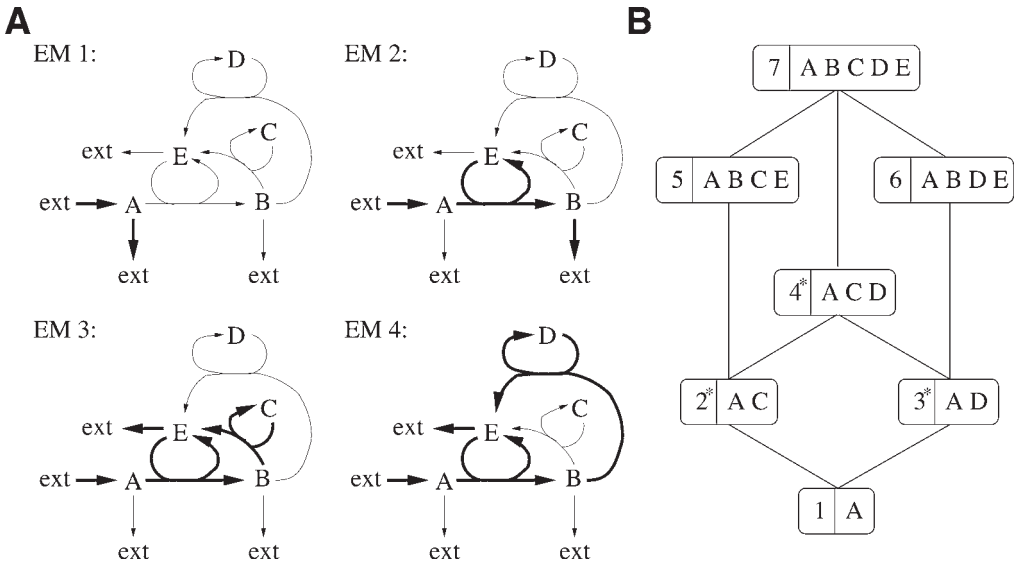


Fig. 2. Comparing elementary modes with organizations in a more complex reaction network with five species. **(A)** The network contains four elementary modes. **(B)** The hierarchy of organizations consists of seven organizations. The starred organizations (organization 2, 3, and 4) contain isolated species that do not react with each other. *See* **Subheading 5.2.** for details.

(just consisting of A and its creation reaction), it is not a steady state organization consisting of elementary modes (**Subheading 4.**). Here, we find that there is no organization containing the elementary mode. Consequently, this mode cannot be present in a steady state of the network (**Subheading 4.**).

5.2. Pathways with Catalysts

A more complex metabolic reaction network is shown in **Fig. 2A**. An external substrate is transformed into metabolite A. With metabolite E acting as a catalyzing enzyme, A can react to form B. Then, B can be transformed into E via two reactions. One is catalyzed by metabolite C, whereas the other by metabolite D. The metabolites A, B, and E are transformed into external product metabolites. Note that in general, each reaction in a metabolic network is implicitely catalyzed by an enzyme. In this example, three metabolites are explicitely modeled as catalysts. The network contains four elementary modes as depicted in **Fig. 2A**. The first mode just uses metabolite A to transform the external substrate metabolite into an external product metabolite. In the second mode, A is transformed into B with the help of enzyme metabolite E, and B is transformed into the external product. The third mode also transforms A to B using E as a catalyst. But additionally, C acts as a second catalyst to transform B into E. Finally, E is transformed into an external product. The fourth mode is similar to the third one with the exception that here, the reaction catalyzed by D is used to transform B into E.

The hierarchy of organizations is shown in **Fig. 2B**. The network contains seven organizations. The smallest one just contains metabolite A. This organization coincides with the first elementary mode, it is a steady state organization. The three organizations above the first organization (2, 3, and 4) all have in common that they contain species that do not participate in any reaction of the organization. As mentioned in **Subheading 4.**, we are not concerned with such organizations here. Organizations 5 and 6 contain exactly the metabolites making up elementary modes 3 and 4. Both are steady state organizations. Here, the sets of metabolites making up the elementary modes are already closed, and hence the smallest organizations containing the modes are already

Table 1
Organizations and Corresponding Elementary Modes

Organization	EMs implying org	Contained EMs
7	–	EM 1, EM 2, EM 3, EM 4
6	EM 2, EM 4	EM 1, EM 2, EM 4
5	EM 2, EM 3	EM 1, EM 2, EM 3
4	–	EM 1
3	–	EM 1
2	–	EM 1
1	EM 1	EM 1

[a]The organization is the smallest enclosing one (col. 2), and all elementary modes contained in the organization (col. 3) for the example network with catalysts.

the very same sets. The smallest organization containing elementary mode 2 is not unique in this example: both organizations 5 and 6 contain the mode and are of equal size (**Subheading 4.**). Such an elementary mode can exist in different steady state network configurations and hence might be of particular importance. The largest organization 7 comprises the whole network. It is a steady state organization combining elementary modes 3 and 4. **Table 1** summarizes the relationship between elementary modes, the smallest organizations containing them, and all modes contained in the organizations.

5.3. Central Sugar Metabolism of E. coli

As a more realistic example we finally analyze a reaction network introduced by Puchalka and Kierzek ***(9)*** modeling the sugar metabolism of *E. coli* including gene expression, signal transduction, transport processes, and enzymatic activities. The main feature of this network is the possibility to model cell growth on different carbon sources, namely glucose, lactose, and glycerol. If different sugars are available in the growth medium, *E. coli* exhibits a diauxic growth behavior by first utilizing exclusively glucose as the single carbon source. Only after depletion of glucose, lactose and to a lesser degree glycerol are metabolized. We concentrate on the case where all carbon sources are available in the medium. Modelling inducers and activators as required for gene transcription, this network has four organizations as discussed in **ref. *10*** in detail. **Figure 3** depicts the hierarchy of organizations. The whole network constitutes an organization, and the three remaining organizations are associated with the uptake of glucose, lactose, and glycerol, respectively.

Computing the elementary modes of the network revealed 550 modes. Determining the smallest organization containing each mode, they can be assigned to the organizations as shown in **Table 2**. Grouping elementary modes according to their enclosing organization helps to deal with the vast number of elementary modes usually found in large networks. In this example, the organizations can give a first hint on the physiological function an elementary mode plays a role in. Whereas organizations specify species compositions required for physiological steady states (or states with increasing species concentrations), the elementary modes within the organizations define the admissible flux distributions for the corresponding state.

6. Conclusion

Combining elementary mode analysis with organization analysis gives a more complete picture of the potential dynamical behavior of metabolic networks. On the one hand, elementary modes represent pathways that can operate at steady state. Because the metabolite set associated with a mode needs not to be closed, single elementary modes are not expected to be observed

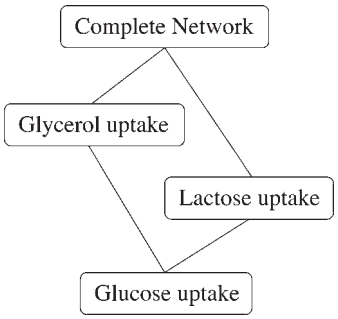


Fig. 3. Hierarchy of organizations for the network modeling the central sugar metabolism of *Escherichia coli*. The lower three organizations are associated with the uptake of glucose, lactose, and glycerol, respectively. The largest organization contains the whole network.

Table 2
Organizations and Corresponding Numbers of Elementary Modes

Organization	EMs implying org	Contained EMs
Complete Network	16	550
Glycerol uptake	37	209
Lactose uptake	325	497
Glucose uptake	172	172

[a]The organization is the smallest enclosing one (col. 2), and number of total elementary modes contained in the organization (col. 3) for the network modelling the central sugar metabolism of *Escherichia coli*.

in feasible system states. Organizations, on the other hand, specify metabolite combinations that are likely to be observed in feasible system states, taking a more global perspective on the system. Such a state can be a steady state or a state in which species have positive production rates. With elementary modes defining the boundaries of the potential steady state behavior of the metabolic network and steady state organizations representing all metabolite compositions enabling feasible steady states, both concepts complement each other. Organizations help to identify all potential steady state metabolite combinations and then elementary modes help to define all admissible steady state flux distributions within the organizational network. And taking the opposite direction, classifying elementary modes according to their enclosing organizations helps to deal with the typical vast number of elementary modes. With organizations also allowing for positive metabolite productions, growth situations (e.g., bacterial growth) can additionally be considered.

References

1. Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420,** 190–193.
2. Schuster, S., Klamt, S., Weckwerth, W., and Pfeiffer, T. (2002) Use of network analysis of metabolic systems in bioengineering. *Bioproc. Biosyst. Eng.* **24,** 363–372.
3. Schuster, S., Klamt, S., Weckwerth, W., and Pfeiffer, T. (2005) Adenine and adenosine salvage pathways in erythrocytes and the role of s-adenosylhomocysteine hydrolase. *FEBS J.* **20,** 5278–5290.
4. Gagneur, J. and Klamt, S. (2004) Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics* **5,** 175.
5. Dittrich, P. and Speroni di Fenizio, P. (2005) Chemical organization theory. *Bull. Math Bid.*, in press.
6. Schuster, S., Fell, D., 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.
7. Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I., and Dandekar, T. (2000) Structural analysis of metabolic networks: elementary flux modes, analogy to petri nets, and application to mycoplasma pneumoniae, in *Proc. Germ. Conf. Bioinf.,* (Bauer, E. -B., Rost, U., Stoye, J., and Vingron, M., eds.), Heidelberg, Germany, pp. 115–120.
8. Fontana, W. and Buss, L. W. (1994) "The arrival of the fittest": towards a theory of biological organization. *B. Math. Biol.* **56,** 1–64.
9. Puchalka, J. and Kierzek, A. M. (2004) Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *Biophys. J.* **86,** 1357–1372.
10. Centler, F., Speroni di Fenizio, P., Matsumaru, N., and Dittrich, P. (2006) Chemical organizations in the central sugar metabolism of *Escherichia Coli. Modeling and Simulation in Science, Engineering and Technology* (accepted).

F. Centler, C. Kaleta, P. S. di Fenizio, and P. Dittrich. *Bioinformatics*, 24(14):1611-1618, Jul 2008.

BIOINFORMATICS **ORIGINAL PAPER**

Vol. 24 no. 14 2008, pages 1611–1618
doi:10.1093/bioinformatics/btn228

Systems biology

Computing chemical organizations in biological networks

Florian Centler[†], Christoph Kaleta, Pietro Speroni di Fenizio[‡] and Peter Dittrich*

Bio Systems Analysis Group, Jena Centre for Bioinformatics (JCB) and Department of Mathematics and Computer Science, Friedrich-Schiller-University Jena, D-07743 Jena, Germany

Received on January 9, 2008; revised on March 6, 2008; accepted on May 7, 2008

Advance Access publication May 14, 2008

Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Novel techniques are required to analyze computational models of intracellular processes as they increase steadily in size and complexity. The theory of chemical organizations has recently been introduced as such a technique that links the topology of biochemical reaction network models to their dynamical repertoire. The network is decomposed into algebraically closed and self-maintaining subnetworks called organizations. They form a hierarchy representing all feasible system states including all steady states.

Results: We present three algorithms to compute the hierarchy of organizations for network models provided in SBML format. Two of them compute the complete organization hierarchy, while the third one uses heuristics to obtain a subset of all organizations for large models. While the constructive approach computes the hierarchy starting from the smallest organization in a bottom-up fashion, the flux-based approach employs self-maintaining flux distributions to determine organizations. A runtime comparison on 16 different network models of natural systems showed that none of the two exhaustive algorithms is superior in all cases. Studying a 'genome-scale' network model with 762 species and 1193 reactions, we demonstrate how the organization hierarchy helps to uncover the model structure and allows to evaluate the model's quality, for example by detecting components and subsystems of the model whose maintenance is not explained by the model.

Availability: All data and a Java implementation that plugs into the Systems Biology Workbench is available from http://www.minet.uni-jena.de/csb/prj/ot/tools.

Contact: dittrich@minet.uni-jena.de

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

1 INTRODUCTION

Biological knowledge of the myriad of processes taking place in living cells is typically represented in network models. Such models detail, how genes regulate each other's activity (de Jong, 2002), how signals detected at the cell surface are propagated within the cell leading to a change in gene expression (Papin *et al.*, 2005), and how external molecules are internalized and interconverted into different metabolites for energy production (Heinrich and Schuster, 1996). As biologists acquire more and more knowledge, these network models not only increase in size, but also first steps are made towards comprehensive whole cell models (Tomita *et al.*, 1999). With classic analysis methods reaching their limits for such networks, the theory of chemical organizations (Dittrich and Speroni di Fenizio, 2007) has been proposed as a novel analysis technique. The main aim is to identify subnetworks that do not create any species outside the subnetwork and that are able to maintain themselves over time. The method cannot only be applied to network models in which the set of species is fixed, but also to constructive dynamical systems in which novel species can appear as the system evolves in time. Signaling systems are an example for such systems as the combinatorial complexity of molecular interactions leads to a vast number of potential species (Blinov *et al.*, 2004). The method has been applied to study a photochemical model of the Martian atmosphere (Centler and Dittrich, 2007), models of HIV infection (Matsumaru *et al.*, 2006a), and a model of *Escherichia coli* (Centler *et al.*, 2007). Furthermore, the evolution of an artificial chemistry was investigated (Matsumaru *et al.*, 2006b) and chemical computing systems have been constructed with the help of organization theory (Matsumaru *et al.*, 2007). In this article, we lay the algorithmic foundation for computing chemical organizations.

Another topology-based method to study metabolic networks is the analysis using elementary modes (Schuster *et al.*, 1999). In contrast to organization theory, only steady states are considered in this approach. However, both methods are closely related and complement each other (Kaleta *et al.*, 2006). Elementary modes are minimal pathways that can operate at steady state. A biologically feasible system state usually comprises a combination of several elementary modes. As organizations embody such feasible system states, elementary modes can be grouped according to the organizations within they are embedded. This allows for the identification of elementary modes that act in concert, possibly providing a common functionality. Furthermore, this classification eases the analysis of the usually vast number of elementary modes (Kaleta *et al.*, 2006).

2 THEORY OF CHEMICAL ORGANIZATIONS

Extending ideas by Fontana and Buss (1994), the theory of chemical organizations (Dittrich and Speroni di Fenizio, 2007) provides a new method to analyze complex reaction networks. A reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$ is defined by a set of reactions $\mathcal{R} \subseteq \mathcal{P}_M(\mathcal{M}) \times \mathcal{P}_M(\mathcal{M})$ occurring among molecular species $\mathcal{M}$,

*To whom correspondence should be addressed.

[†] Present address: Department of Environmental Microbiology, UFZ - Centre for Environmental Research Leipzig-Halle, Germany.

[‡] Present address: Research Institute in Networks & Communications Engineering (RINCE), Dublin City University, Dublin 9, Ireland.

with $\mathcal{P}_M(\mathcal{M})$ denoting the set of all multisets with elements from $\mathcal{M}$. $\langle \mathcal{M}, \mathcal{R} \rangle$ implies an $m \times n$ stoichiometric matrix $\mathbf{S} = (s_{i,j})$ where $s_{i,j}$ is the number of molecules of species $i \in \mathcal{M}$ that is produced in reaction $j \in \mathcal{R}$ (i.e. right hand side minus left hand side); $m = |\mathcal{M}|$, $n = |\mathcal{R}|$.

At its core, chemical organization theory defines an organization $O \subseteq \mathcal{M}$ as a set of molecular species that is (algebraically) closed and self-maintaining. Closure ensures that given the species of an organization, there is no reaction within the reaction network that can create a novel species not yet present in the organization. Formally, $O \subseteq \mathcal{M}$ is closed, if for all reactions $(A \to B) \in \mathcal{R}$ with $A \in \mathcal{P}_M(O)$, then also $B \in \mathcal{P}_M(O)$.

Self-maintenance guarantees that all molecular species that are consumed within the organization O can be recreated simultaneously from organization species at a sufficient rate for their maintenance. A set of species $O \subseteq \mathcal{M}$ is called self-maintaining if a flux vector $v = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n_{\geq 0}$ exists such that the following three conditions are fulfilled: (1) For every reaction $(A \to B) \in \mathcal{R}$ with $A \in \mathcal{P}_M(O)$, its corresponding flux is $v_{A \to B} > 0$. (2) For every reaction $(A \to B) \in \mathcal{R}$ with $A \notin \mathcal{P}_M(O)$, its corresponding flux is $v_{A \to B} = 0$. (3) For every species $i \in O$, its concentration change is non-negative: $(\mathbf{S}v)_i \geq 0$. In other words: if we consider only the subnetwork made up by the species of O [conditions (1) and (2)], we can find a positive flux vector, such that no species of O decays [condition (3)].

The set of organizations, together with the set inclusion $\subseteq$, forms a partially ordered set that can be visualized as a hierarchical structure in a Hasse diagram (see Fig. 3 for an example). Linking to the dynamics, Dittrich and Speroni di Fenizio (2007) have shown that all steady states of the system are instances of organizations, given that the dynamics of the system is described by a set of ordinary differential equations.

Besides these fundamental concepts, the following definitions proved helpful. A set of species $S \subseteq \mathcal{M}$ is called semi-self-maintaining, if all species $s \in S$ that are consumed within the reaction network implied by S are also produced within the same network. This reaction network implied by S consists of all reactions $(A \to B) \in \mathcal{R}$ with $A \in \mathcal{P}_M(S)$. A species $s \in S$ is consumed (produced) if there exists a reaction $(A \to B) \in \mathcal{R}$ with $A \in \mathcal{P}_M(S)$ for which s is contained more times in A than in B (more times in B than in A). Note that each self-maintaining set is semi-self-maintaining (Dittrich and Speroni di Fenizio, 2007). A set of species $S \subseteq \mathcal{M}$ that is closed and semi-self-maintaining is called a semi-organization. A (semi-) organization $O \subseteq \mathcal{M}$ is a connected (semi-) organization, if it is empty, or there exists a species $s \in O$ so that all species of O are connected to s. Two species s_i and $s_j \in O$ are connected to each other, if there exists a sequence of n species $s_1, \ldots, s_n$ with $s_k \in O$ for $k = 1, \ldots, n$, such that s_i and s_1, s_k and s_{k+1} for $k = 1, \ldots, n-1$, and s_n and s_j are directly connected. Two species s_o and $s_p \in O$ are directly connected, if there exists a reaction $(A \to B) \in \mathcal{R}$ with $A \in \mathcal{P}_M(O)$, $s_o \in A \cup B$, and $s_p \in A \cup B$. All input species of the network (i.e. all species that appear as products in reactions with empty reactant side) are defined as being connected to each other. Given a set of organizations $\mathcal{O}$, an organization $O \in \mathcal{O}$ with its corresponding set of reactions $R(O) := \{(A \to B) \in \mathcal{R} \mid A \in \mathcal{P}_M(O) \text{ and } B \in \mathcal{P}_M(O)\}$ is an elementary organization, if there exist no subset of organizations $\mathcal{P} \subseteq \mathcal{O} \setminus O$ such that $R(O) = \bigcup_{P \in \mathcal{P}} R(P)$. In other words, when viewing organizations as sets of reactions, an elementary organization cannot be represented as the union of other organizations. An organization O is called a reactive organization if all species it contains are educt or product in at least one reaction of the set of reactions $R(O)$ that occur within O. Note that all connected organizations with more than one species are also reactive organizations. The opposite is not true, as a reactive organization may contain several connected subnetworks that are disparate.

3 ALGORITHM I: CONSTRUCTIVE APPROACH

To compute the organizations for a given reaction network, one could simply test all species combinations for the properties of closure and self-maintenance in a brute force fashion. However, this approach is only feasible for networks with few species (i.e. less than 30 species) as the number of sets to test equals 2^n, with n being the number of network species. The first more elaborate algorithm to compute organizations determines in a first step all semi-organizations of the reaction network. This is done in a recursive manner: given an already determined semi-organization *so*, the semi-organizations containing *so* are computed in the next step. To find a semi-organization above *so*, the network structure is taken into account to select species that, when added to *so*, are likely to give rise to a larger semi-organization. In this constructive fashion, the hierarchy of semi-organizations is computed from bottom-up. As all organizations are also semi-organizations, the first step delivers all candidate species sets for organizations. Then, in the second step, all semi-organizations have to be identified that are also organizations. The property of self-maintenance is the property distinguishing both organization types. All semi-organizations, for which it can be shown that a flux vector in accordance with the self-maintenance condition exists, are also organizations. This requires to solve a linear programming problem for each semi-organization.

3.1 Step 1: computing semi-organizations

The algorithm starts with the smallest semi-organization and creates the whole hierarchy of semi-organizations from the bottom-up.

3.1.1 The main loop. The main function `computeSemiOrganizations()` is depicted as pseudo code in Figure 1. The set SOsToCheck contains all semi-organizations that have already been found and that still have to be processed. It is initialized with the smallest semi-organization computed by `closure({})`. The function `closure(`*set*`)` computes the smallest closed set that contains the species set set. This is done by iteratively adding all species to set that, according to the reaction rules, can be produced from the species in set. Each addition might enable new reactions. The iteration stops as soon as no new novel species can be created. Taking the closure of the empty set delivers the smallest semi-organization of the network. If the network contains no input species, it is the empty set. If input species are present, the smallest semi-organization contains all input species and their closure.

The central function of the main loop is `SOsDirectlyAbove(so)` (see Fig. 1 for pseudo code). It delivers the set of all semi-organizations that are directly above semi-organization so, that means all semi-organizations that contain so and that do not contain

Function `computeSemiOrganizations`
Input: reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$
Output: set of all semi-organizations in result
result $\leftarrow \emptyset$; SOsToCheck $\leftarrow$ { `closure( {} )` };
while *SOsToCheck* $\neq \emptyset$ **do**
  current $\leftarrow$ `getSmallestSO(`*SOsToCheck*`)`;
  SOsToCheck $\leftarrow$ SOsToCheck $\cup$ `SOsDirectlyAbove(`*current*`)`;
  SOsToCheck $\leftarrow$ SOsToCheck $\setminus$ { current };
  result $\leftarrow$ result $\cup$ { current };
end

Function `SOsDirectlyAbove`
Input: semi-organization so, reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$
Output: set of all semi-orgs. directly above so in result
result $\leftarrow \emptyset$; usableSpecies $\leftarrow \mathcal{M} \setminus$ so ;
foreach *s* $\in$ *usableSpecies* **do**
  result $\leftarrow$ result $\cup$ `SOsDirectlyAbove-Containing(`*so*, { *s* }`)`;
end

Function `SOsDirectlyAboveContaining`
Input: semi-organization so, species set species to be contained in new semi-orgs., reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$
Output: set of all semi-organizations directly above so that contain species in result
result $\leftarrow \emptyset$; closure $\leftarrow$ `closure(`*so* $\cup$ *species*`)`;
if *closure is semi-self-maintaining* **then**
  result $\leftarrow$ { closure } ;
else
  spcsToProduce $\leftarrow$ {$s \in$ closure | s is consumed but not produced in closure } ;
  producingSets $\leftarrow$ `producerSets(`*spcsToProduce*`)`;
  foreach *set* $\in$ *producingSets* **do**
    result $\leftarrow$ result $\cup$ `SOsDirectlyAbove-Containing(`*so, species* $\cup$ *set*`)`;
  end
end

Fig. 1. The constructive approach to compute semi-organizations.

any other semi-organization that contains so[1]. In the Hasse diagram, these are exactly those larger semi-organizations to which so has a connection. In each iteration of the main loop, the smallest semi-organization[2] current is taken from SOsToCheck. All semi-organizations returned by `SOsDirectlyAbove(`*current*`)` are then added to SOsToCheck, and current is removed from the set. The iteration stops when no semi-organization is left in SOsToCheck. In order to avoid processing the same set of species twice, a hash structure can be used to keep track of already processed sets (not shown in pseudo code).

3.1.2 Functions for computing semi-organizations. In order to find a larger semi-organization that contains semi-organization so, it is checked in function `SOsDirectlyAbove(`*so*`)` for each species not in so, whether this species is part of such a larger semi-organization.

The function `SOsDirectlyAboveContaining(`*so*, *speciesSet*`)` computes all semi-organizations that are directly above so and contain the species in speciesSet (see Fig. 1 for pseudo code). First, the closure of the union of so and speciesSet is computed. If the computed closure is semi-self-maintaining, a semi-organization with the desired properties is found and the function returns. If not, those species in the closure are identified that are consumed but not produced. In order to become a semi-organization, these species must be produced somehow. The function `producerSets(`*speciesSet*`)` (see Supplementary Material for pseudo code) returns all possible species combinations that can produce all species in speciesSet. For each such combination, `SOsDirectlyAboveContaining()` is recursively called again. This time, the producer combination is additionally required to be present in the new semi-organizations.

3.2 Step 2: test for self-maintenance

Every semi-organization determined in the first step of the algorithm is a candidate for an organization. For a semi-organization to also be an organization, it must be shown that a flux vector exists in accordance with the self-maintenance condition. Let O be a semi-organization with $m' = |O|$ species implying n' reactions (i.e. all reactants of n' reactions are contained in O). With $\mathbf{S}_O$ being the stoichiometric matrix of this subnetwork, we must determine the existence of a flux vector $v \in \mathbb{R}^{n'}_{\geq 0}$ with $v > \mathbf{0}$ and $\mathbf{S}_O \cdot v \geq \mathbf{0}$. With the solution space of these inequalities forming a convex polyhedral cone in the positive orthant, originating in the point of origin, the problem is equivalent to finding a flux vector v with $v > \mathbf{1}$ and $\mathbf{S}_O \cdot v \geq \mathbf{0}$. This is a linear programming problem that can be solved using the simplex method (Dantzig, 1963). Since only the existence of such a flux vector v is of concern, only the first phase of the simplex method needs to be performed.

3.3 Connected semi-organizations

A simple combinatorial explosion can lead to a vast number of chemical organizations. Consider the reaction network $\langle \{s_1, \ldots, s_n\}, \{\} \rangle$. The network does not contain any reaction. Consequently, all 2^n species combinations are closed and self-maintaining, and hence organizations. However, the organizations contain only isolated network species that do not react with each other. In order to avoid this combinatorial complexity, it is adequate to only consider connected organizations (*cf.* Section 2). In such organizations, all contained species form a single subnetwork in which all species are connected to each other via reactions. In order to compute connected semi-organizations, only the function `SOsDirectlyAbove()` needs to be modified. Now, only those species are added to an already discovered semi-organization

[1] More precisely, *at least* those semi-organizations are delivered. Under certain circumstances, additional semi-organizations are returned that are not *directly* above semi-organization so but rather contain another semi-organization directly above so. Consider for example the reaction network $\langle \{a,b\}, \{a \rightarrow a+b\} \rangle$. The system contains three organizations: $\{a,b\}$ above $\{b\}$ above $\{\}$. Applied to the empty set, the function `SOsDirectlyAbove({})` returns both $\{b\}$ and $\{a,b\}$ here, since function `SOsDirectlyAboveContaining()` first creates the closure of its argument (see Section 3.1.2). For argument $\{a\}$, the closure is already $\{a,b\}$ and hence this semi-organization is returned although not being *directly* above the semi-organization $\{\}$.

[2] If the smallest semi-organization is not unique, the choice is random.

that are directly connected to a member species of that semi-organization. The Supplementary Material contains the pseudo code. The hierarchy of connected organizations can also be used to compute the full organization hierarchy as detailed in the Supplementary Material.

4 ALGORTIHM II: FLUX-BASED APPROACH

The second algorithm takes a somehow opposite strategy as Algorithm I by combining self-maintaining flux distributions until closed sets are discovered. While the first approach operates on species, the flux-based approach operates on reactions. First, the flux based approach computes elementary organizations (Step 1 and 2), which are combined in Step 3 to obtain reactive organizations, that is, organizations where each species participates at least in one reaction within the organization. Organizations containing species that do not react are determined in the final Step 4.

Starting with the condition of self-maintenance, methods from convex analysis can be employed. Given a reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$ and its $m \times n$ stoichiometric matrix $\mathbf{S}$, a flux vector $v \in \mathbb{R}^n_{\geq 0}$ fulfilling the self-maintenance condition must be found to show that a species set is self-maintaining. All such flux vectors lie in a convex polyhedral cone $\mathcal{P}$ in the n-dimensional flux space $\mathbb{R}^n_{\geq 0}$, originating in the point of origin. The cone is defined by the $n+m$ inequalities $v \geq \mathbf{0}$ and $\mathbf{S} \cdot v \geq \mathbf{0}$. These constraints can be transformed into a matrix $\mathbf{A}$ representing the set of spanning vectors V_B or extreme rays of $\mathcal{P}$ (Gagneur and Klamt, 2004). Each point within $\mathcal{P}$ can be written as a linear combination of these extreme rays. Thus, we can compute organizations by searching for combinations of extreme rays whose corresponding set of species fulfills the closure condition.

Given a flux vector v, Algorithm II relies only on the set of reactions that have positive fluxes in v and not on their specific flux values. Hence, we define v^{set} as the set of reaction indices containing all reactions that have positive fluxes in v. Considering the set of extreme rays V_B defining $\mathcal{P}$, V_B^{set} describes the set of reaction sets v_B^{set} corresponding to the extreme rays $v_B \in V_B$. The species set that corresponds to a reaction set v^{set} is denoted by $M(v^{set})$. It contains all reactants and products of the reactions contained in v^{set}.

Step 1: computing extreme rays—To compute the set of extreme rays V_B required for Step 2, we implemented the well-known Schuster algorithm (Schuster *et al.*, 1999) to compute elementary modes. This algorithm computes the extreme rays of a convex cone $\mathcal{P}'$ defined by $v \geq \mathbf{0}$ and $\mathbf{S} \cdot v = \mathbf{0}$. By adding an outflux for each metabolite to the stoichiometric matrix used for the elementary mode computation, the algorithm can also compute the extreme rays of the cone $\mathcal{P}$.

Step 2: computing elementary organizations—The central function in the computation of elementary organizations is `organizationsAbove()` (see Fig. 2 for its pseudo code). Given a self-maintenance flux vector $v \in \mathcal{P}$ and its corresponding reaction set v^{set}, it computes all reactive organizations O that contain $M(v^{set})$ and for which there exists no other organization being a subset of O and a superset of $M(v^{set})$.[3] Hence, the smallest organizations containing $M(v^{set})$ are computed.

[3]More precisely, *at least* those organizations are computed. Under certain circumstances, organizations are also in the result set for which a subset O_s is also an organization and contains $M(v^{set})$. However, in such cases O_s is also contained in the resulting set of organizations.

Function `organizationsAbove`

Input: reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$, set of reaction sets V_B^{set} corresponding to the extreme rays spanning the self-maintenance cone $\mathcal{P}$, reaction set v^{set} corresponding to a self-maintenance flux vector $v \in \mathcal{P}$

Output: set of the smallest reactive organizations containing the reactions of v^{set} in result

```
result ← ∅ ; v^set_Closure ← closure(v^set) ;
if v^set_Closure == v^set then
    result ← { M(v^set) };
else
    select one reaction r with r ∈ (v^set_Closure \ v^set);
    foreach v^set_B ∈ V^set_B with r ∈ v^set_B do
        result = result ∪ organizationsAbove(v^set ∪ v^set_B) ;
    end
end
```

Fig. 2. The central recursive function of the flux-based approach to compute organizations.

First, the closure of the reaction set v^{set}, respectively $M(v^{set})$, is computed. This is done by taking the species set $M(v^{set})$ and iteratively adding all species to the set that can be created by reactions of the network from the species set. The reaction set $v^{set}_{Closure}$ contains all reactions that can take place in the generated closed set of species. If this reaction set is identical to v^{set}, the species set $M(v^{set})$ is an organization. The reaction set $v^{set}_{Closure}$ contains more reactions than v^{set} if either species were added, or $M(v^{set})$ is closed but v^{set} does not contain all reactions that are possible in this set. One such reaction is taken, and all reaction sets $v_B^{set} \in V_B^{set}$ that contain this reaction are consecutively combined with the original reaction set v^{set} and the function is called again recursively. As the initial reaction set v^{set} and the extreme ray reaction sets v_B^{set} correspond to flux vectors fulfilling the self-maintenance condition, also a flux vector v_u fulfilling the self-maintenance property exists for the union $v_u^{set} = v^{set} \cup v_B^{set}$. Hence, all reaction sets that are considered in the recursive function calls are associated with self-maintaining flux vectors. To obtain all elementary organizations, the function `organizationsAbove()` is called for each reaction set $v_B^{set} \in V_B^{set}$ corresponding to one of the extreme rays defining $\mathcal{P}$.

Step 3: computing reactive organizations—Elementary organizations are combined to determine all reactive organizations. This is done by taking all possible combinations of two elementary organizations and calling `organizationsAbove()` for the union of their reaction sets. For every newly discovered organization, this organization must again be combined with each of the elementary organizations, and `organizationsAbove()` must be called again for the reaction set unions.

Step 4: computing all organizations—The organizations we have obtained so far all possess a different set of reactions. Consequently, the final step consists of searching for organizations having the same set of reactions as already discovered ones, but containing different species sets. Hence, we need to determine for all discovered organizations all species sets, that can be added to the organization without changing its set of reactions. This can be done by simply inspecting the reaction list.

5 ALGORITHM III: HEURISTIC APPROACH

The flux-based approach requires the computation of the extreme rays of the cone $\mathcal{P}$ as input. However, instead of starting with the set of all extreme rays V_B, it is feasible to skip the first step of the algorithm by directly starting with a set of elementary organizations. To obtain such a set, a simple heuristic approach is used. A random walk through the reaction network delivers a set of species. After computing the closure of this species set, it is tested whether the closure also fulfills the self-maintenance condition (see Section 3.2). After having determined a sufficient large set of organizations, the associated set of elementary organizations O_{el} is determined. The reaction sets of the organizations in O_{el} are then used as input to the second step of the extreme ray algorithm to compute the complete set of organizations that can be found by combining organizations from O_{el}. The heuristic approach was able to correctly determine the whole set of organizations for all tested networks to which we could also apply the regular flux-based method.

6 IMPLEMENTATION AND PERFORMANCE

The algorithms have been implemented in Java as a command line tool and as an application that integrates into the Systems Biology Workbench (Sauro *et al.*, 2003). Networks can be provided in SBML format (Finney and Hucka, 2003), or in a human friendlier proprietary format. SBML files are read using a Java-based SBML parser from the JigCell project (Vass *et al.*, 2006). The computed hierarchy of organizations is stored in XML format (OrgML) and can be visualized and interactively explored using the accompanying Java application *OrgView*. To solve the linear programming problem for checking for self-maintenance in the constructive approach, we use *lpsolve* (Berkelaar *et al.*, 2005). In order to study the runtime behavior of the deterministic Algorithms I and II, we compute the organizations for 16 different network models of natural systems including a series of 10 networks that increase in size in an iterative fashion created using BioNetGen (Blinov *et al.*, 2004).

The results are shown in Table 1. The complete hierarchy of connected organizations using Algorithm I could be determined for all networks. After 24 h of computing time, Algorithm I did not terminate for the Mars nightside network and networks EGFR 5–9. Algorithm II did not terminate for the Mars night- and dayside networks and networks EGFR 5–10. Generally, computing the hierarchy of connected organizations is much faster than determining the complete hierarchy of organizations. Comparing the runtime of Algorithms I and II shows that no method is superior in all cases. Algorithm II shows weak performance when the number of extreme rays is large, as in the full martian photochemistry models, which contain a large number of cycles. When the extreme rays can be computed quickly, Algorithm II tends to be faster (e.g. EGFR 4). Although there is a tendency that larger networks require longer computation time, the size of the network alone is only a limited indicator for the (expected) runtime. Adding a single reaction can change the number of organizations and the runtime dramatically (e.g. Mars, nightside versus Mars, dayside, Table 1). Additionally, larger networks not necessarily contain more organizations. For example, the largest network EGFR 10 has only 32 organizations, while EGFR 4, a subnetwork of the former, contains 12 288 organizations. However, computing these many organizations using Algorithm I is almost twice as fast as computing the mere 32 organizations of the larger network.

Table 1. Number of regular, connected and elementary organizations and extreme rays in network models, and required runtime of the constructive (Alg. I) and flux-based (Alg. II) approach

Model	Species/ reacts.	Organizations (Alg.I / Alg. II)	Connected orgs. (Alg. I)	Elem. orgs.	Extr. rays
Dry Mars, night	7/15	22 (1s/1s)	7 (1s)	1	0
Dry Mars, day	7/16	6 (1s/1s)	3 (1s)	3	24
Mars, night	31/103	1 088 640 (9.2d/−)	27 (2s)	1	–
Mars, day	31/104	1484 (50s/-)	8 (1s)	7	–
Lambda phage	55/81	7 (1s/2s)	2 (1s)	5	72
Central *E.coli*	92/168	4 (2s/3s)	1 (1s)	1	103
EGFR 1	7/3	48 (2s/1s)	8 (1s)	2	1
EGFR 2	8/5	48 (2s/1s)	8 (1s)	2	2
EGFR 3	10/8	96 (4s/1s)	9 (1s)	2	3
EGFR 4	19/21	12 288 (479s/13s)	16 (1s)	2	5
EGFR 5	37/76	n.a. (−/−)	25 (2s)	2	170
EGFR 6	69/213	n.a. (−/−)	39 (3s)	2*	–
EGFR 7	131/494	n.a. (−/−)	70 (8s)	3*	–
EGFR 8	295/1260	n.a. (−/−)	174 (122s)	5*	–
EGFR 9	356/3088	n.a. (−/−)	72 (863s)	6*	–
EGFR 10	356/3749	32 (831s/−)	13 (1278s)	8*	–

The runtime is computed as the sum of user and system time reported by `memtime` of the constructive and flux-based approach, using a Pentium 4 processor at 1.8 GHz and 1 GB RAM. The four Mars models are taken from Centler and Dittrich (2007), the lambda network was derived from a petri net model by Doi (2005), the *E.coli* model is taken from Centler *et al.* (2007), and the model EGFR 10 of epidermal growth factor receptor (EGFR) is taken from Blinov *et al.* (2006); EGFR 1–EFGR 9 represent steps of smaller networks in the iterative network generation process of BioNetGen leading to EFGR 10. In cases marked '–', the algorithm did not terminate within 24 h. Runtime is rounded up to seconds and also includes the computation of the hierarchical relations between organizations necessary for their representation in a Hasse diagram. In cases marked '*', the heuristic approach was used to estimate the number of elementary organizations. In all these cases, all reactive connected organizations were found. They were computed within the first 10 min of a total runtime of 24 h. In all networks except for Mars, dayside and Lambda phage, the elementary organizations are identical to the reactive connected organizations. For the exceptions, all connected organizations are reactive. n.a., data not available.

6.1 Performance analysis

The parameters network size, number of organizations, and required runtime are only weakly correlated. The main factor determining both the number of organizations and runtime is not the network size, but the specific network topology. For any given number of components it is possible to create a reaction network that either does not contain any semi-organization or in which any species combination forms a semi-organization. This is true even for connected semi-organizations, as detailed in the Supplementary Material. One key element of the constructive algorithm is to find larger semi-organizations by recursivly adding species to semi-organizations that already have been computed. Instead of trying to add species in all possible combinations in brute-force manner, this process is guided by the network topology. For a given species set containing species that are not yet produced from the set, such species are added that enable reactions that lead to the production of the former species, in order to create a semi-self-maintaining species set. If, for example in an extreme case, the network only consists of decay reactions for all species, the algorithm discovers in the first recursion step that no species can be added to the empty species set that would enable a reaction that creates other species.

In this case, the algorithm quickly terminates, independent of network size, returning the empty set as the only semi-organization.

While the number of extreme rays usually explodes with increasing network size (Gagneur and Klamt, 2004), this is not necessarily the case for reactive connected organizations (*cf.* Table 1). As pathways consist of combinations of adjacent reactions, the addition of a reaction to the network increases the number of potential pathways through the system. For organizations however, the addition of a reaction can even reduce their number. This can be exemplified for the number of closed sets of a system. If the network consists of species without any reactions, every species set is a closed set. Adding a reaction reduces the number of closed sets as the set of reactants is no longer closed. Accordingly, the number of organization candidates is reduced.

6.2 Runtime complexity

Organization computation is NP-hard. This can be shown by analyzing a specific subproblem, the question whether a reaction network contains a reactive organization apart from the empty set or not. This problem is NP-complete and thus we conclude that organization computation is NP-hard. The proof is detailed in the Supplementary Material. It is based on a reduction to the 3-SAT problem. A formula F in 3-SAT can be translated into a reaction network for which the existence of a non-empty reactive organization implies the satisfiability of F. Vice versa, if there is no solution such that F evaluates to true, the network does not contain a non-empty reactive organization.

7 APPLICATIONS

To demonstrate the use of organization theory to analyze biochemical reaction network models, we apply the method to a genome-scale metabolic network of *E.coli* by Reed *et al.* (2003). After splitting reversible reactions into explicit forward and backward reactions, the reaction network contains 762 species and 1193 reactions. As the complete hierarchy of organizations could neither be determined by Algorithm I nor Algorithm II, the heuristic approach was used and only reactive organizations are considered. Four different scenarios which differ in input species and decay reactions as shown in Table 2 are studied. The resulting organization hierarchies are depicted in Figure 3.

In the first scenario, exactly those species are defined as input that are also input in the original network by Reed *et al.* (2003). A complex growth medium is modeled (see Table 2 for a list of input species). The heuristic approach computes 70 elementary organizations after 48 h of computing time, creating an organization hierarchy of 88 296 organizations. No organization contains the complete network. However, the three largest organizations contain the pseudo species Biomass, indicating states in which the cell is able to produce biomass required for its survival.

In the second scenario, a decay reaction is added for all network species. By this, all species are assumed to decay spontaneously, simulating flow conditions. To be part of an organization, a species must now be produced at a positive rate from the input species set. While in the first scenario organizations exist which only consist of closed reaction loops, for example a reversible reaction in which one species reacts to another and back, such cycles cannot form organizations in the second scenario. For this scenario, the heuristic

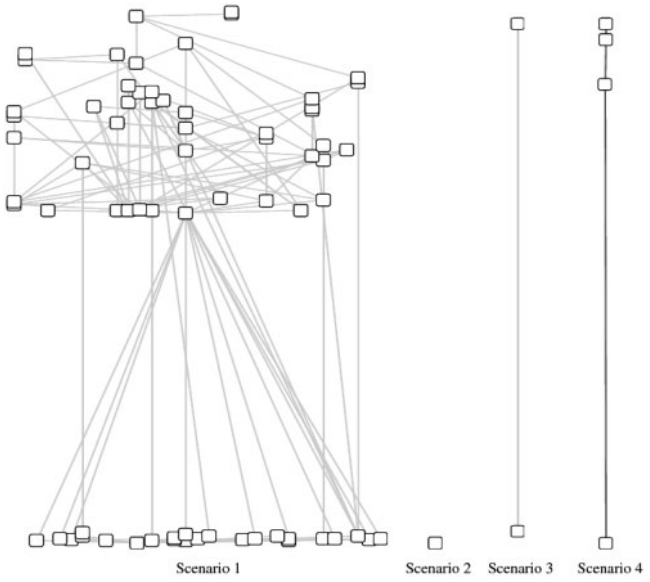


Fig. 3. Hasse diagram of organizations of the genome-scale model of *E.coli* metabolism for the four scenarios as listed in Table 2. Organizations are vertically arranged according to size, with the smallest organization at the bottom (31 species in Scenarios 1, 2 and 4, and 34 species in Scenario 3) and the largest at the top (567 species in Scenario 1 and 547 species in Scenarios 3 and 4). A link between two organizations indicates that the upper organization contains the lower and there is no other organization between them. For the first scenario, only the elementary organizations are shown.

Table 2. Organizations of the genome-scale model of *E.coli* metabolism in four scenarios differing in input and decay reactions

Scenario	Input	Decay	Reactions	Connected orgs.	Elementary orgs.	Reactive orgs.	Largest organization	
							Species	Reactions
1	CM	–	1193	n.a.	≥ 70	n.a.	567	956
2	CM	All	1955	1	1	1	31	77
3	CM + NMS*	All	1958	2	2	2	547	1482
4	CM	All − NMS	1939	4	4	4	547	1464

CM denotes the set of input species given by Reed *et al.* (2003), they are: acetate, CO_2, Fe^{2+}, D-Glucose, glycine, water, H+, K+, D-Lactate, L-Lactate, sodium, ammonia, oxygen, phosphate, sulfate and succinate. NMS denotes the set of non-metabolic species including proteins and tRNA; they are: 3hmrsACP, ACP, acACP, actACP, apoACP, ddcaACP, hdeACP, malACP, myrsACP, octeACP, palmACP, tdeACP, trdox, trdrd, glutrna and trnaglu (for abbrevations see the Supplement of Reed *et al.* (2003)). NMS* denotes the basic non-metabolic species ACP, thioredoxin and tRNA–Glu. Organizations were computed using the heuristic approach, running for 48 h. Except for the first scenario, all organizations were found quickly at the beginning of the computation. Thus, only the results for Scenarios 2, 3 and 4 can be assumed to be complete. n.a., data not available.

approach only identified one organization. It contains the input species and all species that can be created from the input. It is the *closure* of the input species, that means it is the smallest closed species set that contains the input species.

To explain this surprising finding, we analyzed in the first scenario, how organizations are created by the set of organizations to which it has downlinks. Several mechanisms can form the larger organizations. In the simplest case, the union of the downlink organizations already represents the larger organization. In other cases, interactions between species of downlink organizations create new species that are part of the larger organization. And finally, in some cases it is necessary to add novel species to create the larger organization. The closure of the union of downlink organization species then form the larger organization. We call the smallest such sets of additional species creator sets. Among the creator sets of the first scenario are 16 non-metabolic network species representing the acyl carrier protein in several states, tRNA, and thioredoxin, as listed in Table 2. Being a metabolic network model, the model does not contain the *de novo* biosynthesis pathways of these species. They cannot be created from the network.

In the third scenario, we add input reactions for the three non-metabolic species acyl carrier protein, thioredoxin and tRNA–Glu. This ensures that these species are unconditionally available to the system. In this setting, the network contains two organizations. The smallest organization again is the closure of the input species. The second organization contains 547 species including pseudo species Biomass, indicating bacterial growth. Adding the currency metabolites ATP, NAD, coA, mq8 and proteins thioredoxin and ACP, and tRNA–Glu to the small organization is sufficient to create the larger organization.

In the fourth scenario, the unconditional availability of the non-metabolic species is relaxed. Instead of providing them as input, we remove their decay reactions. In this case, we obtain four organizations. The smallest organization is again the closure of the input species. The organization next in size, containing 487 species, is created by adding the currency metabolites ATP, NAD, coA, mq8 and protein thioredoxin. Adding ACP to this organization gives rise to the second largest organization containing 532 species, including pseudo species Biomass. Adding tRNA–Glu finally creates the largest organization encompassing 547 species. This is identical to the largest organization in the third scenario.

Analyzing the four different scenarios reveals the importance of the non-metabolic species. Although being essential for biomass production, their synthesis is not included in the reaction network model. The largest metabolite sets forming self-maintaining networks did not contain all species of the network model. This hints to parts of the network model that need refinement in order to become a truly genome-scale network model of *E.coli* metabolism.

Taking an evolutionary perspective, one could ask what is the smallest set of species that can create the whole metabolism as represented by the largest organization from the input species and their direct derivatives as represented by the smallest organization. We find that such a creator set contains the non-metabolic species and four (non-unique) more species, for example ATP, NAD, coA and mq8. For growth on a complex medium that provides 16 food species (see Table 2), these creator species are enough to create all 547 species present in the largest organization. When assuming that the metabolism has evolved from a small set of metabolite transforming enzymes to a more complex network, the creator sets can give hints on species that were crucial in the expansion process of the metabolism.

If the currency metabolites and the non-metabolic species are assumed to be unconditionally available, the network only contains one organization encompassing the 547 species of the largest organizations in Scenarios 3 and 4. The lack of a richer organizational structure indicates that the metabolism acts as an indivisible unit. The organization hierarchy of four organizations in Scenario 4 is a mere artifact of the incompleteness of the network model, as it does not contain the *de novo* biosynthesis pathways for *all* of its components.

A total of 215 species do not appear in any organization in Scenarios 3 and 4 (see Supplementary Material for a species list). Among them are the 28 substrate deadend species, which only appear as reactants but never as products in reactions. Such deadends indicate incomplete knowledge regarding the metabolism. Besides the obvious incompleteness of the model indicated by the presence of substrate and product deadend metabolites, organization theory provides a second measure of incompleteness. For species that do not appear in any organization, there exists no feasible state in which these species can be maintained, or produced at positive rates under the given growth conditions. Hence, the reaction network model does not explain the synthesis of such species.

8 CONCLUSION

We have presented three algorithms to compute the hierarchy of organizations for a given reaction network. Two of them, the constructive and the flux-based approach, compute the hierarchy of organizations in an exhaustive manner. However, due to the exponential nature of the problem as a network can be constructed in which any species subset is an organization, these algorithms not always finish in reasonable time. In such cases, the heuristic approach can be used to compute at least a significant subset of all organizations. While the constructive approach computes chemical organizations in a bottom-up manner starting from the smallest organization, the flux-based approach combines extreme rays of the self-maintenance cone to find organizations. Testing both algorithms on several network models shows that neither of them is superior in all cases. Not depending on extreme rays, the constructive approach has advantages in networks where the number of extreme rays is very large (e.g. the mars model). When the network contains many species but few extreme rays, the flux-based approach is favorable. Since the number of extreme rays usually grows exponentially with network size, this case might proof rare. However, if a network model is to be analyzed, it cannot be guaranteed beforehand, which algorithm will be more appropriate. If the constructive algorithm is chosen, it is sufficient for many applications to only compute the hierarchy of connected organizations which is usually much faster then computing all organizations.

If both the constructive and the flux-based algorithm fail to compute the organization hierarchy in reasonable time, the proposed heuristic algorithm, which is based on the flux-based approach, can be applied. On current desktop computers, the algorithms can be applied to networks having a species number in the order of 100 species for the deterministic approaches and in the order of 1000 species for the heuristic approach.

Testing the algorithms on a set of example network models demonstrates that neither the runtime nor the number of organizations is strictly correlated with the network size. Both are more dependend on the network topology which is difficult to parameterize.

As an example application of chemical organization theory, we analyze a genome-scale metabolic model of *E.coli*. The analysis unveiled four important aspects of the model. First, the largest organizations in the four analyzed scenarios represent the largest subnetworks that can operate in biologically feasible states, that means steady states and states with increasing metabolite concentrations. However, these subnetworks do not encompass the whole network model. Several species did not appear in any organization. They can serve as starting points in the refinement of the metabolic model to become truly 'genomic-scale'.

Second, three non-metabolic species were identified that proofed essential for the metabolic network: acyl carrier protein, thioredoxin and tRNA–Glu. However, their *de novo* biosynthesis is not accounted for in the model.

Third, only a small set of species is required in addition to the input species to create the metabolic network. In Scenario 3, only the currency metabolites ATP, NAD, Coenzyme A and Ubiquinone 8 need to be added to the smallest organization to create the largest organization in which biomass can be produced. This metabolites are central to the model and might shed light onto the evolution of the metabolism. And fourth, if the non-metabolic and currency metabolites are assumed to be available at all times, the metabolic network model only contains one large organization. The metabolism appears as a indivisible unit that operates as a whole.

Thus, chemical organization theory can on one hand be used to identify parts in the network model that need refinement, and on the other hand it helps to explain model structure and model behavior. In this way organization theory serves as a valuable tool for assessing the quality of whole scale models. By focusing on biologically feasible system states including steady states and states featuring accumulating species, it complements classical flux-based approaches, which focus on steady state flux distributions in metabolic networks.

ACKNOWLEDGEMENTS

We are grateful to Axel von Kamp and Stefan Schuster for helpful discussions on extreme rays and elementary modes. We thank two anonymous reviewers for helpful comments.

Funding: We acknowledge financial support by the BMBF (0312704A), DFG (Di852/4-1/2), and EU (12789).

Conflict of Interest: none declared.

REFERENCES

Berkelaar,M. *et al*. (2005) lp solve: Open source (mixed-integer) linear programming system, version 5.5.

Blinov,M.L. *et al*. (2004) BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics*, **20**, 3289–3291.

Blinov,M.L. *et al*. (2006) A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. *Biosystems*, **83**, 136–151.

Centler,F. and Dittrich,P. (2007) Chemical organizations in atmospheric photochemistries - a new method to analyze chemical reaction networks. *Planet Space Sci*, **55**, 413–428.

Centler,F. *et al*. (2007) Chemical organizations in the central sugar metabolism of Escherichia coli. In Deutsch,A. *et al*.(eds), *Mathematical Modeling of Biological Systems*. Vol. I, pp. 109–123, Birkhäuser, Boston.

Dantzig,G.B. (1963) *Linear Programming and Extensions*. Princeton University Press, Princeton.

de Jong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **9**, 67–103.

Dittrich,P. and Speroni di Fenizio,P. (2007) Chemical organization theory. *Bull. Math. Biol.*, **69**, 1199–1231.

Doi,A. (2005) A hfpn model of bacteriophage lambda. Available at http://www.genomicobject.net/member3/GONET/lambda.html. (last accessed date February 8, 2005).

Finney,A. and Hucka,M. (2003) Systems biology markup language: level 2 and beyond. *Biochem. Soc. Trans.*, **31**, 1472–1473.

Fontana,W. and Buss,L.W. (1994) 'The arrival of the fittest': towards a theory of biological organization. *Bull. Math. Biol.*, **56**, 1–64.

Gagneur,J. and Klamt,S. (2004) Computation of elementary modes: a unifying framework an the new binary approach. *BMC Bioinformatics*, **5**, 175.

Heinrich,R. and Schuster,S. (1996) *The Regulation of Cellular Systems*. Chapman and Hall, New York.

Kaleta,C. *et al*. (2006) Analyzing molecular reaction networks: From pathways to chemical organizations. *Mol. Biotechnol.*, **34**, 117–124.

Matsumaru,N. *et al*. (2006a) Chemical organization theory applied to virus dynamics. *IT - Inf. Technol.*, **48**, 154–160.

Matsumaru,N. *et al*. (2006b) On the evolution of chemical organizations. In Artmann,S. and Dittrich,P. (eds), *Proceedings of the 7th German Workshop on Artificial Life*. Aka, Berlin, pp. 135–146.

Matsumaru,N. *et al*. (2007) Chemical organization theory as a theoretical base for chemical computing. *Int. J. Unconv. Comput.*, **3**, 285–309.

Papin,J.A. *et al*. (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.*, **6**, 99–111.

Reed,J.L. *et al*. (2003) An expanded genome-scale model of Escherichia coli k-12 (ijr904 gsm/gpr). *Genome Biol.*, **4**, R54.

Sauro,H.M. *et al*. (2003) Next generation simulation tools: the systems biology workbench and biospice integration. *OMICS*, **7**, 355–372.

Schuster,S. *et al*. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.

Tomita,M. *et al*. (1999) E-cell: software environment for whole-cell simulation. *Bioinformatics*, **15**, 72–84.

Vass,M.T. *et al*. (2006) The jigcell model builder: a spreadsheet interface for creating biochemical reaction network models. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **03**, 155–164.

Chapter 3

Predicting Phenotypes of Metabolic Networks

In Kaleta et al. [2008] we show how a regulatory network modeled by Boolean logic can be integrated into a metabolic network. This allows us to study the intricate interplay between the regulatory and the metabolic network. Furthermore we are able to analyze the effects of different growth media and gene knockouts on the organism. In Kaleta et al. [2009c] we use the property of steady states and growth states of a reaction network to correspond to chemical organizations in order to validate models of reaction networks. Additionally to the stoichiometric structure we take into account information on kinetics governing the reaction velocities. In an analysis of a large set of curated network models we are able to identify five models with modeling errors and demonstrate an improved performance of chemical organization theory over other stoichiometry based methods.

The Supplemental Materials of Kaleta et al. [2008] and Kaleta et al. [2009c] can be found in Chapter 6 on pages 115 ff. and pages 129 ff., respectively.

BMC Systems Biology

BioMed Central

Research article

Open Access

Phenotype prediction in regulated metabolic networks

Christoph Kaleta[1,2], Florian Centler[1,2,4], Pietro Speroni di Fenizio[1,2,3] and Peter Dittrich*[1,2]

Address: [1]Bio Systems Analysis Group, Department of Mathematics and Computer Science, Friedrich Schiller University Jena, Germany, [2]Jena Centre for Bioinformatics, Jena, Germany, [3]Research Institute in Networks & Communications Engineering (RINCE), Dublin City University, Ireland and [4]Helmholtz Centre for Environmental Research – UFZ, Department of Environmental Microbiology, Leipzig, Germany

Email: Christoph Kaleta - ckaleta@minet.uni-jena.de; Florian Centler - florian.centler@ufz.de; Pietro Speroni di Fenizio - 2008@pietrosperoni.it; Peter Dittrich* - dittrich@minet.uni-jena.de

* Corresponding author

Published: 25 April 2008

BMC Systems Biology 2008, **2**:37 doi:10.1186/1752-0509-2-37

Received: 23 July 2007
Accepted: 25 April 2008

This article is available from: http://www.biomedcentral.com/1752-0509/2/37

Abstract

Background: Due to the growing amount of biological knowledge that is incorporated into metabolic network models, their analysis has become more and more challenging. Here, we examine the capabilities of the recently introduced chemical organization theory (OT) to ease this task. Considering only network stoichiometry, the theory allows the prediction of all potentially persistent species sets and therewith rigorously relates the structure of a network to its potential dynamics. By this, the phenotypes implied by a metabolic network can be predicted without the need for explicit knowledge of the detailed reaction kinetics.

Results: We propose an approach to deal with regulation – and especially inhibitory interactions – in chemical organization theory. One advantage of this approach is that the metabolic network and its regulation are represented in an integrated way as one reaction network. To demonstrate the feasibility of this approach we examine a model by Covert and Palsson (J Biol Chem, 277(31), 2002) of the central metabolism of *E. coli* that incorporates the regulation of all involved genes. Our method correctly predicts the known growth phenotypes on 16 different substrates. Without specific assumptions, organization theory correctly predicts the lethality of knockout experiments in 101 out of 116 cases. Taking into account the same model specific assumptions as in the regulatory flux balance analysis (rFBA) by Covert and Palsson, the same performance is achieved (106 correctly predicted cases). Two model specific assumptions had to be considered: first, we have to assume that secreted molecules do not influence the regulatory system, and second, that metabolites with increasing concentrations indicate a lethal state.

Conclusion: The introduced approach to model a metabolic network and its regulation in an integrated way as one reaction network makes organization analysis a universal technique to study the potential behavior of biological network models. Applying multiple methods like OT and rFBA is shown to be valuable to uncover critical assumptions and helps to improve model coherence.

Background

The analysis of metabolic networks aims at the understanding of metabolic capabilities of organisms to adapt to, and to maintain growth under different external and internal conditions. Various tools exist today to analyze and predict behavior of organisms solely based on metabolic network structure.

Important results have been obtained by applying methods like flux balance analysis [1], modeling by differential equations [2], stochastic simulations [3], or elementary flux mode analysis [4]. While some of these methods concentrate on the network as a whole, others like elementary flux modes decompose it into smaller parts that form functional modules. Chemical organization theory [5] aims at the understanding of reaction networks from both sides. Its basic aim is to identify parts of the network, or more precisely, sets of molecular species, that are likely to coexist on a long time scale without any of the species vanishing or other species appearing anew. This not only encompasses steady states of the network as might be identified by elementary flux mode analysis (see Ref. [6] for the relation between elementary modes and organizations), but also conditions in which there are positive productions of metabolites. Therefore it can be seen as a method that defines a mapping from a reaction network consisting of reactions and metabolites to a set of potential phenotypes [7] of the network as specified by the set of organizations it contains. The theory of chemical organizations has previously been applied to a model of the central sugar metabolism of *E. coli* [8]. It was shown that organizations in the model coincided with known growth phenotypes of *E. coli* under different growth conditions. The growth on each of the carbon sources glucose, lactose, and glycerol could be matched to a specific organization. However, as negative regulation for some genes was ignored because the down-regulation of genes cannot push gene expression levels below the basal level, some organizations represented biologically infeasible system states, for example the simultaneous uptake of all carbon sources.

In this paper, we present an approach to explicitly consider inhibitory regulatory interactions within the analysis of chemical organizations. This allows the more faithful and more precise consideration of biological network models, making the identification of all potential phenotypes of regulated metabolic networks possible. First, the basic concepts of the theory of chemical organizations are presented in the next section. Then, the approach to deal with inhibitory interactions is introduced. The method is then first applied to a subnetwork of a network model of the central metabolism of *E. coli* to predict growth phenotypes. Finally, the method is applied to the complete network model to predict the lethality of gene knockouts.

Theory of Chemical Organizations

The theory of chemical organizations [5] provides a new method to analyze complex reaction networks. Extending ideas by Fontana and Buss [9], one main objective is to determine combinations of network species that are more likely to be present over long periods of (simulation-) time than others. Such sets of species are called organizations. To be an organization, a species set has to fulfill two properties: algebraic closure and self-maintenance. The first property – closure – ensures that given the molecular species of an organization, there is no reaction within the reaction network that could create a novel species not yet present in the organization. The second property – self-maintenance – guarantees that every molecular species that is consumed within the organization can be recreated from organization species at a suffcient rate for its maintenance. The basic concepts required for this paper are summarized now more formally.

Reaction network

Let $\mathcal{M}$ be a set of molecular species, $\mathcal{P}_M(\mathcal{M})$ denotes the set of all multisets with elements from $\mathcal{M}$. A multiset differs from a set in the fact that it can contain the same element more than once. The set of reactions $\mathcal{R}$ occurring among the species $\mathcal{M}$ can then be defined by a relation $\mathcal{R} \subseteq \mathcal{P}_M(\mathcal{M}) \times \mathcal{P}_M(\mathcal{M})$. We call the pair $\langle \mathcal{M}, \mathcal{R} \rangle$ a *reaction network*.

Closed set

A set of species $\mathcal{S} \subseteq \mathcal{M}$ is *closed*, if for all reactions $(\mathcal{A} \to \mathcal{B}) \in \mathcal{R}$ with $\mathcal{A} \in \mathcal{P}_M(\mathcal{S})$, then also $\mathcal{B} \in \mathcal{P}_M(\mathcal{S})$. In other words: if the educts of a reaction are contained in $\mathcal{S}$, then also its products must be in $\mathcal{S}$. There is no reaction in $\mathcal{R}$ that could create any new species not yet in $\mathcal{S}$ from species contained in $\mathcal{S}$.

Self-maintaining set

Given a reaction system $\langle \mathcal{M}, \mathcal{R} \rangle$ with $m = |\mathcal{M}|$ species and $n = |\mathcal{R}|$ reactions, $\mathbf{S} = (m_{i,j})$ be its $m \times n$ stoichiometric matrix, where $m_{i,j}$ is the number of molecules of species i that is produced in reaction j (*i.e.*, right hand side minus left hand side). A set of species $\mathcal{S} \subseteq \mathcal{M}$ is called *self-maintaining* if a flux vector $v = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n_{\geq 0}$ exists such that the following three conditions are fulfilled:

(1) For every reaction $(\mathcal{A} \to \mathcal{B}) \in \mathcal{R}$ with $\mathcal{A} \in \mathcal{P}_M(\mathcal{S})$, its corresponding flux is $v_{\mathcal{A} \to \mathcal{B}} > 0$.

(2) For every reaction $(\mathcal{A}\rightarrow\mathcal{B})\in\mathcal{R}$ with $\mathcal{A}\notin\mathcal{P}_M(\mathcal{S})$, its corresponding flux is $v_{\mathcal{A}\rightarrow\mathcal{B}} = 0$.

(3) For every species $i \in \mathcal{S}$, its concentration change is nonnegative: $(\mathbf{S}v)_i \geq 0$.

In other words: if we consider only the sub-network made up by the species of $\mathcal{S}$ and additionally the species that can be created from $\mathcal{S}$ (but are not in $\mathcal{S}$) (conditions (1) and (2)), we can find a positive flux vector, such that no species of $\mathcal{S}$ decays (condition (3)).

Organization

A set of species $\mathcal{S}\subseteq\mathcal{M}$ that is closed and self-maintaining is called an *organization*.

Balanced organization

An organization $\mathcal{S}\subseteq\mathcal{M}$ is a *balanced organization*, if a flux vector conforming to the self-maintenance condition can still be found, if in requirement 3 of the self-maintenance definition, the greater equal condition is replaced by equality:

(3') For every species $i \in \mathcal{S}$, its concentration change is zero: $(\mathbf{S}v)_i = 0$.

A rigorous link between organizations and the potential dynamics of a reaction system is provided by Theorem 1 from Ref. [5]: Assume that the dynamics is modeled as a "chemical" differential equation system $dx(t)/dt = \mathbf{S}v(x(t))$, then all steady states of the system are instances of organizations. In other words, the species with concentration levels greater than zero in a particular steady state are exactly those species contained in a corresponding organization. Note that organizations do not necessarily contain a steady state, as they can also embody growth in which species have increasing concentrations. The only assumption made for this theorem is that molecules that are present and can react will react sooner or later (formally: $v_{\mathcal{A}\rightarrow\mathcal{B}}(x) > 0$, if and only if for all $i \in \mathcal{A} : x_i > 0$). Note that this assumption differs fundamentally from the assumption made by methods like elementary mode analysis, which assume that each reaction can be switched off independently even if the reactants are present in large concentrations [6].

Computing Organizations

To compute organizations, the convex polyhedral cone $\mathcal{P}$ can be used which is defined by the $n + m$ inequalities $v \geq \mathbf{0}$ and $\mathbf{S} \cdot v \geq \mathbf{0}$ in the flux space $\mathbb{R}^n_{\geq 0}$. This cone contains all self-maintenance flux vectors as described in the self-maintenance definition. In order to find species sets that are self-maintaining and closed, the extreme rays spanning $\mathcal{P}$ are combined in a recursive fashion and the resulting species sets tested. The Supplement contains a detailed description of the algorithm, and outlines a heuristic approach to compute organizations for large network models, for which the runtime of the algorithm exceeds practical limits (see Additional file 1).

Methods

Regulation has not yet been explicitly considered in the analysis using the theory of chemical organizations. The aim of this section is to elaborate a concept that allows us to also account for regulation. This concept makes it possible to also study regulated metabolic networks using organization theory. The basic idea is to map regulatory rules to normal chemical reaction rules. For inhibitory rules and rules using the direction of a reversible flux we introduce pseudo species representing the absence of a species and the direction of a flux, respectively.

Types of Regulation

To examine the effects of regulation on chemical organizations we first need to discuss the general types of regulatory interactions that occur in biological systems in more detail.

Regulation appears on different levels in the cell, being carried out by a variety of biological entities (*e.g.*, small molecules, proteins, RNA) acting on further biological target entities. As we are considered with reaction networks, we focus here on the regulation of reactions. Two different types of regulation have to be considered. The first type of regulation only changes the flux of the regulated reaction slightly. Certain types of autoregulation fall into this category. This kind of regulation does not change the structure of the reaction network and hence does not affect its organizational hierarchy. The second type of regulation is more drastic: it turns a reaction completely off or enables a formerly unavailable reaction. This is the case, for example, when the expression of a protein that catalyzes a reaction is suddenly repressed. As a consequence, the catalyzed reaction is not available to the network anymore once the protein is completely degraded. The induction of uptake pathways is an example for enabling novel reactions. New reaction pathways become suddenly available.

Note that, even though drastic, this kind of categorization of regulation leads to meaningful models, for example when translated into a boolean regulatory network [10] and also generalizes to discretization using more than two levels as used in Ref. [11].

Regulatory interactions do not happen instantly. The time delay between the onset of a regulatory event and its measurable effect in the system can vary between milliseconds (*e.g.*, phosphorylation of proteins in signal cascades [12]) and minutes (*e.g.*, changes in gene expression [13]). However, as we are here interested in the long term behavior of the system, we do not take different time scales of regulation into account.

Modeling Regulatoy Interactions

Several approaches exist to represent and model regulatory interactions [14], for example, boolean logic [10,15,16], stochastic models [17], and differential equations [18]. Whereas some approaches require a very detailed knowledge of the mechanisms and kinetics behind the regulation, the representation of regulatory networks by boolean logic can be useful if the knowledge about the underlying kinetics is limited [11]. In this approach, the state *on* or *off* is assigned to a regulated reaction [15]. We will adopt this notion to model regulatory interactions.

Two types of regulatory events have to be considered: *activation*, in which a species is required in order to perform a certain reaction, and *inhibition*, in which a species inhibits a certain reaction and makes it unavailable. To model this kind of regulation we make use of the properties of reactions being similar to rewriting rules, where the left hand side is being replaced by the right hand side. Taking this approach, the molecules on the left hand side need to be present for the reaction to proceed. Additionally, regulatory events can be triggered not only by the presence or absence of a species, but also by a species being available in excess or not.

Activation

Activation or turning on a reaction by a specific species can be simply modelled by considering this species as a kind of a catalyst. In terms of rewriting rules this approach can be considered as an additional constraint on the presence of certain species for the reaction to proceed. By this, the reaction can only take place when the activating species is present. Being a catalyst, the activating species is not consumed within the reaction it catalyzes.

Let us consider the general case in which species E activates a reaction that transforms a reactant A into product B. In the absence of E, the reaction shall have a zero flux, while the flux shall become positive in the presence of E and A. A reaction $A \rightarrow B$ activated by E becomes:

$$E + A \rightarrow E + B.$$

Note that adding E as a catalyst on the reactant and product side of the reaction equation does not change the stoichiometric matrix **S**. Still, one unit of A is being consumed to produce one unit of B. Therefore any flux vector that guarantees self-maintenance for a set of metabolites including E but without considering E as an activator, will also guarantee self-maintenance when E is added as a catalyst to model activation.

Inhibition

Handling inhibition is more diffcult. If inhibitor I inhibits a reaction, we could add an if-statement to each reaction that guarantees that the reaction is only available when I is absent. But since we intend to model regulation within the language of reactions, such if-statements would not fulfill our requirements. An alternative way to model inhibition is to understand it as another type of activation, that is, as the activation of a reaction by the absence of an inhibitor. For achieving this we have to introduce a pseudo species $\bar{I}$ that represents the absence of inhibitor I. In terms of rewriting rules, such species is just a constraint on the presence of a certain species for a reaction to take place. A reaction $A \rightarrow B$ inhibited by I becomes:

$$\bar{I} + A \rightarrow \bar{I} + B.$$

Only in the absence of I, represented by pseudo species $\bar{I}$, educt A can react to form product B.

Modelling flux-direction dependent regulation

Sometimes it can make sense to define regulatory rules that depend on the *direction* at which one or more reversible reactions operate. Two examples can be found in Covert and Palsson [19], the rule for the pyruvate response regulator (gene *pdhR*) and the rule for the catabolite activator protein (gene *cra*). Given a reversible reaction r: $A \leftrightarrow B$ and the following flux-direction dependent regulatory rule:

"If the flux of reaction r is positive (forward), then activate protein E.",

we map this regulatory rule to the following conventional reaction rules by introducing a pseudo species f_r and its inverse counterpart $\overline{f_r}$:

$$r_f: A + f_r \rightarrow B + f_r \qquad (1)$$

$$r_b : B + \overline{f_r} \rightarrow A + \overline{f_r} \quad (2)$$

$$r_E : A + f_r \rightarrow A + f_r + E \quad (3)$$

Because in consistent organizations (see below) either f_r or $\overline{f_r}$ is present (but not both), the reaction rules assure that either the forward r_f or the backward reaction rule r_b can be active. Here, in Eq. (3), we use the presence of the molecular species A and f_r as an indicator for the activity of reaction r_f. Conversely, we can use the presence of the molecular species B and $\overline{f_r}$ as an indicator for the activity of the backward reaction r_b.

If necessary, we can add a pseudo species like f_r to every reversible reaction rule r in order to obtain an indicator for the direction in which it operates. Then, these indicators can even be combined to represent a more complex rule. Note that in Covert and Palsson [19], different flux-directions are logically combined to indicate a surplus of a molecular species (*e.g.*, *PYR* for the regulation of *pdhR*; and *FDP* or *F6P* for the regulation of *cra*).

Consistent organization

Introducing pseudo species causes a problem, as now two network species represent the same molecular entity. When computing the organizations of such a network, some organizations might exist that contain both or neither of the two species. In both cases, the presence of the species is not clearly defined. Either the presence and absence is indicated simultaneously, or no statement is made at all. Consequently, we only consider those organizations in the remaining of this paper, in which for all species it is clearly defined whether they are present or not. We call such organizations consistent.

Consistent organization

An organization $O \subseteq \mathcal{M}$ is called a *consistent organization*, if for all species $S \in \mathcal{M}$ for which there exists a pseudo species $\overline{S} \in \mathcal{M}$ that indicates its absence, either S or $\overline{S}$ is contained in the organization.

In passing we note that this approach allows one to model more than two states of a molecule, for example different phosphorylation states.

Modeling Boolean Logic

There are few cases where a reaction is regulated by a single molecule alone. In most cases regulation is more complex, for example if the availability of a reaction is being determined by the interaction of a set of proteins. In such cases we need to model regulation by a set of boolean functions. This section presents an approach to account for such functions on the level of regulation (see also [20]).

All binary boolean functions can be reduced to either *AND* or *OR*, and the negation *NOT*. The construction of a negation has been outlined above. In principle, it would be suffcient to present a method to implement *AND* or *OR*. However, we present methods for both to ease the process of converting regulation logics to chemical reactions.

First we consider the *AND* function. A typical regulation incorporating an *AND* function is the required presence of two activators to perform a reaction. If we consider activator $E1$ and activator $E2$ to be necessary for a reaction converting educt A into product B we can write:

$$E1 + E2 + A \rightarrow E1 + E2 + B.$$

Next, the *OR* function is considered. An example for this case is a reaction transforming educt A into product B that can alternatively be activated by two activators $E1$ and $E2$. The presence of one of these activators is suffcient to perform the reaction. In this case, the reaction is split into two parts; one that accounts for the presence of activator $E1$, and one that accounts for the presence of activator $E2$:

$$E1 + A \rightarrow E1 + B$$

$$E2 + A \rightarrow E2 + B.$$

Taking these two basic functions, it is possible to model all regulatory relationships that can be represented by boolean rules in metabolic networks [20].

Example: Regulatory switch

As an example for the presented procedure, we analyze a simple reaction network comprising – apart from inflow and outflow – two reactions forming a switch as depicted in Figure 1(A). The product of one reaction inhibits the other reaction and vice versa. Additionally, an inhibitor I can shut down both reactions. Thus, we have a simple *AND* function that requires for both reactions that both I *and* $P1$, respectively $P2$, are absent. A model without regulation would contain only reactions transforming A to $P1$ and $P2$, the influx to A and the outflux from the products: $\mathcal{R}' = \{\emptyset \rightarrow A, A \rightarrow P1, A \rightarrow P2, P1 \rightarrow \emptyset, P2 \rightarrow \emptyset\}$.

The boolean expressions describing the regulation are:

$$A \rightarrow P1 \text{ if } \neg I \ \ \neg P2$$

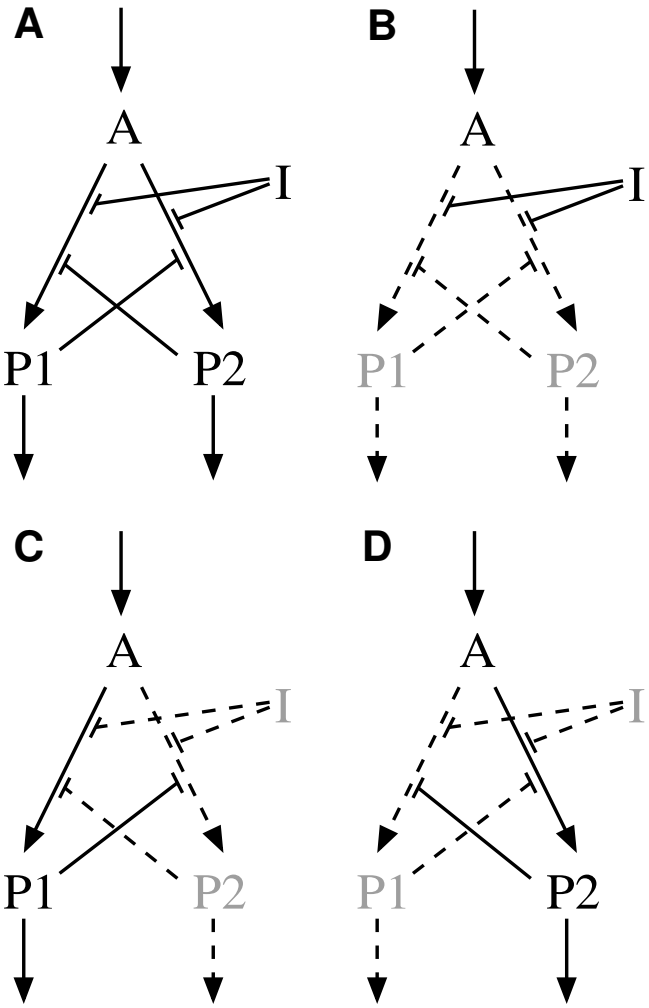


Figure 1
A simple switch. Regulatory switch network (**A**) and the reaction networks belonging to its three consistent organizations (**B**, **C**, and **D**). Absent species appear in gray. Inactive reactions and interactions are dashed. Panel **B** represents Organization 10 = {*A*, *I*} where inhibitor *I* represses both reactions from *A* to *P*1 and *P*2. Panels **C** and **D** represent Organization 11 = {*A*, *P*1} and Organization 12 = {*A*, *P*2} where one pathway is active, either over *P1* or *P*2.

$$A \to P2 \text{ if } \neg I \ \ \neg P1.$$

Applying the presented procedure, these expressions are transformed into chemical reactions. The resulting reaction network contains the following reactions:

$$\mathcal{R} = \{$$

$$\varnothing \to A, \tag{1a}$$

$$\overline{P2} + \bar{I} + A \to \overline{P2} + \bar{I} + P1, \tag{2a}$$

$$\overline{P1} + \bar{I} + A \to \overline{P1} + \bar{I} + P2, \tag{3a}$$

$$P1 \to \varnothing \tag{4}$$

$$P2 \to \varnothing \ \}. \tag{5}$$

The network contains 16 organizations as listed in Table 1.

Three organizations are consistent: $O_{10} = \{A, I, \overline{P1}, \overline{P2}\}$, $O_{11} = \{A, \bar{I}, P1, \overline{P2}\}$, and $O_{12} = \{A, \bar{I}, \overline{P1}, P2\}$. In the remaining organizations it is at least for one species not clearly defined whether it is present or not. Taking Organization 2 for example, the presence of *A* and *I* is determined with *A* present and *I* absent, but there is no information concerning species *P*1 and *P*2. In Organization 6, inhibitor *I* is present and absent at the same time. Figure 1(B,C,D)) depicts the reaction networks belonging to the three consistent organizations. They represent the three states of the switch. In Organization 10, inhibitor *I* is present and shuts down reactions (2) and (3), turning the switch off. In the other two consistent organizations *I* is not present and there is either a flux through reaction

Table 1: A simple switch. All organizations of the regulatory switch network (Figure 18.1). Three organizations are consistent: 10, 11, and 12 (marked bold).

Org.	Species	Real Species
1	A	-
2	$A, \bar{I}$	-
3	A, I	-
4	$A, \overline{P2}$	-
5	$A, \overline{P1}$	-
6	$A, I, \bar{I}$	-
7	$A, I, \overline{P1}$	-
8	$A, I, \overline{P2}$	-
9	$A, \overline{P1}, \overline{P2}$	-
10	$A, I, \overline{P1}, \overline{P2}$	A, I
11	$A, \bar{I}, P1, \overline{P2}$	$A, P1$
12	$A, \bar{I}, \overline{P1}, P2$	$A, P2$
13	$A, \bar{I}, I, \overline{P1}, \overline{P2}$	-
14	$A, \bar{I}, I, P1, \overline{P2}$	-
15	$A, \bar{I}, \overline{P1}, P1, \overline{P2}, P2$	-
16	$A, \bar{I}, I, \overline{P1}, P1, \overline{P2}, P2$	-

(2) (Organization 11) or through reaction (3) (Organization 12). They represent the two other states of the switch.

Results

We apply the method to a model of the central metabolism of *E. coli* by Covert and Palsson [19]. The authors used the model to study the effects of regulation on flux balance analysis. The regulatory network is defined by a set of boolean functions. There are 73 enzymes, which catalyze 113 reactions. Of these reactions, 43 are regulated by 16 regulatory proteins and therefore controlled by logic statements. The unregulated proteins are assumed to be present in the cell at all times. We add an inflow for all of them in our analysis. To incorporate the regulation into the reaction network, we add the proteins that catalyze reactions explicitly as catalysts in the reactions as described above. Following the introduced protocol, the regulatory logic is incorporated by introducing pseudo species and adapting the reactions accordingly. The activity of several genes is described by boolean statements. Appropriate chemical reactions are added to model this gene regulation. We analyze two variants of the network model by Covert and Palsson [19]: a simplified *core network* to study wildtype growth phenotypes on different carbon sources, and the *complete network* for predicting knockout experiments. Both networks are listed in the Supplement (see Additional file 1) and provided as SBML files (see Additional files 2 and 3).

The core network model

For studying growth on different carbon sources including diauxic shift, the network is reduced to the set of reactions that lead from external glucose, lactose, and glycerol to pyruvate via glycolysis. Additionally, the pentose-phosphate pathway reactions and the reactions leading from glucose-6-phosphate to this pathway are removed. The resulting network comprises 49 reactions of the original network. The considered part of the network does not contain any ATP production. However, ATP is used up by some reactions, for example in glucose uptake. Therefore, ATP is provided as input. Furthermore, UTP, NAD, NADP, Ubiquinone, and external hydrogen ions are necessary for other uptake and transformation reactions and cannot be provided by this part of the network. These species are added as input as well. To model growth, an outflow is added for every biomass precursor, as in the original network. Since we consider proteins as being active only when they are produced, an outflow for every protein is added, modeling degradation. In order to model different growth media and conditions, self-replicator reactions for external glucose, lactose, glycerol, and oxygen are added of the form $M \rightarrow 2\ M$. These reactions ensure that a constant supply of the respective species is available, whenever it is considered to be present. Using self-replicator reactions, all $2^4 = 16$ different growth conditions can be modeled in a single network and can be simultaneously considered in one analysis.

The final model comprises 95 species (including 15 pseudo species representing the absence of a species) and 168 reactions. The complete list of network reactions can be found in the Supplement.

The complete network model

For predicting the lethality of gene knockouts we use the complete network model of the regulated central metabolism of *E. coli* by Covert and Palsson [19]. Depending on the availability of oxygen and the different carbon sources in the growth medium, influxes are added for the respective external species. The currency metabolites HEXT, PI, ADP, ATP, NAD, NADH, Q, QH2, NADP, NADPH, FAD, FADH, UTP, and COA are considered to be unconditionally available in the cell. Input reactions are added for all these species. Without the influxes of external carbon sources and oxygen, the network contains 227 species and 468 reactions.

Growth Phenotypes on Different Carbon Sources

The core network model contains 16 consistent organizations. They are listed in Table 2 (see Supplement for a graphical representation). The consistent organizations coincide with the 16 possible growth conditions. The smallest organization, Organization 1, just contains the input metabolites plus the products of the hydrolyzation of ATP, ADP, and phosphate. When analyzing the genes that are active in this organization, we find that the response regulators for glucose, lactose, and glycerol are active, indicating that the respective carbon sources are not present. Due to the absence of oxygen, the aerobic response regulators ArcA and Fnr are also active.

In Organization 2, external oxygen is available. Consequently, the aerobic response regulators ArcA and Fnr are absent here. This is the only difference between Organizations 1 and 2.

Glucose uptake

The first organization that utilizes an external carbon source is Organization 3, which contains the reactions for glucose uptake. Consequently, the metabolites of the central metabolism are present in this organization. When examining the proteins of the organization, we find that the glucose response regulator Mlc is absent. The organization next in size is Organization 4. Here, lactose is additionally in the medium. Although the repressor of the *lac* gene, LacI, is absent in the organization, no uptake reactions for external lactose are contained in the organization. The lactose permease LacY, a product of the *lac* genes, is missing. As glucose is available in the medium, the lactose uptake system is not induced by the presence

Table 2: Growth phenotypes of the core model. Consistent organizations in the core network model of the regulated central metabolism of *E. coli*, ordered by size.

Consistent Organiza.	Species	Growth medium	Uptake
1	Input metabolites, ADP, PI, ArcA, Fnr, GalR, GalS, GlpR, LacI, Mlc, PykF, Ubiquitous proteins	-	-
2	Input metabolites, ADP, O2, O2xt, PI, GalR, GalS, GlpR, LacI, Mlc, PykF, Ubiquitous proteins	O2	-
3	Input metabolites, Glycolysis metabolites, ADP, G1P, GLC, GLCxt, LCTSxt, NADH, PI, PPI, UDPG, ArcA, Crr, FadR, Fnr, Food, GalP, GalR, GalS, GlpR, LacI, Pgk, PtsGHI, PykF, Ubiquitous proteins	GLC	GLC
4	Input metabolites, Glycolysis metabolites, ADP, G1P, GLC, GLCxt, LCTSxt, NADH, O2, O2xt, PI, PPI, UDPG, ArcA, Crr, FadR, Fnr, Food, GalP, GlpR, Pgk, PtsGHI, PykF, Ubiquitous proteins	GLC, LCTS	GLC
5	Input metabolites, Glycolysis metabolites, ADP, G1P, GLC, GLCxt, GLxt, LCTSxt, NADH, PI, PPI, UDPG, ArcA, Crr, FadR, Fnr, Food, GalP, GalR, GalS, LacI, Pgk, PtsGHI, PykF, Ubiquitous proteins	GLC, GL	GLC
6	Input metabolites, Glycolysis metabolites, ADP, G1P, GLC, GLCxt, GLxt, LCTSxt, NADH, O2, O2xt, PI, PPI, UDPG, ArcA, Crr, FadR, Fnr, Food, GalP, Pgk, PtsGHI, PykF, Ubiquitous proteins	GLC, GL, LCTS	GLC
7	Input metabolites, Glycolysis metabolites, ADP, G1P, GLC, GLCxt, NADH, PI, PPI, UDPG, Crr, FadR, Food, GalP, GalR, GalS, GlpR, LacI, Pgk, PtsGHI, PykF, Ubiquitous proteins	GLC, O2	GLC
8	Input metabolites, Glycolysis metabolites, ADP, G1P, GLC, GLCxt, NADH, O2, O2xt, PI, PPI, UDPG, Crr, FadR, Food, GalP, GlpR, Pgk, PtsGHI, PykF, Ubiquitous proteins	GLC, LCTS, O2	GLC
9	Input metabolites, Glycolysis metabolites, ADP, G1P, GLC, GLCxt, GLxt, NADH, PI, PPI, UDPG, Crr, FadR, Food, GalP, GalR, GalS, LacI, Pgk, PtsGHI, PykF, Ubiquitous proteins	GLC, GL, O2	GLC
10	Input metabolites, Glycolysis metabolites, ADP, G1P, GLC, GLCxt, GLxt, NADH, O2, O2xt, PI, PPI, UDPG, Crr, FadR, Food, GalP, Pgk, PtsGHI, PykF, Ubiquitous proteins	GLC, GL, LCTS, O2	GLC
11	Input metabolites, Glycolysis metabolites, ADP, G1P, GL, GL3P, GLxt, NADH, NADPH, O2, O2xt, PI, PPI, QH2, UDPG, ArcA, Crr, Fnr, Food, GalP, GalR, GalS, GlpABC, GlpF, GlpK, LacI, Mlc, Pgk, PtsGHI, PykF, Ubiquitous proteins	GL	GL
12	Input metabolites, Glycolysis metabolites, ADP, G1P, GL, GL3P, GLxt, NADH, NADPH, PI, PPI, QH2, UDPG, Crr, Food, GalP, GalR, GalS, GlpD, GlpF, GlpK, LacI, Mlc, Pgk, PtsGHI, PykF, Ubiquitous proteins	GL, O2	GL
13	Input metabolites, Glycolysis metabolites, Lactose derivatives, ADP, G1P, GLC, LCTS, LCTSxt, NADH, PI, PPI, UDPG, ArcA, Crr, Fnr, Food, GalE, GalK, GalM, GalP, GalT, GlpR, LacY, LacZ, Mlc, Pgk, PtsGHI, PykF, Ubiquitous Proteins	LCTS	LCTS
14	Input metabolites, Glycolysis metabolites, Lactose derivatives, ADP, G1P, GLC, LCTS, LCTSxt, NADH, O2, O2xt, PI, PPI, UDPG, ArcA, Crr, Fnr, Food, GalE, GalK, GalM, GalP, GalT, LacY, LacZ, Mlc, Pgk, PtsGHI, PykF, Ubiquitous Proteins	GL, LCTS	LCTS
15	Input metabolites, Glycolysis metabolites, Lactose derivatives, ADP, G1P, GLC, GLxt, LCTS, LCTSxt, NADH, PI, PPI, UDPG, Crr, Food, GalE, GalK, GalM, GalP, GalT, GlpR, LacY, LacZ, Mlc, Pgk, PtsGHI, PykF, Ubiquitous proteins	LCTS, O2	LCTS
16	Input metabolites, Glycolysis metabolites, Lactose derivatives, ADP, G1P, GLC, GLxt, LCTS, LCTSxt, NADH, O2, O2xt, PI, PPI, UDPG, Crr, Food, GalE, GalK, GalM, GalP, GalT, LacY, LacZ, Mlc, Pgk, PtsGHI, PykF, Ubiquitous Proteins	GL, LCTS, O2	LCTS

For brevity, pseudo species indicating the absence of a species are not listed. A list of abbrevations can be found in the Supplement. A species followed by 'xt' denotes its extra-cellular form. "Ubiquitous proteins" include the proteins that are considered ubiquitously present in the cell and therefore are not listed separately. They are: Eno, Fba, Fbp, GalU, GapA, Glk, GpmA, GpmB, GpsA, PfkA, PfkB, Pgi, Pgm, PykA, and TpiA. "Input metabolites" denotes the metabolites provided as input to the system: HEXT (external hydrogen), Q (Ubiquinone), ATP, UTP, NAD, and NADP. "Glycolysis metabolites" denotes the metabolites of the glycolysis: G6P, F6P, FDP, T3P2, T3P1, 13PDG, 3PG, 2PG, PEP, and PYR. "Lactose derivatives" denotes the derivatives of lactose in the central metabolism: GAL1P, GLAC, UDPGAL, bDGLAC, bDGLC.

of external lactose. This effect is known as inducer exclusion. The metabolite required for upregulation of the *lac* genes is not taken up by the cell. Organization 5 represents a similar case in which glycerol is available in the growth medium but not taken up. All external carbon sources and oxygen are available in Organization 10, but the cell is still exclusively utilizing glucose. Organizations 6 to 9 represent further input combinations defining growth conditions with external glucose available. The availability of oxygen does not change the reactions in the part of the central metabolism that is considered in the core model.

Lactose uptake

In Organization 13, lactose is the exclusive external carbon source. Consequently, LacI is absent as it is bound by allolactose, a derivative of lactose. Hence, it cannot repress the genes necessary for lactose uptake and utilization. We find the corresponding gene products present in this organization, namely LacZ and LacY. Additionally,

derivatives of lactose, for example galactose, are contained in this organization. These metabolites are created in the pathway leading from lactose to the central metabolism. Another diauxic shift effect can be observed in Organization 14. Here, external lactose and glycerol are present as carbon sources, but as in the case with glucose and lactose, only lactose is taken up. Organizations 15 and 16 represent further growth conditions in which lactose is taken up. Once again, the availability of oxygen does not change the reactions in the modeled part of the central metabolism.

Glycerol uptake

Glycerol is the exclusive external carbon source in Organization 11. As all proteins necessary for glycerol uptake are present, glycerol is taken up. For glycerol uptake, three different enzymes catalyze the reaction from glycerol-3-phosphate to dihydroxyacetone-phospate, a metabolite of glycolysis. One of these enzymes, glycerol-3-phosphate-dehydrogenase, is constitutively expressed in the model. The other two proteins, glycerol-3-phosphate kinases GlpABC and GlpD are specific for anaerobic, respectively aerobic growth conditions. Therefore, GlpABC is present and GlpD absent in Organization 11, where no oxygen is available. When oxygen is available as in Organization 12, GlpD is present and GlpABC absent.

Predicting Gene Knockout Experiments

Knockout experiments are performed using the complete network model. Gene knockouts are modeled by deleting all reactions in which the corresponding protein takes part as educt or product. The set of consistent organizations is determined for each knockout experiment using a heuristic approach detailed in the Supplement. The lethality of a knockout can be predicted by the existence of organizations containing all biomass precursor metabolites. If such an organizations is not found, the knockout is predicted to be lethal. We use organization theory (OT) and an adapted version of organization theory (aOT, see below) to predict the same gene knockouts as Covert and Palsson [19], who used regulatory flux balance analysis (rFBA) for gene knockout predictions. Reference [19] is also our source for *in vivo* data and predictions by flux balance analysis (FBA) and rFBA. The results are summarized in Table 3. Out of 116 experiments, the predictions by FBA are correct in 97 cases (83,6%). The predictions by rFBA are correct in 106 cases (91,4%) and improve the results of FBA in nine cases. Unmodified OT predicts the lethality of knockouts correctly in 101 cases (87,1%), while aOT predicts 106 cases (91,4%) correctly as rFBA. The additional model-specific assumptions made by aOT are taken from Covert and Palsson [19] and will be described in detail now.

Assumption that accumulation of mass is lethal

In two cases, OT predicts a lethal knockout to be nonlethal (*rpiA*, and *rpiA* + *rpiB* on glucose). The self-maintenance property allows for the accumulation of internal metabolites, while in rFBA, only steady states are considered, and any accumulation of metabolites is regarded as lethal. In these two cases, the organizations containing all biomass precursors contain metabolites with positive productions. (Note, that all species except the pseudo species indicating the absence of species decay in the network model. Hence, all organizations are indeed balanced organizations. However, accumulation of metabolites occur, if the decay reactions, which are not present in the original network, are removed.) Hence, OT predicts the knockout to be nonlethal while rFBA predicts it to be lethal as no steady state exists. In aOT we restrict our analysis to organizations that are balanced (*i.e.*, internal metabolites are not allowed to accumulate). With this model specific assumption aOT classifies the two knockout experiments correctly.

Assumption that secreted molecules have no effect

Further three incorrect predictions by OT (*ackA* and *pta* on acetate, and *ppc* on glycerol) yield deeper insights into the differences between chemical organization theory and regulatory flux balance analysis, namely in the treatment of by the cell secreted molecules.

In the case of acetate uptake, there are two pathways that enable the utilization of this carbon source as depicted in Figure 2(A). One pathway leads directly from acetate to acetyl-CoA, and the other takes the route via acetyl phosphate. The first pathway is catalyzed by the acetyl-CoA synthethase (gene *acs*). According to the model, the corresponding gene is only transcribed if no carbon source is available or at most acetate or formate, or both. The second pathway is catalyzed by acetate kinase A (gene *ackA*) and phosphotransacetylase (gene *pta*). If one of these genes is knocked out, the first pathway can still support the central metabolism, given that acetate is the exclusive external carbon source. In this case, chemical organization theory predicts both knockouts as lethal, which is not the case *in vivo* and correctly predicted by rFBA. The reason for this discrepancy is that in any network containing the biomass precursor metabolite pyruvate, this metabolite will be secreted by the cell. Therefore, such a network also comprises the external form of pyruvate which is an inhibitor for the only remaining uptake reaction for acetate. Consequently, there exists no organization containing all biomass precursor metabolites when acetate is the exclusive carbon source in the growth medium and the second pathway is knocked out. Because the presence of metabolites is not explicitly considered in rFBA, this inhibition is not detected by rFBA. However, because the knockout is nonlethal in *in vivo* experiments, the levels of secreted

Table 3: Comparison of knockout predictions.

	glc	gl	suc	ac	rib	glc (-O_2)	Dual Substr.	Reference
aceA	+/+/+/+		+/+/+/+	-/-/-/-		+/+/+/+		[30]
aceB				-/-/-/-				[31]
aceEF	-/+/-/-		-/+/-/-	+/+/+/+		+/+/+/+	+/+/+/+ (glc-ac)	[32]
ackA				**+/+/+/-**				[33]
ackA + pta + acs				-/-/-/-				[33]
acnA	+/+/+/+	+/+/+/+	+/+/+/+	+/+/+/+		+/+/+/+		[31,34]
acnB	+/+/+/+	+/+/+/+	+/+/+/+	-/+/+/+		+/+/+/+		[34]
acnA + acnB	-/-/-/-	-/-/-/-	-/-/-/-	-/-/-/-		-/-/-/-		[34]
acs				+/+/+/+				[33]
adh	+/+/+/+					-/+/+/+		[35]
cyd	+/+/+/+							[36]
cyo	+/+/+/+							[36]
eno	-/-/-/-	-/-/-/-	-/-/-/-				+/+/+/+ (gl-suc)	[37]
fbaA	-/+/+/+							[38]
fbp	+/+/+/+	-/-/-/-	-/-/-/-	-/-/-/-				[39]
frdA	+/+/+/+		+/+/+/+	+/+/+/+		+/+/+/+		[30]
fumA				-/+/-/-		+/+/+/+		[31]
gap	-/-/-/-	-/-/-/-	-/-/-/-				+/+/+/+ (gl-suc)	[37]
glk	+/+/+/+							[38]
glk + pfkA	+/+/+/+							[38]
glk + pts	-/-/-/-							[38]
gltA	-/-/-/-			-/-/-/-				[34]
gnd	+/+/+/+							[38]
icd	-/-/-/-			-/-/-/-				[34]
mdh	+/+/+/+	+/+/+/+	+/+/+/+			+/+/+/+		[40]
ndh	+/+/+/+	+/+/+/+						[41]
nuo	+/+/+/+	+/+/+/+						[41]
pfl						+/+/+/+		[42]
pgi	+/+/+/+	+/-/-/-	+/-/-/-					[38]
pgi + gnd	-/-/-/-							[38]
pgi + zwf	-/-/-/-							[38]
pgk	-/-/-/-	-/-/-/-	-/-/-/-				+/+/+/+ (gl-suc)	[37]
pgl	+/+/+/+							[38]
ppc	-/+/-/-	**-/+/-/+**	+/+/+/+				+/+/+/+ (gl-suc) +/+/+/+ (glc-suc)	[38,40]
pta				**+/+/+/-**				[33]
pts	+/+/+/+							[38]
pykA	+/+/+/+							[38]
pykA + pykF	+/+/+/+							[38]
pykF	+/+/+/+							[38]
rpiA	**-/+/-/+**				+/+/+/+		+/+/+/+ (glc-rib)	[43]
rpiA + rpiB	**-/-/-/+**				-/+/+/+		+/+/+/+ (glc-rib)	[43]
rpiB	+/+/+/+				+/+/+/+		+/+/+/+ (glc-rib)	[43]
rpiR + rpiA	+/N/+/+				+/N/+/+		+/N/+/+ (glc-rib)	[43]
sdhABCD	+/+/+/+		-/-/-/-	-/-/-/-		+/+/+/+		[30]
sucAB- lpd	-/+/+/+		-/+/+/+	-/+/+/+		+/+/+/+	+/+/+/+ (glc-suc)	[30,32]
tpi	-/+/+/+	-/-/-/-	-/-/-/-	-/-/-/-			+/+/+/+ (glc-suc) +/+/+/+ (glc-gl)	[37,44]
zwf	+/+/+/+							[38]

Comparing *in vivo* knockout experiment results with predictions made by FBA, rFBA, OT and by aOT. A '+' indicates growth, a '-' no growth of the mutants on the indicated substrate(s). For cases denoted as 'N', data was not available. Results and predictions are derived from *in vivo*/FBA/rFBA/(a)OT. *In vivo* data and references, FBA and rFBA predictions are taken from Ref. [19]. In five instances, predictions made by OT deviate from rFBA predictions (bold cases). See text for discussion. The growth medium contained glucose (glc), glycerol (gl), succinate (suc), acetate (ac), or ribose (rib). Anaerobic condition is denoted by '-O_2'.

pyruvate might not be suffcient to have an effect on the expression of *acs*. Or, the cell switches its uptake from acetate to pyruvate until it is depleted and then switches back to acetate again.

The incorrect prediction of the knockout of *ppc* on glycerol as nonlethal can be explained by the same argument. Gene *ppc* codes for the phosphoenolpyruvate carboxylase, which supplies the citric acid cycle with oxaloacetate

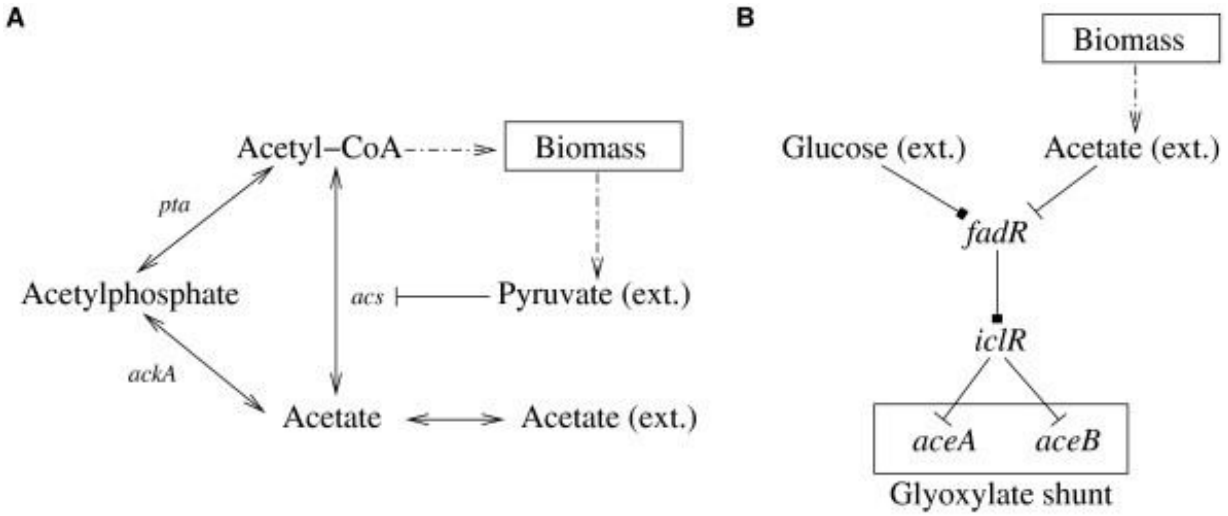


Figure 2
Illustration of knockout experiments. Panel **A**: Illustration of the sub-network that explains why deletion of *pta* or *ackA* is wrongly predicted as lethal by OT. According to the model, the alternative route via *acs* is inhibited by the presence of external pyruvate, which is always excreted when biomass is produced. Thus, there cannot be an uptake of external acetate. *In vivo*, however, the excreted pyruvate is negligible. Panel **B**: Illustration of the sub-network that explains why deletion of *ppc* is wrongly predicted as viable by OT. The glyoxylate shunt is activated if no external glucose but external acetate is present. According to the model, acetate is always excreted when biomass is produced. Therefore, the glyoxylate shunt is always upregulated, if glucose is not contained in the growth medium. Arrows indicate metabolic reactions, squares indicate activation and T-shaped lines inhibition. Dotted arrows indicate schematic reactions, which abstract a set of metabolic reactions.

(OA). When *ppc* is knocked out, the only alternative for OA production is the glyoxylate shunt, consisting of the isocitrate lyase (gene *aceA*) and the malate synthase A (gene *aceB*). However, the glyoxylate shunt is only active if *E. coli* grows on acetate or fatty acids as the sole carbon source [21]. Hence, the knockout of *ppc* on a glycerol growth medium is lethal *in vivo*, as the glyoxylate shunt is not activated. In the model, the glyoxylate shunt is regulated by the fatty acid and acetate response regulators IclR and FadR as depicted in Figure 2(B). Gene *fadR* is only expressed if external glucose, or no acetate is available in the growth media. If activated, FadR leads to an upregulation of *iclR*, which then leads to a downregulation of *aceA* and *aceB*, inactivating the glyoxylate shunt. If acetate is available in the growth medium, *fadR* is not expressed, and thus *aceA* and *aceB* are expressed at high levels, activating the glyoxylate shunt. Any organization containing the biomass precursor metabolite acetyl-CoA also contains acetate, which is secreted. Hence, any organization containing the biomass precursors also contains the external form of acetate, which activates the glyoxylate shunt.

Even though considerable amounts of acetate secretion on glycerol media has only been reported in high density cultures [22], OT indicates the possibility for *E. coli* to grow on glycerol if *ppc* is knocked out. Our result suggests that the activation of the glyoxylate shunt would be enough. There have been reports of *E. coli* strains growing on glucose media that were also able to grow with *ppc* knocked out, which is usually lethal. This was facilitated by a mutation that lead to an upregulation of the glyoxylate shunt [23]. Thus, the results of OT might be biological feasible in this case.

The problems encountered in the regulation by secreted species can be resolved by introducing a pseudo species representing the secreted version of a species. Therefore, in aOT, a metabolite that can be secreted is represented by two species: one represents the metabolite at high concentration (*e.g.*, when externally supplied), the other species represents the metabolite at low concentration (*e.g.*, when secreted). These modifications allow aOT to correctly predict the three cases discussed above. Note that the underlying assumption of this modification is also made by Covert and Palsson [19].

Discussion

By transforming the boolean formalism that represents the regulation of a metabolic network into reaction rules, we were able to demonstrate how chemical organization theory can be applied to regulated metabolic networks. Using a model of the central metabolism of *E. coli* [19],

each of the 16 wildtype growth scenarios were correctly predicted down to the presence of each protein. Each external condition could be directly mapped to a single organization implying a distinct qualitative state of the network (*i.e.*, a set of molecular species present).

Knockout experiments

Without specific assumptions, organization theory (OT) was able to predict the lethality of knockout experiments correctly in 101 out of 116 cases (87.1%). With model specific assumptions, "adapted" organization theory (aOT) was able to predict the lethality of knockout experiments correctly in 106 cases (91.4%), achieving the same performance as rFBA [19]. When comparing the performance of OT and rFBA with respect to knockout predictions, it must be noted that the model used for the comparison was co-developed with rFBA. Specific assumptions were made that were deliberatly not made in the pure organization analysis as OT shall provide an universal analysis technique applicable to general biological network models (*e.g.*, [24]). In this light, we consider the performance of organization theory as competitive.

Moreover, there are cases in other models in which predictions by OT are more accurate than those by FBA. This occurs, for example, due to the fact that FBA only incompletely takes co-factors into account. These co-factors are molecules that are necessary for some reactions to proceed. They can interact through various means with the substrates and products of a reaction. In the (unregulated) metabolic model by Reed et al. [25], we identified 10 cases in which OT correctly classifies a knockout as lethal while (r)FBA does not. The reason for this difference is that the identified co-factors are necessary for producing biomass metabolites (i.e., metabolites that appear on the left hand side of the biomass producing reaction), but they are not considered as biomass metabolites (see Supplement for details).

Influence of model specific assumptions

In our analysis of the model by Covert and Palsson [19], we found five cases in which OT predictions differ from rFBA predictions. All of these cases can be resolved by aOT in a straightforward way by taking assumptions into account also made by rFBA in [19]. In particular, the deviation between OT and (r)FBA has uncovered two critical aspects:

First, (r)FBA only considers steady states. A system state with accumulating metabolites is regarded as lethal. In OT, accumulating metabolites are explicitly allowed to also cover system states related to growth. To adopt the steady state assumption in OT however, one simply can restrict the analysis to balanced organizations in aOT. However, only considering steady states is not necessarily the best "natural" choice. As models usually are not complete, the biological system might contain pathways that are not modeled but can take care of overproduced metabolites. Also, certain molecular species accumulate in the cell at certain time points, for example in different phases of the cell cycle. Hence, states with positive productions of certain species are not necessarily lethal.

Second, in OT only the presence or absence of metabolites is considered. Hence, even smallest concentrations of species will potentially trigger further responses. When secreted metabolites have much lower concentrations than external metabolites supplied by the growth medium, this can lead to wrong predictions as secreted metabolites shall not trigger further regulation (cf. Figure 2). However, the problem can easily be resolved by introducing two species representing the metabolite at a high and at a low concentration, respectively. Alternatively, we could use the positive flow from the carbon source as a signal for the high concentration of the external metabolite (as it has been done by Covert and Palsson [19]). Note however, that with the unmodified OT we found a phenotype that indicates how to bypass the *in vivo* lethality of a knockout in one case.

Identification of all potential qualitative phenotypes

Organization theory provides a rigorous link between an organization and the potential dynamics of the reaction system (cf. Theorem 1 in Ref. [5]): if there is a steady state, then the species with positive concentrations must be an organization. Thus, we can guarantee that there is no other species combination that can give rise to a stationary state. Note that the species sets our method identifies contain also proteins, so that the organizations we obtain describe not only which metabolites are present but also which proteins are active in any possible steady state.

A related issue has been raised by Shlomi et al. [26]: The choice of a flux vector producing biomass in the FBA phase of rFBA leaves open a whole range of possible flux vectors in the space of possible solutions of rFBA. Thus, also the outcome of the experiments might depend on this choice. In contrast, OT takes all possible fluxes into account. If there are several qualitative phenotypes (i.e. sets of species with positive concentrations) consistent with the regulatory rules, several organizations will be found (see Methods).

Application to large-scale models

The largest organization does not necessarily encompass the whole system. Thus when analyzing larger models, OT allows the exact prediction of those parts of a model that can give rise to a steady state. Parts missing in such a state can then be refined. For example in the case of knockout experiments, we can determine which part of a network is

still available. Such knowledge is important when trying to reduce a metabolic model for a specific purpose, for example to increase the production of a certain metabolite. Even though the presented method can also be used for genome scale networks, it remains to be seen if all of the regulated organization in such a system could still be found. The bottleneck in the computation is the determination of all organizations of a regulated network before checking if they are consistent. Their number can grow exponentially with network size, in which case computation is primarily constrained by the available memory. This problem might be circumvented by an approach that partitions the solution space of the self-maintainance condition and searches those partitions in parallel for organizations (see Supplement for algorithmic details). However, such an approach has not yet been implemented. Currently, the analysis of a pure (non-regulated) metabolic network of genome scale by [25] is possible with the heuristic approach (see Supplement). If the number of organizations in such a network is small, all of them can be found with the heuristic approach [27]. Whether and how this scales to regulated genome-scale models is an open issue.

Other Approaches for Integrated Network Analysis

In recent years other approaches for the integration of regulatory networks into stoichiometric analysis have been proposed: Steady-state regulatory flux balance analysis (SR-FBA) [26] builds on rFBA and allows the flux analysis within one step. Boolean regulatory rules are integrated into the FBA approach by using mixed-integer linear programming. Thus, fluxes that obey the regulatory constraints can be computed directly without prior determination of the state of the regulatory network as in rFBA. In contrast to our approach SR-FBA focuses, like FBA, on specific fluxes through the network. Thus a possible flux for each possible reaction among the molecules like in organization theory (by the definition of closure), is not required for a feasible flux vector in SR-FBA. Likewise problems with co-factors like in FBA occur (see Discussion and Supplement).

Another approach, a matrix formalism to analyze regulatory systems, has been proposed by Gianchandani et al. [28]. Similar to our approach, they formulate the regulatory network by representing it as reactions in the stoichiometric matrix. Then they analyze the integrated network by using extreme pathway analysis [29]. The main difference to our work is that modelling based on the stoichiometric matrix requires a flux through the regulatory network. Thus inhibitors and activators are consumed upon interaction and are not modelled as catalysts. Even though the authors only analyzed a pure regulatory network, an integration into a metabolic network seems to be possible. Interestingly, the concept of a functional state of the resulting network, *i.e.*, the state when all external inputs are defined, resembles the organizations we found. However, those states are restricted to the regulatory part of the network, since the closure of the participating molecules is not taken into account. Another aspect of this concept that has not yet been analyzed is to which extent this method can also be applied to larger networks, for example those we analyzed in this work. The integration of the regulatory network into a metabolic network using a flux through the network might further increase the combinatorial explosion of the number of extreme pathways. Nonetheless, this approach is valuable for identifying underlying pathways in a regulatory network, a prospect which has not yet been analyzed in connection with OT.

Conclusion

Because OT does not rely on kinetics, it can serve as a first step to analyze the potential behavior of regulated systems. The analysis delivers all potential network phenotypes described by the sets of molecular species that can coexist over a long time (cf. Theorem 1 in Ref. [5]). The further analysis of the network can then focus on interesting phenotypes. Taking the other direction, it is possible to validate *in silico* network models. All phenotypes of interest observed *in vivo* should have corresponding organizations in the network model.

When regulation is considered in metabolic networks, the presented approach offers the advantage that both the metabolism and its regulation are modeled within one single framework: chemical reaction rules forming a network. The unification comes at the expense of introducing a set of pseudo species to represent the absence of species. This allows one to model and consider inhibitory interactions within the framework of organization theory. Using an appropriate user interface, the pseudo species can be easily hidden.

Authors' contributions

CK and PD introduced the idea of pseudo species indicating the absence of metabolites. All authors contributed to the presented new concepts and procedures building on the framework of organization theory. Software implementation, data preparation, and analysis was done by CK and FC. All authors read and approved the final manuscript.

Additional material

Additional file 1

Supplement including the analyzed network models and a descritption of the employed algorithms to compute organizations.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-37-S1.pdf]

Additional file 2

The analyzed core model in SBML format.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-37-S2.xml]

Additional file 3

The analyzed complete model in SBML format.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-37-S3.xml]

Acknowledgements

Financial support by the *Federal Ministry of Education and Research* (BMBF) grant 0312704A, the EU (ESIGNET, grant 12789), and the *German Research Foundation* (DFG) grant Di852/4-x is greatly acknowledged.

References

1. Varma A, Palsson B: **Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use.** *Bio/Technology* 1994, **12:**994-998.
2. Teusink B, Passarge J, Reijenga CA, Esgalhado E, Weijden CC van der, Schepper M, Walsh MC, Bakker BM, van Dam K, Westerhoff HV, Snoep JL: **Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry.** *Eur J Biochem* 2000, **267(17):**5313-5329.
3. Conrado RJ, Mansell TJ, Varner JD, DeLisa MP: **Stochastic reaction-diffusion simulation of enzyme compartmentalization reveals improved catalytic effciency for a synthetic metabolic pathway.** *Metab Eng* 2007, **9(4):**355-363.
4. Schuster S, Fell DA, Dandekar T: **A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks.** *Nat Biotechnol* 2000, **18(3):**326-332.
5. Dittrich P, Speroni di Fenizio P: **Chemical organization theory.** *B Math Biol* 2007, **69(4):**1199-1231.
6. Kaleta C, Centler F, Dittrich P: **Analyzing molecular reaction networks: from pathways to chemical organizations.** *Mol Biotechnol* 2006, **34(2):**117-123.
7. Bochner BR: **New technologies to assess genotype-phenotype relationships.** *Nat Rev Genet* 2003, **4(4):**309-314.
8. Centler F, Speroni di Fenizio P, Matsumaru N, Dittrich P: **Chemical organizations in the central sugar metabolism of *Escherichia Coli*.** In *Mathematical Modeling of Biological Systems Volume I*. Edited by: Deutsch A, Brusch L, Byrne H, de Vries G, Herzel HP. Birkhäuser, Boston; 2007:109-123.
9. Fontana W, Buss LW: **'The Arrival of the Fittest': Towards a Theory of Biological Organization.** *B Math Biol* 1994, **56:**1-64.
10. Covert MW, Schilling CH, Palsson B: **Regulation of gene expression in flux balance models of metabolism.** *J Theor Biol* 2001, **213:**73-88.
11. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER: **A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles.** *Plant Cell* 2004, **16(11):**2923-2939.
12. Segall JE, Manson MD, Berg HC: **Signal processing times in bacterial chemotaxis.** *Nature* 1982, **296(5860):**855-857.
13. Hargrove JL, Hulsey MG, Beale EG: **The kinetics of mammalian gene expression.** *Bioessays* 1991, **13(12):**667-674.
14. de Jong H: **Modeling and simulation of genetic regulatory systems: a literature review.** *J Comput Biol* 2002, **9:**67-103.
15. Thomas R: **Boolean formalization of genetic control circuits.** *J Theor Biol* 1973, **42(3):**563-585.
16. Kauffman SA: *The Origins of Order: Self-Organization and Selection in Evolution* Oxford University Press; 1993.
17. Turner TE, Schnell S, Burrage K: **Stochastic approaches for modelling in vivo reactions.** *Comput Biol Chem* 2004, **28(3):**165-178.
18. Mendes P, Kell D: **Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation.** *Bioinformatics* 1998, **14(10):**869-883.
19. Covert MW, Palsson B: **Transcriptional regulation in constraints-based metabolic models of Escherichia coli.** *J Biol Chem* 2002, **277(31):**28058-28064.
20. Matsumaru N, Centler F, Speroni di Fenizio P, Dittrich P: **Chemical Organization Theory as a Theoretical Base for Chemical Computing.** *Int J Unconv Comp* 2007, **3(4):**285-309.
21. Maloy SR, Nunn WD: **Genetic regulation of the glyoxylate shunt in Escherichia coli K-12.** *J Bacteriol* 1982, **149:**173-180.
22. Korz DJ, Rinas U, Hellmuth K, Sanders EA, Deckwer WD: **Simple fed-batch technique for high cell density cultivation of Escherichia coli.** *J Biotechnol* 1995, **39:**59-65.
23. Sauer U, Eikmanns BJ: **The PEP-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria.** *FEMS Microbiol Rev* 2005, **29(4):**765-794.
24. Novère NL, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M: **BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems.** *Nucleic Acids Res* 2006:D689-D691.
25. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).** *Genome Biol* 2003, **4(9):**R54.
26. Shlomi T, Eisenberg Y, Sharan R, Ruppin E: **A genome-scale computational study of the interplay between transcriptional regulation and metabolism.** *Mol Syst Biol* 2007, **3:**101.
27. Centler F, Kaleta C, di Fenizio PS, Dittrich P: **Computing Chemical Organizations in Biological Networks.** *Bioinformatics (accepted)* 2008.
28. Gianchandani EP, Papin JA, Price ND, Joyce AR, Palsson BO: **Matrix formalism to describe functional states of transcriptional regulatory systems.** *PLoS Comput Biol* 2006, **2(8):**e101.
29. Schilling CH, Letscher D, Palsson BO: **Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective.** *J Theor Biol* 2000, **203(3):**229-248.
30. Creaghan IT, Guest JR: **Succinate dehydrogenase-dependent nutritional requirement for succinate in mutants of Escherichia coli K12.** *J Gen Microbiol* 1978, **107:**1-13.
31. Cronan JE, LaPorte D: **Tricarboxylic acid cycle and glyoxylate bypass.** In *Escherichia coli and Salmonella: cellular and molecular biology Volume 1*. Edited by: Neidhardt FC, III RC, Ingraham JL, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M, Umbarger HE. Washington, D.C: ASM Press; 1996:206-216.
32. Langley D, Guest JR: **Biochemical genetics of the alpha-keto acid dehydrogenase complexes of Escherichia coli K12: isolation and biochemical properties of deletion mutants.** *J Gen Microbiol* 1977, **99(2):**263-276.
33. Kumari S, Tishel R, Eisenbach M, Wolfe AJ: **Cloning, characterization, and functional expression of acs, the gene which encodes acetyl coenzyme A synthetase in Escherichia coli.** *J Bacteriol* 1995, **177(10):**2878-2886.
34. Gruer MJ, Bradbury AJ, Guest JR: **Construction and properties of aconitase mutants of Escherichia coli.** *Microbiology* 1997, **143(Pt 6):**1837-1846.
35. Cunningham PR, Clark DP: **The use of suicide substrates to select mutants of Escherichia coli lacking enzymes of alcohol fermentation.** *Mol Gen Genet* 1986, **205(3):**487-493.
36. Calhoun MW, Oden KL, Gennis RB, de Mattos MJ, Neijssel OM: **Energetic effciency of Escherichia coli: effects of mutations in components of the aerobic respiratory chain.** *J Bacteriol* 1993, **175(10):**3020-3025.

37. Irani MH, Maitra PK: **Properties of Escherichia coli mutants deficient in enzymes of glycolysis.** *J Bacteriol* 1977, **132(2):**398-410.
38. Fraenkel DG: **Glycolysis.** In *Escherichia coli and Salmonella: cellular and molecular biology Volume 1*. Edited by: Neidhardt FC, III RC, Ingraham JL, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M, Umbarger HE. Washington, D.C: ASM Press; 1996:189-198.
39. Fraenkel DG, Horecker BL: **Fructose-1, 6-diphosphatase and acid hexose phosphatase of Escherichia coli.** *J Bacteriol* 1965, **90(4):**837-842.
40. Courtright JB, Henning U: **Malate dehydrogenase mutants in Escherichia coli K-12.** *J Bacteriol* 1970, **102(3):**722-728.
41. Tran QH, Bongaerts J, Vlad D, Unden G: **Requirement for the proton-pumping NADH dehydrogenase I of Escherichia coli in respiration of NADH to fumarate and its bioenergetic implications.** *Eur J Biochem* 1997, **244:**155-160.
42. Mat-Jan F, Alam KY, Clark DP: **Mutants of Escherichia coli deficient in the fermentative lactate dehydrogenase.** *J Bacteriol* 1989, **171:**342-348.
43. Sørensen KI, Hove-Jensen B: **Ribose catabolism of Escherichia coli: characterization of the rpiB gene encoding ribose phosphate isomerase B and of the rpiR gene, which is involved in regulation of rpiB expression.** *J Bacteriol* 1996, **178(4):**1003-1011.
44. Anderson A, Cooper R: **Gluconeogenesis in Escherichia coli The role of triose phosphate isomerase.** *FEBS Lett* 1969, **4:**19-20.

BIOINFORMATICS **ORIGINAL PAPER** Vol. 25 no. 15 2009, pages 1915–1922
doi:10.1093/bioinformatics/btp332

Systems biology

Using chemical organization theory for model checking

Christoph Kaleta[1,2,3,†], Stephan Richter[1,2,†] and Peter Dittrich[1,2,*]

[1]Bio Systems Analysis Group, Jena Centre for Bioinformatics (JCB), [2]Department of Mathematics and Computer Science and [3]Department of Bioinformatics, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

Received on July 22, 2008; revised on May 4, 2009; accepted on May 17, 2009

Advance Access publication May 25, 2009

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: The increasing number and complexity of biomodels makes automatic procedures for checking the models' properties and quality necessary. Approaches like elementary mode analysis, flux balance analysis, deficiency analysis and chemical organization theory (OT) require only the stoichiometric structure of the reaction network for derivation of valuable information. In formalisms like Systems Biology Markup Language (SBML), however, information about the stoichiometric coefficients required for an analysis of chemical organizations can be hidden in kinetic laws.

Results: First, we introduce an algorithm that uncovers stoichiometric information that might be hidden in the kinetic laws of a reaction network. This allows us to apply OT to SBML models using modifiers. Second, using the new algorithm, we performed a large-scale analysis of the 185 models contained in the manually curated BioModels Database. We found that for 41 models (22%) the set of organizations changes when modifiers are considered correctly. We discuss one of these models in detail (BIOMD149, a combined model of the ERK- and Wnt-signaling pathways), whose set of organizations drastically changes when modifiers are considered. Third, we found inconsistencies in 5 models (3%) and identified their characteristics. Compared with flux-based methods, OT is able to identify those species and reactions more accurately [in 26 cases (14%)] that can be present in a long-term simulation of the model. We conclude that our approach is a valuable tool that helps to improve the consistency of biomodels and their repositories.

Availability: All data and a JAVA applet to check SBML-models is available from

`http://www.minet.uni-jena.de/csb/prj/ot/tools`

Contact: dittrich@minet.uni-jena.de

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

1 INTRODUCTION

Reaction networks are widely used to model biological systems at various levels, including the molecular level (Le Novére *et al.*, 2006), the cellular level (Wodarz and Nowak, 1999), the ecological scale (Hofbauer and Sigmund, 1998) and the level of social interactions (Dittrich *et al.*, 2003). Because of the growing number of models, methods for their verification are needed. Approaches that can be used for this purpose are flux balance analysis (FBA; Varma and Palsson, 1994), elementary mode analysis (Schuster *et al.*, 1999), extreme pathway analysis (Schilling *et al.*, 2000) and chemical organization theory (OT; Dittrich and Speroni di Fenizio, 2007). These methods have in common that they allow deriving constraints to dynamic behavior from the stoichiometric structure of a network. This stoichiometric structure is defined by the number of molecules of educts as well as the products each reaction consumes and produces. Thus, stoichiometry-based methods do not require a precise knowledge of the underlying reaction kinetics, which are often partly or totally unknown.

Here, we will use OT. An important property of chemical organizations is that every steady state and growth state[1] of a network corresponds to a chemical organization (Dittrich and Speroni di Fenizio, 2007, and Supplementary Material). These states we call the limit behavior of a model. However, this property is fulfilled only if a reaction network meets a condition formulated by Feinberg and Horn (1974): each reaction has a non-zero flux if and only if all of its educts have a positive concentration. Using this property, OT has already been applied to the prediction of growth phenotypes (Centler *et al.*, 2007) and the outcome of knockout experiments (Kaleta *et al.*, 2008), as well as in the design of chemical programs to solve NP-complete problems (Matsumaru *et al.*, 2007).

In a recent work, we used OT to assess the quality of a genome-scale reaction network of *Escherichia coli* by identifying species and reactions that could not be present in the limit behavior of the model during simulation (Centler *et al.*, 2008). We concluded that these species and reactions hint at missing knowledge as they were mostly part of pathways starting from or ending in dead-end species. Here, we want to extend this approach in two directions. First, we present a method for more accurately predicting the limit behavior of a reaction network if information on reactions kinetics is available. If modeled in Systems Biology Markup Language (SBML, Hucka *et al.*, 2003), the velocity of a reaction depends on the concentration of its educts, products and modifiers. A modifier is a species whose concentration affects the reaction velocity, but whose concentration itself is not changed by this reaction. Some modifiers, as for example catalysts or activators, are required to be present

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[1]As growth state we define a situation where some species accumulate. An example is exponential growth in which, for instance, the overall amount of DNA increases given that there is a continuous supply (inflow) of nutrients.

for a non-zero reaction velocity. However, if we want a reaction to fulfill the Feinberg condition, such modifiers need to be added on its educt and product sides. Hence, since information necessary for the analysis using OT can be hidden in the kinetic laws, we present an algorithm for extracting this information. Second, using this approach, we demonstrate how knowledge of the organizational structure of a reaction network and thus of its limit behavior can help to uncover modeling inconsistencies. These inconsistencies are represented by species as well as by reactions that belong to no organization, indicating either incomplete knowledge, compounds missing from the specified growth media or modeling errors.

This work is structured as follows. In Section 2, we give a short outline of OT and present an algorithm that modifies the stoichiometric structure in a reaction network such that the Feinberg condition is fulfilled. We use this algorithm in Section 3 to demonstrate how these modifications affect the organizational structure of a model of the extracellular signal related-kinase (ERK)/Wnt-signaling pathway. In Section 4, we use our approach to find inconsistencies in a large-scale analysis of the models of the BioModels Database (Le Novére *et al.*, 2006) and compare our results with those obtained by other stoichiometric analysis techniques. Finally, we conclude in Section 5.

2 METHODS

2.1 Chemical OT

We define a reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$ by a set of molecular species $\mathcal{M}$ and a set of reaction rules $\mathcal{R}$. A reaction rule $\rho \in \mathcal{R}$ is defined by the stoichiometric coefficients $l_{i,\rho}$ and $r_{i,\rho}$ denoting the left- and right-hand sides of a reaction rule, respectively. Given a reaction rule $\rho \in \mathcal{R}$, we denote the set of reactant species and set of product species by $\mathrm{LHS}(\rho) := \{i \in \mathcal{M} | l_{i,\rho} > 0\}$ and $\mathrm{RHS}(\rho) := \{i \in \mathcal{M} | r_{i,\rho} > 0\}$, respectively. With $\mathbf{N} = (n_{i,\rho}) = (r_{i,\rho} - l_{i,\rho})$, we denote the stoichiometric matrix of $\langle \mathcal{M}, \mathcal{R} \rangle$. W.l.o.g. we assume $\mathbf{v}_\rho \geq 0$; hence a reversible reaction has two entries in v.

Given a set $A \subseteq \mathcal{M}$, its set of reaction rules $\mathcal{R}_A = \{\rho \in \mathcal{R} | \mathrm{LHS}(\rho) \subseteq A\}$, and the corresponding stoichiometric matrix $\mathbf{N}_A$, we say that A is closed if for all reaction rules $\rho \in \mathcal{R}_A$, $\mathrm{RHS}(\rho) \in A$. Thus, we call A closed if there is no reaction with educts from A producing a species not in A. A is self-maintaining if there exists a strictly positive flux vector $\mathbf{v}' \in \mathbb{R}_{>0}^{|\mathcal{R}_A|}$ such that all species in A are produced at a non-negative rate, that is, $\mathbf{N}_A \mathbf{v}' \geq 0$ (Dittrich and Speroni di Fenizio, 2007). A set A that is closed and self-maintaining is called an organization (Fontana and Buss, 1994). An organization is called reactive if each of its species participates in at least one reaction of that organization. Elementary organizations are reactive organization that cannot be generated as union of other reactive organizations (Centler *et al.*, 2008).

Because organizations may share the same species, the set of organizations together with the set inclusion $\subseteq$ form a partially ordered set that can be visualized in a Hasse diagram, providing a hierarchical view of the network under consideration: organizations are vertically arranged by size, with small organizations at the bottom. Two organizations are connected by a line, if the upper contains the lower organization and no other organization exists between them. For simplicity, only species appearing for the first time, i.e. which are not element of a lower organization, are displayed.

2.2 Analyzing reaction networks with modifiers

In this section, we introduce an algorithm that allows application of OT to reaction network models containing modifiers. As an example, we use a phosphorylation cycle, a typical motive found in signaling networks (Fig. 1**A**). The network consists of seven molecular species $\mathcal{M} = \{A, B, M1, \ldots M5\}$ and three reactions $\mathcal{R} = \{R1, R2, R3\}$.

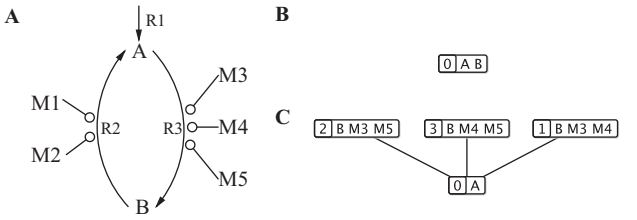


Fig. 1. (**A**) Example network (phosphorylation cycle) with seven species and three reaction rules. (**B** and **C**) Hasse diagrams of elementary organizations of the unprocessed and processed networks, respectively. Only species appearing for the first time in each organization are displayed.

For the reactions $\mathcal{R} = \{ R1 : \emptyset \to A,\ R2 : B \to A,\ R3 : A \to B \}$, we assume the following kinetic laws (omitting rate constants and units):

$$v_{R1} = 1$$
$$v_{R2} = [B](1 + [M1] + [M2])$$
$$v_{R3} = [A]([M3][M4] + [M3][M5] + [M4][M5])$$

This model can be formulated in SBML, with $M1, M2$ being modifiers of reaction $R2$ and $M3, M4, M5$ being modifiers of reaction $R3$, while not appearing as reactants.

Our algorithm consists of two steps: first, we examine the kinetic law of each reaction to detect minimal sets of modifiers that are necessary for that reaction to have a positive flux. Then, we use this information to adapt a reaction's set of reactants in order to more faithfully reflect the algebraic structure of the network used for computation of chemical organizations.

2.2.1 Step 1: identifying sets of essential modifiers In this first step, we identify all minimal supporting modifier sets of each reaction. Given a reaction $\rho \in \mathcal{R}$, a *minimal supporting modifier set* (*supporting set*, for short) is defined as a minimal set of modifiers that need positive concentrations (while all others are absent) to allow reaction ρ to have a positive flux. If at least one of these modifiers is additionally set to a zero concentration, the flux of the reaction is constrained to zero. There might be several possibly overlapping supporting sets. With respect to a certain reaction, a modifier is called *essential* if it is contained in all supporting sets of the reaction.

Determination of supporting sets: to decide whether a set of modifiers is a supporting set for a particular reaction, we follow a straightforward approach. If a set of modifiers is a supporting set, a positive concentration of only these modifiers allows a non-zero flux, while a positive concentration of only a proper subset of these modifiers constrains the flux to zero. Following this idea, we implemented FormulaChecker, which tries to compute the velocity of each reaction in terms of modifier concentrations. All variables in the kinetic law that represent undefined parameters or educt or product species are not further resolved; i.e. they are treated as symbols. The modifiers we want to test to determine whether they belong to a supporting set are also treated as symbols. The remaining modifiers are set to zero concentration. Function calls are resolved by application of their respective parameters, if necessary. Applying FormulaChecker can lead to two different results for the reaction velocity:

(1) The result is zero. In this case the tested modifier set is not a supporting set. Let $\{M3\}$, for example, be the set to be checked in $R3$. Setting the concentrations of the remaining modifiers to zero results in $v_{R_3} = 0$. Thus, $\{M3\}$ is not a supporting set of $R3$. This also applies to the sets $\{M4\}$ and $\{M5\}$.

(2) The result is non-zero. Thus, it might be a constant only depending on parameters, or a formula, dependent on variables. Checking $\{M3, M4\}$ in $R3$ yields the kinetic law $v_{R_3} = [A]([M3][M4])$. Since we know that $\{M3\}$, $\{M4\}$ and $\{M5\}$ do not represent supporting sets, $\{M3, M4\}$

has to be a supporting set. In contrast, if we check the empty set in $R2$ by setting $M1$ and $M2$ to zero values in the kinetic law, we obtain $v_{R_2} = [B]$. In consequence, neither $\{M1\}$ nor $\{M2\}$ represent supporting sets of R2; the supporting set is the empty set, and no further tests are required.

Finding all supporting sets: in order to find all supporting sets of a reaction, the algorithm analyzes the power set of the reaction's set of modifiers to ensure that all supporting sets are found. The sets are checked in increasing size order, trying to avoid testing the whole power set of modifiers. If we find that a set of modifiers is a minimal supporting set, we do not need to test any of its supersets.

Looking at $R3$ in the example, after the empty set, all single-modifier sets are checked. We find that neither $M3$ nor $M4$ nor $M5$ allow a positive flux if standing alone. In the next step all two-element sets are tested. Since all these sets allow a positive flux of $R3$, but none of the smaller ones, we conclude that $\{M3, M4\}$, $\{M3, M5\}$ and $\{M4, M5\}$ are the supporting sets. In consequence, we do not have to test the superset $\{M3, M4, M5\}$.

2.2.2 Step 2: adapting the reactions In the second step, each reaction possessing at least one supporting set is processed. For each supporting set the reaction is duplicated and the modifiers of the supporting set are added as catalysts to the duplicate reaction. Finally, the original reaction is removed from the model. In order to preserve the dynamics of the original model in the processed model, the kinetic law of each of the duplicate reactions is divided by the number of derived reactions, i.e. the number of supporting sets. The duplicate reactions get new names of the form `[old_reaction_name] variant [number].`

For our example, we obtain the following set of reaction rules $\mathcal{R} = \{R1, R2, R3_{\text{variant 1}}, R3_{\text{variant 2}}, R3_{\text{variant 3}}\}$ with

$$R1 : \rightarrow A$$

$$R2 : B \rightarrow A$$

$$R3_{\text{variant 1}} : A + M3 + M4 \rightarrow B + M3 + M4$$

$$R3_{\text{variant 2}} : A + M3 + M5 \rightarrow B + M3 + M5$$

$$R3_{\text{variant 3}} : A + M4 + M5 \rightarrow B + M4 + M5$$

For a more detailed outline of the processing of the kinetic laws, see the Supplementary Material.

2.2.3 Example application Applying the algorithm to our example, we can see several effects of the processing of the kinetic laws (see Figs 1B and C for the Hasse diagrams of elementary organizations). Two trends are superimposed. First, some organizations vanish, including the organization solely containing A and B in the unprocessed network. In the processed network, a reaction still converts B to A. In order to replenish B, one pair of the modifiers *M3, M4* and *M5* is necessary. Thus, $\{A, B\}$ does not fulfill the self-maintenance condition in the processed network. Second, some organizations appear for the first time, as in the case of the organization containing A in the processed network. In the original network, the set $\{A\}$ was not closed since $R3$ unconditionally produced B from A.

3 ORGANIZATIONAL STRUCTURE OF THE ERK/WNT-SIGNALING PATHWAY

In order to demonstrate the utility of the incorporation of kinetic laws into the analysis with OT, we analyze the model *BIOMD149*[2] from the BioModels database (Le Novére *et al.*, 2006) containing an integrated ERK and Wnt/β-catenin signaling pathway (Figure 2).

[2]We abbreviate the official name of the BioModels by reducing the number to three digits. The original name of the model is *BIOMD0000000149*.

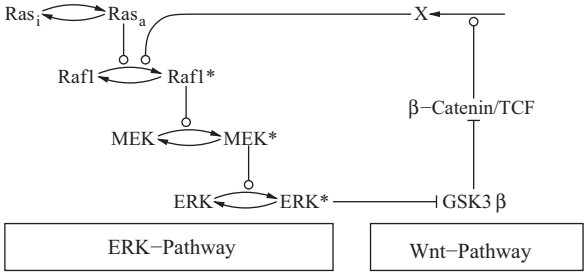


Fig. 2. Simplified representation of the reaction network from *BIOMD149* (Kim *et al.*, 2007) combining the ERK- and Wnt-signaling pathways. The Wnt signal, serving as input to both pathways, is not shown. Lines with circles represent essential modifiers identified with the presented approach. Lines ending in orthogonal bars indicate inhibition.

This model is based on the work of Kim *et al.* (2007), who described a positive feedback loop between these two pathways important in the development of some cancer. The positive feedback loop works through a yet unknown mechanism modeled by a species called 'molecule X'. The transcription of this molecule is modeled to be upregulated by a complex of β-catenin and T-cell factor (TCF). The availability of β-catenin is regulated by active glycogen synthase kinase 3β (GSK-3β), which in turn is inactivated by phosphorylated ERK. According to the model, X upregulates the signaling through the ERK-pathway. The rates of phosphorylation of the different levels of the ERK-pathway are modeled with kinetic laws. Thus, a high concentration of phosphorylated Raf increases the rate of phosphorylation of MEK, which in turn increases the rate of phosporylation of ERK.

Without the processing of the kinetic laws the network contains 384 reactive organizations generated from the union of 11 elementary organizations. After processing, the network contains 150 reactive organizations generated from the union of 18 elementary organizations. Thus, the number of reactive organizations declines, while the number of elementary organizations increases. Figures 3 and 4 depict the Hasse diagram of elementary organizations of both networks. The Hasse diagram of the unprocessed network (Fig. 3) displays a very simple structure. The smallest organization already contains X. From the kinetic law of the production reaction of X, it can be determined that a positive concentration of the complex β-catenin/TCF is required for a non-zero flux of this reaction. But this is not taken into account since this constraint is modeled through the modifiers of the reaction and not on the level of substrates and educts as required by the Feinberg condition. Consequently, the different levels of the ERK-signaling pathway are also present independent of each other. This can be observed by the presence of the corresponding phosphorylated and dephosphorylated proteins directly above the smallest organization in the Hasse diagram.

From a simulation perspective, the reactive organizations of the original network would indicate a state of the network where, for example, MEK and MEK* as well as the input species could be constantly present (Fig. 3, organization 4). However, by examining the kinetic laws of the phosphorylation from MEK to MEK*, we find that this reaction has a flux of zero if the species Raf1* is not present. Thus, only the dephosphorylation of MEK would have a positive flux, finally using up all MEK*. After

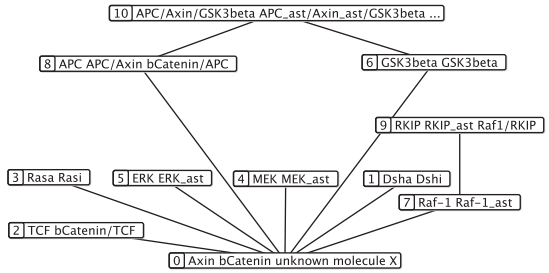


Fig. 3. Hasse diagram of elementary organizations of *BIOMD149* without processing of the kinetic laws. Only species appearing for the first time in each organization are shown. For example organization 9 contains the species displayed in the nodes corresponding to organization 0, 7 and 9. Not all species in organization 10 are displayed. A list of abbreviations can be found in the Supplementary Material. Phosphorylated forms of a protein are denoted by the suffix '_ast'. Active/Inactive forms by the suffix 'a'/'i'.

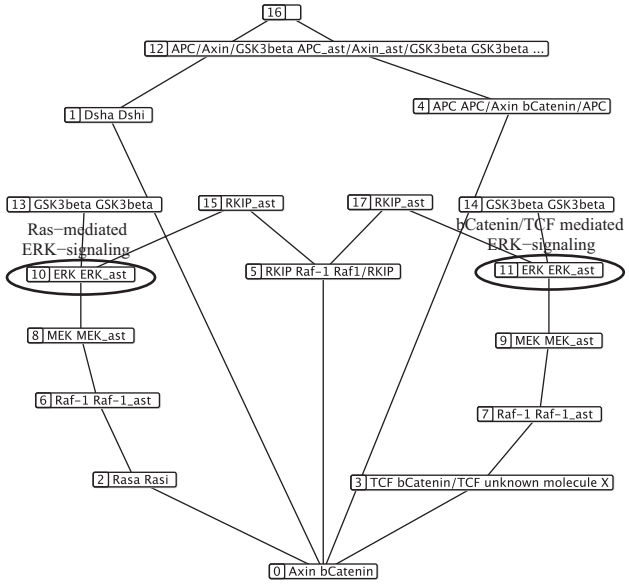


Fig. 4. Hasse diagram of elementary organizations of *BIOMD149* after the processing of the kinetic laws. Only species appearing for the first time in each organization are shown. Not all species in organization 12 are displayed. Naming follows the same conventions as in Figure 3. The different pathways for upregulation of the ERK-signaling pathway are indicated. In comparison to Figure 3, we find, for example, the node corresponding to organization 6 above the node corresponding to organization 2 (corresponding to the nodes labeled 7, respectively, 3 in Fig. 3). This corresponds to the conclusion that a positive concentration of Rasa and Rasi is required for the presence of Raf1 and Raf1* in the limit behavior. Comparison with Figure 3 shows that this conclusion can be drawn only if the kinetic laws are processed.

processing of the corresponding kinetic law, Raf1* is identified as an essential modifier and added as a catalyst to the reaction, as seen in the Hasse diagram of the processed network (Fig. 4). The organization containing the species MEK* and MEK (Fig. 4, organization 8) is situated above the organization containing Raf1* (Fig. 4, organization 6).

From this perspective, the processing of the kinetic laws can be seen as adding mechanistic detail to the reactions. Thus, when we find Raf1* necessary for the phosphorylation of MEK to MEK*, the addition of the modifier Raf1* as catalyst corresponds to the complex formation between Raf1* with MEK prior to phosphorylation. The approach to consider kinetic laws in OT can be seen as refinement of the reactions of a model making use of the additional information present in kinetic laws. Even though OT does not explicitly require the kinetic laws of a reaction network, knowledge about them can be used to better predict the limit behavior of a reaction network. Conversely, in the sense of the Feinberg condition, the underlying mechanisms are modeled more accurately on the stoichiometric level of the network if this approach is used.

In agreement with the results of Kim *et al.* (2007), we find an alternative route for the activation of the ERK-pathway, indicated by the organizations 3, 7, 9 and 11 in Figure 4. Through the action of the complex β-catenin/TCF, the transcription of X is upregulated and, thus, bypasses the activation of Raf by Ras. A constant activation of β-catenin/TCF, for example through a mutation, can result in a decoupling from any signal and consequently lead to a constant upregulation of the ERK-signaling pathway, as is often found in cancer (Kim *et al.*, 2007). In the unprocessed network, we do not obtain these results.

4 LARGE-SCALE ANALYSIS OF BIOMODELS

In order to demonstrate the utility of our approach, we analyze the models of the 11th release[3] of the BioModels database (Le Novére *et al.*, 2006). This database contains 185 manually curated models of biological networks in SBML format.

SBML allows species to be defined as external. Thus, their concentration is assumed constant. For the computation of chemical organizations, we add an inflow and outflow reaction of the form $\emptyset \to s$ and $s \to \emptyset$ for each external species s. For all except 3 models, we were able to compute the reactive organizations using the deterministic algorithms for organization computation [see Centler *et al.* (2008) for algorithmic details]. For the remaining three models (*BIOMD014*, *BIOMD019* and *BIOMD049*), a heuristic based on a random walk strategy to determine organizations (Centler *et al.*, 2008) needed to be applied. Since we wanted to identify species appearing in no organization and each of these models did contain an organization encompassing the entire species set, computation of the complete set of organization was not necessary for these models.

A total of 172 models contained a non-empty organization. In the remaining 13 models only the empty organization was found, since they contained neither reactions nor species. An overview of the number of reactive organizations is given in Table 2 in the Supplementary Material. While 77 models contained only a single reactive organization, the highest number of organizations was found in *BIOMD175*, with 319 248 reactive organizations. An overview of the distribution of the number of organizations can be found in Figure 5.

Species participating in none of the reactions can drastically increase the number of organizations in a network. Thus, we computed only the reactive organizations in each network and omitted species participating in no reaction (in 24 models) from the analysis. In 31 models some species did not appear in

[3]The BioModels Database is updated in releases whereby models are corrected or added. We downloaded the models used in this work on October 20, 2008.

any reactive organization. A first analysis showed that this set contains many models where such behavior was intended. Thus, in several models the concentration of some species was set to a non-zero value at a given time point (e.g. $t=0$). To take into account this short-time behavior, we added an inflow reaction for each such species. Doing this, we found that only five models with species absent from any reactive organization remained: *BIOMD044, BIOMD093, BIOMD094, BIOMD143* and *BIOMD151* (Table 1). By analyzing the reactions in which the missing species participated and comparing the SBML models to their description in the corresponding publications, we found the potential inconsistencies. We identified all these inconsistencies as actual modeling errors.

4.1 Resolving network inconsistencies

In three of the five models, *BIOMD093, BIOMD094* and *BIOMD143*, we identified reactions that were set to irreversible despite their kinetic laws producing negative fluxes in the course of the simulation, as described in the corresponding publications. Thus, they were indeed reversible and we modified them accordingly. Repeating the analysis, we found all species present in the reactive organizations of *BIOMD093*. In *BIOMD094*, missing species remained. However, this was an intended behavior since a gene knockout was modeled (Yamada *et al.*, 2003).

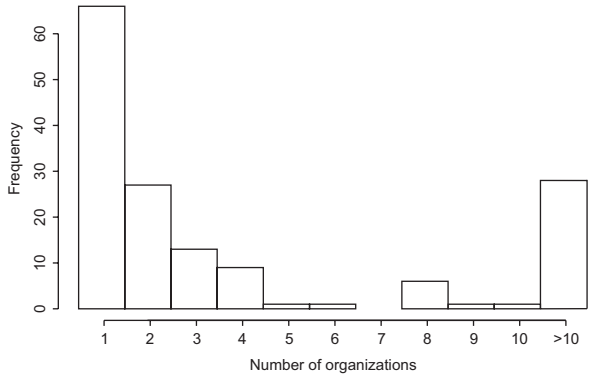


Fig. 5. Histogram of the number of reactive organizations in the models of the BioModels Database. Please note that this number includes six models with more than 1000 organizations (listed below >10 organizations).

In *BIOMD143,* we still found some species absent after we had changed reactions with negative fluxes in the simulation to be reversible. This model describes the oscillatory metabolism of activated neutrophils (Olsen *et al.*, 2003). A simplified and decompartmentalized version of the relevant reactions is depicted in Figure 6. The species absent from the reactive organizations are hydrogen from cytoplasm and phagosome. The model contains only reactions consuming these two species. During simulation even negative concentrations of both species appear. The reason for the consumption of these species is inconsistent modeling of the stoichiometry of the reactions and an inconsistent kinetic law. Cytoplasmatic and phagosomal hydrogen are consumed together with superoxide (O_2^-) to produce hydrogen peroxide (H_2O_2). In the course of the disposal of H_2O_2 by ferric peroxidase in the phagosome, additional four protons from melatonin (MLTH) are consumed to produce the initial form of ferric peroxidase. With the exception of ferric peroxidase and free radicals of melatonin (MLT), all species are consumed without producing equivalent products. Thus, the disposal of H_2O_2 by ferric peroxidase consumes oxygen and protons. The model contains an inflow for NADPH and O_2. Oxidation of NADPH by oxygen or free radicals of melatonin can produce superoxide and melatonin, respectively. Thus, there is a constant inflow of NADPH and oxygen that can replenish the consumed species. However, the kinetic law of the production of superoxide from O_2^- and hydrogen does not depend on the concentration of hydrogen in the model. Together with a zero initial concentration of hydrogen, the simulation of the model leads to a negative concentration of this species. Making the rate law dependent on the concentration of hydrogen resolves the problem of negative concentration of hydrogen. Additionally, either removing the inconsistencies in the stoichiometry or adding an inflow for

Table 1. Selected results from the large-scale analysis

Model	Description	Species/ Reactions	Reactive Orgs.	First Step		Second Step	
				OT (spec./rea.)	FBM (rea.)	OT (spec./rea.)	FBM (rea.)
BIOMD037	Sporulation control network in *P. polycephalum* (Goldbeter, 1991)	12/14 (14)	1 (2)	3/6	10	12/14	14
BIOMD044	Model of intracellular calcium oscillations (Borghans *et al.*, 1997)	7/8 (8)	2 (2)	3/4	5	6/7	7
BIOMD093	JAK/STAT signal transduction pathway (Yamada *et al.*, 2003)	34/48 (48)	5 (3)	11/16	30	31/43	43
BIOMD094	JAK/STAT signal transduction pathway (Yamada *et al.*, 2003)	34/47 (47)	2 (3)	5/5	27	24/24	40
BIOMD143	Oscillatory metabolism of activated neutrophils (Olsen *et al.*, 2003)	20/20 (20)	1 (1)	4/4	4	7/5	5
BIOMD149	Crosstalk between Wnt and ERK Pathways (Kim *et al.*, 2007)	28/39 (39)	150 (384)	28/39	39	–	–
BIOMD151	IL-6 signal transduction in hepatocytes (Singh *et al.*, 2006)	68/114 (114)	80 (96)	49/71	111	19/14	112

See Supplementary Material for the entire table. The five models in which inconsistencies have been identified are shaded in light gray. The first 4 columns give general details about the models. Numbers in brackets indicate the number of reactions of the original network that can increase through processing of the kinetic laws. The number of species remains constant. The fourth column gives the number of reactive organizations in the modified and (in brackets) the original network. In the fifth and sixth columns species and reactions that can be present in the limit behavior of the processed network are given. OT denotes the predictions by OT, and FBM the predictions by flux-based methods. In some cases, FBM identifies more reactions to be present in the limit behavior than OT. These cases are shaded in dark gray. The seventh and eighth columns give the same numbers when inflow reactions for species with an event setting their concentration to a positive value at a certain time point are added. In cases where the original network already contained all species, those numbers are omitted.

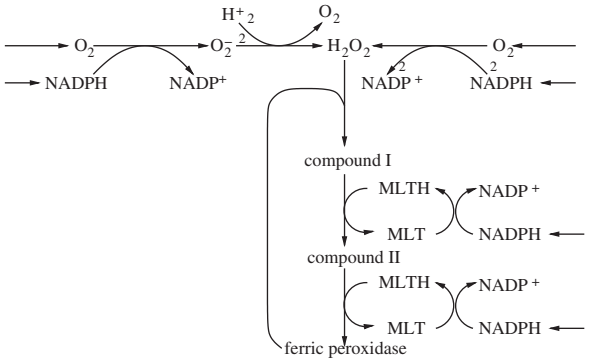


Fig. 6. Simplified representation of the reactions of *BIOMD143*. As a result of inconsistent stoichiometries hydrogen and oxygen are consumed in the course of detoxification of hydrogen peroxide. There is only an inflow of oxygen, and the consumption of hydrogen does not depend upon its concentration. Consequently, a simulation leads to a negative concentration of hydrogen.

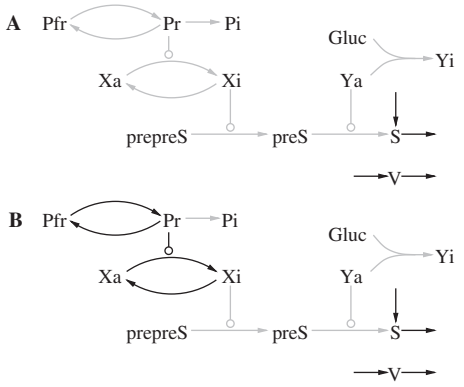


Fig. 7. Reaction network from *BIOMD037* modeling the sporulation control in *P.polycephalum* from Marwan (2003). Lines ending in circles indicate essential modifiers for a reaction. Light gray reactions cannot have a positive flux in the limit behavior, according to **A** OT and **B** flux-based methods. Abbreviations: Pr, active photoreceptor; Pi, inactive photoreceptor; (pre/prepre)S, sporulation signal (and precursors); Ya/i, active/inactive glucose receptor; Gluc, glucose; Xa/i, active/inactive signal transducer.

hydrogen allows positive concentrations of this species during simulation.

The reasons for the missing species in *BIOMD044* are very similar. Here, a species is modeled to serve as a pseudo-substrate to a reaction that could have been modeled without substrate. The kinetic law governing the reaction does not depend upon the concentration of this substrate. Since it is not produced by any other reaction, negative concentrations appear in the course of the simulation. Replacing the respective reaction by an inflow reaction resolves the problem.

In *BIOMD151* almost all species are absent from reactive organizations. This network represents an integrated model of the JAK/STAT and ERK-signaling pathways regulated by IL-6 in hepatocytes (Singh *et al.*, 2006). A detailed analysis of the model and the set of ordinary differential equations presented in Singh *et al.* (2006) showed that a complex formation step was missing, such that the signal from IL-6 could not be transmitted to the subsequent signaling pathways. Only the complex dissociation reaction was present. During simulation it had a negative flux, mimicking the complex formation reaction. Adding the missing step produced a model in which all species appeared in a reactive organization.

4.2 Comparison with flux-based methods

Next, we will compare our results with those obtained with flux-based methods, including FBA (Varma and Palsson, 1994), elementary mode analysis (Schuster *et al.*, 1999) and extreme pathway analysis (Schilling *et al.*, 2000). These methods can be used to check whether a certain reaction can be present in a steady-state flux obeying the irreversibility constraint. Thus, they can predict whether a reaction can be present in the limit behavior of a reaction network. In FBA this can be done directly, while elementary mode analysis and extreme pathway analysis return a set of vectors spanning the solution space of the steady-state condition. However, since OT also takes into account growth states, in which some species accumulate, the steady-state condition is adapted accordingly (details can be found in the Supplementary Material). Furthermore, since we only want to know whether a reaction can appear in any steady state or growth state, we do not need to apply these methods directly, but can use a linear programming approach similar to FBA, outlined in the Supplementary Material.

We compared the predictions of flux-based methods to those of OT for the models of the BioModels Database. With OT, we identified 31 models where some reactions did not appear in any reactive organization. The same 31 models are identified using flux-based methods. However, when analyzing the predicted set of available reactions in detail, we found differences in 25 of the 31 models. Due to the definition of self-maintenance, the set of available reactions is a subset of those predicted by flux-based methods. Thus, in all 25 cases, flux-based methods found reactions present in the limit behavior that indeed could not maintain a positive flux in a long-term simulation.

The reason for this difference closely follows a concept presented in Kaleta *et al.* (2006): a steady-state flux in a network uses some species that cannot be produced at a positive rate. In this flux, these species might be interconverted into each other or act as catalysts. Further assume that there is a reaction steadily draining some of the unproducible species. Thus, they will finally vanish. In consequence, this steady-state flux cannot be part of any steady state of the complete network. If a particular reaction is present only in such steady-state fluxes, it is predicted to be present in the limit behavior of a reaction network by flux-based methods, while OT correctly identifies it as absent since it correctly takes into account the drain of the unproducible species. We will outline this concept in more detail using *BIOMD037*, a model of the sporulation control network in *Physarum polycephalum* by Marwan (2003) (Fig. 7). While OT predicts 8 of the 12 reactions to be absent from the limit behavior (Fig. 7A), flux-based methods identify only four such reactions (Fig. 7B). The differentially predicted reactions account for the interconversion of Pfr to Pr and Xi to Xa. Flux-based methods find a flux where the conversion of Pfr to Pr and vice versa is in equilibrium. However, this does not take into account that there is also a reaction irreversibly converting Pr to Pi. Thus, a non-zero concentration of Pr will be depleted by the conversion into Pi. In consequence, there is no reactive organization containing Pfr and Pi.

Additionally, we find an interesting case in the interconversion of Xa to Xi and vice versa. The conversion of Xi to Xa requires the presence of Pr. Flux-based methods identify an equal flux of both reactions as a feasible flux, since Pr acts only as a catalyst. However, the analysis using OT shows that such a flux also requires the presence of Pr. Thus, both species cannot persist in the limit behavior since Pr, required for the reaction of Xi to Xa, will vanish over time. Since Xa is steadily converted to Xi, only this species would finally remain. This demonstrates how our approach takes the kinetic laws into account which is not possible using flux-based methods.

In two of the models in which we identified inconsistencies, *BIOMD094* and *BIOMD151*, predictions for the presence of reactions in the limit behavior between OT and flux-based methods differ. In *BIOMD151*, OT predicts nine reactions to be present, while flux-based methods identify 112 of the 114 overall reactions. As outlined above, flux-based methods can predict only the same or a larger set of reactions to be present in the limit behavior. Thus, the search for inconsistencies is simplified by reducing the size of the system to analyze if OT is used. This is also corroborated by three models in the uncurated branch of the BioModels Database containing inconsistencies. In all three models, flux-based methods predict more reactions to be present in the limit behavior than OT (see Supplementary Material for further details).

5 CONCLUSIONS

In this work, we demonstrated that information hidden in kinetic laws affects the results obtained from chemical organization theory (OT). We presented an approach that is able to uncover this information. This approach enabled us to refine the chemical organizations in 41 of the 185 models (22%) of the BioModels Database. The Hasse diagram of organizations of the processed model of a combined ERK/Wnt-signaling pathway took into account the different levels of phosphorylation in the signaling cascade, while the set of organizations of the unprocessed network did not. Furthermore, the Hasse diagram of organizations demonstrated several possible pathways for constant upregulation of this pathway, an important event in carcinogenesis consistent with the results of Kim *et al.* (2007).

Analyzing the 185 models of the BioModels Database, we checked the behavior of the models during long-term simulation (limit behavior). Thus, we found 31 models where several species could not persist in a long-term simulation. Furthermore, we identified five models in which some species could not be present at all during simulation. This was due to inconsistent reversibility constraints in two models, negative concentrations of some species during simulation in other two models and a missing reaction in the fifth model. In the non-curated branch of the BioModels Database, we identified the models with modeling errors. Comparing the set of species present and the reactions having a non-zero flux in the limit behavior, we found OT able to predict those sets more accurately in 25 models (14%) compared with flux-based methods like FBA, elementary mode analysis and extreme pathway analysis. These models account for 81% of the models in which the set of species and reactions present in the limit behavior of the model did not encompass the entire set of species and reactions. In five of the 8 models of both branches of the BioModels Database in which we detected modeling errors, OT made more accurate predictions in comparison to flux-based methods.

These results demonstrate that OT is a valuable tool in three important aspects of network design and analysis. First, when this approach is used to extract additional information from the kinetic laws of the reactions, the set of organizations corresponds to the potential steady state and growth states of a reaction network. Thus, important information about the dynamic structure of a reaction network can be uncovered. Second, OT can be used in an iterative fashion to assist in model building by identifying inconsistencies that need to be resolved. Third, OT more faithfully identifies parts of a network whose maintenance is not yet explained than flux-based methods. Thus, it is of particular interest for identifying gaps due to missing knowledge in large-scale metabolic networks as documented in Centler *et al.* (2008). In consequence, it can be beneficial for methods aiming to remove such inconsistencies (Kumar *et al.*, 2007; Reed *et al.*, 2006). In the other direction, our approach could be extended by these methods to automatically propose changes in order to remove inconsistencies. However, computational constraints currently prohibit the application of our deterministic algorithms to very large networks (e.g. more than 500 reactions). An approximation can be used for networks of this size, but the results require manual checking. A more efficient algorithm that will enable the application of OT to genome-scale networks is in development.

ACKNOWLEDGEMENTS

The authors would like to thank four anonymous referees for very helpful comments. Furthermore, we thank Sandro Lambeck for helpful information on the Wnt/β-catenin pathway.

Funding: German Research Foundation (Grant Di852/4-2); the European Union, NEST-project ESIGNET no. 12789.

Conflict of Interest: none declared.

REFERENCES

Borghans,J.M. *et al.* (1997) Complex intracellular calcium oscillations. A theoretical exploration of possible mechanisms. *Biophys. Chem.*, **66**, 25–41.

Centler,F. *et al.* (2007) Chemical organizations in the central sugar metabolism of *Escherichia coli*. In Deutsch,A. (eds) *Mathematical Modeling of Biological Systems*, Vol. I. Birkhäuser, Boston, pp. 109–123.

Centler,F. *et al.* (2008) Computing chemical organizations in biological networks. *Bioinformatics*, **24**, 1611–1618.

Dittrich,P. and Speroni di Fenizio,P. (2007) Chemical organization theory. *Bull. Math. Biol.*, **69**, 1199–1231.

Dittrich,P. *et al.* (2003) On the formation of social order – modeling the problem of double and multi contingency following Luhmann. *JASSS*, **6**.

Feinberg,M. and Horn,F.J.M. (1974) Dynamics of open chemical systems and the algebraic structure of the underlying reaction network. *Chem. Eng. Sci.*, **29**, 775–787.

Fontana,W. and Buss,L.W. (1994) 'The arrival of the fittest': toward a theory of biological organization. *Bull. Math. Biol.*, **56**, 1–64.

Goldbeter,A. (1991) A minimal cascade model for the mitotic oscillator involving cyclin and CDC2 kinase. *Proc. Natl Acad. Sci. USA*, **88**, 9107–9111.

Hofbauer,J. and Sigmund,K. (1998) *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge.

Hucka,M. *et al.* (2003) The Systems Biology Markup Language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

Kaleta,C. *et al.* (2006) Analyzing molecular reaction networks: from pathways to chemical organizations. *Mol. Biotechnol.*, **34**, 117–124.

Kaleta,C. *et al.* (2008) Phenotype prediction in regulated metabolic networks. *BMC Syst. Biol.*, **2**, 37.

Kim,D. *et al.* (2007) A hidden oncogenic positive feedback loop caused by crosstalk between Wnt and ERK pathways. *Oncogene*, **26**, 4571–4579.

Kumar,V.S. *et al.* (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, **8**, 212.

Le Novére,N. *et al.* (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.*, **34**, D689–D691.

Marwan,W. (2003) Theory of time-resolved somatic complementation and its use to explore the sporulation control network in *Physarum polycephalum*. *Genetics*, **164**, 105–115.

Matsumaru,N. *et al.* (2007) Toward organization-oriented chemical programming: a case study with the maximal independent set problem. In Dressler,F. and Carreras,I. (eds) *Advances in Biologically Inspired Information Systems*, Vol. 69 of *Studies in Computational Intelligence*, Springer, Berlin, pp. 147–163.

Olsen,L.F. *et al.* (2003) A model of the oscillatory metabolism of activated neutrophils. *Biophys. J.*, **84**, 69–81.

Reed,J.L. *et al.* (2006) Systems approach to refining genome annotation. *Proc. Natl Acad. Sci. USA*, **103**, 17480–17484.

Schilling,C.H. *et al.* (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, **203**, 229–248.

Schuster,S. *et al.* (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.

Singh,A. *et al.* (2006) Modeling regulatory mechanisms in IL-6 signal transduction in hepatocytes. *Biotechnol. Bioeng.*, **95**, 850–862.

Varma,A. and Palsson,B.O. (1994) Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technology*, **12**, 994–998.

Wodarz,D. and Nowak,M.A. (1999) Specific therapy regimes could lead to long-term immunological control of HIV. *Proc. Natl Acad. Sci. USA*, **96**, 14464–14469.

Yamada,S. *et al.* (2003) Control mechanism of JAK/STAT signal transduction pathway. *FEBS Lett.*, **534**, 190–196.

Chapter 4

Pathway Analysis in Genome-scale Metabolic Networks

In Kaleta et al. [2009b] we introduce elementary flux patterns, a new concept that allows to analyze fluxes through subsystems of large metabolic networks. We apply this concept to the verification of elementary modes in a previously studied model of central metabolism of *E. coli* and in the study of unconventional pathways. In Kaleta et al. [2009a] we use elementary flux pattern analysis to detect pathways in a genome-scale metabolic model that stand in contrast to the widely held believe in biochemistry that the conversion of even-chain fatty acids into glucose is not possible in humans.

The Supplemental Materials of Kaleta et al. [2009b] and Kaleta et al. [2009a] can be found in Chapter 6 on pages 140 ff. and pages 148 ff., respectively.

C. Kaleta, L. F. de Figueiredo, and S. Schuster. *Genome Res.*, 2009.

Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns

Christoph Kaleta, Luís Filipe de Figueiredo and Stefan Schuster

Genome Res. published online June 18, 2009
Access the most recent version at doi:10.1101/gr.090639.108

C. Kaleta, L. F. de Figueiredo, and S. Schuster. *Genome Res.*, 2009.

Methods

Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns

Christoph Kaleta,[1,2,3] Luís Filipe de Figueiredo,[1] and Stefan Schuster[1]

[1]*Department of Bioinformatics, Friedrich Schiller University Jena, D-07743 Jena, Germany;* [2]*Bio Systems Analysis Group, Jena Centre for Bioinformatics (JCB) and School of Mathematics and Computer Science, Friedrich Schiller University Jena, D-07743 Jena, Germany*

Elementary modes represent a valuable concept in the analysis of metabolic reaction networks. However, they can only be computed in medium-size systems, preventing application to genome-scale metabolic models. In consequence, the analysis is usually constrained to a specific part of the known metabolism, and the remaining system is modeled using abstractions like exchange fluxes and external species. As we show by the analysis of a model of the central metabolism of *Escherichia coli* that has been previously analyzed using elementary modes, the choice of these abstractions heavily impacts the pathways that are detected, and the results are biased by the knowledge of the metabolic capabilities of the network by the user. In order to circumvent these problems, we introduce the concept of elementary flux patterns, which explicitly takes into account possible steady-state fluxes through a genome-scale metabolic network when analyzing pathways through a subsystem. By being similar to elementary mode analysis, our concept now allows for the application of many elementary-mode-based tools to genome-scale metabolic networks. We present an algorithm to compute elementary flux patterns and analyze a model of the tricarboxylic acid cycle and adjacent reactions in *E. coli.* Thus, we detect several pathways that can be used as alternative routes to some central metabolic pathways. Finally, we give an outlook on further applications like the computation of minimal media, the development of knockout strategies, and the analysis of combined genome-scale networks.

[Supplemental material is available online at http://www.genome.org. All data and an application to compute elementary flux patterns are available online from http://hades.bioinf.uni-jena.de/~m3kach/EFPA/.]

In functional genomics and metabolic engineering, metabolic pathway analysis has proved to be a very useful methodology (Carlson et al. 2002; Schwender et al. 2004; Feist and Palsson 2008; Trinh et al. 2009). Elementary modes (Schuster et al. 2000) are a central concept in this field. An elementary mode represents a minimal set of reactions that can operate at steady state with all reactions proceeding in their appropriate direction (Schuster et al. 2000) and, hence, can be considered as a formal definition of a metabolic pathway. Elementary modes have been used in many areas of biotechnology, such as assessing network flexibility (Stelling et al. 2002), finding pathways with optimal yields for certain metabolic species (Schuster et al. 2002a; Krömer et al. 2006), finding possible targets for the engineering of metabolic networks (Klamt 2006), and analyzing the effect of such an engineering (Carlson et al. 2002; Schwender et al. 2004). Due to the growing availability of genome-scale metabolic networks (Duarte et al. 2004, 2007; Borodina and Nielsen 2005; Thiele et al. 2005; Feist et al. 2006, 2007; Jamshidi and Palsson 2007; Oh et al. 2007) and the comprehensive analysis conducted on them (for review, see Feist and Palsson 2008), it becomes desirable to apply elementary mode analysis in such networks.

However, the principal problem encountered when trying to compute elementary modes in larger metabolic networks is that their number is growing exponentially with network size (Klamt and Stelling 2002; Schuster et al. 2002b; Acuña et al. 2009). Thus, they become difficult to analyze or even impossible to enumerate because of constraints in memory or computation time. Although there have been recent efforts to port the algorithms for the computation of elementary modes to larger networks by means of parallelization (Klamt et al. 2005) or improvements of the existing algorithms (von Kamp and Schuster 2006; Terzer and Stelling 2008), none of these algorithms permits the analysis of elementary modes in genome-scale metabolic networks.

In consequence, elementary mode analysis is applied to smaller networks containing reactions of interest rather than the entire known system. The remainder of the system is modeled using abstractions like exchange fluxes and external metabolites. Exchange fluxes correspond to the production or consumption of a species by a large set of reactions of the remaining model. External species, in contrast, are considered to be buffered by reactions of the complete system. Hence, they are excluded from the steady-state condition. However, as we show in this study, there are three important drawbacks involved in the introduction of such abstractions (cf. Liebermeister et al. 2005).

First, the approach is usually biased by the modeler's knowledge of the network. For instance, glycolysis and pentose phosphate pathways are usually considered the principal routes for the supply of metabolites from glucose in the growth media to the tricarboxylic acid (TCA) cycle. Thus, the Entner–Doudoroff pathway—which represents an alternative route for the production of pyruvate in several bacteria—is often ignored even though it is of importance in some conditions (Fischer and Sauer 2003; Li et al. 2006). In consequence, some of the possible pathways of a large network through a subnetwork are not found by elementary mode

[3]Corresponding author.
E-mail christoph.kaleta@uni-jena.de; fax 49-3641-946452.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.090639.108.

analysis (Fig. 1A). Second, the aforementioned abstractions might not be able to take into account the dependencies between the production and consumption of metabolites that constitute the interface of the subnetwork to the remaining system. This can give rise to elementary modes that obey the steady-state condition within the subnetwork but are not part of any steady-state flux through the entire network (Fig. 1B,C). Third, by only focusing on a small part of the known network, the integration of a pathway through this subnetwork into a pathway on the scale of the entire system is not straightforward, and the interdependencies between the pathways of several subsystems cannot be analyzed (Fig. 1D).

The concept of "elementary flux patterns" that we introduce in this study circumvents these problems by taking into account the possible fluxes through the entire network, when analyzing the steady-state fluxes through a subnetwork. An elementary flux pattern is defined as a set of reactions within a subsystem of a larger network that represents the basic routes of each steady-state flux of the larger network through the subnetwork. Thus, flux modes in a subsystem can be determined that are feasible in the context of the entire genome-scale system. Through their definition, elementary flux patterns allow a consistent application of concepts from elementary mode analysis to genome-scale metabolic networks without the drawbacks that arise by the definition of external species or exchange fluxes. Furthermore, each elementary flux pattern can be mapped to at least one elementary mode in the complete system, even allowing the user to analyze pathways on the genome scale.

This article is sectioned into three main parts. First, we will introduce elementary flux patterns and an algorithm to compute them. Subsequently, we apply this concept to a genome-scale metabolic network of *Escherichia coli*. Then we discuss our results and give an outlook on further applications of elementary flux patterns.

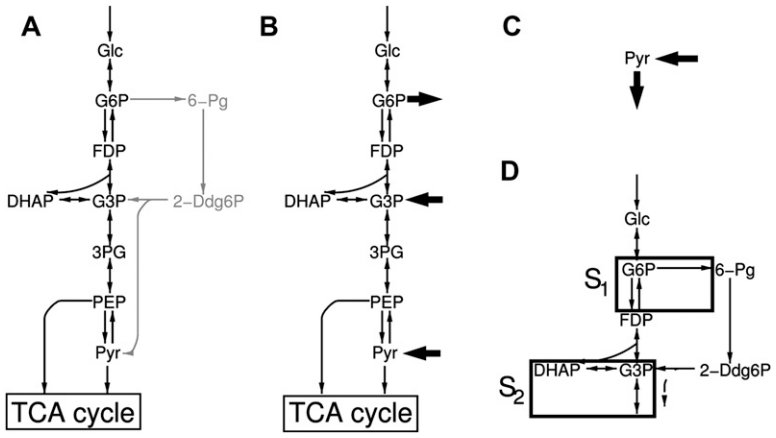


Figure 1. Examples for problems in elementary mode analysis. (*A*) Condensed network of glycolysis (dark reactions) and Entner–Doudoroff pathway (gray reactions). The pentose phosphate pathway has been omitted for clarity. In most analyses of glycolysis, the Entner–Doudoroff pathway is not considered. Hence, not all pathways from glucose into the TCA cycle are found, and the wrong conclusion might be drawn that only the pentose–phosphate pathway (not shown) can be used to bypass the knockout of one of the enzymes converting G6P to FDP. In vivo the knockout of the corresponding reactions is partially bypassed by a flux through the Entner–Doudoroff pathway (Fischer and Sauer 2003). (*B*) Modeling the Entner–Doudoroff pathway by adding an outflow of G6P and an inflow of G3P and Pyr only partially resolves the problem (thick arrows). Such an approach is often used in elementary mode analysis to avoid the consideration of some pathways in detail (e.g., the outflow of succinyl-CoA from the TCA cycle in Schuster et al. 1999, analyzed in Results). However, this can lead to fluxes like in *C*, which are not part of any feasible pathway when the entire network in *A* is considered. This is because the coupling of the influx of Pyr with the outflow reactions of G6P and the inflow reaction of G3P are neglected in *B*. (*D*) Elementary mode analysis does not allow one to analyze the dependencies between the subsystems S_1 and S_2 unless the reactions connecting them are taken into account. Thus, it is not possible to deduce that a zero flux from G6P to FDP in S_1 would imply that there cannot be a positive flux from DHAP to G3P in S_2. A list of abbreviations can be found in Supplemental material S2.

Methods

Next, we will formally introduce the concepts central to this study. We start by giving a short introduction to elementary mode analysis. This is followed by a definition of elementary flux patterns and the outline of an algorithm to compute them. Subsequently, we compare our method to other approaches for pathway analysis in genome-scale metabolic networks.

A metabolic network comprising n reactions and m metabolites is defined by the $m \times n$ stoichiometric matrix $\mathbf{M}$. An entry M_{ij} of $\mathbf{M}$ is negative if species i is an educt of reaction j and positive if it is a product. Since elementary flux patterns are defined within a subsystem of k reactions of the entire system, we assume for simplicity that the k first columns of $\mathbf{M}$ (i.e., the k first reactions) constitute the subsystem.

Elementary modes

Elementary modes represent minimal sets of reactions that can operate at steady state with all reactions proceeding in thermodynamically feasible directions (Schuster et al. 2000). They are minimal in the sense that there is no subset of reactions that could also operate at steady state.

Formally, an elementary mode is a flux vector $\mathbf{v}$ of length n that assigns a flux to each reaction, such that

$$\mathbf{M} \cdot \mathbf{v} = 0. \quad (1)$$

Furthermore, $\mathbf{v}$ has to obey the thermodynamic constraints of the reactions; that is, if reaction i is irreversible, v_i has to be non-negative. To simplify the analysis, we adopt the widely used procedure of splitting reversible reactions into forward and backward directions. Thus, the irreversibility constraint becomes

$$\mathbf{v} \geq 0. \quad (2)$$

A flux mode $\mathbf{v}$ is called elementary if there exists no flux mode $\mathbf{v}'$ that uses a proper subset of the reactions of $\mathbf{v}$ (Schuster et al. 2000).

Through the concept of elementarity, the set of elementary modes of a reaction network and their superposition describe all of the possible steady states of this network. That is, each steady state can be written as a non-negative linear combination of elementary modes. As mentioned above, the analysis with elementary modes can be further simplified by the introduction of external species. External species are considered to be buffered by reactions outside of the model. Thus, the steady-state condition can be relaxed by removing the rows corresponding to external species from the stoichiometric matrix $\mathbf{M}$ in condition 1. This describes also the major problem when defining

external species. Through removing species from **M**, all information on the dependencies between their production and consumption by reactions of the remaining system is lost.

Elementary flux patterns

Flux patterns are defined as sets of reactions in a subsystem of k reactions of interest in a large metabolic network. They correspond to all possible routes that a steady-state flux of the entire network can take through the subsystem. To simplify the analysis, only elementary routes are considered. Hence, a flux pattern is called elementary if it cannot be derived as a combination of at least two other flux patterns. For an example of elementary flux patterns in a subsystem of the TCA cycle and some adjacent reactions, see Figure 2.

Formally, a flux pattern s is a set of indices i with $1 \le i \le k$ that fulfills the following conditions (please note that we assume that the first k reactions in **M** represent the subsystem):

$$\mathbf{v} \geq 0 \tag{3}$$

$$\mathbf{M} \cdot \mathbf{v} = 0 \tag{4}$$

$$\forall i \in s \,:\, v_i > 0 \tag{5}$$

$$\forall j \in \{1..k\} \backslash s \,:\, v_j = 0. \tag{6}$$

Thus, a flux pattern is a set of reactions, or, more precisely, a set of reaction indices, of the subnetwork that is part of a steady-state flux **v** of the entire network. We require that the indices s of **v** are nonzero, while the remaining are zero. Given the set of elementary flux patterns S of a system, that is, a set of sets of reaction indices, we call a flux pattern $s \in S$ elementary if

$$\nexists s'_{i_1}, \ldots, s'_{i_l} \subseteq S \backslash s : \bigcup_{1 \le k \le l} s'_{i_k} = s \text{ with } i_1, \ldots, i_l \in \{1, \ldots, |S|\}. \tag{7}$$

Thus, we call s elementary if there exists no set of flux patterns (not including s) whose union is equal to s. Please note that this definition is less restrictive than that in the concept of elementary modes, as the elementarity of a flux mode requires that the nonzero indices of one elementary mode cannot be a subset of the nonzero indices of another elementary mode. An analogous statement for flux patterns does not hold. More details on this difference are given in the next section.

In conditions 3 to 6 no statement is made about the relationship between the reactions of the subsystem. Thus, in contrast to elementary mode analysis, reactions need not interface with each other through substrates or products. In consequence, it is possible to analyze the dependencies between different subsystems like in Figure 1D without needing to add all reactions connecting both to a combined subsystem.

Furthermore, the empty set also fulfills the flux pattern condition. However, the presence of the empty flux pattern indicates only that the zero flux O^n is a valid, although trivial, solution of Equations 1 and 2; hence, it is of no interest here.

Through the splitting of reversible reactions into two irreversible reactions, spurious cycles of the forward and back direction of the reversible reaction occur. If required, they can be removed in a post-processing step.

Comparison of elementary modes and elementary flux patterns

Elementary flux patterns are tightly coupled to fluxes within the entire system. As demonstrated in Supplemental material S5.2, each elementary flux pattern is part of at least one elementary mode of the complete system. Following a procedure outlined in Supplemental material S5.3, this elementary mode can be obtained. Additionally, alternative pathways can be computed by constraining some of the reactions of such an elementary mode to zero. Furthermore, by slightly adapting the constraints in the formulation of the integer linear program used for the computation of the k-shortest elementary modes (de Figueiredo et al. 2009), it is possible to enumerate all elementary modes containing a given flux pattern (data not shown). However, the latter approach requires optimizing an integer linear program with as many integer variables as reactions in the system. This procedure is computationally very demanding.

Besides their definition in subnetworks of metabolic models, the most obvious difference between elementary modes and elementary flux patterns is that the former are defined as a vector and the latter as a set of indices. An elementary mode represents a particular flux distribution in a network, in which flux proportions are considered (although it is indeterminate

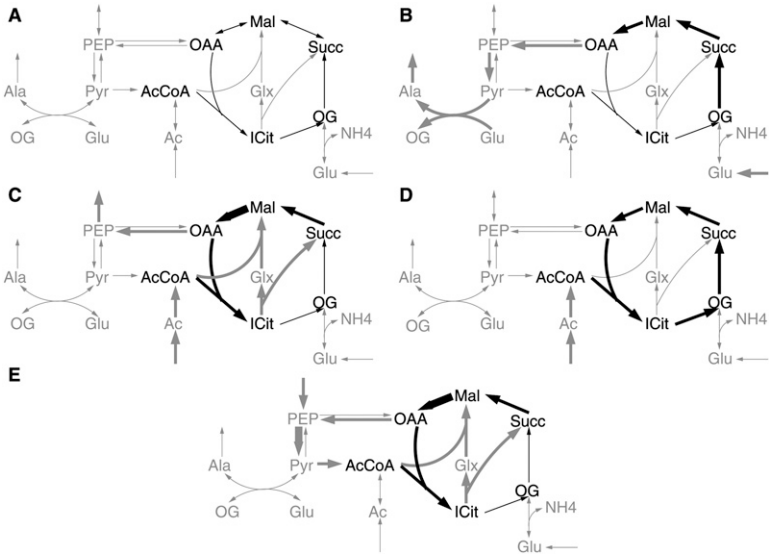


Figure 2. Elementary flux patterns in a condensed model of the TCA cycle and adjacent reactions. (*A*) Model of the entire network. The TCA cycle is chosen as a subsystem (dark reactions). (*B–D*) Elementary flux patterns of the system are indicated by thick black arrows. The associated pathway through the entire system is indicated by thick gray arrows. The thickness of the arrows corresponds to the relative flux through each reaction. (*E*) Alternative pathway through the entire system using the same reactions of the flux pattern depicted in *C*. While the pathway in *C* corresponds to the glyoxylate cycle that can be used for growth on fatty acids, *E* corresponds to the phosphoenolpyruvate-glyoxylate cycle used as a catabolic pathway during growth on low glucose concentrations (Fischer and Sauer 2003). A list of abbreviations can be found in Supplemental material S2.

with respect to scaling). In contrast, a flux pattern can correspond to several flux proportions within the genome-scale system. Hence, only the binary pattern is considered. However, most of the applications of elementary modes only require the set of nonzero indices of the elementary modes (Gagneur and Klamt 2004), called the "activity set" in Nuño et al. (1997). Thus, a reduction to mere index sets does not represent an obstacle in many applications of elementary modes.

Furthermore, if the subnetwork encompasses the entire network ($k = n$), each elementary flux pattern corresponds to an elementary mode of the network (see Supplemental material S5.1 for a proof).

Another difference with elementary mode analysis can be found in the splitting of reversible reactions into irreversible forward and backward directions. It has been shown that, besides spurious cycles consisting of forward and backward steps and the doubling of entirely reversible elementary modes, the set of elementary modes does not change by splitting reversible reactions (Gagneur and Klamt 2004). In principle, this statement also holds for elementary flux patterns. Replacing the positivity condition in Equation 5 by a nonzero condition for reversible reactions of the subsystem would make a splitting unnecessary. However, this would introduce ambiguities since elementary flux patterns are defined as sets of reactions, that is, as binary patterns. In consequence, each reversible reaction would also be modeled as a single index, and it would not be clear which direction of a reversible reaction is used. Thus, splitting reversible reactions simplifies the analysis.

Computation of elementary flux patterns

The elementary flux patterns of a subsystem can be computed by iteratively solving a mixed-integer linear program (MILP) that returns an elementary flux pattern. By consecutively adding additional constraints, it is assured that a new elementary flux pattern is always returned. If all elementary flux patterns have been found, the MILP becomes infeasible and the iteration is stopped. For a detailed outline of the procedure, see the Appendix.

Computational complexity

Next, we want to comment on the computational complexity of the problem of finding elementary flux patterns. As outlined in Supplemental material S4, the computation time of the algorithm presented in the Appendix is polynomial in the size of the entire system and exponential in the size of the subsystem. This effort pays in that a comprehensive view on the metabolic capabilities pertaining to the subsystem (embedded in the whole system) is obtained. The runtime complexity of the computation of elementary flux patterns is similar to that of fixed-parameter algorithms (Downey and Fellows 1998), a method by which an NP-hard problem is tackled by confining the combinatorial explosion to subproblems such as in weighted cluster editing (Böcker et al. 2008). This result is of central importance since it demonstrates that the application to larger and larger systems is not the limiting factor in the computation of elementary flux patterns.

Implementation

The algorithm has been implemented as a command-line version and a graphical user interface (GUI) in Java. The GUI enables the selection of reactions for the subnetwork as well as the analysis of elementary flux patterns. Both programs accept reaction networks in the widely used Systems Biology Markup Language (Hucka et al. 2003) as well as in a proprietary, more human-readable, format. The GUI can interface with the Systems Biology Workbench (SBW) (Sauro et al. 2003). Thus, it can be called from any SBW-compliant application. This allows an easy integration into a wide variety of tools such as network design, simulation, and further analysis that are available for SBW. The linear programs and the mixed-integer linear programs are solved using the open-source Clp and Cbc solvers from the COIN-OR project (Lougee-Heimer 2003). Cbc implements a parallelized MILP solver. Hence, the multi-processor architecture of current computer systems can be fully exploited. The running time of the algorithm for a selected list of subsystems in two genome-scale metabolic networks is given in Table 1.

Comparison to other genome-scale pathway analysis methods

Next, we want to compare the concept of elementary flux patterns to other methods that allow pathway analysis in genome-scale metabolic networks.

Flux balance analysis

Flux balance analysis (FBA) consists of the search for a flux distribution in a reaction network that optimizes a given objective function and obeys certain constraints on reaction fluxes (Varma and Palsson 1994). A variant of FBA is called flux minimization (Holzhütter 2004). FBA has been extensively used to study metabolic networks (for review, see Raman and Chandra 2009) and has seen many extensions to take into account additional information like regulatory rules (Covert et al. 2001; Shlomi et al. 2007) and reaction kinetics (Covert et al. 2008). Like elementary flux pattern analysis, FBA can be readily used to find a pathway producing a certain metabolite. However, our method is better suited for the exhaustive enumeration of pathways in a subsystem. Furthermore, FBA meets the problem that usually most of the reactions of a computed flux are used for balancing of cofactors like ATP,

Table 1. List of subsystems of two genome-scale networks for which elementary flux patterns (EFPs) have been computed

Model	Subsystem[a]	No. of reactions	No. of EFPs	Computation time (sec)[b]
E. coli (Feist et al. 2007)	Purine and pyrimidine biosynthesis	33	22	113
1972 species, 3559 reactions	Glycolysis and gluconeogenesis	27	33	96
	Central metabolism (Schuster et al. 1999)	33	83	2042
S. cerevisiae (Duarte et al. 2004)	Tyrosine, tryptophan, and phenylalanine metabolism	34	30	29
1177 species, 1939 reactions	Glycolysis and gluconeogenesis	30	80	312

[a]The function of the subsystem as annotated in the corresponding model.
[b]Computation time was measured on an Intel Core 2 Quad Q9300 machine with 4096 MB RAM running Linux Kernel 2.6.25 and Java Hotspot VM version 1.6.0. COIN-OR Cbc version 2.0 has been used to solve the mixed-integer linear programs.

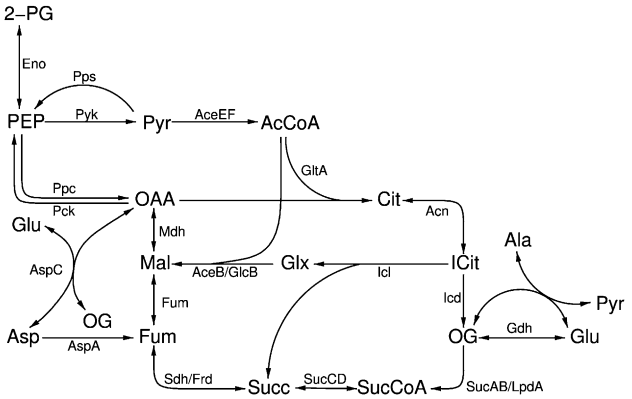


Figure 3. Scheme of part of the central metabolism of *E. coli*, which is here studied as a subnetwork of a genome-scale network. A list of abbreviations can be found in Supplemental materials S2 and S3.

NADH, or NADPH, and, hence, it is not straightforward to extract the underlying pathway used for the conversion of a source species into a target species. In elementary flux pattern analysis, this does not occur unless cofactor balancing reactions are explicitly considered in the subsystem.

Furthermore, the determination of a global pathway using the reactions of a flux pattern in a subsystem uses a linear programming method similar to FBA. Thus, elementary flux pattern analysis is, in a sense, a combination of elementary mode analysis and FBA.

Flux variability analysis

Flux variability analysis extends flux balance analysis by allowing one to determine the individual minimal and maximal fluxes of reactions given that the optimal value of an objective function is maintained (Mahadevan and Schilling 2003). Thus, it is possible to identify which reactions might be used by an optimal flux. However, it is not possible to obtain a comprehensive overview on all possible pathways but only on a subset of all reactions that can be used to achieve an optimal value of the objective function (Mahadevan and Schilling 2003).

Flux coupling analysis

Flux coupling analysis (Burgard et al. 2004) extends the concept of enzyme subsets (Pfeiffer et al. 1999). It allows one to detect global dependencies in the use of reactions at steady state by determining how fluxes through pairs of reactions are coupled to each other. As outlined in Supplemental material S7, the application of flux coupling analysis can be seen as a special case of elementary flux pattern analysis in all subsystems containing only two reactions. In consequence, results from flux coupling analysis can similarly be obtained by using elementary flux pattern analysis. Our method can be seen as a generalization of flux coupling analysis since it not only examines the dependencies between pairs of reactions, but also between any subset of reactions within a subsystem.

Results

We analyzed a model of the tricarboxylic acid cycle (TCA cycle), the glyoxylate shunt, and associated reactions in *E. coli* that has been previously analyzed using elementary modes (Schuster et al. 1999). We started from the genome-scale metabolic model of *E. coli* from Feist et al. (2007) and defined the subsystem as the set of reactions that were used for the elementary mode analysis by Schuster et al. (1999). We modified the genome-scale model by adding an inflow for glucose and other basic compounds that are necessary for the production of biomass (see Supplemental material S1 for a complete list). The subsystem is depicted in Figure 3. The final network contains 1972 species and 3559 reactions.

In the first step, we did not incorporate into the subsystem the output fluxes used in the elementary mode analysis. Elementary mode analysis requires these fluxes to guarantee a steady-state flux corresponding to the production of a certain species. Such fluxes correspond to the assumption that a certain species can be consumed by reactions outside of the subnetwork without using any further flux through the subnetwork. Since elementary flux pattern analysis takes into account the entire network, we do not need to add such abstract reactions.

Elementary modes of the system

In a first step, we analyzed the elementary modes that have been found by Schuster et al. (1999) using the method outlined in Supplemental material S5.4. Thus, we checked for each elementary mode whether it could be part of a steady-state flux through the entire system. Ten out of the 16 elementary modes given in Schuster et al. (1999) are part of such a steady-state flux, while all six elementary modes producing the external species succinyl-CoA are not. Even though there are four reactions in the complete network consuming succinyl-CoA, all of them need additional species that can only be produced at a positive rate using intermediates of the TCA cycle. Since the TCA cycle is part of the subsystem, each elementary mode producing succinyl-CoA needs an additional flux through the subsystem in order to metabolize succinyl-CoA. This is confirmed by an overview of the reactions consuming succinyl-CoA, given in Table 2.

As an example, succinyl-CoA is consumed in lysine synthesis, in which oxaloacetate is also used. Later in that pathway, succinate is released. The subsystem model used by Schuster et al. (1999) cannot properly take into account the coupling between the fluxes of succinyl-CoA consumption, oxaloacetate consumption, and succinate regeneration on that route. Most interestingly, entry number 3 in Table 2 does not require any additional substrate besides succinyl-CoA. However, if succinyl-CoA is consumed during propionate utilization, additionally, oxaloacetate is converted into pyruvate. Oxaloacetate cannot be reconverted into pyruvate unless the reactions of the subsystem are used. Thus, even in the case of propionate utilization, the degradation of succinyl-CoA requires an additional flux through the subsystem.

Elementary flux patterns of the system

In a next step, we computed the elementary flux patterns of the subsystem. To be able to analyze the production and consumption

Table 2. List of reactions consuming succinyl-CoA in the network of Feist et al. (2007)

Number	Reaction	Functional assignment
1	Arg + SuccCoA → CoA + Sucarg	Arginine degradation
2	Hom + SuccCoA → CoA + Suchms	Methionine synthesis
3	SuccCoA → MmCoA	Propionate utilization
4	H_2O + SuccCoA + Thdp → CoA + Sl2a6o	Lysine synthesis

A list of abbreviations can be found in Supplemental materials S2 and S3.

of amino acids as it has been done by Schuster et al. (1999), the outflow reactions for the biomass forms of alanine, aspartate, and glutamate present in the genome-scale network were added to the subsystem. As mentioned above, there is no such outflow reaction for succinyl-CoA, since this species can only be further metabolized using additional fluxes from the subsystem. Additionally, we added a reaction allowing the transport of alanine from the cytosol to the periplasmic space. In the model of Feist et al. (2007), species are drained from the model only in their extracellular forms. Since the model did not contain any mechanism transporting alanine to the extracellular space, such a reaction needed to be added. This system gives rise to 83 elementary flux patterns, of which eight produce one of the amino acids. In contrast, Schuster et al. (1999) found only 16 elementary modes. This shows the arbitrariness involved in the definition of exchange reactions with the remaining system necessary for elementary mode analysis. As we will outline in the following, there are several additional intermediates of the subsystem besides 2-phosphoglycerate that can be produced from glucose. Additionally, there are several pathways bypassing some of the reactions in the subsystem. This leads to many additional pathways through the subsystem in comparison to those detected by Schuster et al. (1999).

Sometimes elementary flux patterns are cryptic when considering only the reactions of the subsystem they contain. In such a case, an analysis of the genome-scale elementary modes associated to each elementary flux pattern is necessary. This can be accompanied by constraining the fluxes of some reactions to zero to investigate alternative elementary modes to which an elementary flux pattern corresponds. One example is the reaction of malate to oxaloacetate. There are two enzymes catalyzing this reaction, malate dehydrogenase (Mdh) and malate:quinone oxidoreductase (Mqo). Since we only took into account the reactions used in Schuster et al. (1999), we added only the reaction of Mdh to the subsystem. However, elementary modes can also use Mqo not present in the subsystem. Indeed, we find many elementary flux patterns producing malate and consuming oxaloacetate without the intermediate action of Mdh.

Entry points into the TCA cycle

Even though we considered glycolysis as the principal pathway for producing TCA cycle intermediates, we found many pairs of elementary flux patterns that differ only in the routes producing these intermediates. In each such pair, one elementary flux pattern uses 2-phosphoglycerate as the entry point, while the other uses pyruvate. There are several pathways capable of producing pyruvate from glucose without the use of glycolysis. One is the Entner–Doudoroff pathway (Fig. 4B), which is used by some bacteria as an alternative pathway to glycolysis, even though it has a lower ATP yield. In *E. coli*, it was found that a knockout of the glycolysis enzyme phosphoglucose isomerase resulted in an up-regulation of the Entner–Doudoroff pathway when grown on glucose (Fischer and Sauer 2003). In consequence, 30% of the glucose was metabolized through the Entner–Doudoroff pathway.

The principal pathway connecting glycolysis to the TCA cycle proceeds via the conversion of pyruvate to acetyl-CoA and carbon dioxide. However, the consumption of acetyl-CoA does not allow a positive production rate of any of the species of the TCA cycle (Weinman et al. 1957; Schuster and Fell 2007). Thus, TCA cycle intermediates need to be replenished using an additional inflow from, for example, glycolysis or malate from the glyoxylate bypass. In this context, we found an elementary flux pattern that is capable of producing phosphoenolpyruvate from acetyl-CoA via the glyoxylate bypass. This is an interesting case since the network is supplied with glucose as carbon source. However, a pathway different from glycolysis is used to produce acetyl-CoA from glucose. An example for such a pathway is the production of pyruvate by the Entner–Doudoroff pathway and the subsequent reaction of pyruvate and coenzyme A to formate and acetyl-CoA catalyzed by the pyruvate formate lyase. Within the subsystem, acetyl-CoA is subsequently converted into phosphoenolpyruvate using the glyoxylate shunt.

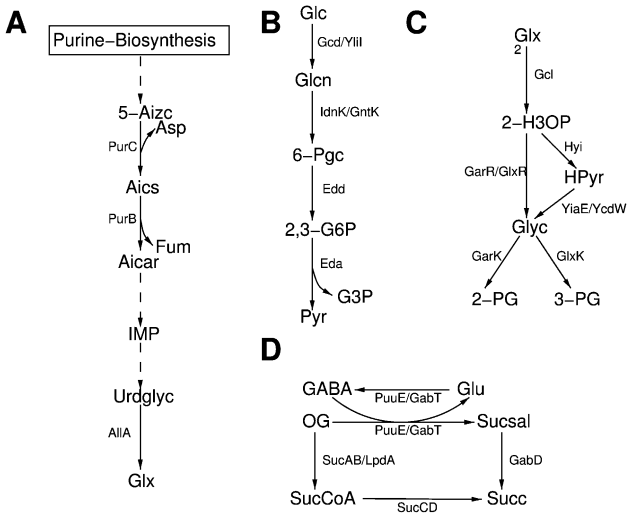


Figure 4. Alternative pathways in the central metabolism of *E. coli*. Dashed arrows represent condensed reactions. (*A*) Alternative glyoxylate producing pathway; (*B*) Entner–Doudoroff pathway; (*C*) glycerate pathway; (*D*) GABA-shunt. The lower pathway represents the standard route from oxoglutarate to succinate in the TCA cycle, and the upper pathway depicts the GABA shunt. A list of abbreviations can be found in Supplemental materials S2 and S3.

The fourth entry point necessitates the production of acetyl-CoA and glyoxylate by sources outside of the subsystem. More details about this pathway are given in the next section.

An alternative source for glyoxylate

While isocitrate lyase is used in the glyoxylate bypass to produce glyoxylate, we found several elementary flux patterns that use up glyoxylate, which is produced without isocitrate lyase present. Most interestingly whenever we encountered such a case, aspartate aminotransferase was operative. Constraining the flux of this reaction to zero prevented glyoxylate production through the alternative pathway. Aspartate aminotransferase produces aspartate and oxoglutarate from glutamate and oxaloacetate. Oxoglutarate can be aminated into glutamate, counterbalancing the consumption of glutamate by this reaction. Aspartate is essential for a reaction in the synthesis of the purine base inosine. Using several reactions, this compound is subsequently converted into glyoxylate via the intermediate of inosine-monophosphate and urate. Some central reactions of this pathway are depicted in Figure 4A.

The glycerate pathway

Some elementary flux patterns containing isocitrate lyase, the first enzyme of the glyoxylate bypass, do not contain the malate synthase catalyzing the formation of malate from glyoxylate and acetyl-CoA. These elementary flux patterns use the glycerate pathway

(Hansen and Hayashi 1962) that condenses two glyoxylates to 2-hydroxy-3-oxopropanoate and produces either 2-phosphoglycerate or 3-phosphoglycerate (Fig. 4C). An interesting aspect of this pathway is that it needs only three steps to produce 3-phosphoglycerate from glyoxylate. Thus, it might be of importance during gluconeogenesis from acetyl-CoA since the common glucogenic route via malate needs five reactions for the same conversion.

The GABA shunt

In some elementary flux patterns that contain most of the reactions of the TCA cycle, the reactions from oxoglutarate to succinyl-CoA and further to succinate are missing, while oxoglutarate is produced and succinate is consumed from sources outside the subsystem. In these cases, reactions belonging to a pathway known in plants as the gamma-aminobutyric acid shunt (GABA shunt) are used. This pathway produces succinate from oxoglutarate via the intermediate of succinic semialdehyde (Fig. 4D). While GABA has been primarily considered in the context of acid resistance in *E. coli* (Richard and Foster 2003), a recent study suggested a role beyond stress response in plants and proposed the GABA shunt to be an integral part of the TCA cycle (Fait et al. 2008). Thus, this pathway might also be of importance as a bypass for a part of the TCA cycle in *E. coli*. However, its role as such a bypass has, to our knowledge, not yet been investigated in detail.

Amino–acid–producing elementary flux patterns

We found eight elementary flux patterns producing amino acids (Fig. 5). A single elementary flux pattern produces alanine (Fig. 5A), while aspartate is produced by three elementary flux patterns (Fig. 5B–D) and glutamate by four elementary flux patterns (Fig. 5E–H).

The elementary flux pattern producing alanine contains only two reactions, the production of alanine from pyruvate and glutamate as well as the outflow of extracellular alanine. Hence, the substrates of the reaction producing alanine can be replenished from sources outside the subsystem. This is corroborated by our previous finding that pyruvate, which is aminated into alanine, can be produced without using glycolysis. Furthermore, glutamate can be produced from oxoglutarate using one of several transaminase reactions not included in the subsystem.

The number of elementary flux patterns producing glutamate are due to the possible entry points of species into the subsystem. These are acetyl-CoA alone or in conjunction with either 2-phosphoglycerate, pyruvate, or glyoxylate (via the alternative glyoxylate-producing pathway). The route via glyoxylate involves the alternative pathway of glyoxylate production depicted in Figure 4A. Thus the action of the aspartate aminotransferase is also necessary.

For aspartate, only three of the entry points are used by the elementary flux patterns. Producing aspartate, using acetyl-CoA, and the glyoxylate bypass is a flux pattern, but does not fulfill the elementarity condition. An explanation is in order.

Each pathway drawing species from the TCA cycle with acetyl-CoA alone entering the cycle needs to include the glyoxylate bypass. This pathway requires the malate synthase and the fumarase to replenish oxaloacetate. Additionally, the aspartate transaminase is necessary to produce aspartate from oxaloacetate. The pathway using glyoxylate (from the alternative glyoxylate producing pathway) and acetyl-CoA as substrates also requires the malate synthase and the fumarase. As depicted in Figure 6, the fumarase is needed in order to reconvert fumarate consumed by the alternative glyoxylate-producing pathway in Figure 4A into aspartate. Any flux pattern producing aspartate using acetyl-CoA as entry point into the subsystem would be a union of two elementary flux patterns. The first is the elementary flux pattern containing the reactions of the subsystem corresponding to the alternative glyoxylate-producing pathway; and the second, the aspartate-producing elementary flux pattern in Figure 6.

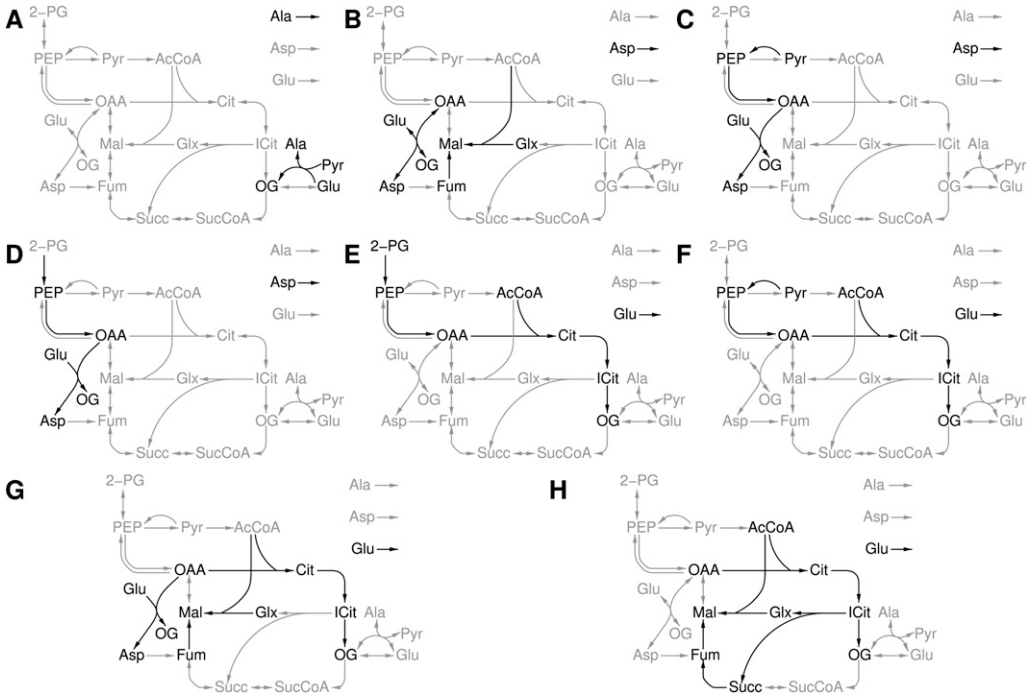


Figure 5. Amino acid producing elementary flux patterns. (*A*) Alanine; (*B–D*) aspartate; (*E–H*) glutamate. Black reactions belong to the elementary flux pattern; gray reactions are the remaining reactions of the subsystem. A list of abbreviations can be found in Supplemental materials S2 and S3.

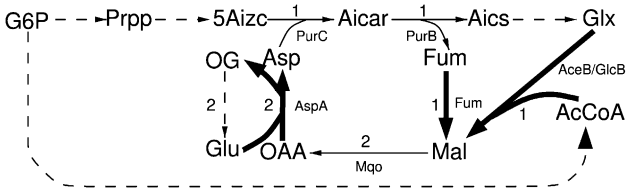


Figure 6. Elementary flux pattern belonging to an aspartate-producing pathway. (Bold arrows) The reactions of the elementary flux pattern (i.e., they belong to the subsystem). The species entering the subsystem are acetyl-CoA and glyoxylate via the alternate glyoxylate-producing pathway presented in Figure 4A. (Dashed lines) Condensed reaction. Selected fluxes are given. A list of abbreviations can be found in Supplemental materials S2 and S3.

Another interesting aspect arises from the observation that, only for glutamate, no reaction from the subsystem producing this amino acid is necessary. However, the production of oxoglutarate appears in any flux pattern producing glutamate. Thus, while the reactions of the subsystem are necessary to produce oxoglutarate, the conversion into glutamate can also be performed by reactions outside the subsystem.

Discussion

Here we have introduced elementary flux pattern analysis as a new concept for the investigation of pathways in genome-scale metabolic networks. In contrast to elementary modes, which represent an important tool for the analysis of metabolic networks, they more accurately depict the metabolic capabilities of a subsystem integrated into a genome-scale model and offer several important advantages.

First, the modeling of the interaction of the subsystem with the entire model is not biased by the knowledge of pathways into and out of the subsystem by the modeler, which might be limited to pathways seen under standard conditions. Thus, instead of 16 pathways through the TCA cycle and some adjacent reactions found with elementary mode analysis in Schuster et al. (1999), we detected 83 possible routes. We started with 2-phosphoglycerate as the principal species produced from glucose that served as input to the subsystem. However, our analysis found four possible input points, including a previously unknown pathway producing glyoxylate from glucose-6-phosphate without the intermediate action of one of the essential enzymes of the glyoxylate bypass. Moreover, we identified a pathway that is similar to the GABA shunt in plants and may serve to bypass some of the reactions in the TCA cycle.

Second, constraints imposed by the stoichiometric structure of the entire network upon possible fluxes in the subsystem are properly taken into account. In a network analyzed in Schuster et al. (1999), we found that such constraints lead to the problem that six of the 16 reported elementary modes were not part of any steady-state flux through the entire system. Generally, referring to the questions posed in the title of this article, when metabolic subsystems are combined to larger networks, usually some of the combinations of elementary modes and, thus, pathways drop out. This is because of some metabolites that were external in the subsystems and now have to fulfill the steady-state condition since they became internal. This example shows that the set of pathways in the whole metabolic network is smaller than the "sum" of the pathway sets on the local scale.

Third, the possibility to analyze the interaction between different subsystems of a metabolic network represents a promising avenue of further research in order to gain a better understanding of the intricate structure of metabolism.

Fourth, most of the tools building on elementary mode analysis only necessitate the sets of reactions of the elementary modes. Hence, these methods can also use elementary flux patterns as a base, now allowing their application to genome-scale metabolic networks. These applications include the development of gene-knockout strategies for strain improvement and the analysis of the robustness of metabolic networks, as outlined next.

In order to prove the utility of elementary flux patterns, we chose the detection of unconventional metabolic pathways in the central metabolism of *E. coli* as the main focus of our work. The knowledge of such pathways is not only of theoretical interest but also of importance in many fields of biotechnology, for instance, in the analysis of gene knockout experiments and metabolic flux analysis (Wittmann 2007). In the analysis of gene knockout experiments, our method can help to identify alternative pathways that might be used to bypass a knockout. In metabolic flux analysis, our method can help to improve the stoichiometric models used for the calculation of intra-cellular reaction fluxes from labeling experiments. These models usually only include a specific part of the entire metabolism. Hence, they face the same problem as elementary mode analysis, by potentially not taking into account all of the possible routes into and out of the subsystem. In consequence, refining the subsystem after studying its connection to the entire known metabolism using elementary flux patterns can help to improve the calculated fluxes. This can be of special importance if the cell is subject to extreme conditions resulting in a redirection of fluxes from pathways seen under standard conditions to alternative pathways (Fischer and Sauer 2003; Wittmann et al. 2007).

Finally, we want to outline some further applications of elementary flux patterns.

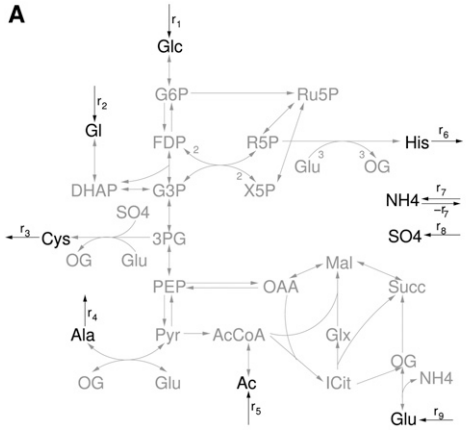
Determining minimal media

Elementary flux patterns can be used to compute the composition of minimal media, that is, the set of metabolites minimally required for the production of a desired product (Fig. 7). Note that some (but not necessarily all) elementary flux patterns correspond to minimal media (Fig. 7). This is of special interest for the determination of growth media required for the synthesis of complex metabolites like antibiotics (Tollnick et al. 2004). One focus could be the analysis of the different proposed growth media with respect to efficiency of the production of the antibiotic and the cost for the production of the metabolites used in the medium. Furthermore, if the network contains a reaction that indicates which metabolites are essential for the growth of the cell, it is possible to compute all minimal growth media.

Computing minimal cut sets

Minimal cut sets correspond to minimal sets of reactions that need to be removed from a system in order to suppress any steady-state flux performing a certain function (Klamt 2006). Such a function can be, for instance, the production of a side-metabolite by a target reaction in order to increase the yield of a desired product. A scenario for the application of elementary flux patterns in this context is outlined in Figure 8. Due to the size of the system in which elementary flux patterns can be computed, they now allow computing minimal cut sets even in genome-scale metabolic networks.

A



B

EFP	Inflow	AA
$\{r_1\}$	Glc	-
$\{r_2\}$	Gl	-
$\{r_5\}$	Ac	-
$\{r_7, -r_7\}$	NH4	-
$\{-r_7, r_9\}$	Glu	-
$\{r_4, r_9\}$	Glu	Ala
$\{r_1, r_4, r_7\}$	Glc, NH4	Ala
$\{r_2, r_4, r_7\}$	Gl, NH4	Ala
$\{r_4, r_5, r_7\}$	Ac, NH4	Ala
$\{r_3, r_8, r_9\}$	Glu, SO4	Cys
$\{r_1, r_3, r_7, r_8\}$	Glc, NH4, SO4	Cys
$\{r_2, r_3, r_7, r_8\}$	Gl, NH4, SO4	Cys
$\{r_3, r_5, r_7, r_8\}$	Ac, NH4, SO4	Cys
$\{r_6, r_9\}$	Glu	His
$\{r_6, r_7, r_9\}$	Glu, NH4	His
$\{r_2, r_6, r_7\}$	Gl, NH4	His
$\{r_5, r_6, r_7\}$	Ac, NH4	His
$\{r_1, r_6, r_7\}$	Glc, NH4	His

Figure 7. Determining minimal media for the production of amino acids. (*A*) Network under consideration for computation of elementary flux patterns. Dark reactions belong to the subsystem. (*B*) Elementary flux patterns of the subsystem. The inflow reaction that is used and the amino acid that can be produced from this input medium are indicated. Examples of minimal and nonminimal media are provided by the synthesis of histidine, which can be produced from glutamate alone as well as from ammonia and glutamate, respectively. The latter elementary flux pattern is more efficient in terms of glutamate consumed. A list of abbreviations can be found in Supplemental material S2.

Determining the robustness of metabolic networks

In a previous study, we used elementary modes to define a general measure for the susceptibility of reaction networks to knockouts (Behre et al. 2008). However, this approach builds on elementary modes, and, thus, the robustness is only determined for elementary modes of a subnetwork that might not reflect all the potential pathways as detected by elementary flux pattern analysis. Hence, elementary flux patterns allow for a more realistic assessment of the robustness.

Analysis of host–pathogen interactions

The cost for the computation of elementary flux patterns only scales polynomially with the size of the underlying genome-scale network. Thus, it is possible to analyze large-scale networks that are made up of several genome-scale metabolic networks. As such, it is possible, for instance, to define a subsystem that contains metabolic reactions of a host and vital reactions of a parasite in order to determine the interplay between the metabolisms of both organisms (Raghunathan et al. 2009). In this context, an interesting application is to determine which combinations of exchange reactions with the medium of the parasite in the host can be impaired in order to harm the parasite.

In summary, the concept of elementary flux patterns opens up an entire new avenue for the analysis of genome-scale metabolic networks. It allows for the incorporation of all the information available in a genome-scale metabolic network when analyzing a specific subsystem. This is of central importance, since a comprehensive knowledge about the, for instance, 3359 reactions and 1972 species in the model of the metabolism of *E. coli* by Feist et al. (2007) is difficult to achieve.

Acknowledgments

We thank Peter Dittrich, whose ideas initially led to the development of the concept of elementary flux patterns. We thank three anonymous referees for very helpful comments. We acknowledge financial support from the German Ministry for Research and Education (BMBF) to C.K. within the framework of the Forsys Partner initiative and from the Fundação Calouste Gulbenkian, Fundação para a Ciência e a Tecnologia (FCT) and Siemens SA Portugal (Ph.D. grant SFRH/BD/32961/2006) to L.F.F..

Appendix

As in the main text, we assume that the first k columns (i.e., reactions) of the stoichiometric matrix $\mathbf{M}$ correspond to the reactions of the subsystem. We will outline the algorithm for the computation of elementary flux patterns by first demonstrating how a linear program can be formulated that allows to test whether a set of reactions fulfills the flux pattern condition (conditions 3 to 6). Then we proceed by integrating this linear program into a mixed-integer linear program that allows us to enumerate all elementary flux patterns.

Each flux pattern s is part of at least one flux $\mathbf{v} \in \mathbb{R}^n$ through the entire system that fulfills two conditions. First, $\mathbf{v}$ needs to balance all species—that is, it needs to be at steady state; and second, $\mathbf{v}$ needs to obey the irreversibility of some reactions. Since we split reversible reactions into two irreversible forward and backward steps, this is equivalent to the condition that all fluxes need to be non-negative. In terms of a linear program with the variables $\mathbf{v}$, this translates into the constraints:

$$(\mathrm{LP}\,1)\ \mathbf{M} \cdot \mathbf{v} = 0$$

$$(\mathrm{LP}\,2)\ \mathbf{v} \geq 0.$$

Furthermore, we require that $\mathbf{v}$ has nonzero entries for the reactions of the flux pattern s and zero entries for the remaining reactions of the subsystem. Thus, we add

$$(\mathrm{LP}\,3)\ \forall i \in s : v_i \geq 1$$

$$(\mathrm{LP}\,4)\ \forall j \in \{1, \ldots, k\} \backslash s : v_j = 0$$

as additional constraints. Note that we require that $v_i \geq 1$ because it is not possible to formulate the constraint $v_i > 0$ in a linear program. Since the constraints do not impose any upper bounds on $\mathbf{v}$,

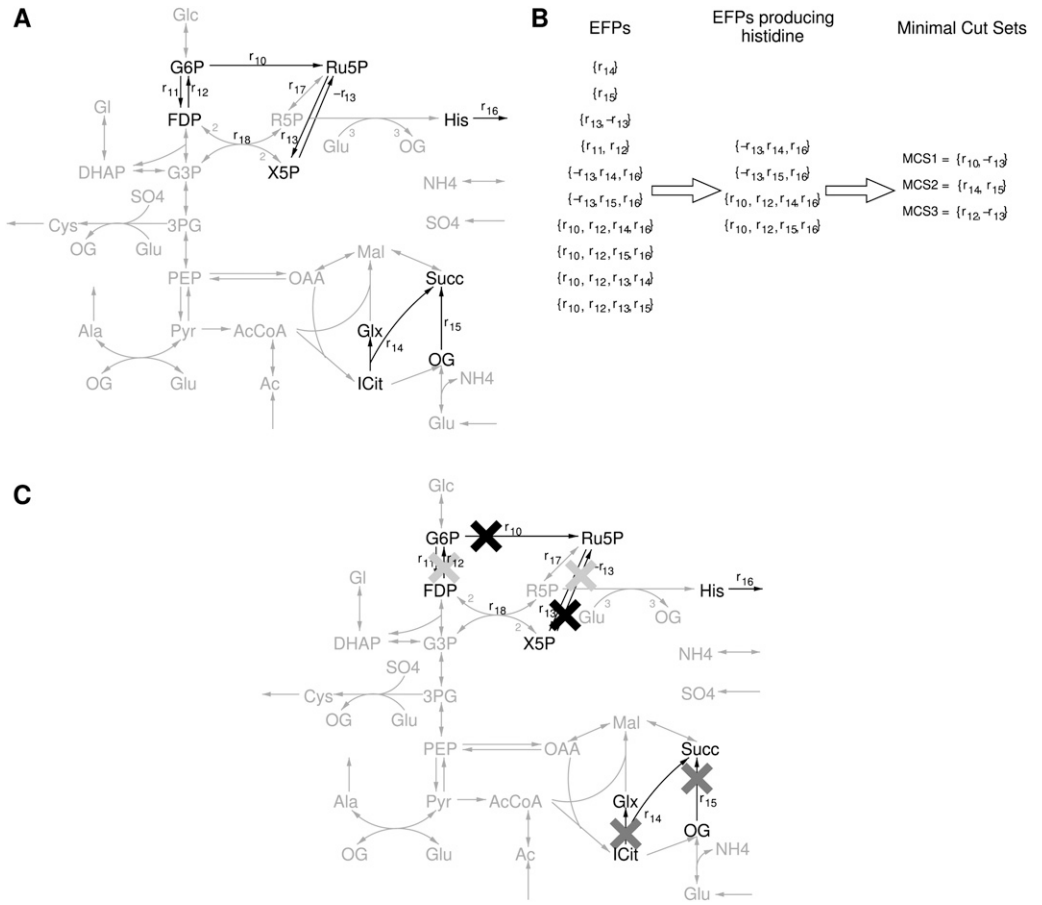


Figure 8. Scenario for the application of minimal cut sets. (*A*) Reaction network under consideration, which is supplied with acetate and glutamate. It is assumed that the production of histidine should be prevented. Histidine is produced from the central metabolism intermediate ribose-5 phosphate (Umbarger 1978). Assuming that it is not possible to entirely knock out the reactions r_{17} and r_{18} that directly produce ribose-5 phosphate, the reactions r_{10}–r_{15} are defined as alternative targets (black reactions). (*B*) In order to determine minimal cut sets, r_{10}–r_{15} are added to the subsystem containing the outflow reaction of histidine. From the elementary flux patterns, the minimal cut sets can be obtained. (*C*) Minimal sets of reactions that need to be knocked out to suppress the production of histidine: (black crosses) MCS1; (dark gray crosses) MCS2; (light gray crosses) MCS3. Subsequently, the minimal cut sets can be ranked according to different criteria, like side effects or required effort to knock out the corresponding genes, to determine the optimal knockout strategy. A list of abbreviations can be found in Supplemental material S2.

the constraint $v_i > 0$ is equal to $v_i \geq 1$. To fulfill the flux pattern condition, we only need to test whether **v** exists. Thus, we only need to check the feasibility of the linear program, and no objective function is required.

In order to find all elementary flux patterns, we need to combine constraints LP 1 and LP 2 with a mixed-integer linear program. Doing this, we first need to introduce a mapping from **v** to a binary variable $\mathbf{b} \in \{0, 1\}^k$ indicating the reactions used by **v** in the subsystem. $b_i = 1$ indicates that reaction i of the subsystem is used, and $b_i = 0$ indicates the contrary. Thus, in addition to (LP 1) and (LP 2), we add the constraint:

$$(\text{MILP 1})\ \forall i \in \{1, \ldots, k\} : b_i \leq v_i \leq c \cdot b_i$$

with a sufficiently large constant c. If $b_i = 0$, the lower and upper bounds of this constraint are 0, hence v_i is constrained to zero. In the other direction, $b_i = 1$ implies that $1 \leq v_i \leq c$. Hence, v_i is larger than 1 and smaller than c. In consequence, c has to be chosen either as the maximal velocity of reaction i or as a general maximal reaction velocity of the network under consideration. **b** indicates the reactions the flux vector **v** is using in the subsystem and, hence, corresponds to a flux pattern. Thus, we introduce the mapping:

$$\Theta(\mathbf{b}) = \{i | b_i = 1\}$$

from **b** to the encoded flux pattern. The idea of the MILP is to iteratively search for elementary flux patterns until no further flux pattern can be found. This can be achieved by iteratively solving the MILP and removing, each time, the set of previously found elementary flux patterns S from the solution space by adding an additional constraint. We can exclude previously found elementary flux patterns by requiring that each new flux pattern cannot be written as a combination of previously found elementary flux patterns. In order to implement this constraint in the MILP, we first need to reformulate it. This can be done by taking a closer look at the reactions contained in $\Theta(\mathbf{b})$. If we find that each reaction $r \in \Theta(\mathbf{b})$ is also contained in a previously found elementary flux pattern $s' \in S$ that is a subset of $\Theta(\mathbf{b})$, this implies that $\Theta(\mathbf{b})$ is a combination of elements from S. Since the contrary also holds, we need to ensure that $\Theta(\mathbf{b})$ contains at least one r that is not an element of any previously found elementary flux pattern that is

a subset of $\Theta(\mathbf{b})$. This can be achieved by introducing an additional set of binary variables $\mathbf{h} \in \{0, 1\}^k$ with $h_i = 1$ indicating that reaction i is an element of $\Theta(\mathbf{b})$ and not an element of any elementary flux pattern $s' \in S$ that is a subset of $\Theta(\mathbf{b})$. In consequence, h_i can only equal 1 if b_i does so. Thus, we add the constraint:

$$(\text{MILP } 2)\ \forall i \in \{1, \ldots, k\} : b_i - h_i \geq 0.$$

Furthermore, we want for each $h_i = 1$ that reaction i is not an element of any s' that is a subset of $\Theta(\mathbf{b})$. For each s', we can count the number of common elements with $\Theta(\mathbf{b})$ by the sum $\sum_{i \in s'} b_i$. This sum is equal to the number of elements $|s'|$ in s' if and only if s' is a subset of $\Theta(\mathbf{b})$. Thus, we add the constraint:

$$(\text{MILP } 3)\ \forall s' \in S : \sum_{i \in s'} (b_i + h_i) \leq |s'|.$$

In consequence, if s' is a subset of $\Theta(\mathbf{b})$, all h_i with $i \in s'$ are constrained to zero. Conversely, each $h_r = 1$ indicates a reaction r that is not an element of any elementary flux pattern in S that is a subset of $\Theta(\mathbf{b})$. In order to ensure that we find at least one reaction $r \in \Theta(\mathbf{b})$ that fulfills this condition, we add

$$(\text{MILP } 4)\ \sum_{i=1}^{k} h_i \geq 1.$$

Thus, we ensure that $\Theta(\mathbf{b})$ is not a combination of previously found elementary flux patterns. To guarantee that we find only flux patterns that are elementary, the objective function for the MILP is the minimization of $\sum_{i=1}^{k} b_i$, the number of reactions of the flux pattern $\Theta(\mathbf{b})$ [for the proof of elementarity of $\Theta(\mathbf{b})$, see Supplemental material S6].

In each iteration, the MILP returns a new elementary flux pattern. Finally, we obtain no further solution if all elementary flux patterns have been found. For a condensed list of the constraints of the MILP, see Supplemental material S6.

References

Acuña V, Chierichetti F, Lacroix V, Marchetti-Spaccamela A, Sagot MF, Stougie L. 2009. Modes and cuts in metabolic networks: Complexity and algorithms. *Biosystems* **95:** 51–60.

Behre J, Wilhelm T, von Kamp A, Ruppin E, Schuster S. 2008. Structural robustness of metabolic networks with respect to multiple knockouts. *J Theor Biol* **252:** 433–441.

Böcker S, Briesemeister S, Bui QBA, Truß A. 2008. A fixed-parameter approach for weighted cluster editing. In *Proceedings of Asia-Pacific Bioinformatics Conference (APBC 2008)*, volume 5, pp. 211–220. Series on Advances in Bioinformatics and Computational Biology. Imperial College Press, London, UK.

Borodina I, Nielsen J. 2005. From genomes to in silico cells via metabolic networks. *Curr Opin Biotechnol* **16:** 350–355.

Burgard AP, Nikolaev EV, Schilling CH, Maranas CD. 2004. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res* **14:** 301–312.

Carlson R, Fell D, Srienc F. 2002. Metabolic pathway analysis of a recombinant yeast for rational strain development. *Biotechnol Bioeng* **79:** 121–134.

Covert MW, Schilling CH, Palsson BØ. 2001. Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* **213:** 73–88.

Covert MW, Xiao N, Chen TJ, Karr JR. 2008. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics* **24:** 2044–2050.

de Figueiredo LF, Podhorski A, Rubio A, Beasley JE, Schuster S, Planes FJ. 2009. Calculating the *k*-shortest elementary flux modes in metabolic networks. In *Proceedings MATHMOD 09 Vienna—full papers CD volume* (eds. I Troch and F Breitenecker), pp. 736–747. http://www.mathmod.at/index.php?id=115.

Downey RG, Fellows MR. 1998. Parameterized complexity. In *Monographs in computer science*, pp. 489–516. Springer, New York.

Duarte NC, Herrgård MJ, Palsson BØ. 2004. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* **14:** 1298–1309.

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ. 2007. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci* **104:** 1777–1782.

Fait A, Fromm H, Walter D, Galili G, Fernie AR. 2008. Highway or byway: The metabolic role of the GABA shunt in plants. *Trends Plant Sci* **13:** 14–19.

Feist AM, Palsson BØ. 2008. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol* **26:** 659–667.

Feist AM, Scholten JCM, Palsson BØ, Brockman FJ, Ideker T. 2006. Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Mol Syst Biol* **2:** 2006.0004. doi: 10.1038/msb4100046.

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ. 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3:** 121. doi: 10.1038/msb4100155.

Fischer E, Sauer U. 2003. Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *Eur J Biochem* **270:** 880–891.

Gagneur J, Klamt S. 2004. Computation of elementary modes: A unifying framework and the new binary approach. *BMC Bioinformatics* **5:** 175. doi: 10.1186/1471-2105-5-175.

Hansen RW, Hayashi JA. 1962. Glycolate metabolism in *Escherichia coli*. *J Bacteriol* **83:** 679–687.

Holzhütter HG. 2004. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur J Biochem* **271:** 2905–2922.

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, et al. 2003. The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* **19:** 524–531.

Jamshidi N, Palsson BØ. 2007. Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Syst Biol* **1:** 26. doi: 10.1186/1752-0509-1-26.

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

Klamt S, Stelling J. 2002. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep* **29:** 233–236.

Klamt S, Gagneur J, von Kamp A. 2005. Algorithmic approaches for computing elementary modes in large biochemical reaction networks. *Syst Biol (Stevenage)* **152:** 249–255.

Krömer JO, Wittmann C, Schröder H, Heinzle E. 2006. Metabolic pathway analysis for rational design of L-methionine production by *Escherichia coli* and *Corynebacterium glutamicum*. *Metab Eng* **8:** 353–369.

Li M, Ho PY, Yao S, Shimizu K. 2006. Effect of *lpdA* gene knockout on the metabolism in *Escherichia coli* based on enzyme activities, intracellular metabolite concentrations and metabolic flux analysis by ^{13}C-labeling experiments. *J Biotechnol* **122:** 254–266.

Liebermeister W, Baur U, Klipp E. 2005. Biochemical network models simplified by balanced truncation. *FEBS J* **272:** 4034–4043.

Lougee-Heimer R. 2003. The Common Optimization INterface for Operations Research: Promoting open-source software in the operations research community. *IBM J Res Develop* **47:** 57–66.

Mahadevan R, Schilling CH. 2003. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* **5:** 264–276.

Nuño JC, Sánchez-Valdenebro I, Pérez-Iratxeta C, Meléndez-Hevia E, Montero F. 1997. Network organization of cell metabolism: Monosaccharide interconversion. *Biochem J* **324:** 103–111.

Oh YK, Palsson BØ, Park SM, Schilling CH, Mahadevan R. 2007. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* **282:** 28791–28799.

Pfeiffer T, Sánchez-Valdenebro I, Nuño JC, Montero F, Schuster S. 1999. Metatool: For studying metabolic networks. *Bioinformatics* **15:** 251–257.

Raghunathan A, Reed J, Shin S, Palsson B, Daefler S. 2009. Constraint-based analysis of metabolic capacity of *Salmonella typhimurium* during host–pathogen interaction. *BMC Syst Biol* **3:** 38. doi: 10.1186/1752-0509-3-38.

Raman K, Chandra N. 2009. Flux balance analysis of biological systems: Applications and challenges. *Brief Bioinform* **10:** 435–449.

Richard HT, Foster JW. 2003. Acid resistance in *Escherichia coli*. *Adv Appl Microbiol* **52:** 167–186.

Sauro HM, Hucka M, Finney A, Wellock C, Bolouri H, Doyle J, Kitano H. 2003. Next generation simulation tools: The Systems Biology Workbench and BioSPICE integration. *OMICS* **7:** 355–372.

Schuster S, Fell D. 2007. Modelling and simulating metabolic networks. In *Bioinformatics: From genomes to therapies* (ed. T Lengauer), Vol. 2, pp. 755–806. Wiley-VCH, Weinheim, Germany.

Schuster S, Dandekar T, Fell DA. 1999. Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol* **17:** 53–60.

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

Schuster S, Klamt S, Weckwerth W, Pfeiffer T. 2002a. Use of network analysis of metabolic systems in bioengineering. *Bioprocess Biosyst Eng* **24:** 363–372.

Schuster S, Pfeiffer T, Moldenhauer F, Koch I, Dandekar T. 2002b. Exploring the pathway structure of metabolism: Decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics* **18:** 351–361.

Schwender J, Goffman F, Ohlrogge JB, Shachar-Hill Y. 2004. Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. *Nature* **432:** 779–782.

Shlomi T, Eisenberg Y, Sharan R, Ruppin E. 2007. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol Syst Biol* **3:** 101. doi: 10.1038/msb4100141.

Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED. 2002. Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420:** 190–193.

Terzer M, Stelling J. 2008. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics* **24:** 2229–2235.

Thiele I, Vo TD, Price ND, Palsson BØ. 2005. Expanded metabolic reconstruction of *Helicobacter pylori* (iIT341 GSM/GPR): An in silico genome-scale characterization of single- and double-deletion mutants. *J Bacteriol* **187:** 5818–5830.

Tollnick C, Seidel G, Beyer M, Schügerl K. 2004. Investigations of the production of cephalosporin C by *Acremonium chrysogenum*. *Adv Biochem Eng Biotechnol* **86:** 1–45.

Trinh CT, Wlaschin A, Srienc F. 2009. Elementary mode analysis: A useful metabolic pathway analysis tool for characterizing cellular metabolism. *Appl Microbiol Biotechnol* **81:** 813–826.

Umbarger HE. 1978. Amino acid biosynthesis and its regulation. *Annu Rev Biochem* **47:** 532–606.

Varma A, Palsson BØ. 1994. Metabolic flux balancing: Basic concepts, scientific and practical use. *Technology (NY)* **12:** 994–998.

von Kamp A, Schuster S. 2006. METATOOL 5.0: Fast and flexible elementary modes analysis. *Bioinformatics* **22:** 1930–1931.

Weinman EO, Srisower EH, Chaikoff IL. 1957. Conversion of fatty acids to carbohydrate; application of isotopes to this problem and role of the Krebs cycle as a synthetic pathway. *Physiol Rev* **37:** 252–272.

Wittmann C. 2007. Fluxome analysis using GC-MS. *Microb Cell Fact* **6:** 6. doi: 10.1186/1475-2859-6-6.

Wittmann C, Weber J, Betiku E, Krömer J, Böhm D, Rinas U. 2007. Response of fluxome and metabolome to temperature-induced recombinant protein synthesis in *Escherichia coli*. *J Biotechnol* **132:** 375–384.

Received December 19, 2008; accepted in revised form June 8, 2009.

C. Kaleta, L. F. de Figueiredo, R. Guthke, and S. Schuster. *Biochem. J.*, 2009, in preparation.

Systematic determination of gluconeogenic pathways from fatty acids in humans

Christoph KALETA[*,†,1], Luís F. de FIGUEIREDO[*], Reinhard GUTHKE[‡], Stefan SCHUSTER[*]
[*] Department of Bioinformatics, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany
[†] Bio Systems Analysis Group, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany
[‡] Systems Biology/Bioinformatics Group, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute[3], D-07743 Jena, Germany

The question whether glucose can be produced from fatty acids in humans has a long standing tradition in biochemistry. An important pathway allowing this conversion in plants, some bacteria, fungi and nematodes is the glyoxylate bypass (in conjunction with the tricarboxylic acid cycle and gluconeogenesis). However, the corresponding enzymes have not yet been found in mammals, which led to a general consent and textbook dogma that gluconeogenesis from fatty acids is impossible in humans. We re-investigated this question using a recently introduced method for pathway analysis in large-scale metabolic networks, elementary flux patterns. Analyzing a genome-scale model of human metabolism we found pathways on which gluconeogenesis from fatty acids is indeed feasible in humans. Investigating these pathways in detail we identified essential reactions and alternative routes that can be taken. Previous works support some of the routes we detected, but have been largely ignored by the scientific community. On the proposed pathways 4 moles of acetyl-CoA are converted into one mole of glucose, and two moles of carbon dioxide. Even though the existence of the acetoacetate decarboxylase, an enzyme essential for these pathways, has not yet been clarified in humans, there is clear evidence that acetone produced from decarboxylation of acetoacetate can be used for gluconeogenesis. The discovery of several hitherto unreported pathways in acetone metabolism further underlines the potential of elementary flux pattern analysis in refining and extending our knowledge on the metabolic capabilities of living organisms.

Key words: Gluconeogenesis, fatty acids, genome-scale metabolic networks, elementary flux patterns.

It is well-known that excess sugar in the human diet can readily be converted both into glycerol and fatty acids and, thus, into lipids and triglycerides. Thus, a question that biochemistry students are often asked in their exams is whether fatty acids are gluconeogenic in humans,[2] that is, whether they can be converted back into glucose. The expected answer is "No". This summarizes the result of a debate that lasted more than five decades. Starting in the late nineteenth century the question whether carbohydrates could be produced from fatty acids was an important topic of research. Especially a possible link to diabetes spurred the discussion. However, it was not until the 1950s that such a conversion could be monitored directly using ^{14}C labeled fatty acids (1). It was found that part of the label arrives at glucose, proving that there is a connected route from acetyl-CoA to glucose. However, as shown mathematically (1, 2), there cannot be any sustained conversion at steady-state along the TCA cycle due to stoichiometric constraints. In particular, oxaloacetate would not be balanced. A possible route that does allow this conversion in prokaryotes (3), plants (4), fungi (5) and nematodes (6), the glyoxylate shunt, was reported in 1957 (3). It produces an additional oxaloacetate, thus balancing this compound. However, the corresponding enzymes have not been found in humans (7). Due to the work of Weinman *et al.* (1) and the absence of the enzymes of the glyoxylate shunt in humans, there is a general consent that carbohydrates cannot be produced from fatty acids in humans although the opposite conversion is feasible. In consequence, this statement can be found throughout prominent biochemistry textbooks (8–10).

Recently we re-investigated this question in a small model of the central metabolism of humans (2). To this aim we used the concept of elementary modes (11, 12), which allows to detect all feasible metabolic pathways in a reaction network, and were able to corroborate the results of Weinman *et al.* (1). A disadvantage of elementary modes is that they cannot be applied to large-scale metabolic networks such as genome-scale networks (13). To overcome these limitations, the concept of elementary flux patterns has been introduced (14, 15). Similar to elementary modes, elementary flux patterns consider possible pathways through a small-scale system. However, in contrast to elementary modes, this small-scale system is embedded into a large-scale system by allowing only for pathways through the small system that are part of a pathway in the large system (Figure 1). This allows, besides from being applicable to much larger networks, a more reliable analysis of the metabolic capabilities of the subsystem as compared to elementary modes (14, 15). Furthermore, all potential pathways in the large system are taken into account and, hence, elementary flux patterns allow to detect all possible routes for the conversion of one compound into another in large-scale metabolic networks.

Theoretical and simulation studies have turned out to be very useful also for other issues in biochemistry (16–18). Thus, equipped with elementary flux patterns, we want to answer the question whether carbohydrates can be produced from fatty acids within a genome-scale model of human metabolism (19).

Abbreviations used: TCA cycle, tricarbolic acid cycle; CoA, coenzyme A.

[1]To whom correspondence should be addressed (email Christoph.Kaleta@uni-jena.de)

[2]We are concerned with even-chain fatty acids here. The β-oxidation of odd-chain fatty acids yields propionate, which is gluconeogenic. However, odd-chain fatty acids are only found in very small amounts in humans.

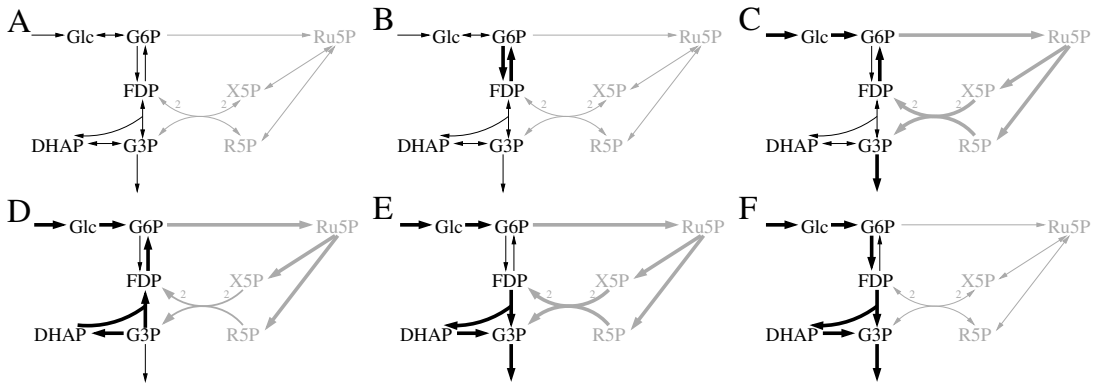


Figure 1: Examples of elementary flux patterns. **A** Example network of condensed upper glycolysis and pentose phosphate pathway with corresponding elementary flux patterns. Reactions of the subsystem are drawn in black. **B-F** Elementary flux patterns of the system. Dark thick arrows correspond to the reactions of each flux pattern in the subsystem. Gray thick arrows indicate the reactions used by an elementary mode of the entire system using the reactions of the flux pattern in the subsystem. Individual flux ratios have been omitted for clarity. A list of abbreviations can be found in Supplemental Material B.

METHODS

Elementary modes are a frequently used concept in pathway analysis (2, 11, 12, 20–23, for a review see (24)). Elementary modes represent minimal sets of reactions that can operate at steady state with all reactions proceeding in their appropriate direction (12). The reaction set is minimal in the sense that there is no subset of reactions that could also operate at steady state. Even though there have been improvements in existing algorithms (25–27), the analysis with elementary modes or the related extreme pathways (28) is constrained to medium size networks (25, 27). Thus, normally only a subnetwork of the entire known (whole-cell) system is considered. The interface to the remaining system is modeled using abstractions like exchange reactions, which correspond to inflow or outflow reactions, and external metabolites, which need not be balanced in the elementary modes. However, these abstractions cannot fully take into account all the metabolic capabilities of the entire system and the results depend heavily on their choice. In consequence, elementary mode analysis not necessarily finds all pathways through the subsystem or might even detect pathways that are not part of any steady-state flux through the entire system (14, 15).

To circumvent these problems, elementary flux patterns have been introduced (14, 15). Elementary flux patterns are defined as the basic routes of each steady-state flux of a large-scale metabolic network through a predefined subsystem, corresponding to specific reactions of interest within this model. Through their definition, each elementary flux pattern is associated to at least one elementary mode in the complete system and therefore represents specific parts of a feasible metabolic pathway (14, 15). Following an approach outlined in (15) this elementary mode can be obtained. Thus, even though a subsystem has to be predefined, it can be iteratively enlarged. On the other hand it is possible to scale pathway analysis, by considering an overall pathway and focusing on certain parts where alternative routes can be analyzed in detail. For an example of a reaction network and elementary flux patterns in a subsystem see Figure 1.

Formal definition of elementary flux patterns will be given next.

For a metabolic network with n reactions among m compounds the $m \times n$ stoichiometric matrix $\mathbf{N}$ indicates the consumption and production of each compound in each reaction. An entry n_{ij} of $\mathbf{N}$ is negative if compound i is an educt of reaction j and positive, if it is a product.

For simplicity we assume that the columns in $\mathbf{N}$ are ordered such that the first k columns represent the reactions of the subnetwork. A set s of reaction indices i with $1 \leq i \leq k$ is a flux pattern if we can find a flux vector $\mathbf{v} \in \mathbb{R}^n$ which fulfills the following conditions

$$\mathbf{v} \geq 0 \tag{1}$$

$$\mathbf{N} \cdot \mathbf{v} = 0 \tag{2}$$

$$\text{For all } i \text{ included in } s\text{: } v_i > 0 \tag{3}$$

$$\text{For all } j,\ 1 \leq j \leq k, \text{ not included in } s\text{: } v_j = 0 \tag{4}$$

Thus, a flux pattern is a set of reactions, or more precise, a set of reaction indices, of the subnetwork that is part of a steady-state flux $\mathbf{v}$ of the entire network. We require that the fluxes v_i which belong to the flux pattern s (i.e., $i \in s$) are positive, while the remaining are zero.

Given the set of elementary flux patterns S of a system, that is, a set of sets of reaction indices, we call a flux pattern $s \in S$ elementary if it cannot be written as the union of other flux patterns (not including s). For a more formal definition see (14, 15).

Pathway detection using elementary flux patterns

Elementary flux patterns can be used to elucidate all possible pathways consuming a certain compound and producing another. This process builds upon a successive expansion of the subsystem under study to reactions that belong to alternative pathways. It will be outlined by way of a small example network comprising glycolysis and the pentose phosphate pathway (Figure 2). Within this system we want to find all pathways producing ribose-5 phosphate (R5P), a precursor in histidine and nucleotide syntheses (29), from glucose (Glc).

Thus, we start with a subsystem encompassing the inflow reaction of Glc and the outflow of R5P (Figure 2A). We find two elementary flux patterns. One of them only contains the inflow reaction of Glc and the other the inflow of Glc as well as the outflow of R5P. The first elementary flux pattern indicates the existence of a pathway consuming Glc at steady state with-

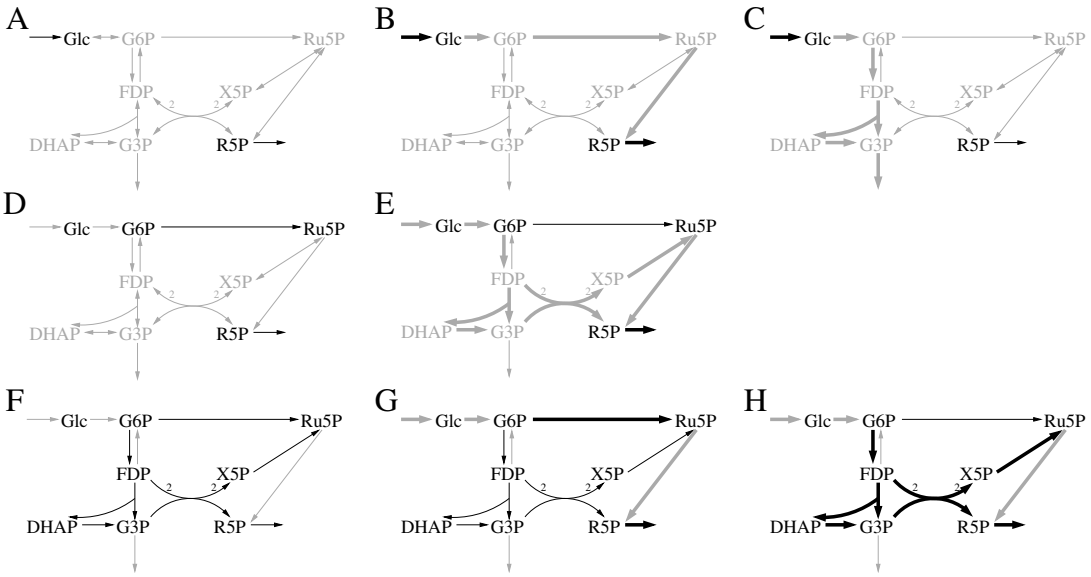


Figure 2: Schematic representation of the iteration process for searching pathways. Dark arrows correspond to reactions belonging to the subsystems. **A**, **D** and **F** Subsystems used in the different iteration steps. **B**, **C** and **E** Selected elementary flux patterns (dark thick arrows) and the reactions used by an associated elementary mode in the remaining system (gray thick arrow). If only one direction of a reversible reaction belongs to a subsystem, the reverse direction is omitted for clarity. **G** and **H** Elementary flux patterns of the final system producing R5P with the associated pathways through the entire system. A list of abbreviations can be found in Supplemental Material B.

out using the outflow of R5P. This corresponds to the glycolytic pathway producing glycerol-3 phosphate (G3P) which is subsequently drained from the system (Figure 2C). The other flux pattern corresponds to a pathway producing R5P (Figure 2B). Since we found no elementary flux pattern containing the outflow of R5P and not containing the inflow of Glc we can conclude that the inflow reaction is required for the production of R5P. Otherwise we would have obtained a second elementary flux pattern containing only the outflow of R5P.

Next, we need to determine an elementary mode through the entire system using exactly the reactions of the elementary flux pattern in the subsystem. Note that such an elementary mode can be obtained using linear programming not requiring the enumeration of all elementary modes of the entire system (15). This elementary mode corresponds to a first pathway for the production of R5P from Glc. From this initial pathway it is possible to deduce reactions that are essential for the conversion of R5P to Glc (see Supplemental Material C for more details). Thus, we find that in addition to the inflow of Glc, the conversion of Glc to glucose-6 phosphate (G6P) and the conversion of ribulose-5 phosphate (Ru5P) to R5P are required for production of R5P at steady state. The knowledge of essential reactions can simplify the analysis in two ways. First, we not need include essential reactions into the subsequent subsystems since every pathway producing R5P will use them anyway. Second, if we find several sequences of essential reactions the task of searching for pathways can be split into sub-tasks. Each sub-task then consists in the search for a pathway connecting a product of a sequence of essential reactions and the educt of the next sequence of essential reactions.

In order to determine reactions that belong to alternative pathways, we include the reactions of the first detected pathway into the subsystem of the next step. As noted above, we need not add essential reactions and, hence, the subsystem of the second step consists of two reactions (Figure 2B): the conversion of G6P to Ru5P and the outflow of R5P. This subsystem gives rise to two elementary flux patterns. One contains the conversion of G6P to Ru5P (not shown) and the other the outflow of R5P (Figure 2E). Thus, the second elementary flux pattern corresponds to a pathway for the production of R5P. Besides the essential reactions this flux pattern also uses reactions that do not belong to the subsystem. Hence, we have identified reactions that belong to an alternative route. These reactions are subsequently added to the subsystem of the third step (Figure 2F). In this subsystem we find eight elementary flux patterns, two of which contain the outflow of R5P (Figure 2G and 2H). Determining the elementary modes associated to these elementary flux patterns, we find that both use only reactions that are either essential for the pathway or belong to the subsystem. Hence, we have identified all pathways producing R5P from Glc. These pathways correspond to the elementary modes associated to each of the two flux patterns containing the outflow of R5P.

RESULTS

From the BiGG database (`http://bigg.ucsd.edu/`) we downloaded a fully compartmentalized genome-scale model of human metabolism that has been presented in (19). For the analysis we split reversible reactions into irreversible forward and backward directions. Additionally, we added an inflow of the compounds NH_4, Fe^{2+}, Fe^{3+}, H^+, K^+, Ca^{2+}, Na^+, phosphate, O_2, water, and Cl that are necessary for the cell to survive. To simulate growth on even-chain fatty acids, we added an influx of cytosolic acetyl-CoA. Finally, we added an outflow of cytosolic glucose 6-phosphate in order to simulate gluconeogenesis. Thus, we obtained a network containing 3673 reactions and 3188 metabolites.

Step 1: Determining an initial pathway. In a first step we de-

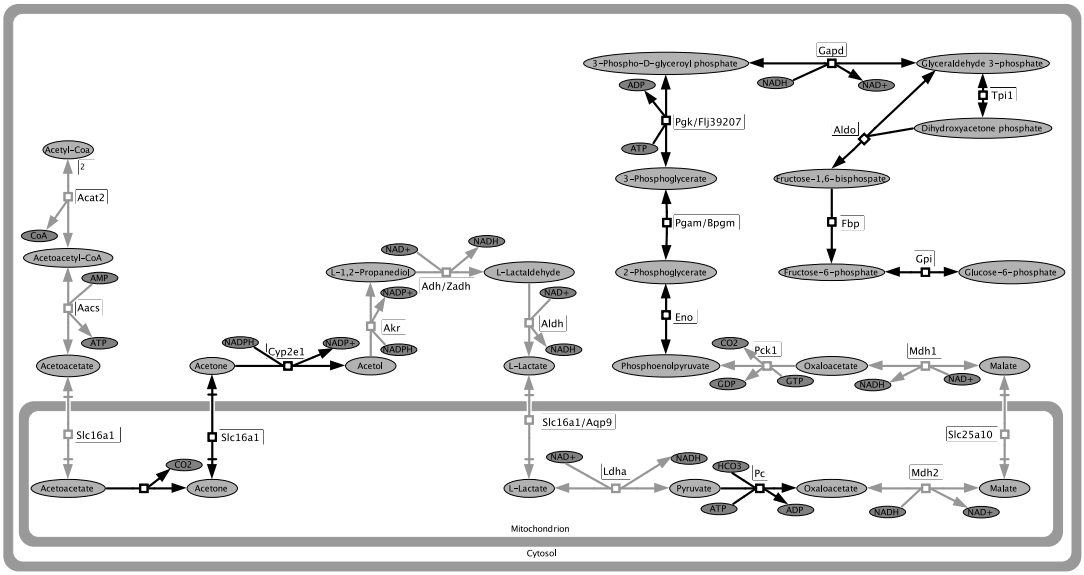


Figure 3: Pathway for the conversion of fatty acids into glucose. Essential reactions of the pathway are drawn in black, non-essential reactions are drawn in gray. A list of abbreviations as well as the participating proteins in each reaction can be found in Supplemental Materials A and B.

termined whether the model could produce glucose 6-phosphate from acetyl-CoA. Thus, as outlined in Methods we computed the elementary flux patterns of a subsystem that only contained the inflow reaction of cytosolic acetyl-CoA and the outflow of cytosolic glucose 6-phosphate. This system gives rise to 2 elementary flux patterns. The first flux pattern only contains the inflow of acetyl-CoA. This represents the capability of the model to metabolize acetyl-CoA, for instance, by conversion into CO_2 along the TCA cycle. The second elementary flux pattern contains both the inflow and the outflow reactions. Thus, our first conclusion is that, in opposition to common biochemical dogma, there is indeed a pathway that can produce glucose from fatty acids in humans. This first step in the analysis is of importance, since the model could have included pathways that allow the production of gluconeogenic precursors from sources other than acetyl-CoA. In this case we would have found that the second elementary flux pattern contains only the outflow of glucose 6-phosphate.

Following the procedure outlined in Methods, we next identified a pathway for the conversion of fatty acids to glucose by examining a genome-scale elementary mode that contained the second elementary flux pattern of the above system. We found a pathway that contains 26 reactions (Figure 3). It uses a route which involves the formation of ketone bodies through the degradation of acetone towards pyruvate, via acetol and L-1,2-propanediol.

Step 2: Determining alternative routes. Next we identified essential reactions of the detected pathway. We identified 12 such reactions (Figure 3). Analyzing these reactions in detail we found that they belong to a linear sequence of three pathways, one terminating with the outflow of glucose-6 phosphate. Following the approach outlined in Methods it was thus possible to split the search for pathways into three distinct tasks. These tasks correspond to the search for pathways in the conversion of

(1) Cytosolic or mitochondrial acetyl-CoA into mitochondrial acetoacetate

(2) Cytosolic acetol into mitochondrial pyruvate

(3) Mitochondrial oxaloacetate into cytosolic phosphoenolpyruvate

Since (1) and (3) correspond to well-known pathways, ketogenesis and gluconeogenesis from pyruvate, respectively, we next focused on the elucidation of possible metabolic pathways in (2).

Conversion of acetol to pyruvate

We determined the alternative routes of the conversion of cytosolic acetol to either cytosolic or mitochondrial pyruvate using the approach outlined in Methods. The subsystem in the final iteration contains 22 reactions and gives rise to 114 elementary flux patterns of which 22 contain the production of glucose 6-phosphate (Figure 4; see Table D in the Supplement Material for a list of pathways). The first branching of the pathways appear in the conversion of acetol to methylglyoxal or to L-1,2-propanediol. Methylglyoxal is either directly converted into pyruvate or over the intermediates D-lactaldehyde and D-lactate. Additionally, a route via lactoyl-glutathione is possible. The pathway via D-lactaldehyde was up to date unreported in the context of acetone metabolism (30). The alternative route via L-1,2-propanediol is followed by a conversion into L-lactaldehyde. This compound can be either converted into methylglyoxal or L-lactate. In the former case, the above mentioned routes can be taken and, in the latter case, L-lactate is converted into pyruvate in either mitochondrion or cytosol. In the cytosol the conversion to pyruvate can proceed by reduction of either NAD^+or ferricytochrome.

Most of the reactions in acetone metabolism oxidize NADPH or reduce NAD^+. Overall 1 to 3 NADPH are oxidized and 0 to 4 $NAD^+$are reduced (see Supplemental Material D for a detailed list). Special emphasis can be put on the oxidation of methylglyoxal to pyruvate. In this case one $NADP^+$can be reduced such that the net balance in NADPH is only -1 (pathways 7 and 14, Table 1). Pathways not using this reaction have a net balance of a least 2 NADPH oxidized. Thus, using these pathways always requires a replenishment of NADPH. We will outline the impacts

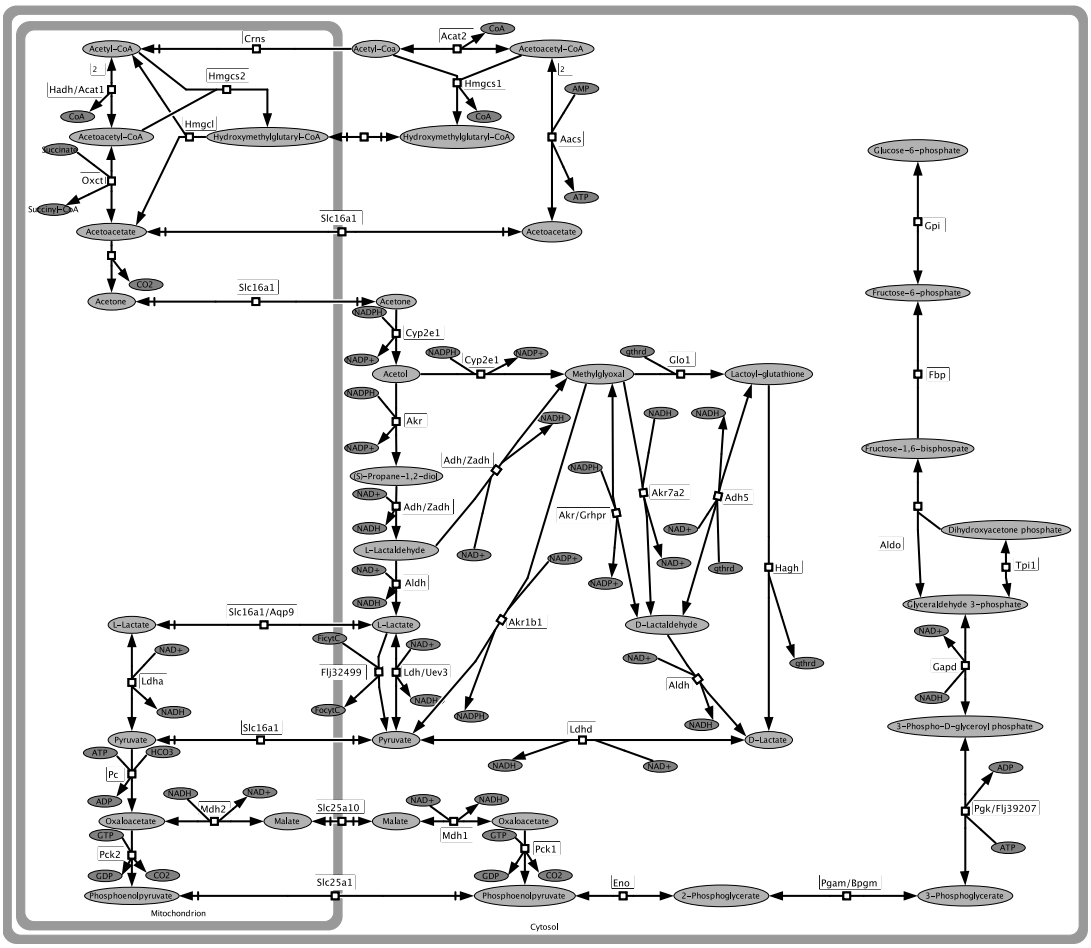


Figure 4: Gluconeogenic pathways from fatty acids. The pathways from acetyl-CoA to acetone correspond to the major pathways in ketogenesis. Only two gluconeogenic pathways from pyruvate to glucose 6-phosphate are shown. Import of acetyl-CoA into the mitochondrion proceeds via carnitine transporters. The dissociation of hydroxymethylglutaryl-CoA to acetoacetyl-CoA and acetyl-CoA can also proceed in the peroxysom. Export of phosphoenolpyruvate from the mitochondrion necessitates antiport of citrate. Aqp9, Akr and one form of Acat2 are noted as irreversible in the model, but are lumped together with reversible reactions. A list of enzymes can be found in Supplemental Material A. Abbreviations: gthrd-reduced glutathione, focytC-ferrocytochrome, ficytC-ferricytochrome.

#	**Reactions**	**Enzyme**	$\frac{\mathbf{NADPH}}{\mathbf{NADH}}$
7	NADPH + Acetol$_c$ → NADP$^+$ + 12ppd-S$_c$	Akr	-1/+2
	NAD$^+$ + 12ppd-S$_c$ → NADH + Lald-L$_c$	Adh/Zadh	
	NAD$^+$ + Lald-L$_c$ → NADH + Mthgxl$_c$	Adh/Zadh	
	NADP$^+$ + Mthgxl$_c$ → NADPH + Pyr$_c$	Akr1b1	
	Pyr$_c$ → Pyr$_m$	Slc16a1	
14	NADPH + Acetol$_c$ → NADP$^+$ + Mthgxl$_c$	Cyp2e1	-1/0
	NADP$^+$ + Mthgxl$_c$ → NADPH + Pyr$_c$	Akr1b1	
	Pyr$_c$ → Pyr$_m$	Slc16a1	

Table 1: Two most efficient pathways for the conversion of acetol into pyruvate in terms of oxidized NADPH. The second column indicates the reactions and the fourth column the ratio of consumed NADPH to produced NADH in each pathway. This number also includes the oxidized NADPH in the conversion of acetone to acetol. Indices indicate the localization of the metabolite: c - cytosol, m - mitochondrion. A full list of pathways can be found in Supplemental Material D. A list of abbreviations as well as the proteins catalyzing each reaction can be found in Supplemental Material A and B.

of this requirements in the Discussion.

An important point is the import of pyruvate into mitochondria. As presented in Figure 4 pyruvate can pass the membrane to the mitochondrion either via direct import or via conversion to lactate, import and subsequent reconversion to pyruvate. The latter mechanism is called lactate shuttle (31). However, we will exclude it from our analysis since it appears to play no significant role in pyruvate import into the mitochondrion (32).

DISCUSSION

An overview of the pathways for conversion of fatty acids into glucose is given in Figure 4. Besides the pathways in acetone degradation, also the ketogenic routes producing acetoacetate from acetyl-CoA and two gluconeogenic routes from pyruvate are displayed.

Acetoacetate decarboxylase in mammals

An essential reaction in the presented pathways is the decarboxylation of acetoacetate into acetone. This reaction is noted as non-enzymatic in the human model of (19). The corresponding enzyme, the acetoacetate decarboxylase, has been known in bacteria for decades (33). In contrast, there have been only few reports on the presence of such an enzyme in mammals. The supposed protein has been characterized in catalytic activity (34) and inhibitors (35). However, to our knowledge, it has never been purified, sequenced, nor are there known homologues. Since 1986 there has been no published work presenting further evidence and, thus, it remains to be investigated whether this reaction is catalyzed by a protein in humans (cf. (30)). Nevertheless, as we will outline in the following, there is a constant decarboxylation of acetoacetate into acetone. Furthermore, in ketogenic diets, drastically increased levels of acetone have been observed (36).

Evidence of gluconeogenesis from fatty acids

Owing to our results we further searched for evidence that supported the pathways we detected. A route via acetone as a potential pathway for gluconeogenesis from fatty acids had been proposed already in the pioneering work by Weinman (1). This route was further substantiated by the reports of enzymatic conversion of acetoacetate into acetone in animals (34, 35). Ongoing research on acetone metabolism led to the discovery of many further potentially gluconeogenic routes (37–39), covering some of the pathways presented in this work. However, these findings seem to have been largely ignored by the scientific community as can be seen from the statement about the non-existence of gluconeogenic routes from fatty acids in humans in prominent biochemistry textbooks (8–10). That the presence of the acetoacetate decarboxylase has not yet been confirmed in humans might play a role in that.

Thus, we next want to outline evidence supporting a role of fatty acids in gluconeogenesis.

Gluconeogenesis becomes important when the glucose level in the body can not be sustained any more by the glycogen store in the liver. Such a situation arises, for instance, during starvation and fasting. During starvation the glycogen supply of the liver is sufficient for up to 1 day in humans (40). When glycogen stores are depleted, cells start to break down proteins to supply amino acids for gluconeogenesis. Additionally the glycerol component of lipids can be converted into glucose while fatty acids serve as principal metabolite to fuel the TCA cycle. However, there is a limit to protein degradation. In a first phase mainly muscular protein is degraded. In prolonged starvation this protein store is not sufficient and also proteins essential for the maintenance of principal body functions are broken down to serve for gluconeogenesis (40). In this context, the central role of the presented pathways can be seen in the replenishment of TCA cycle intermediates, anaplerosis, in order to provide metabolites for gluconeogenesis and to maintain energy production.

The principal fuel of the brain under normal conditions is glucose. In starvation, since lipids cannot pass the blood/brain-barrier, the brain reduces its energy requirements and starts to use ketone bodies as principal fuel (41). The ketone bodies, acetoacetate, β-hydroxybuturate and acetone are produced from acetyl-CoA. The primary site of production is the mitochondrial matrix of liver cells (Figure 4 contains pathways for the production of acetoacetate). The increase of the production of ketone bodies leads to their accumulation and hence to rising levels of acetoacetate which is constantly decarboxylated into acetone. Thus, (42) noted that up to 37% of acetoacetate production were converted into acetone in 21 days fasted humans. Furthermore, since acetone is also excreted via urine and breath, it is important to note that only 2-30% of acetone is disposed through these pathways (42). The remainder could account for up to 11% of gluconeogenesis during starvation (42). This is also corroborated by a recent study in which it was concluded that lactate, pyruvate, glycerol, oxoglutarate equivalents, and alanine could not fully account for renal gluconeogenesis (40). Gluconeogenesis from acetone was suggested as an explanation for this difference (40). Additionally, starvation was found to activate cytochrome P450 (Cyp2e1) activity, an enzyme essential in the presented pathways (43).

Furthermore, other situations associated to an increased ketone body production, such as diabetic ketoacidosis, are associated to high levels of acetone (44, 45).

Another important facet of these pathways arises from evolutionary considerations. (46) found that increased survival time of obese rats during starvation correlated with the activity of acetone metabolism. Even though ketone bodies are the principal fuel for the brain during starvation, there is a constant requirement for glucose. Especially in humans this requirement is increased, since they have the largest brain size compared to the body size. Hence, the use of fatty acids for gluconeogenesis in humans is of evolutionary advantage since it allows to compensate for the increased rate of gluconeogenesis required during starvation.

Finally, it is important to note that methylglyoxal, one of the possible intermediates of the pathways presented in Figure 4, is toxic (47). Thus, depending on the route taken, side-effects of the usage of these pathways may occur if levels of methylglyoxal are too high.

Energetic requirements for gluconeogenesis from fatty acids

An important aspect related to the presented pathways are energetic requirements that constrain its capacity. An experimental investigation of the gluconeogenic role of acetone showed differences in its utilization between species. (48) found that death by starvation arises prior to the depletion of the lipid stores in obese rats. Furthermore, (49) demonstrated net synthesis of glucose from acetone in murine hepatocytes, while there was no net synthesis in perfused rat liver (50). Instead, net synthesis from acetone could be demonstrated in either cases when also other gluconeogenic substrates were given (49). This can be explained by the increased requirement for NADPH in the conversion of acetone to pyruvate. Each possible pathway requires the oxidation of at least two mole of NADPH (see Figure 4). However, there are only two pathways which reduce the net balance in NADPH to -1 by the direct conversion of methylglyoxal to pyruvate (path-

ways 7 and 14 in Table 1). The strong need for NADPH stands in contrast to the decreased capability of the cell to supply this compound in case of glycolysis not being the principal source of TCA-cycle intermediates (51). There are two major ways for the production of NADPH from $NADP^+$in the cytosol: the pentose-phosphate pathway and the malic enzyme. The activity of both pathways is reduced during starvation, because they involve routes which would further deplete gluconeogenic metabolites (51). In consequence, the amount of NADPH available represents the rate limiting factor in acetone metabolism (46, 52).

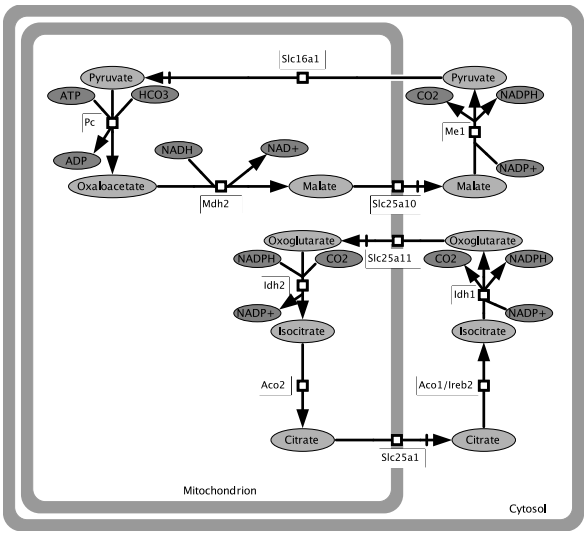


Figure 5: Two pathways for the production of cytosolic NADPH from mitochondrial sources. The upper pathway depicts the transfer of electrons from mitochondrial NADH to cytosolic NADPH. The lower pathway depicts a transfer of electrons from mitochondrial NADPH to cytosolic NADPH. Please note that the second pathway involves the transport of oxoglutarate and citrate over the mitochondrial membrane and necessitates the antiport of L-malate and phosphoenolpyruvate, respectively. A list of enzymes can be found in Supplemental Material A.

Since cytosolic NADPH producing reactions are impaired during starvation, NADPH has to be produced from mitochondrial sources. There are two known pathways of NADPH replenishment from mitochondria (51)(Figure 5). One involves the transport of mitochondrial NADPH into the cytosol and another the transfer of electrons from mitochondrial NADH to cytosolic $NADP^+$. The direct transport from the mitochondrion involves a citrate:oxoglutarate shuttle. Both metabolites are intermediates of the TCA cycle and would potentially reduce mitochondrial fatty acid oxidation when used for this shuttle system. The other pathway of NADPH replenishment uses a route via cytosolic malic enzyme which decarboxylates malate to pyruvate. Both metabolites are intermediates of gluconeogenesis from acetone. Hence, each oxidized NADPH would require an additional cycle in the conversion of pyruvate to malate. Each of these cycles involves the hydrolization of one ATP to ADP and the oxidation of one NADH. Especially pathways 7 and 14 (Table 1) in the conversion of acetol to pyruvate are of interest since they only require a net-consumption of one mole NADPH per mole pyruvate produced. Thus, only a single pyruvate-malate cycle would be necessary to balance NADPH consumption. This also explains the differences seen between species. Since there are several possible pathways in acetone metabolism, there might be species-differences in the principal pathways which are used for acetone degradation. Pathways 7 and 14 would reduce the requirements for NADPH production during gluconeogenesis.

Furthermore, we investigated the energetic requirements of the presented pathways in terms of ATP consumption under the assumption that all other metabolites, including NADPH and NADH need to be balanced. Details on the calculation can be found in Supplemental Material E. We found that the presented pathways consume in between 6 and $23\frac{1}{2}$ mole ATP for the production of one mole glucose from 4 mole acetyl-CoA. The most efficient pathway in terms of ATP consumption proceeds via the oxidation of methylglyoxal to pyruvate and additionally only needs to balance the consumption of 1 mole NADPH (pathway 7 in Table 1). However, here it is important to note that mitochondrial NADH concentration is already high during ketogenesis (53). This concentration would be further increased by import of NADH produced in the cytosol. Thus, pathway 14 might be favorable since no cytosolic NADH is produced. Using this pathway the cost for the synthesis of one mole glucose is increased to 16 ATP.

New pathways in acetone metabolism

Due to the analysis of alternative pathways for the conversion of acetol to pyruvate we identified a new intermediate in acetone metabolism not considered in (30). This metabolite, D-lactaldehyde, can be produced from either methylglyoxal or lactoyl-glutathione, an intermediate in the conversion of methylglyoxal to D-lactate. All of these reactions involve the oxidation of either one NADH or NADPH. The conversion of methylglyoxal to D-lactaldehyde is either catalyzed by an aldo-keto reductase or by a glyoxylate reductase. The catalysis by the aldo-keto reductase can proceed either via oxidation of NADPH or NADH. Both reactions are irreversible. In contrast, the glyoxylate reductase which uses NADPH as co-factor is noted as reversible in the model. Even though aldo-keto reductase is an important enzyme in the disposal of methylglyoxal, it is only weakly expressed in the liver (54). Consequently this pathway might be of greater importance in other tissues. Along with this new intermediate in acetone metabolism, we were able to identify 14 new pathways in acetone metabolism (see Supplemental Material D for an overview). This further underlines the potential of elementary flux pattern analysis in the discovery of previously unknown pathways.

Comparison to other pathway finding techniques

Other pathway detection techniques are elementary mode analysis (11), flux balance analysis (55) and graph-based methods (56–58). Elementary mode analysis can only be applied to predefined networks of small or medium size and, thus, cannot be used in networks covering the metabolism of an entire organism. The step for the determination of an initial gluconeogenic route can be considered as an approach similar to flux balance analysis. However, since flux balance analysis only returns a single pathway, it cannot systematically enumerate all possible routes. Finally, graph-based methods can be applied to genome-scale networks or databases covering pathways of several organisms, but face the problem that they only find connected routes. That is, for many cases they do not detect a route of metabolic conversions (2). However, a direct comparison is difficult, since these methods are based on reactions present in the KEGG database (59) which misses some of the essential reactions of the presented pathways.

CONCLUSION

Using elementary flux pattern analysis we were able to identify several pathways converting fatty acids into glucose in humans. Over decades, it has been textbook knowledge that such a pathway does not exist in mammals (8–10). This dogma is somewhat surprising in the light of earlier works discussing a possible role of fatty acids in gluconeogenesis (37, 38). Some of the routes leading from acetyl-CoA to glucose via acetol and methylglyoxal or propanediol detected here had been proposed earlier (37), without attracting much attention in the biochemical literature ever since, let alone in textbooks. Additionally to the eight pathways for the degradation of acetone reported in (30) we were able to detect 14 new routes, proceeding through one new intermediate. This further underlines the potential of elementary flux patterns in pathway discovery.

The crux of the presented pathways is the acetoacetate decarboxylase whose presence in mammals cannot be taken for certain until it has been characterized in terms of sequence or structure. However, there is clear evidence that acetoacetate is converted to acetone in situations of increased ketone body production, which might even be entirely due to a non-enzymatic mechanism. During starvation, for instance, up to 37% of the acetoacetate production is converted into acetone which is subsequently metabolized to pyruvate (42). Furthermore, increased survival time during starvation has been associated to the activity of the acetone metabolism in rats (46) and enzymes of acetone metabolism are activated during starvation (43). Finally, common gluconeogenic substrates cannot account for renal glucose production during prolonged starvation (40).

As outlined in the Discussion, the flux through the presented pathways is much lower than the flux through the glyoxylate shunt in plants or fungi. It is mainly limited by the amount of NADPH available. This is exacerbated by a reduced capability of the body to produce NADPH during starvation (51). However, two of the presented pathways only necessitate one mole of NADPH per mole of acetone converted into pyruvate and a cycle through the malic enzyme represents a way to replenish NADPH pools using intermediates of these pathways. An additional limit may arise from low maximal activities of several enzymes of the presented pathways.

Still, there are two important points that remain to be investigated. First, the existence of the acetoacetate decarboxylase needs to be clarified. Second, due to differential energetic requirements of the presented pathways, there is a need to elaborate which of them are preferentially taken in humans. Especially a link to ketogenic diets that are used to ease the effects of seizure represent an interesting avenue of further research since the presented pathways might play a role in the observed effects (36, 60).

These results underline the potential of elementary flux patterns for the analysis of the metabolic capabilities of reaction networks on the genome scale. They also show the usefulness of whole-cell models because such networks allow for a comprehensive analysis of whether a given substrate can be transformed into a given product at steady state.

ACKNOWLEDGMENTS

The authors would like to thank Michael Ristow and Jörn Behre for helpful discussions.

FUNDING

Financial support from the German Ministry for Research and Education (BMBF) to C.K. within the framework of the Forsys Partner initiative (grant FKZ 0315285E) and from the Fundação Calouste Gulbenkian, Fundação para a Ciência e a Tecnologia (FCT) and Siemens SA Portugal (PhD grant SFRH/BD/32961/2006) to L.F.F. is gratefully acknowledged. Furthermore we would like to acknowledge the BMBF within the program HepatoSys (grant FKZ 0313078).

REFERENCES

1. Weinman, E. O., Srisower, E. H. and Chaikoff, I. L. (1957) Conversion of fatty acids to carbohydrate; application of isotopes to this problem and role of the krebs cycle as a synthetic pathway. Physiol. Rev. **37**, 252–272
2. de Figueiredo, L. F., Schuster, S., Kaleta, C. and Fell, D. A. (2009) Can sugars be produced from fatty acids? a test case for pathway analysis tools. Bioinformatics **25**, 152–158
3. Kornberg, H. L. and Madsen, N. B. (1957) Synthesis of c4-dicarboxylic acids from acetate by a glyoxylate bypass of the tricarboxylic acid cycle. Biochim. Biophys. Acta **24**, 651–653
4. Kornberg, H. L. and Beevers, H. (1957) A mechanism of conversion of fat to carbohydrate in castor beans. Nature **180**, 35–36
5. Sandeman, R. A. and Hynes, M. J. (1989) Isolation of the faca (acetyl-coenzyme a synthetase) and acue (malate synthase) genes of aspergillus nidulans. Mol. Gen. Genet. **218**, 87–92
6. Liu, F., Thatcher, J. D., Barral, J. M. and Epstein, H. F. (1995) Bifunctional glyoxylate cycle protein of caenorhabditis elegans: a developmentally regulated protein of intestine and muscle. Dev. Biol. **169**, 399–414
7. Kondrashov, F. A., Koonin, E. V., Morgunov, I. G., Finogenova, T. V. and Kondrashova, M. N. (2006) Evolution of glyoxylate cycle enzymes in metazoa: evidence of multiple horizontal transfer events and pseudogene formation. Biol. Direct **1**, 31
8. Stryer, L., Berg, J. and Tymoczko, J. (2002) Biochemistry vol. 5th of *Trends in Biochemical Sciences* W.H. Freeman and Company 5th edn.
9. Voet, D. J., Voet, J. G. and Pratt, C. W. (2005) Fundamentals of Biochemistry: Life at the Molecular Level Wiley 2nd edn.
10. Lehninger, A., Nelson, D. L. and Cox, M. M. (2008) Lehninger Principles of Biochemistry W. H. Freeman 5th edn.
11. Schuster, S., Dandekar, T. and Fell, D. A. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends Biotechnol. **17**, 53–60
12. 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
13. Klamt, S. and Stelling, J. (2002) Combinatorial complexity of pathway analysis in metabolic networks. Mol. Biol. Rep. **29**, 233–236
14. Kaleta, C., de Figueiredo, L. F. and Schuster, S. (2009) Detecting metabolic conversions in genome-scale metabolic networks on the basis of elementary flux patterns in subnetworks in Proceedings of the Vienna Conference on Mathematical Modelling, vol. 2 of *ARGESIM Reports* (Troch, I. and Breitenecker, F., eds.) pp. 748 – 759
15. Kaleta, C., de Figueiredo, L. F. and Schuster, S. (2009) Can the whole be less than the sum of its parts? pathway analysis in genome-scale metabolic networks using elementary flux patterns Genome Res. Revised version submitted
16. Ni, T. C. and Savageau, M. A. (1996) Application of biochemical systems theory to metabolism in human red blood cells. signal propagation and accuracy of representation. J. Biol. Chem. **271**, 7927–7941
17. Covert, M. W. and Palsson, B. Ø. (2002) Transcriptional regulation in constraints-based metabolic models of escherichia coli. J. Biol. Chem. **277**, 28058–28064
18. Aledo, J. C. and del Valle, A. E. (2004) The atp paradox is the expression of an economizing fuel mechanism. J. Biol. Chem. **279**, 55372–55375
19. Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R. and Palsson, B. Ø. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc. Natl. Acad. Sci. U. S. A. **104**, 1777–1782

20. Schwender, J., Goffman, F., Ohlrogge, J. B. and Shachar-Hill, Y. (2004) Rubisco without the calvin cycle improves the carbon efficiency of developing green seeds. Nature **432**, 779–782
21. Carlson, R. and Srienc, F. (2004) Fundamental *Escherichia coli* biochemical pathways for biomass and energy production: creation of overall flux states. Biotechnol. Bioeng. **86**, 149–162
22. Vijayasankaran, N., Carlson, R. and Srienc, F. (2005) Metabolic pathway structures for recombinant protein synthesis in *Escherichia coli*. Appl. Microbiol. Biotechnol. **68**, 737–746
23. Helms, V. (2008) Principles of Computational Cell Biology: From Protein Complexes to Cellular Networks Wiley-Vch 1st edn.
24. Trinh, C. T., Wlaschin, A. and Srienc, F. (2009) Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. Appl. Microbiol. Biotechnol. **81**, 813–826
25. Klamt, S., Gagneur, J. and von Kamp, A. (2005) Algorithmic approaches for computing elementary modes in large biochemical reaction networks. IEE Proc. Syst. Biol. **152**, 249–255
26. von Kamp, A. and Schuster, S. (2006) Metatool 5.0: fast and flexible elementary modes analysis. Bioinformatics **22**, 1930–1931
27. Terzer, M. and Stelling, J. (2008) Large-scale computation of elementary flux modes with bit pattern trees. Bioinformatics **24**, 2229–2235
28. Schilling, C. H., Letscher, D. and Palsson, B. Ø. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. J. Theor. Biol. **203**, 229–248
29. Umbarger, H. E. (1978) Amino acid biosynthesis and its regulation. Annu. Rev. Biochem. **47**, 532–606
30. Kalapos, M. P. (2003) On the mammalian acetone metabolism: from chemistry to clinical implications. Biochim. Biophys. Acta **1621**, 122–139
31. Brooks, G. A., Dubouchaud, H., Brown, M., Sicurello, J. P. and Butz, C. E. (1999) Role of mitochondrial lactate dehydrogenase and lactate oxidation in the intracellular lactate shuttle. Proc. Natl. Acad. Sci. U. S. A. **96**, 1129–1134
32. Yoshida, Y., Holloway, G. P., Ljubicic, V., Hatta, H., Spriet, L. L., Hood, D. A. and Bonen, A. (2007) Negligible direct lactate oxidation in subsarcolemmal and intermyofibrillar mitochondria obtained from red and white rat skeletal muscle. J. Physiol. **582**, 1317–1335
33. Davies, R. (1943) Studies on the acetone-butanol fermentation: 4. acetoacetic acid decarboxylase of cl. acetobutylicum (by). Biochem. J. **37**, 230–238
34. van Stekelenburg, G. J. and Koorevaar, G. (1972) Evidence for the existence of mammalian acetoacetate decarboxylase: with special reference to human blood serum. Clin. Chim. Acta **39**, 191–199
35. López-Soriano, F. J., Alemany, M. and Argilés, J. M. (1985) Rat acetoacetic acid decarboxylase inhibition by acetone. Int. J. Biochem. **17**, 1271–1273
36. Beisswenger, B. G. K., Delucia, E. M., Lapoint, N., Sanford, R. J. and Beisswenger, P. J. (2005) Ketosis leads to increased methylglyoxal production on the Atkins diet. Ann. N. Y. Acad. Sci. **1043**, 201–210
37. Argilés, J. M. (1986) Has acetone a role in the conversion of fat to carbohydrates in mammals? Trends Biochem. Sci. **11**, 61–63
38. Landau, B. R. and Brunengraber, H. (1987) The role of acetone in the conversion of fat to carbohydrate Trends Biochem. Sci. **12**, 113–114
39. Hetenyi, G. and Ferrarotto, C. (1985) Gluconeogenesis from acetone in starved rats. Biochem. J. **231**, 151–155
40. Owen, O. E., Smalley, K. J., D'Alessio, D. A., Mozzoli, M. A. and Dawson, E. K. (1998) Protein, fat, and carbohydrate requirements during starvation: anaplerosis and cataplerosis] Am. J. Clin. Nutr. **68**, 12–34
41. Morris, A. A. M. (2005) Cerebral ketone body metabolism. J. Inherit. Metab. Dis. **28**, 109–121
42. Reichard, G. A., Haff, A. C., Skutches, C. L., Paul, P., Holroyde, C. P. and Owen, O. E. (1979) Plasma acetone metabolism in the fasting human. J. Clin. Invest. **63**, 619–626
43. Johansson, I., Lindros, K. O., Eriksson, H. and Ingelman-Sundberg, M. (1990) Transcriptional control of cyp2e1 in the perivenous liver region and during starvation. Biochem. Biophys. Res. Commun. **173**, 331–338
44. Rao, G. S. (1983) Diagnostic potential of breath analysis in oral disease and hygiene. Clin. Chem. **29**, 1692
45. Owen, O. E., Trapp, V. E., Skutches, C. L., Mozzoli, M. A., Hoeldtke, R. D., Boden, G. and Reichard, G. A. (1982) Acetone metabolism during diabetic ketoacidosis. Diabetes **31**, 242–248
46. Coleman, D. L. (1980) Acetone metabolism in mice: increased activity in mice heterozygous for obesity genes. Proc. Natl. Acad. Sci. U. S. A. **77**, 290–293
47. Thornalley, P. J. (1996) Pharmacology of methylglyoxal: formation, modification of proteins and nucleic acids, and enzymatic detoxification–a role in pathogenesis and antiproliferative chemotherapy. Gen. Pharmacol. **27**, 565–573
48. Cherel, Y., Robin, J. P., Heitz, A., Calgari, C. and Maho, Y. L. (1992) Relationships between lipid availability and protein utilization during prolonged fasting. J. Comp. Physiol. [B] **162**, 305–313
49. Kalapos, M. P., Mandl, J., Bánhegyi, G., Antoni, F. and Garzó, T. (1994) Net glucose production from acetone in isolated murine hepatocytes. the effect of different pretreatments of mice. Int. J. Biochem. **26**, 1069–1079
50. Gavino, V. C., Somma, J., Philbert, L., David, F., Garneau, M., Bélair, J. and Brunengraber, H. (1987) Production of acetone and conversion of acetone to acetate in the perfused rat liver. J. Biol. Chem. **262**, 6735–6740
51. Thurman, R. G. and Kauffman, F. C. (1979) Factors regulating drug metabolism in intact hepatocytes. Pharmacol. Rev. **31**, 229–251
52. Bondoc, F. Y., Bao, Z., Hu, W. Y., Gonzalez, F. J., Wang, Y., Yang, C. S. and Hong, J. Y. (1999) Acetone catabolism by cytochrome p450 2e1: studies with cyp2e1-null mice. Biochem. Pharmacol. **58**, 461–463
53. Kalapos, M. P. (1999) Possible physiological roles of acetone metabolism in humans. Med. Hypotheses **53**, 236–242
54. Jagt, D. L. V. and Hunsaker, L. A. (2003) Methylglyoxal metabolism and diabetic complications: roles of aldose reductase, glyoxalase-i, betaine aldehyde dehydrogenase and 2-oxoaldehyde dehydrogenase. Chem. Biol. Interact. **143-144**, 341–351
55. Varma, A. and Palsson, B. Ø. (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110. Appl. Environ. Microbiol. **60**, 3724–3731
56. Croes, D., Couche, F., Wodak, S. J. and van Helden, J. (2005) Metabolic pathfinding: inferring relevant pathways in biochemical networks. Nucleic Acids Res. **33**, W326–W330
57. Croes, D., Couche, F., Wodak, S. J. and van Helden, J. (2006) Inferring meaningful pathways in weighted metabolic networks. J. Mol. Biol. **356**, 222–236
58. Rahman, S. A., Advani, P., Schunk, R., Schrader, R. and Schomburg, D. (2005) Metabolic pathway analysis web service (pathway hunter tool at cubic). Bioinformatics **21**, 1189–1193
59. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in kegg. Nucleic Acids. Res. **34**, D354–D357
60. Kalapos, M. P. (2007) Possible mechanism for the effect of ketogenic diet in cases of uncontrolled seizures. the reconsideration of acetone theory. Med Hypotheses **68**, 1382–1388

Chapter 5

Discussion

This thesis presents a collection of six papers dedicated to the study of metabolic networks using stoichiometric analysis techniques. Instead of using the different methods independently from each other the aim was to integrate them in order to allow for a more comprehensive analysis of the networks under study.

The work presented in Chapter 2 is dedicated to the analysis of the relationship between elementary modes [Schuster et al., 1999] and chemical organizations [Dittrich and Speroni di Fenizio, 2007]. A connection between both concepts is not straightforward since the former deals with pathways and the latter with sets of compounds that are likely to coexist over long periods of (simulation) time. Furthermore, the self-maintenance condition required for a set of compounds to be an organization allows for states in which some molecules accumulate (growth states), while elementary mode analysis only considers steady states. Yet, the solution space of the steady-state condition is a subspace of the space of all flux vectors that are a feasible solution of the self-maintenance condition. Additionally, it is possible to adapt a reaction network such that the elementary modes of the adapted network correspond exactly to the extreme rays of the solution space of the self-maintenance condition of the original network (cf. Chapter 2). Starting from this relationship, elementary mode analysis and chemical organization theory can benefit from each other in two important points.

First, determining the organizations to which an elementary mode belongs allows one to deduce more information about the steady states in which an elementary mode can appear, since the potential qualitative steady states of a reaction network correspond to its chemical organizations [Dittrich and Speroni di Fenizio, 2007]. On the one hand, an elementary mode might be enclosed in at least one chemical organization. Since, the number of elementary modes is growing exponentially with network size [Klamt and Stelling, 2002], the task of the analysis of elementary modes can be simplified by assigning them to the smallest (balanced) reactive organization in which they are enclosed. Empirical results show that the number of reactive organizations is usually much smaller than the number of elementary modes. Thus, since chemical organizations often coincide with specific biological functions as demonstrated in this work, it is possible to get a first hint at the function of the elementary modes contained in them. Furthermore, it is possible to group elementary modes according to the organizations in which they appear. On the other hand, if an elementary mode is not enclosed in any chemical organization, it can be deduced that it is not part of any steady state of the network. Hence, chemical organizations can help to identify feasible elementary modes, that is, elementary modes that can be present at steady state.

Second, the assignment of elementary modes to chemical organizations does not necessitate an *a priori* computation of the chemical organizations of a reaction network. Instead it is possible to compute these assignments along with the chemical organizations of the network, leading to an elementary mode based algorithm for the computation of chemical organizations. The basis of this algorithm is the observation that each steady-state flux of a reaction network is a positive linear combination of elementary modes. By adapting the reaction network for the computation of elementary modes it is possible to obtain a set of elementary modes that correspond to the extreme rays of the solution space of the self-maintenance condition. By searching for combinations of these extreme rays additionally fulfilling the closure condition it is possible to compute chemical organizations (cf.

Chapter 2).

The process of the computation of chemical organizations is split into three steps. In the first step, the set of extreme rays is used as input to an algorithm that computes all elementary organizations. These organizations are the building blocks of the set of organizations of a network, that is, they cannot be written as set union of other organizations. In the second step, the set of elementary organizations is used as input for a similar algorithm in order to find all reactive organizations. These are chemical organizations in which each compound participates in at least one reaction. In general, organizations can contain compounds that do not participate in any reaction. In consequence, several chemical organizations might use the same set of reactions but differ in the set of compounds which are contained in them. In consequence, the third step of the algorithm consists in checking for each reactive organization whether some compounds can be added without enabling additional reactions.

If the computation of elementary modes is not possible for a particular network, a constructive algorithm that employs a different search strategy can be employed. This strategy consists in, starting from the smallest organization, searching above previously found semi-organizations for new semi-organizations. These are a different types of organizations in which the compound set needs to fulfill the semi-selfmaintenance condition additionally to the closure condition. This condition requires that every compound of a compound set that is consumed is also produced. For the search strategy, compounds are added to existing semi-organizations and it is afterward tested which compounds need to be added additionally in order to fulfill the closure and the semi-selfmaintenance condition. Finally, if all semi-organizations have been found it can be tested for each if it fulfills the self-maintenance condition. Thus, all organizations can be found since the set of chemical organizations is a subset of the set of semi-organizations.

In general, both algorithms for the computation of chemical organizations are competitive. As shown in the second part of Chapter 2 and the Supplemental

Material of Kaleta et al. [2009c] (pages 129 ff.) in some cases the elementary mode based algorithm for the computation of chemical organizations terminates faster, while in some cases the constructive algorithm is the only with which a result can be obtained since the number of elementary modes is too large.

The concept of closure is the most important difference between chemical organization theory and other stoichiometry based network analysis techniques. Especially in comparison to elementary mode analysis it allows to derive additional information from the stoichiometric structure of a reaction network. First, it is possible to map a set of compounds to a set of reactions that might be performed from these compounds. Using elementary modes, in contrast, one implicitly assumes that enzymes can be arbitrarily turned *on* and *off*. As discussed above, this leads to elementary modes that are not part of any steady state. Second, the concept of closure allows to bypass the problem seen in elementary mode analysis that a pathway is only detected if mass-flow is involved [Gianchandani et al., 2006, Behre and Schuster, 2009]. This is exemplified by a signaling cascade that receives two inputs (Figure 5.1). The transfer of information along a signaling cascade usually does not involve a concurrent transfer of atoms. If the signal is transferred by subsequent cycles of phosphorylations (Figure 5.1), elementary mode analysis reveals only the cycles of phosphorylations on the different levels of the cascade (Figure 5.1C). In contrast, the elementary organizations exactly depict the different levels in the signaling cascade (Figure 5.1D). Additionally it is taken into account that the phosphorylation of a protein on one level of the cascade requires a phosphorylated protein on a previous level. Thus, in contrast to elementary mode analysis, also compounds playing a catalyzing role are appropriately taken into account.

Chapter 3 focuses on the analysis of phenotypes of reaction networks, that is, the sets of chemical organizations they give rise to. While the central point of the first part was to elaborate on the similarities and differences between chemical organizations and flux-based methods like elementary modes, the second part fo-

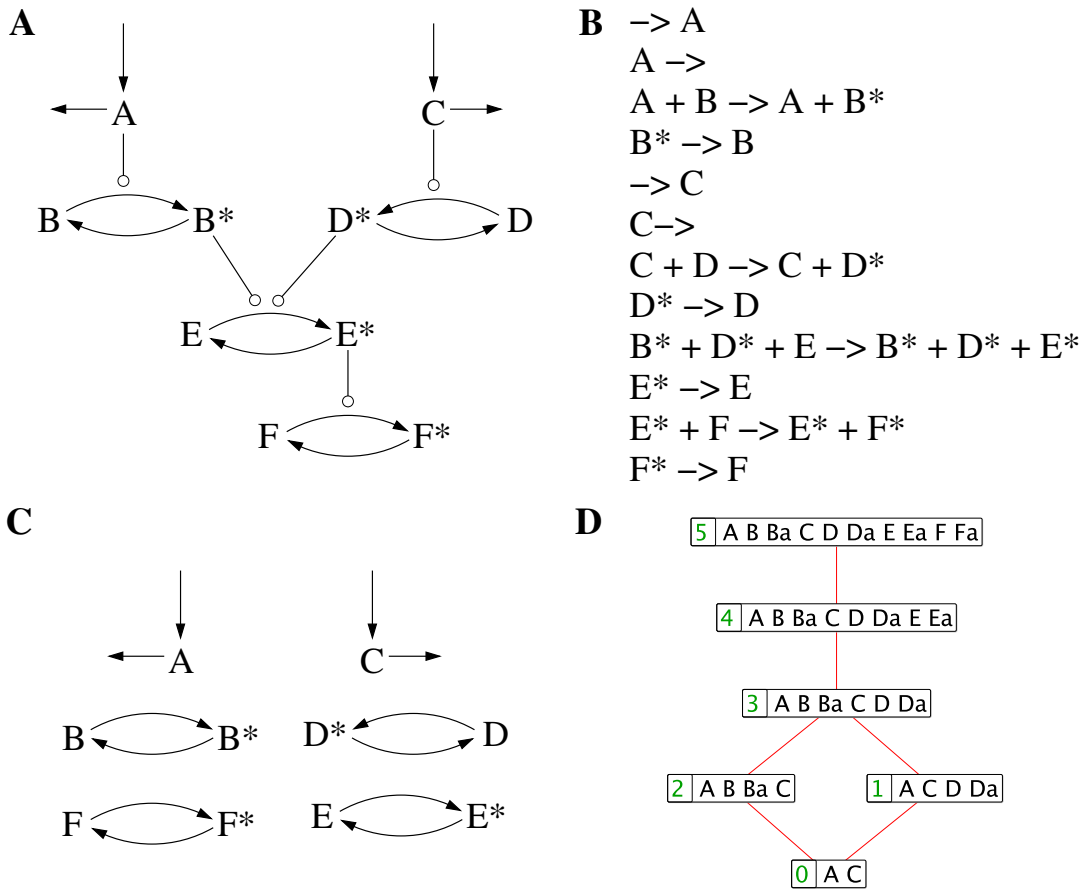


Figure 5.1: Elementary modes and chemical organizations in a signaling cascade. **A** Signaling cascade. Lines ending in circles denote activation of a reaction. Names ending with "*" indicate the active (phosphorylated) form of a protein. **B** Transformation of the network into reaction rules. **C** Six elementary modes of the system. Apparently the elementary modes of the system do not capture the information flow through the network. **D** Hasse diagram of reactive organizations of the network. Phosphorylated forms are denoted by the suffix "a". In contrast to the set of elementary modes, the topology of the network is captured. For instance, $D*$ can only be present at steady state if C is present. If this was not the case, the (spontaneous) dephosphorylation of $D*$ would finally reduce any concentration of this form of the protein to zero. In contrast there is an elementary mode containing the phosphorylation and dephosphorylation of D independently of the presence of C.

cused on the comparison of the results that can be obtained from both methods. Thus, we analyzed the chemical organizations of a genome-scale metabolic network of *Escherichia coli*, the chemical organizations of the models of the BioModels Database [Novère et al., 2006], and an integrated regulatory and metabolic network of the central metabolism of *E. coli*.

The genome-scale metabolic model of *E. coli* by Reed et al. [2003] has already

been extensively studied (see Feist and Palsson [2008] for a review). For metabolic networks of such size, the set of chemical organizations cannot be enumerated directly, but a heuristic has to be used. We performed our analysis in four steps. First, we used the heuristic algorithm for the computation of organizations and aborted the process after we had found more than 88000 organizations. From these we extracted 70 elementary organizations. We found that many of the elementary organizations are able to produce biomass, a pseudo-compound whose presence indicates that the cell is viable. This is concurrent with results previously obtained for this network [Reed et al., 2003]. Second, we added an outflow for each compound. This modification simulates growth, corresponding to the requirement that the cell has to be able to reproduce all of its constituents to survive. However, in this setting we only found a single organization corresponding to the compounds that can be directly produced from the growth medium. This organization is not able to produce biomass. In contrast, flux balance analysis indicates the viability of the organism associated to the metabolic network. However, our analysis demonstrated that all except few species would finally vanish in a simulation of the model and, hence, indicated that the organism is indeed not viable. Analyzing the reasons for this difference we found that the metabolic model contains non-metabolic compounds which are proteins or tRNAs whose replenishment is not accounted for. These compounds do not cause problems when flux balance analysis is used, since they act as co-factors and are not consumed. In the third step, after adding an inflow for the non-metabolic compounds, we found two organizations (Figure 5.2). The first contains compounds that can be produced from the growth medium. The second contains a reaction producing biomass. Thus, as cannot be deduced using flux balance analysis, reactions producing the contained tRNAs and proteins need to be added to the model in order to explain growth. Furthermore, we determined the compounds besides enzymes that need to be present in the cell such that all of the compounds in the second organization can be produced from the growth medium. We identified ADP, NAD,

coenzyme A, and menaquinone as one set of such compounds. These four compounds can hence be considered as a kind of autocatalytic metabolites since their presence alone allows for the regeneration of all constituents of the cell, including themselves, from the growth medium (Figure 5.2). These results might give a hint on compounds that were of central importance at the very beginning of the development of metabolism. Further support for this assumption was presented by Kun et al. [2008] who found that the same compounds are autocatalytic in genome-scale metabolic networks of other organisms. In a fourth step we confirmed the results we obtained in the third step since the removal of the outflow for all non-metabolic compounds gave the same results even though the Hasse diagram of organizations was more complex.

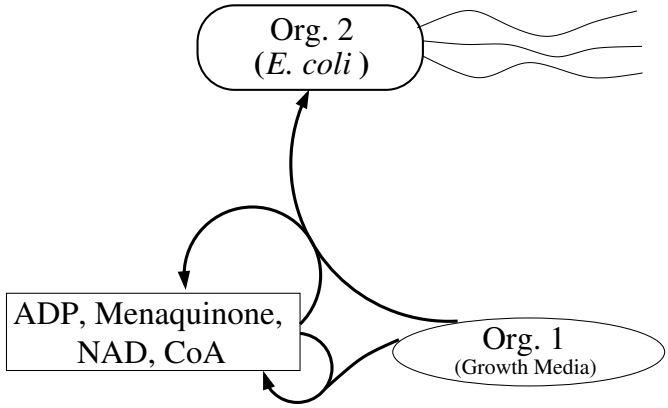


Figure 5.2: Schematic representation of the role of ADP, NAD, menaquinone and coenzyme A (CoA) in the metabolism of *E. coli*. The system is provided with an input in form of a growth medium. ADP, NAD, menaquinone and CoA can recreate themselves from the growth media (lower reaction). Additionally, they are able to produce all the constituents of the cell (upper reaction). Note that organization 2 in principle corresponds to *E. coli* since all the biomass metabolites required for a survival of the cell are contained within this organization. Furthermore, organization 2 additionally contains the growth media as well as ADP, NAD, menaquinone and CoA.

Furthermore, we extended our analysis by comparing the predictions of the lethality of gene knockout experiments between flux balance analysis and chemical organization theory (see Supplemental Material of Kaleta et al. [2009c], pages 115 ff.). Here we found that chemical organization theory made correct predictions in ten additional cases. The differences are due to compounds that act as catalysts in the production of biomass but are not included in the reaction producing biomass

themselves.

In the first part of Chapter 3 we more closely examined the interplay between the environment, the regulatory network and the metabolic network of *E. coli*. For this analysis we translated a set of Boolean regulatory rules taken from Covert and Palsson [2002] into a set of reactions and combined it with a metabolic network. The regulatory network also contains rules involving the absence of compounds, for instance, in the case of an inhibition. Thus, we introduced pseudo-compounds that correspond to the absence of a compound. To avoid organizations that contain both or none of the two representations of a compound, we only allowed for *consistent organizations* in which either of both is present. In a first setting we analyzed a reduced model of the central metabolism of *E. coli* with a focus on 16 different environmental conditions involving supply of different carbon sources and oxygen. For each input-scenario we found a distinct chemical organization corresponding to the phenotype of the cell under this condition. Thus, we were also able to see the diauxic shift, the effect that *E. coli* only takes up a preferential carbon source if several carbon sources are provided. For instance, if lactose and glucose are available only glucose is taken up. After depletion of glucose the cells start to import lactose. In a second setting we analyzed the outcome of gene knockout experiments and compared the results to those that can be obtained using regulatory flux balance analysis (rFBA, Covert et al. [2001]). In a first analysis we found that the predictions by rFBA were more accurate than those by organization theory. As reason for this differences we identified several assumptions implicitly build into rFBA. For instance, rFBA differentiates between compounds that are excreted and those that are provided *ab initio* with the growth media, even if they are identical. This leads to the inactivation of essential pathways in the analysis using chemical organization theory due to the excretion of compounds exerting regulatory effects, while rFBA does not take into account such an excretion. Also differentiating between compounds in the growth media and those that are excreted as well as taking into another additional assumption, the

predictions by chemical organization theory match those of rFBA.

Owing to these results our next work focused on the large-scale analysis of a repository of curated models of reaction networks, the BioModels Database [Novère et al., 2006]. Our aim was to identify models with inconsistencies. This analysis is based on the assumption that, if provided with an input of compounds whose concentration is set to a non-zero value at a specific time-point of the simulation, all compounds should appear in an organization. Compounds that do not appear in any organization cannot be present at positive concentration at all. In this analysis we also took into account that almost all of the models of the BioModels Database contain detailed information on reaction kinetics. In some cases reaction kinetics are not consistent with the stoichiometries of the reactions in the sense of the Feinberg condition [Feinberg and Horn, 1974](*cf.* Chapter 3). That is, even though all of the educts of a reaction are present the flux through this reaction might still be constrained to zero if another, for instance, catalyzing compound is absent. In the Systems Biology Markup Language [SBML, Hucka et al., 2003], which is used as means of representation of reaction networks in the BioModels Database, such compounds are called modifiers. We presented an algorithm that compares the kinetic laws of the reactions of a model to their stoichiometries. If this method identifies sets of modifiers whose presence is required for a non-zero flux through a reaction, its stoichiometry is adapted accordingly. For instance, a modifier that corresponds to a catalyst is added on the educt and product side of the reaction. We found that this process changes the organizational structure of 41 of the 185 models (22%) of the BioModels Database. We analyzed these changes in more detail in a model of the Wnt/ERK-signaling pathway and demonstrated that they more accurately reflect the dependencies in the corresponding signaling cascades. Furthermore we identified five models in which some compounds do not appear in any organization. In each model we demonstrated that the missing compounds are due to modeling errors. Two models contain reactions whose kinetics produce negative fluxes despite the reactions being defined as ir-

reversible, two models produce negative concentrations for some compounds and in the last model a reaction is missing. Furthermore we compared our results to those that can be obtained using flux-based methods like the aforementioned elementary mode analysis, extreme pathway analysis and flux balance analysis. We adapted these methods such that they also could take into account situations in which some compounds accumulate. Such, we found that the same five erroneous models are identified by flux-based methods since they will always find a reaction missing from any feasible flux if the entire compound set is not an organization (*cf.* Chapter 3). However, the compounds and reactions that can be present at non-zero concentration during a simulation are more accurately identified by chemical organization theory in two of these models. Hence, the search for the reasons of inaccuracies is simplified.

Meanwhile, an alternative approach for checking reaction networks that builds upon a stoichiometry check has been proposed [Gevorgyan et al., 2008]. By this approach, reactions that do not fulfill the conservation of mass can be detected. Such reaction, for instance $A \rightarrow A+B$, either produce or consume mass. However, while any reaction should obey the conservation of mass, this is often not the case for many networks due to simplifications that are used in modelling. For instance, protons are often not balanced and inflow reactions of the form $\emptyset \rightarrow A$ are commonly used. Both simplifications give rise to cases in which the conservation of mass is not fulfilled. Likewise, when applying the approach by [Gevorgyan et al., 2008] to the same version of the BioModels Database as analyzed in Chapter 3, 124 out of the 185 models are detected to be inconsistent. This is a large number of models that need to be checked and preliminary results indicate that in most cases the network is not incorrect but only simplifications like those discussed above are used.

While chemical organization theory allows a view on the potential steady states and growth states of a reaction network, elementary mode analysis allows to determine precisely which pathways can be used at steady state. However, while the

analysis using chemical organization theory can, to a limited extent, be applied to genome-scale metabolic networks, this is not the case for elementary modes. Thus, the subsequent focus of my work was to develop a theoretical framework that allows to circumvent these limitations of the concept of elementary modes. The problem has two important facets. First, the number of elementary modes is usually growing exponentially with network size [Klamt and Stelling, 2002]. Thus, it is impossible to efficiently enumerate all of them in a genome-scale metabolic network with current methods, even though methods have been proposed to enumerate a subset of elementary modes in large metabolic networks [Beasley and Planes, 2007, Acuña et al., 2009, de Figueiredo et al., 2009]. Second, even if it were possible to enumerate all elementary modes, their number makes it impossible to store or analyze them. For instance, the number of extreme pathways [Schilling et al., 2000], which are a subset of the elementary modes, of a genome-scale model of human metabolism [Duarte et al., 2007] is estimated at 10^{29} [Yeung et al., 2007]. Due to these difficulties, elementary mode analysis usually focuses on a specific part of the known metabolism of an organism and models the remaining system using abstractions like external metabolites and exchange reactions. However, as shown, these abstractions bear the problem that they drastically influence the elementary modes that are found.

The concept of elementary flux patterns introduced in the first part of Chapter 4 circumvents these problems. Elementary flux patterns relate fluxes or pathways within a subsystem to fluxes through the entire network. This concept connects methods aimed at the analysis of single fluxes like flux balance analysis and methods aimed at the exhaustive enumeration of all pathways like elementary mode analysis. Similar to elementary modes, elementary flux patterns are computed in a subsystem of a large-scale metabolic network. However, in contrast to elementary modes, the remaining network is not modeled using external compounds or exchange reactions. Instead it is explicitly taken into account since each elementary flux pattern is part of a steady-state flux through the entire network that

uses reactions of the subsystem. Thus, in contrast to elementary mode analysis, all routes through the subsystem which are indeed part of a pathway through the entire network are found. Furthermore, the concept of elementary flux patterns does not constrain the topology of the reactions of the subsystem. Thus, it is possible to analyze the dependencies between two subsystems which do not interface to each other through common substrates (Figure 5.3). In elementary mode analysis this would require an addition of all reactions connecting both subsystems to a combined subsystem. As described, the analysis of disconnected subsystems can be helpful, for instance, when trying to detect all minimal compound sets required for the production of complex metabolites like antibiotics.

In the first part of Chapter 4 we analyzed a model that has been studied in Schuster et al. [1999] by integrating it into a genome-scale metabolic model of *E. coli*. We found that of the 16 elementary modes reported, six are not part of any steady-state flux through the entire network. Furthermore, we found that there are indeed 83 different pathways. As one reason for this difference we found that instead of a single intermediate of the subsystem that can be produced from glucose as considered in the original work there are three additional intermediates (Figure 5.4): pyruvate, acetyl-CoA and glyoxylate. Along with the intermediates that can be produced from glucose we also found a new pathway for the production of glyoxylate. Furthermore, we detected a GABA-shunt like pathway that allows to bypass some of the reactions in the TCA cycle and has been suggested as an important metabolic route in plants recently [Fait et al., 2008].

Finally, in the second part of Chapter 4, we demonstrated how elementary flux patterns can be used to detect all possible pathways converting one compound into another. This process builds upon an iterative expansion of the subsystem. It starts from an initial pathway by separating it into essential steps and steps for which alternative routes exist. The alternative routes are identified by iteratively adding reactions to the subsystem that connect sequences of essential reactions. We applied this method to an example famous in biochemistry – the conversion of

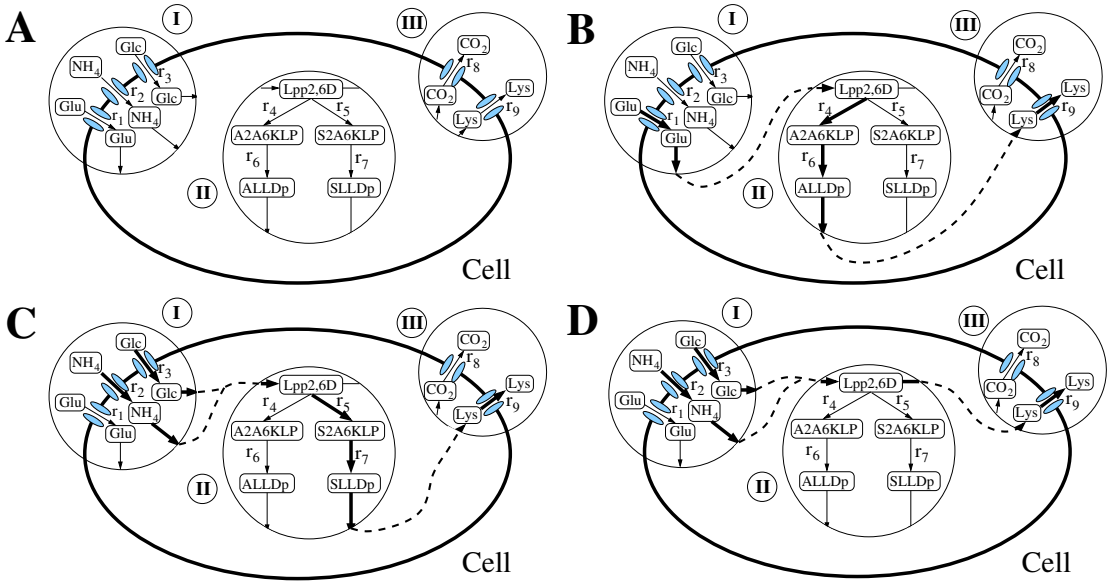


Figure 5.3: Example for the application of elementary flux patterns in a disconnected subsystem (drawn schematically). Reactions for which an index is given belong to the subsystem. Bold arrows indicate reactions belonging to an elementary flux pattern. Dotted lines schematically indicate the pathway through the remaining system used by each elementary flux pattern. **A** The subsystem is made up by three parts: (I) Uptake reactions; (II) A branching point in lysine biosynthesis; (III) Excretion reactions. **B** Computing the elementary flux patterns of this subsystem we find that glutamate alone is sufficient for the production of lysine as it can serve as carbon and nitrogen-source at the same time. **C** In contrast, if glutamate is not taken up, glucose and either glutamate or ammonia are required. Furthermore, the possible routes at the branching point of lysine biosynthesis in (II) can be seen in the elementary flux patterns depicted in **B**-**D**. Here either the route via ALLDp (B) or SLLDp (C) can be used. **D** Some bacteria possess an additional route from Lpp2,6D [Schrumpf et al., 1991] which has not been added to the subsystem. Thus, we even find a flux pattern that contains the production of lysine without using any reaction from (II). Abbreviations: A2A6KLP, acetyl-2-amino-6-keto-L-pimelate; ALLDp, acetyl-L,L-diaminopimelate; Glc, glucose; Lpp2,6D, L-piperideine-2,6-dicarboxylate; Lys, L-lysine; S2AKLP, succinyl-2-amino-6-keto-L-pimelate; SLLDp, succinyl-L,L-diaminopimelate.

even-chain fatty acids to glucose in mammals. While it has long been known that glucose can be converted into fatty acids, the question whether the reverse is also possible has long remained unsolved. This topic has been extensively discussed in the first half of the last century since a connection to diabetes was assumed. In 1957 Kornberg and Madsen found the glyoxylate shunt, a pathway that allows such a conversion in various species. However, the corresponding enzymes were not detected in mammals. Even though there exists a connected route from acetyl-CoA to glucose through the TCA cycle, Weinman et al. [1957] demonstrated mathematically that no net production of glucose from acetyl-CoA is possible

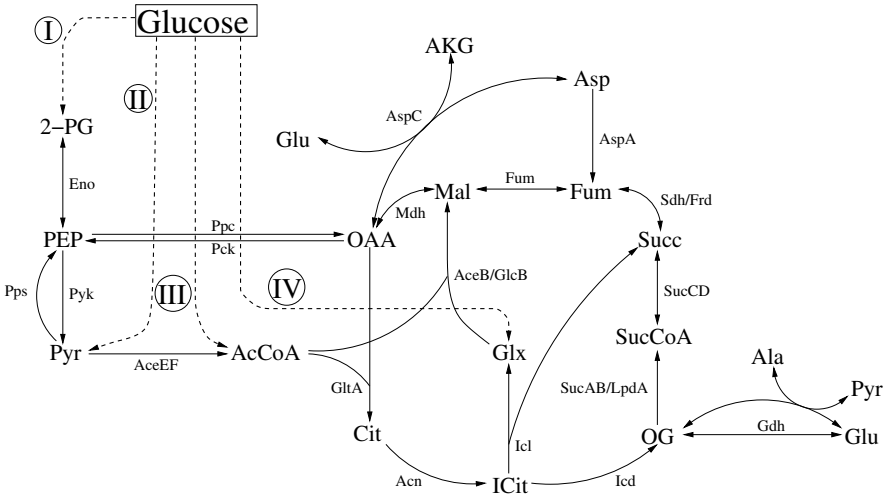


Figure 5.4: Entry points into a subsystem of central metabolism: **(I)** 2-phosphoglycerate from Glycolysis; **(II)** pyruvate from the Entner-Doudoroff-pathway; **(III)** acetyl-CoA from the Entner-Doudoroff-pathway and the pyruvate formate lyase; **(IV)** glyoxylate from an alternative glyoxylate producing pathway running through nucleotide synthesis. In the work of Schuster et al. [1999] it was assumed that (I) represents the only inflow into this system. Dashed arrows correspond to schematic reactions. A list of abbreviations can be found in the Supplemental Material of Kaleta et al. [2009b], pages 140 ff.

along this route (Figure 5.5). Owing to these results it was long assumed that there is no gluconeogenic pathway from fatty acids in humans.

We re-investigated this problem using a genome-scale metabolic model of human metabolism [Duarte et al., 2007] and indeed found a pathway for gluconeogenesis from acetyl-CoA. Using the concept of elementary flux patterns we were able to identify essential reactions and alternative routes in acetone metabolism that can be taken. Analyzing the essential reactions we found that one step of the pathway might not be catalyzed by an enzyme: the conversion of acetoacetate into acetone. Nevertheless, there is incontestable evidence that during situations of increased ketone body production significant amounts of acetoacetate are converted into acetone of which only small amounts are exhaled or excreted [Reichard et al., 1979]. In a literature search we also found previous reports on possible routes for gluconeogenesis from fatty acids via acetone [Weinman et al., 1957, Argilés, 1986, Landau and Brunengraber, 1987] which seem to have been largely ignored by the scientific community since prominent biochemistry text books still state the non-existence of a gluconeogenic route from fatty acids in mammals [Stryer et al.,

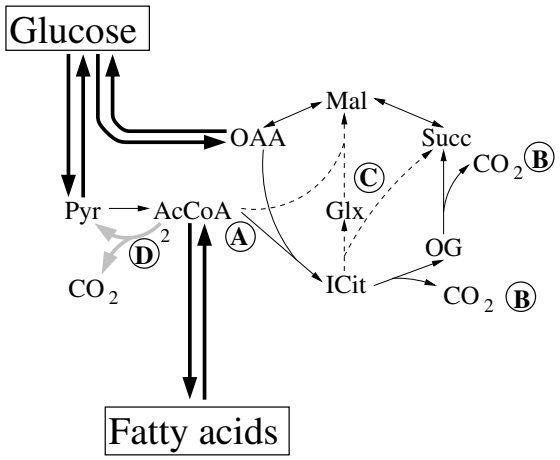


Figure 5.5: Classical scheme of the interconversion between glucose and fatty acids in humans. **A** The product of β-oxidation of even-chain fatty acids, acetyl-CoA, can enter the TCA cycle by reaction with oxaloacetate to citrate. However, in order to replenish the oxaloacetate consumed in this reaction, the TCA cycle has to be used resulting in the release of two carbon atoms in form of carbon dioxide (**B**). Hence, while two carbon atoms enter the TCA cycle in form of acetyl-CoA two others are lost and a net production of oxaloacetate which would be required for gluconeogenesis is not possible. **C** In some bacteria, plants, fungi and nematodes the carbon releasing steps of the TCA cycle can be bypassed using the glyoxylate shunt (dashed arrows). In consequence, one mole of oxaloacetate can be produced from two mole of acetyl-CoA, allowing for gluconeogenesis from fatty acids. However, this pathway is not present in humans. **D** Schematic representation of the newly detected gluconeogenic pathways via ketogenesis and acetone metabolism in human (thick gray arrow). In contrast to the TCA cycle only one carbon is lost during conversion of two acetyl-CoA to pyruvate. A list of abbreviations can be found in the Supplemental Material of Kaleta et al. [2009b], pages 140 ff.

2002, Voet et al., 2005, Lehninger et al., 2008]. We also determined the energetic requirements of these pathways and found that the availability of NADPH, which is only produced in small amounts during ketogenesis, is a potential rate limiting factor. However, a set of reactions using intermediates of the presented pathway transferring electrons from mitochondrial NADH to cytosolic NAPD+ might compensate for this decreased NADPH production capability. We hope that our "re-discovery" of a gluconeogenic route from fatty acids changes the view of the metabolic capabilities of humans and think that it is of special importance for a further understanding of effects seen during the increased production of ketone bodies such as in starvation, diabetes and when using ketogenic diets to ease the effects of seizure.

Chapter 6

Conclusion and Prospects

This work demonstrated how the integration of different techniques for the stoichiometric analysis of reaction networks can help to further our understanding of such systems. It opens up many avenues for the further study of reaction networks. One central point is the development of a deterministic approach for the computation of chemical organizations in genome-scale metabolic networks. Using an approach similar to the algorithm for the computation of elementary flux patterns it might be possible to enumerate all elementary organizations in genome-scale reaction networks. From the elementary organizations, all chemical organizations can be obtained. Such an approach can simplify the analysis of large-scale metabolic networks by identifying parts of the network that cannot appear at steady state like it has already been done for a genome-scale network of *E. coli* in Chapter 3. As described, in comparison to other stoichiometric analysis techniques like flux balance analysis, predictions by chemical organization theory are more accurate. Along these lines it is also of interest to investigate whether the structure of the Hasse diagram of chemical organizations in *E. coli* can also be observed in other organisms. If this is the case, our suggestion that this structure is related to the evolution of metabolism would be substantiated.

Furthermore, the concept of elementary flux patterns has been presented as a method to circumvent problems seen in the analysis of elementary modes in the

context of genome-scale metabolic networks. As a first application we used this concept for the detection of previously unknown metabolic pathways. The knowledge and further investigation of such pathways can help to extend our understanding not only of the known but also of the potential metabolic capabilities of organisms. Such a knowledge is of importance when analyzing effects seen under nonstandard conditions [Fischer and Sauer, 2003, Wittmann et al., 2007]. Other promising applications of the concept of elementary flux patterns include to port methods building on elementary mode analysis to elementary flux patterns. Areas of application include the computation of minimal cut sets [Klamt and Gilles, 2004, Klamt, 2006], the analysis of the robustness of metabolic networks [Wilhelm et al., 2004, Behre et al., 2008] and the analysis of integrated genome-scale metabolic networks. Especially the last application is of certain interest. Following reconstruction of the first genome-scale metabolic model of human [Duarte et al., 2007], recent efforts have concentrated on the reconstruction of tissue-specific metabolic models [Shlomi et al., 2008]. Integrating several such models with appropriately defined boundaries between each other, elementary flux pattern analysis allows to study the interaction between different tissues on the level of metabolism. Furthermore, integrated genome-scale metabolic networks allow the study of host-pathogen interactions. Such, it might be possible to identify all metabolic pathways connecting host and pathogen and devise strategies that impair pathways that are essential for the parasite but can be bypassed in the host.

Another important aspect is the alignment of the concepts presented and extended in this work with large-scale biological data similar to the analysis of knockout experiments in Chapter 3. One example is the analysis of the correlation between gene and protein expression data with chemical organizations and elementary flux patterns as it has already been done for elementary modes [Schwartz et al., 2007].

Bibliography

V. Acuña, F. Chierichetti, V. Lacroix, A. Marchetti-Spaccamela, M.-F. Sagot, and L. Stougie. Modes and cuts in metabolic networks: Complexity and algorithms. *Biosystems*, 95(1):51–60, Jan 2009.

J. M. Argilés. Has acetone a role in the conversion of fat to carbohydrates in mammals? *Trends Biochem. Sci.*, 11:61–63, 1986.

J. E. Beasley and F. J. Planes. Recovering metabolic pathways via optimization. *Bioinformatics*, 23(1):92–98, Jan 2007.

J. Behre and S. Schuster. Modeling signal transduction in enzyme cascades with the concept of elementary flux modes. *J. Comput. Biol.*, 16(6):829–844, Jun 2009.

J. Behre, T. Wilhelm, A. von Kamp, E. Ruppin, and S. Schuster. Structural robustness of metabolic networks with respect to multiple knockouts. *J. Theor. Biol.*, 252(3): 433–441, Jun 2008.

A. P. Burgard and C. D. Maranas. Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng.*, 74(5):364–375, Sep 2001.

R. Carlson, D. Fell, and F. Srienc. Metabolic pathway analysis of a recombinant yeast for rational strain development. *Biotechnol. Bioeng.*, 79(2):121–134, Jul 2002.

F. Centler, P. Speroni di Fenizio, N. Matsumaru, and P. Dittrich. Chemical organizations in the central sugar metabolism of *Escherichia coli*. In A. Deutsch, L. Brusch, H. Byrne, G. de Vries, and H.-P. Herzel, editors, *Mathematical Modeling of Biological Systems, Volume I*, pages 109–123, Boston, 2007. Birkhäuser.

F. Centler, C. Kaleta, P. S. di Fenizio, and P. Dittrich. Computing chemical organizations in biological networks. *Bioinformatics*, 24(14):1611–1618, Jul 2008.

A. Cornish-Bowden and M. L. Cárdenas. Systems biology may work when we learn to understand the parts in terms of the whole. *Biochem. Soc. Trans.*, 33(Pt 3):516–519, Jun 2005.

M. W. Covert and B. Ø. Palsson. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.*, 277(31):28058–28064, Aug 2002.

M. W. Covert, C. H. Schilling, and B. Ø. Palsson. Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.*, 213(1):73–88, Nov 2001.

M. W. Covert, N. Xiao, T. J. Chen, and J. R. Karr. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics*, 24(18):2044–2050, Sep 2008.

L. F. de Figueiredo, A. Podhorski, A. Rubio, J. E. Beasley, S. Schuster, and F. J. Planes. Calculating the k-shortest elementary flux modes in metabolic networks. In I. Troch and F. Breitenecker, editors, *Proceedings MATHMOD 09 Vienna - Full Papers CD Volume*, pages 736 – 747, 2009.

P. Dittrich and P. Speroni di Fenizio. Chemical organization theory. *Bull. Math. Biol.*, 69(4):1199–1231, 2007.

N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.*, 104(6):1777–1782, Feb 2007.

J. S. Edwards and B. Ø. Palsson. Metabolic flux balance analysis and the in silico analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics*, 1:1, 2000.

A. Fait, H. Fromm, D. Walter, G. Galili, and A. R. Fernie. Highway or byway: The metabolic role of the GABA shunt in plants. *Trends Plant Sci.*, 13(1):14–19, Jan 2008.

M. Feinberg and F. J. M. Horn. Dynamics of open chemical systems and the algebraic structure of the underlying reaction network. *Chem. Eng. Sci.*, 29(3):775–787, 1974.

A. M. Feist and B. Ø. Palsson. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat. Biotechnol.*, 26(6):659–667, Jun 2008.

A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, 3:121, 2007. doi: 10.1038/msb4100155.

D. A. Fell and J. R. Small. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem J.*, 238(3):781–786, Sep 1986.

E. Fischer and U. Sauer. Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *Eur. J. Biochem.*, 270(5):880–891, Mar 2003.

W. Fontana and L. W. Buss. 'The arrival of the fittest': Toward a theory of biological organization. *Bull. Math. Biol.*, 56:1–64, 1994.

A. Gevorgyan, M. G. Poolman, and D. A. Fell. Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics*, 24(19):2245–2251, Oct 2008.

E. P. Gianchandani, J. A. Papin, N. D. Price, A. R. Joyce, and B. Ø. Palsson. Matrix formalism to describe functional states of transcriptional regulatory systems. *PLoS Comput. Biol.*, 2(8):e101, Aug 2006. doi: 10.1371/journal.pcbi.0020101.

T. Handorf, O. Ebenhöh, and R. Heinrich. Expanding metabolic networks: Scopes of compounds, robustness, and evolution. *J. Mol. Evol.*, 61(4):498–512, Oct 2005.

M. J. Herrgård, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blüthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. L. Novère, P. Li, W. Liebermeister, M. L. Mo, A. P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasić, D. Weichart, R. Brent, D. S. Broomhead, H. V. Westerhoff, B. Kirdar, M. Penttilä, E. Klipp, B. Ø. Palsson, U. Sauer, S. G. Oliver, P. Mendes,

J. Nielsen, and D. B. Kell. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.*, 26(10): 1155–1160, Oct 2008.

M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, 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, J. Wang, and S. B. M. L. Forum. The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, Mar 2003.

A. R. Joyce and B. Ø. Palsson. Predicting gene essentiality using genome-scale in silico models. *Methods Mol. Biol.*, 416:433–457, 2008.

C. Kaleta, F. Centler, and P. Dittrich. Analyzing molecular reaction networks: From pathways to chemical organizations. *Mol. Biotechnol.*, 34(2):117–123, Oct 2006.

C. Kaleta, F. Centler, P. S. di Fenizio, and P. Dittrich. Phenotype prediction in regulated metabolic networks. *BMC Syst. Biol.*, 2:37, 2008. doi: 10.1186/1752-0509-2-37.

C. Kaleta, L. F. de Figueiredo, R. Guthke, and S. Schuster. Systematic determination of gluconeogenic pathways from fatty acids in humans. *Biochem. J.*, 2009a. In preparation.

C. Kaleta, L. F. de Figueiredo, and S. Schuster. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res.*, 2009b. In print.

C. Kaleta, S. Richter, and P. Dittrich. Using chemical organization theory for model checking. *Bioinformatics*, 25(15):1915–1922, Aug 2009c.

H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, Nov 2002.

K. R. Kjeldsen and J. Nielsen. In silico genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnol. Bioeng.*, 102(2): 583–597, Feb 2009.

S. Klamt. Generalized concept of minimal cut sets in biochemical networks. *Biosystems*, 83(2-3):233–247, 2006.

S. Klamt and E. D. Gilles. Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2):226–234, Jan 2004.

S. Klamt and J. Stelling. Combinatorial complexity of pathway analysis in metabolic networks. *Mol. Biol. Rep.*, 29(1-2):233–236, 2002.

H. L. Kornberg and N. B. Madsen. Synthesis of C4-dicarboxylic acids from acetate by a glyoxylate bypass of the tricarboxylic acid cycle. *Biochim. Biophys. Acta*, 24(3): 651–653, Jun 1957.

J. O. Krömer, C. Wittmann, H. Schröder, and E. Heinzle. Metabolic pathway analysis for rational design of L-methionine production by *Escherichia coli* and *Corynebacterium glutamicum*. *Metab. Eng.*, 8(4):353–369, Jul 2006.

A. Kun, B. Papp, and E. Szathmry. Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. *Genome Biol*, 9(3):R51, 2008.

B. R. Landau and H. Brunengraber. The role of acetone in the conversion of fat to carbohydrate. *Trends Biochem. Sci.*, 12:113–114, 1987.

J. M. Lee, E. P. Gianchandani, and J. A. Papin. Flux balance analysis in the era of metabolomics. *Brief. Bioinform.*, 7(2):140–150, Jun 2006.

A. Lehninger, D. L. Nelson, and M. M. Cox. *Lehninger Principles of Biochemistry*. W. H. Freeman, 5th edition, June 2008.

E. R. Lindahl. Molecular dynamics simulations. *Methods Mol. Biol.*, 443:3–23, 2008.

R. Mahadevan and C. H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.*, 5(4):264–276, Oct 2003.

N. Matsumaru, T. Lenser, T. Hinze, and P. Dittrich. *Advances in Biologically Inspired Information Systems*, volume 69 of *Studies in Computational Intelligence*, chapter Toward Organization-Oriented Chemical Programming: A Case Study with the Maximal Independent Set Problem, pages 147–163. 2007.

F. Matthäus, C. Salazar, and O. Ebenhöh. Biosynthetic potentials of metabolites and their hierarchical organization. *PLoS Comput. Biol.*, 4(4):e1000049, Apr 2008. doi: 10.1371/journal.pcbi.1000049.

R. A. Notebaart, F. H. J. van Enckevort, C. Francke, R. J. Siezen, and B. Teusink. Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics*, 7:296, 2006. doi: 10.1186/1471-2105-7-296.

N. L. Novère, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, J. L. Snoep, and M. Hucka. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.*, 34(Database issue):D689–D691, Jan 2006.

Y.-K. Oh, B. Ø. Palsson, S. M. Park, C. H. Schilling, and R. Mahadevan. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *J. Biol. Chem.*, 282(39):28791–28799, Sep 2007.

B. Papp, B. Teusink, and R. A. Notebaart. A critical view of metabolic network adaptations. *HFSP J.*, 3(1):24–35, Feb 2009.

J. L. Reed, T. D. Vo, C. H. Schilling, and B. Ø. Palsson. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.*, 4(9):R54, 2003.

G. A. Reichard, A. C. Haff, C. L. Skutches, P. Paul, C. P. Holroyde, and O. E. Owen. Plasma acetone metabolism in the fasting human. *J. Clin. Invest.*, 63(4):619–626, Apr 1979.

C. H. Schilling, D. Letscher, and B. Ø. Palsson. Theory for the systemic definition of

metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, 203(3):229–248, Apr 2000.

B. Schrumpf, A. Schwarzer, J. Kalinowski, A. Pühler, L. Eggeling, and H. Sahm. A functionally split pathway for lysine synthesis in corynebacterium glutamicium. *J. Bacteriol.*, 173(14):4510–4516, Jul 1991.

R. Schuetz, L. Kuepfer, and U. Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.*, 3:119, 2007. doi: 10.1038/msb4100162.

S. Schuster, T. Dandekar, and D. A. Fell. Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, 17(2):53–60, Feb 1999.

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(3):326–332, Mar 2000.

S. Schuster, S. Klamt, W. Weckwerth, and T. Pfeiffer. Use of network analysis of metabolic systems in bioengineering. *Bioproc. Biosyst. Eng.*, 24:363–372, 2002.

S. Schuster, T. Pfeiffer, and D. A. Fell. Is maximization of molar yield in metabolic networks favoured by evolution? *J. Theor. Biol.*, 252(3):497–504, Jun 2008.

J.-M. Schwartz, C. Gaugain, J. Nacher, A. de Daruvar, and M. Kanehisa. Observing metabolic functions at the genome scale. *Genome Biol.*, 8(6):R123, Jun 2007.

J. Schwender, F. Goffman, J. B. Ohlrogge, and Y. Shachar-Hill. Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. *Nature*, 432 (7018):779–782, Dec 2004.

T. Shlomi, Y. Eisenberg, R. Sharan, and E. Ruppin. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol. Syst. Biol.*, 3:101, 2007. doi: 10.1038/msb4100141.

T. Shlomi, M. N. Cabili, M. J. Herrgård, B. Ø. Palsson, and E. Ruppin. Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, 26(9):1003–1010, Sep 2008.

J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912): 190–193, Nov 2002.

L. Stryer, J. Berg, and J. Tymoczko. *Biochemistry - 5th Edition*, volume 5th of *Trends in Biochemical Sciences*. W.H. Freeman and Company, 2002. ISBN 0-7167-4684-0.

P. F. Suthers, M. S. Dasika, V. S. Kumar, G. Denisov, J. I. Glass, and C. D. Maranas. A genome-scale metabolic reconstruction of *Mycoplasma genitalium*, iPS189. *PLoS Comput. Biol.*, 5(2):e1000285, Feb 2009. doi: 10.1371/journal.pcbi.1000285.

K.-Y. Tsai and F.-S. Wang. Evolutionary optimization with data collocation for reverse engineering of biological networks. *Bioinformatics*, 21(7):1180–1188, Apr 2005.

A. Varma and B. Ø. Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.*, 60(10):3724–3731, Oct 1994a.

A. Varma and B. Ø. Palsson. Metabolic Flux Balancing: Basic concepts, scientific and practical use. *Bio/Technology*, 12(10):994–998, October 1994b.

B. D. Ventura, C. Lemerle, K. Michalodimitrakis, and L. Serrano. From in vivo to in silico biology and back. *Nature*, 443(7111):527–533, Oct 2006.

D. J. Voet, J. G. Voet, and C. W. Pratt. *Fundamentals of Biochemistry: Life at the Molecular Level*. Wiley, 2nd edition, March 2005. ISBN 0471214957.

E. O. Weinman, E. H. Srisower, and I. L. Chaikoff. Conversion of fatty acids to carbohydrate; application of isotopes to this problem and role of the Krebs cycle as a synthetic pathway. *Physiol. Rev.*, 37(2):252–272, Apr 1957.

H. V. Westerhoff and B. O. Palsson. The evolution of molecular biology into systems biology. *Nat. Biotechnol.*, 22(10):1249–1252, Oct 2004.

T. Wilhelm, J. Behre, and S. Schuster. Analysis of structural robustness of metabolic networks. *Syst. Biol. (Stevenage)*, 1(1):114–120, Jun 2004.

C. Wittmann, J. Weber, E. Betiku, J. Krömer, D. Böhm, and U. Rinas. Response of fluxome and metabolome to temperature-induced recombinant protein synthesis in *Escherichia coli*. *J. Biotechnol.*, 132(4):375–384, Dec 2007.

M. Yeung, I. Thiele, and B. . Palsson. Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics*, 8(1):363, 2007. doi: 10.1186/1471-2105-8-363.

Supplemental Materials

F. Centler, C. Kaleta, P. S. di Fenizio, and P. Dittrich. *Bioinformatics*, 24(14):1611-1618, Jul 2008.
-Supplemental Material-

BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–7

Computing Chemical Organizations in Biological Networks - Supplementary Material

F. Centler*, C. Kaleta, P. Speroni di Fenizio, and P. Dittrich†

Bio Systems Analysis Group, Jena Centre for Bioinformatics (JCB), and Department of Mathematics and Computer Science, Friedrich-Schiller-University Jena, D-07743 Jena, Germany

1 ALGORITHM I - CONSTRUCTIVE APPROACH

1.1 Computing producing species sets

In the constructive approach, the function `producerSets(`*speciesSet*`)` computes all species combinations that produce the species contained in speciesSet. Its pseudo code is detailed below. In the first step, for each species s in speciesSet, a set of species sets productionSets_s is generated. This set contains all species combinations that can produce s. The sets are computed by inspecting all reactions. For each reaction in which s is produced (having a positive stoichiometric coefficient), the set of the reaction educts forms a producing set. In the second step, all possible combinations of species sets from the sets productionSets are generated. Each combination contains exactly one set of productionSets_s for each species s in speciesSet.

Function `producerSets`

Input: set of species to produce speciesSet, reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$
Output: set of all species sets that can produce all species in speciesSet in result

```
result ← ∅ ;
foreach s ∈ speciesSet do
    productionSets_s ← ∅ ;
    foreach reaction ∈ R do
        if s has positive stoichiometric coefficient in reaction
        then
            productionSets_s ← productionSets_s ∪ {
            educts(reaction) } ;
        end
    end
end
repeat
    current ← ∅;
    foreach s ∈ speciesSet do
        select a set setProducingS from productionSets_s;
        current ← current ∪ setProducingS;
    end
    result ← result ∪ { current } ;
until all possible set combinations have been considered;
```

1.2 Computing Connected Semi-Organizations

For computing the connected organizations using the constructive approach, only the function `SOsDirectlyAbove()` needs to be modified:

Function `ConnectedSOsDirectlyAbove`

Input: semi-organization so, reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$
Output: set of all connected semi-organizations directly above so in result

```
result ← ∅ ; usableSpeciesSets ← ∅ ;
if so = {} then
    usableSpeciesSets ← ∪_{s∈M}{{s}};
else
    foreach reaction ∈ R with educts(reaction) ⊈ so
    do
        if ∃ s ∈ so with s ∈ educts(reaction) ∪
        products(reaction) then
            usableSpeciesSets ← usableSpeciesSets ∪ {
            educts(reaction) \ so } ;
        end
    end
end
foreach set ∈ usableSpeciesSets do
    result ← result ∪ SOsDirectlyAbove-
    Containing(so, set);
end
```

1.3 Creating the Complete Hierarchy of Organizations from Connected Organizations

The connected organizations can be used to construct all organizations of the network. They can be viewed as a basis for the complete hierarchy of organizations. If the network does not contain input species, every organization is a combination of connected or basis organizations. If n is the number of basis organizations, $\sum_{i=0}^{n} \binom{n}{i} = 2^n$ different set combinations exist. However, not every combination of basis organizations gives an organization. For example, consider the simple reaction network containing three species and one reaction $\langle \{a, b, c\}, \{a + b \rightarrow c\} \rangle$. Species $\{a\}$ and $\{b\}$ are two connected organizations. As such they are part of the basis, but their combination $\{a, b\}$ lacks the properties of closure and self-maintenance and hence is not an organization. Consequently, to obtain all organizations from the connected organizations, set unions of all combinations of basis

*Current address: Department of Environmental Microbiology, UFZ - Centre for Environmental Research Leipzig-Halle, Germany

†to whom correspondence should be addressed

organizations have to be considered and tested for the organization properties.

If no input species are defined for the reaction network, the basis organizations are exactly the connected organizations. In the presence of input species, the basis is larger. Firstly, again all connected organizations are basis organizations. Secondly, the inflow reactions of the input species must be removed from the network. The connected organizations of the resulting network are additionally basis organizations. This step is required to find connected subnetworks that are not connected to input species. In this case, not all set union combinations must be tested, since all organizations contain at least the input species. The whole procedure can be summarized in four steps:

1. For the given reaction network, compute the set of connected organizations O_{init}.
2. Remove all inflow reactions and compute the set of connected organizations for the modified network $O_{withoutinput}$.
3. The set of basis organizations is $O_{basis} = O_{init} \cup O_{withoutinput}$.
4. Make set unions of all possible combinations of organizations from O_{basis} such that exactly one organization from O_{init} is contained in every combination. (If a combination of organizations from O_{init} is already an organization, it is already an element of O_{init}.) Test the species set of each combination for the closure and self-maintenance property. With $|O_{init}| = m$ and $|O_{withoutinput}| = n$, there are $m \cdot \sum_{i=0}^{n} \binom{n}{i} = m \cdot 2^n$ species sets to be tested.

To show that this procedure is sufficient to create all organizations, we need to proof that any organization is a combination of basis organizations. For this purpose, networks with and without input species will be discussed separately.

Networks without input. If the network has no input species, the basis organizations are exactly the connected organizations. Taking any organization O, we find that it is either connected or not. In the former case, it is a basis organization. In the later case, it consists of two or more parts that are not connected to each other. When inspecting each isolated part separately, we find that each part is closed and self-maintaining. In other words, each part is an organization. Even more, each part is a connected organization and hence a basis organization. Therefore, the unconnected organization O is equal the set union of these basis organizations.

Networks with input. Again, taking any organization O of the network, we find that it is either connected or not. If it is connected, it is already a basis organization. If not, we again inspect the isolated parts of the organization. Like in the case without input species, all parts are closed, self-maintaining, and connected. Some parts contain input species and others not. Recall that in the presence of input species, all input species are present in all organizations. Hence, the union of all isolated parts that contain at least one input species will be an organization (and contained in O_{init}). Parts without input species are only organizations in the absence of input species, and hence contained in $O_{withoutinput}$. We find that all isolated parts of organization O can be associated to basis organizations in O_{init} and $O_{withoutinput}$. Consequently, O is equal the set union of these basis organizations.

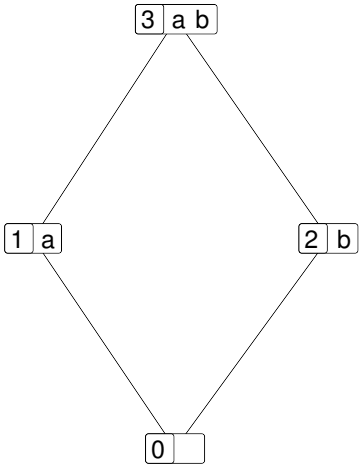


Fig. 1. For the chemical reaction network $\langle\{a, b\}, \{a + b \to 2a + 2b\}\rangle$, all four organizations are connected and hence part of the basis. Since Organization 3 is the union of Organizations 1 and 2, the set of basis organizations is not minimal.

We conclude that all organizations are created using the described procedure. It must be noted that the basis organizations do not form a basis for all organizations that is minimal. Consider the reaction network consisting of two species and one reaction $\langle\{a, b\}, \{a + b \to 2a + 2b\}\rangle$. This system contains four organizations as depicted in Figure 1. All four organizations are connected and therefore basis organizations. However, Organization 3 is the set union of Organizations 1 and 2, and hence would not be required in the basis. In this sense, the set of basis organizations does not form a basis for all organizations that is minimal.

2 NETWORK SIZE VS. NUMBER OF (CONNECTED) SEMI-ORGANIZATIONS

The number of semi-organizations of a reaction network of n species lies between one and 2^n. This is even true for connected semi-organizations. The number of semi-organizations depends on the network topology, making a prediction of the number of semi-organizations for a given network of size n difficult.

We demonstrate how networks can be constructed featuring the extreme numbers of semi-organizations one and 2^n for a given number of species n.

For the first case, we create a reaction network $\langle\mathcal{M}, \mathcal{R}\rangle$ containing a decay reaction for all of the n species:

$$\begin{aligned} \mathcal{M} &= \{s_1, \ldots, s_n\} \\ \mathcal{R} &= \{s_1 \to \emptyset, \cdots, s_n \to \emptyset\}. \end{aligned}$$

It is obvious that the empty set is the only semi-organization of this network.

If the reaction network does not contain any reactions, any of the 2^n possible species sets is trivially a semi-organization:

$$\begin{aligned}\mathcal{M} &= \{s_1, \ldots, s_n\} \\ \mathcal{R} &= \{\}.\end{aligned}$$

If we add an autocatalytic reaction of the form $\{s_i, s_j\} \rightarrow \{s_i, s_i, s_j, s_j\}$ for any species pair (s_i, s_j), any species set is even a connected semi-organization:

$$\begin{aligned}\mathcal{M} &= \{s_1, \ldots, s_n\} \\ \mathcal{R} &= \{\{s_i, s_j\} \rightarrow \{s_i, s_i, s_j, s_j\} | i, j = 1, \ldots, n\}.\end{aligned}$$

3 RUNTIME COMPLEXITY OF ORGANIZATION COMPUTATION

Organization computation is NP-hard. This can be shown by analyzing a specific subproblem, the question whether a reaction network contains a reactive organization apart from the empty set or not. This problem is NP-complete and thus we conclude that organization computation is NP-hard. To show this, we reduce the 3-SAT problem to the mentioned subproblem of organization computation.

Given a boolean formula F in 3-CNF (conjunctive normal form) with n boolean variables b_1, ..., b_n, and l clauses, the 3-SAT problem poses the question, whether an assignment for the boolean variables exist so that F is evaluated to true. As implied by 3-CNF, formula F is made up by clauses connected with "AND" operators. Each clause consists of three literals that are connected with "OR" operators.

We reduce 3-SAT to our subproblem by constructing a reaction network for which the existence of a non-empty reactive organization implies a positive answer to the 3-SAT problem and vice versa. The system consists of l clause species c_1, ..., c_l with each species corresponding to one clause of the 3-SAT problem. Additionally, we introduce species x_1, ..., x_n representing the true, and species $\overline{x_1}$, ..., $\overline{x_n}$ representing the false state of the boolean variables b_1, ..., b_n of F. A clause is evaluated to true if at least one of the literals it contains evaluates to true. Thus we add for each clause species a production reaction such that it is produced if one of its literals is evaluated to true. That means, the clause $(b_1 \vee \neg b_2 \vee b_3)$ is translated into three clause reactions that produce the species c_1 representing this clause: $x_1 \rightarrow x_1 + c_1$, $\overline{x_2} \rightarrow \overline{x_2} + c_1$, and $x_3 \rightarrow x_3 + c_1$. To check whether all clauses evaluate to true, we introduce the species e. This species is produced with stoichiometry $3 \cdot l$ in a master reaction whenever all clause species are present: $c_1 + \cdots + c_l \rightarrow 3 \cdot l\ e$. Next we add e as an educt to all clause reactions, thus $x_1 \rightarrow x_1 + c_1$ becomes $e + x_1 \rightarrow x_1 + c_1$. The stoichiometry of $3 \cdot l$ in e production balances the consumption in the clause reactions. Additionally, we introduce a reaction $x_i + \overline{x_i} \rightarrow \emptyset$ for all n species corresponding to the boolean variables to assure that only one form of a variable can be present at a time. Now, if there is a flux through the master reaction, e is produced at a positive rate. For this to happen, all clause species must be produced at positive rates, and hence, at least one of the clause reactions proceeds for each clause. These reactions are all catalyzed by a variable species x_i, and there can only be one form of a species (either the form "true" or "false") at a time. Therefore, there exists an assignment to the boolean variables in F, such that each of its clauses evaluates to true. By coupling the master reaction to the clause reactions by adding the species e, we require a flux producing all clause species at the same time. Consequently, there is no assignment to the boolean variables b_i such that F evaluates to true, if there is no reactive organization containing at least one reaction. However, there still might be organizations that do neither contain the "true" nor the "false" form for some species. In this case, F can already be evaluated to be true without assigning a specific value to the corresponding variables b_i.

From the construction process it can be directly derived, that each assignment to the variables satisfying F implies at least one reactive organization in the reaction network. Hence, the question whether there exists a reactive organization apart from the trivial one is NP-complete, since this problem can be simulated on a nondeterministic turing machine. After guessing a set of species, it can be checked in polynomial time whether these species enable at least one reaction and fulfill the closure as well as the self-maintenance condition or not.

While it can be shown, that organization computation is NP-hard, the runtime complexity of the computation of elementary modes has not yet been determined. Counting the number of elementary modes of a given system is $\#P$-complete (Chierichetti et al., 2007). This is of interest since the computation of elementary modes of a modified system is a subroutine of the flux based approach for organization computation.

4 LIST OF NOT PRODUCIBLE METABOLITES

In Scenarios 3 and 4, a total of 215 metabolites did not appear in any organization. Hence, they are not present and cannot be produced in any feasible state of the metabolic network under the given growth conditions. Metabolite abbreviations are taken from the supplementary material of Reed et al. (2003). Deadend metabolites only appearing as substrates (^{s}) or products (^{p}) are marked.

Abbr.	Metabolite	Abbr.	Metabolite
23doguln^s	2,3-dioxo-l-gulonate	cbl1(e)	cob(i)alamin (external)
25dkglcn^s	2,5-diketo-d-gluconate	cechddd	cis-3-(3-carboxyethyl)-3,5-cyclohexadiene-1,2-diol
26dap-m(e)	meso-2,6-diaminoheptanedioate (external)	cenchddd	cis-3-(3-carboxyethenyl)-3,5-cyclohexadiene-1,2-diol
2ddglcn	2-dehydro-3-deoxy-d-gluconate	chol	choline
2ddglcn(e)	2-dehydro-3-deoxy-d-gluconate (external)	chol(e)	choline (external)
2dh3dgal	2-dehydro-3-deoxy-d-galactonate	cinnm^s	trans-cinnamate
2dhglcn^s	2-dehydro-d-gluconate	cit(e)	citrate (external)
2dhguln	2-dehydro-l-gulonate	crn	l-carnitine
2pglyc^s	2-phosphoglycolate	crn(e)	l-carnitine (external)
3dgulnp^s	3-keto-l-gulonate-6-phosphate	crncoa	carnitinyl-coa
3dhguln^p	3-dehydro-l-gulonate	csn(e)	cytosine (external)
3hcinnm	3-hydroxycinnamic acid	ctbt	crotonobetaine
3hcinnm(e)	3-hydroxycinnamic acid (external)	ctbtcoa	crotonobetainyl-coa
3hpppn	3-(3-hydroxy-phenyl)propionate	cyan^s	cyanide
3hpppn(e)	3-(3-hydroxy-phenyl)propionate (external)	cynt	cyanate
4ahmmp^s	4-amino-5-hydroxymethyl-2-methylpyrimidine	cynt(e)	cyanate (external)
4h2opntn	4-hydroxy-2-oxopentanoate	cys-l(e)	l-cysteine (external)
4mhetz^s	4-methyl-5-(2-hydroxyethyl)-thiazole	dad-2(e)	deoxyadenosine (external)
5dh4dglc	5-dehydro-4-deoxy-d-glucarate	dann	7,8-diaminononanoate
5prdmbz	n1-(5-phospho-alpha-d-ribosyl)-5,6-dimethylbenzimidazole	dcyt(e)	deoxycytidine (external)
8aonn	8-amino-7-oxononanoate	dgsn(e)	deoxyguanosine (external)
aacald^p	aminoacetaldehyde	dhcinnm	2,3-dihydroxicinnamic acid
acac	acetoacetate	dhpppn	3-(2,3-dihydroxyphenyl)propanoate
acac(e)	acetoacetate (external)	din(e)	deoxyinosine (external)
acgam6p	n-acetyl-d-glucosamine 6-phosphate	dmbzid^s	5,6-dimethylbenzimidazole
acgam(e)	n-acetyl-d-glucosamine (external)	dms^p	dimethyl sulfide
acmana	n-acetyl-d-mannosamine	dms(e)^p	dimethyl sulfide (external)
acmana(e)	n-acetyl-d-mannosamine (external)	dmso^s	dimethyl sulfoxide
acmanap	n-acetyl-d-mannosamine 6-phosphate	dmso(e)	dimethyl sulfoxide (external)
acnam	n-acetylneuraminate	dtbt	dethiobiotin
acnam(e)	n-acetylneuraminate (external)	duri(e)	deoxyuridine (external)
acon-t^s	trans-aconitate	dxyl^s	1-deoxy-d-xylulose
aconm^p	e-3-carboxy-2-pentenedioate 6-methyl ester	f1p	d-fructose 1-phosphate
adocbi	adenosyl cobinamide	fcl-l	l-fuculose
adocbip	adenosyl cobinamide phosphate	fru	d-fructose
adocbl	adenosylcobalamin	fru(e)	d-fructose (external)
agdpcbi	adenosine-gdp-cobinamide	fruur	d-fructuronate
agpc(ec)	acyl-glycerophosphocholine (E. coli)	fuc-l	l-fucose
alltn	allantoin	fuc-l(e)	l-fucose (external)
alltn(e)	allantoin (external)	fuc1p-l^p	l-fucose 1-phosphate
alltt	allantoate	fuc1p-l(e)	l-fucose 1-phosphate (external)
altrn	d-altronate	g3pc	sn-glycero-3-phosphocholine
amob	s-adenosyl-4-methylthio-2-oxobutanoate	g3pi^s	sn-glycero-3-phospho-1-inositol
amp(e)^p	amp (external)	g3ps^s	glycerophosphoserine
ap4a^s	p1,p4-bis(5'-adenosyl) tetraphosphate	g6p(e)	d-glucose 6-phosphate (external)
ap5a^s	p1,p5-bis(5'-adenosyl) pentaphosphate	gal(e)	d-galactose (external)
apoacp^s	apoprotein [acyl carrier protein]	galct-d	d-galactarate
arab-l	l-arabinose	galct-d(e)	d-galactarate (external)
arab-l(e)	l-arabinose (external)	galctn-d	d-galactonate
arbt6p^s	arbutin 6-phosphate	galctn-d(e)	d-galactonate (external)
asp-l(e)	l-aspartate (external)	galt1p	galactitol 1-phosphate
bbtcoa	gamma-butyrobetainyl-coa	galt(e)	galactitol (external)
betald^s	betaine aldehyde	galur	d-galacturonate
btcoa	butanoyl-coa	galur(e)	d-galacturonate (external)
btn	biotin	gam(e)	d-glucosamine (external)
btnso^s	d-biotin d-sulfoxide	gbbtn	gamma-butyrobetaine
but	butyrate (n-c4:0)	gbbtn(e)^p	gamma-butyrobetaine (external)
but(e)	butyrate (n-c4:0) (external)	glcr	d-glucarate
cbi	cobinamide	glcr(e)	d-glucarate (external)
cbl1	cob(i)alamin	glcur	d-glucuronate

Abbr.	Metabolite	Abbr.	Metabolite
glcur(e)	d-glucuronate (external)	ocdca	octadecanoate (n-c18:0)
gln-l(e)	l-glutamine (external)	ocdca(e)	octadecanoate (n-c18:0) (external)
glyb	glycine betaine	op4en	2-oxopent-4-enoate
glyb(e)	glycine betaine (external)	pac	phenylacetic acid
glyc3p(e)	glycerol 3-phosphate (external)	pacald	phenylacetaldehyde
gp4g^{s}	p1,p4-bis(5'-guanosyl) tetraphosphate	pc(ec)^{s}	phosphatidylcholine (E. coli)
gsn(e)	guanosine (external)	peamn^{s}	phenethylamine
hdca(e)	hexadecanoate (n-c16:0) (external)	phaccoa^{p}	phenylacetyl-coa
hkndd	2-hydroxy-6-oxonona-2,4-diene-1,9-dioate	pmcoa	pimeloyl-coa
hkntd	2-hydroxy-6-ketononatrienedioate	pnto-r(e)	(r)-pantothenate (external)
hqn^{p}	hydroquinone	pppn	phenylpropanoate
idp	idp	pppn(e)	phenylpropanoate (external)
inost^{p}	myo-inositol	rbl-l	l-ribulose
itp	itp	rdmbzi	n1-(alpha-d-ribosyl)-5,6-dimethylbenzimidazole
lcts	lactose	rib-d	d-ribose
lcts(e)	lactose (external)	rib-d(e)	d-ribose (external)
mal-l(e)	l-malate (external)	rml	l-rhamnulose
malt	maltose	rmn	l-rhamnose
malt6p^{p}	maltose 6'-phosphate	rmn(e)	l-rhamnose (external)
malt(e)	maltose (external)	sbt-d(e)	d-sorbitol (external)
malthp	maltoheptaose	seln^{s}	selenide
malthx	maltohexaose	selnp^{p}	selenophosphate
malthx(e)	maltohexaose (external)	ser-d	d-serine
maltpt	maltopentaose	ser-d(e)	d-serine (external)
maltpt(e)	maltopentaose (external)	spmd(e)	spermidine (external)
malttr	maltotriose	suc6p	sucrose 6-phosphate
malttr(e)	maltotriose (external)	sucr(e)	sucrose (external)
maltttr	maltotetraose	tag6p-d	d-tagatose 6-phosphate
maltttr(e)	maltotetraose (external)	tagur	d-tagaturonate
man6p(e)	d-mannose 6-phosphate (external)	tartr-l	l-tartrate
man(e)	d-mannose (external)	tartr-l(e)	l-tartrate (external)
mana	d-mannonate	taur	taurine
melib	melibiose	taur(e)	taurine (external)
melib(e)	melibiose (external)	tcynt^{p}	thiocyanate
met-d^{p}	d-methionine	thm(e)	thiamin (external)
met-d(e)	d-methionine (external)	tma^{p}	trimethylamine
met-l(e)	l-methionine (external)	tma(e)^{p}	trimethylamine (external)
mi1p-d^{s}	1d-myo-inositol 1-phosphate	tmao^{s}	trimethylamine n-oxide
mnl(e)	d-mannitol (external)	tmao(e)	trimethylamine n-oxide (external)
nac(e)	nicotinate (external)	tre(e)	trehalose (external)
nad(e)	nicotinamide adenine dinucleotide (external)	tsul	thiosulfate
nmn(e)	nmn (external)	tsul(e)	thiosulfate (external)
no2	nitrite	ttdca(e)	tetradecanoate (n-c14:0) (external)
no2(e)	nitrite (external)	urdglyc	(-)-ureidoglycolate
no3	nitrate	xu5p-l	l-xylulose 5-phosphate
no3(e)	nitrate (external)	xyl-d	d-xylose
o2-^{s}	superoxide anion	xyl-d(e)	d-xylose (external)

REFERENCES

F. Chierichetti, V. Lacroix, A. Marchetti-Spaccamela, M.-F. Sagot, and L. Stougie. Modes and cuts in metabolic networks: Complexity and algorithms. Technical Report Ext. rep. 2007-01, Technische Universiteit Eindhoven, Eindhoven, 2007.

J. L. Reed, T. D. Vo, C. H. Schilling, and B. O. Palsson. An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). *Genome Biol*, 4(9):R54, 2003. doi: 10.1186/gb-2003-4-9-r54. URL http://dx.doi.org/10.1186/gb-2003-4-9-r54.

Phenotype Prediction in Regulated Metabolic Networks
-Supplementary Material-

Christoph Kaleta, Florian Centler, Pietro Speroni di Fenizio, Peter Dittrich

April 3, 2008

Bio Systems Analysis Group
Department of Mathematics and Computer Science
Friedrich-Schiller-University Jena
D-07737 Jena, Germany
Jena Centre for Bioinformatics (JCB), Jena, Germany

Contents

1 Organizations in the Core Network Model of *E. coli*

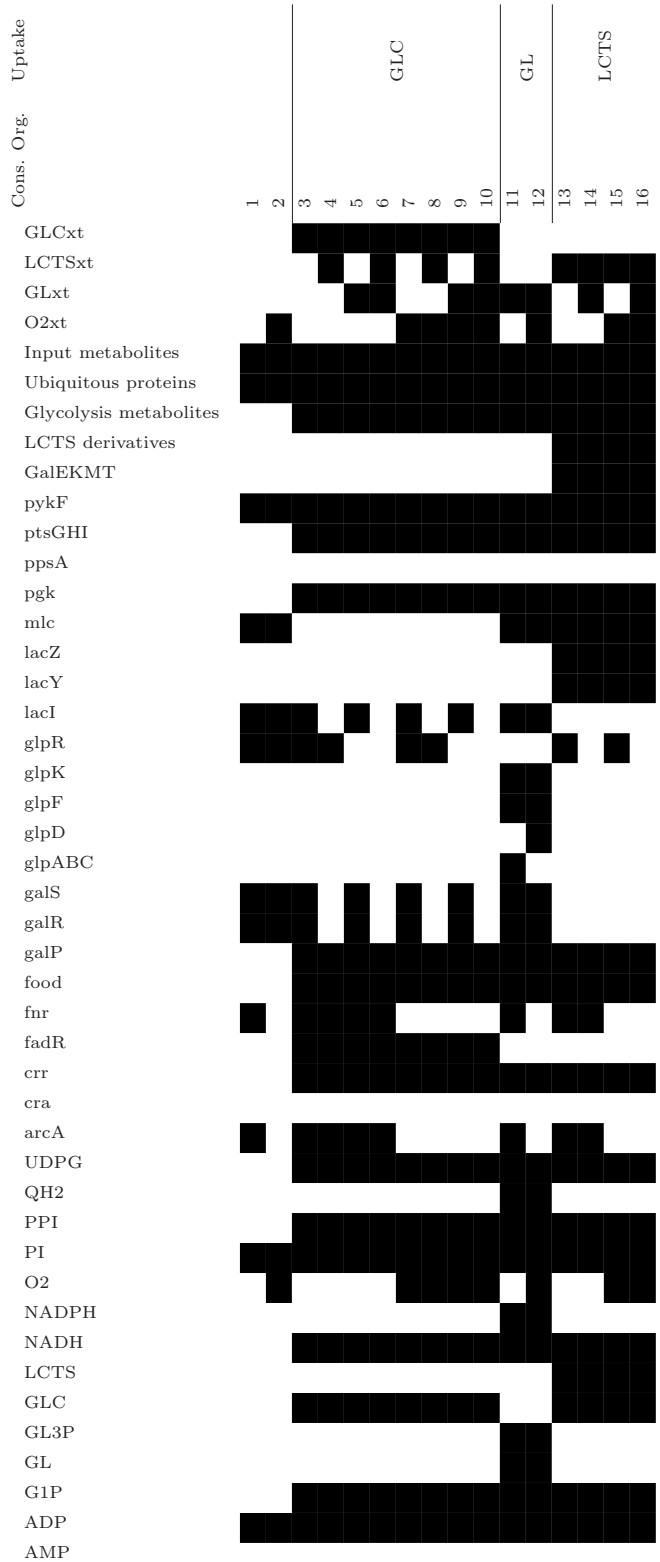


Table 1: Consistent organizations in the core network model of the regulated central metabolism of *E. coli*, ordered by size. The presence of a metabolite in an organization is indicated by a black box. A list of abbrevations can be found in this Supplement. A species followed by 'xt' denotes its extra-cellular form. "Ubiquitous proteins" include the proteins that are considered ubiquitously present in the cell and therefore are not listed separately. They are: Eno, Fba, Fbp, GalU, GapA, Glk, GpmA, GpmB, GpsA, PfkA, PfkB, Pgi, Pgm, PykA, and TpiA. "Input metabolites" denotes the metabolites provided as input to the system: HEXT (external hydrogen), Q (Ubiquinone), ATP, UTP, NAD, and NADP. "Glycolysis metabolites" denotes the metabolites of the glycolysis: G6P, F6P, FDP, T3P2, T3P1, 13PDG, 3PG, 2PG, PEP, and PYR. "Lactose derivatives" denotes the derivatives of lactose in the central metabolism: GAL1P, GLAC, UDPGAL, bDGLAC, bDGLC.

2 The Regulated *E. coli* Network

Species and reactions marked with '*' make up the core network model. The original network can be found in Covert and Palsson [1]. Due to the incomplete set of reactions, the species responsible for the regulation of *cra* and *pdhR* are not contained in the core network model. They are marked with '†'.

2.1 Metabolites

Abbr.	Metabolite	Abbr.	Metabolite
$13PDG^*$	1,3-bis-Phosphoglycerate	$LACxt$	External lactate
$2PG^*$	2-Phosphoglycerate	$LCTS^*$	Lactose
$3PG^*$	3-Phosphoglycerate	$LCTSxt^*$	External Lactose
AC	Acetate	MAL	Malate
$ACCOA$	Acetyl-CoA	NAD^*	Nicotinamide adenine dinucleotide
$ACTP$	Acetyl-phosphate	$NADH^*$	Nicotinamide adenine dinucleotide red.
$ACxt$	External acetate	$NADP^*$	Nicotinamide adenine dinucleotide phosphate
ADP^*	Adenosine diphosphate	$NADPH^*$	Dihydronicotinamide adenine dinucleotide phosphate reduced
AKG	a-Ketoglutarate	$O2^*$	Oxygen
AMP^*	Adenosine monophosphate	$O2xt^*$	External Oxygen
ATP^*	Adenosine triphosphate	OA	Oxaloacetate
$bDGLAC^*$	b-D-Galactose	PEP^*	Phosphoenolpyruvate
$bDGLC^*$	b-D-Glucose	PI^*	Phosphate (inorganic)
$Biomass$	Cell biomass	$PIxt$	External phosphate
CIT	Citrate	PPI^*	Pyrophosphate
$CO2$	Carbon dioxide	PYR^*	Pyruvate
$CO2xt$	External carbon dioxide	$PYRxt$	External pyruvate
COA	Coenzyme A	Q^*	Ubiquinone
$D6PGC$	D-6-Phosphate-gluconate	$QH2^*$	Ubiquinol
$D6PGL$	D-6-Phosphate-glucono-delta-lactone	$R5P$	Ribose 5-phosphate
$E4P$	Erythrose 4-phosphate	RIB	Ribose
ETH	Ethanol	$RIBxt$	External ribose
$ETHxt$	External ethanol	$RL5P$	Ribulose 5-phosphate
$F6P^*$	Fructose 6-phosphate	$S7P$	sedo-Heptulose
FAD	Flavin adenine dinucleotide	$SUCC$	Succinate
$FADH$	Flavin adenine dinucleotide reduced	$SUCCOA$	Succinate CoA
FDP^*	Fructose 1,6-diphosphate	$SUCCxt$	External succinate
FOR	Formate	$T3P1^*$	Glyceraldehyde 3-phosphate
$FORxt$	External Formate	$T3P2^*$	Dihydroxyacetone phosphate
FUM	Fumarate	$UDPG^*$	UDP Glucose
$G1P^*$	Glucose 1-phosphate	$UDPGAL^*$	UDP Galactose
$G6P^*$	Glucose 6-phosphate	UTP^*	Uridine triphosphate
$GAL1P^*$	Galactose 1-Phosphate	$X5P$	Xylulose-5-phosphate
GL^*	Glycerol	$food^*$	carbon source present in medium
$GL3P^*$	Glycerol 3-phosphate	$LactateUP$	lactate uptake activated
$GLAC^*$	Galactose	$CraCondNeg^\dagger$	*cra* regulation
w GLC^*	a-D-Glucose	$ftktA^\dagger$	*cra* regulation
$GLCxt^*$	External glucose	$ftktB^\dagger$	*cra* regulation
GLX	Glyoxylate	$ftalA^\dagger$	*cra* regulation
$Oxid-$	superoxid radicals	$ftalB^\dagger$	*cra* regulation
$Radicals$		$fpgi^\dagger$	*cra* regulation
$GLxt^*$	External glycerol	$PdhRCondNeg^\dagger$	*pdhR* regulation
$HEXT^*$	External H+	$fdctA^\dagger$	*pdhR* regulation
$ICIT$	Isocitrate	$fdcuA^\dagger$	*pdhR* regulation
LAC	D-Lactate	$fdld^\dagger$	*pdhR* regulation

2.2 Genes and Proteins

Gene	Protein	Gene	Protein
aceA	Isocitrate lyase	*pgm**	Phosphoglucomutase
aceB	Malate synthase A	*pntAB*	Pyridine nucleotide transhydrogenase
aceEF, *lpdA*	Pyruvate dehydrogenase	*ppa*	Inorganic pyrophosphatase
ackA	Acetate kinase A	*ppc*	Phosphoenolpyruvate carboxylase
acnA	Aconitase A	*ppsA**	Phosphoenolpyruvate synthase
acnB	Aconitase B	*pta*	Phosphotransacetylase
acs	Acetyl-CoA synthetase	*pykA**	Pyruvate Kinase II
adhE	Acetaldehyde dehydrogenase	*pykF**	Pyruvate Kinase I
adk	Adenylate kinase	*rbsK*	Ribokinase
atpABC− DEFGHI	F0F1-ATPase	*rpe*	Ribulose phosphate 3-epimerase
cydAB	Cytochrome oxidase bd	*rpiA*	Ribose-5-phosphate isomerase A
cyoABCD	Cytochrome oxidase bo3	*rpiB*	Ribose-5-phosphate isomerase B
dld	D-Lactate dehydrogenase 1	*sdhABCD*	Succinate dehydrogenase complex
*eno**	Enolase	*sfcA*	Malic enzyme (NAD)
*fba**	Fructose-1,6-bisphosphatate aldolase	*sucAB*, *lpdA*	2-Ketoglutarate dehyrogenase
*fbp**	Fructose-1,6-bisphosphatase	*sucCD*	Succinyl-CoA synthetase
fdnGHI	Formate dehydrogenase-N	*talA*	Transaldolase A
fdoIHG	Formate dehydrogenase-O	*talB*	Transaldolase B
frdABCD	Fumarate reductase	*tktA*	Transketolase I
fumA	Fumarase A		
fumB	Fumarase B	*tktB*	Transketolase II
fumC	Fumarase C		
*galE**	UDP-glucose 4-epimerase	*tpiA**	Triosphosphate Isomerase
*galK**	Galactokinase	*zwf*	Glucose 6-phosphate-1-dehydrogenase
*galM**	Aldose 1-epimerase (mutorotase)	*focA*	Formate transport
		*ptsGHI**, *crr**	Glucose transport
*galT**	Galactose-1-phosphate uridylyltransferase	*galP**	Glucose transport (low affinity)
*galU**	UDP-glucose-1-phosphate uridylyltransferase	*glpF**	Glycerol transporter
*gapA**	Glyceraldehyde-3-phosphate dehydrogenase-A complex	*lacY**	Lactose permease
*glk**	Glucokinase	*pitAB*	Phosphate transport
*glpABC**	Glycerol-3-phosphate dehydrogenase (anaerobic)	*rbsABCD*	Ribose transport
*glpD**	Glycerol-3-phosphate dehydrogenase (aerobic)	*dctA*	Succinate transport
*glpK**	Glycerol kinase	*dcuA*	Succinate transport
gltA	Citrate synthase	*dcuB*	Succinate transport
gnd	6-Phosphogluconate dehydrogenase (decarboxylating)	*dcuC*	Succinate efflux
*gpmA**	Phosphoglycerate mutase 1	*arcA**	Aerobic/Anaerobic response regulator
*gpmB**	Phosphoglycerate mutase 2	*cra**(fruR)	Catabolite activator protein
*gpsA**	Glycerol-3-phosphate-dehydrogenase-[NAD(P)+]		
icdA	Isocitrate dehydrogenase	*dcuR*	Dicarboxylate response regulator
*lacZ**	Beta-galactosidase (LACTase)	*dcuS*	Dicarboxylate response sensor
maeB	Malic enzyme (NADP)	*fadR**	Fatty acid/Acetate response regulator
mdh	Malate dehydrogenase	*fnr**	Aerobic/Anaerobic response regulator
ndh	NADH dehydrogenase II	*galR**	Galactose operon repressor
nuoABEF− GHIJKLMN	NADH dehydrogenase I	*galS**	Galactose operon repressor
pckA	Phosphoenolpyruvate carboxykinase	*glpR**	Glycerol response regulator
*pfkA**	Phosphofructokinase	*iclR*	Fatty acid/Acetate response regulator
*pfkB**	Phosphofructokinase B	*lacI**	Lactose operon repressor
pflAB	Pyruvate formate lyase 1	*mlc**	Glucose response regulator
pflCD	Pyruvate formate lyase 2	*pdhR*	Pyruvate response regulator

Gene	Protein	Gene	Protein
*pgi**	Phosphoglucose isomerase	*rbsR*	Ribose response regulator
*pgk**	Phosphoglycerate kinase	*rpiR*	Ribose response regulator
pgl	6-Phosphogluconolactonase		

2.3 Pseudo Species

$\overline{G6P}^*$, $\overline{ACxt}$, $\overline{GLAC}^*$, $\overline{GLCxt}^*$, $\overline{GLxt}^*$, $\overline{GalR}^*$, $\overline{GalS}^*$, $\overline{GlpR}^*$, $\overline{IclR}$, $\overline{LACxt}$, $\overline{LCTSxt}^*$, $\overline{O2xt}^*$, $\overline{PYR}$, $\overline{PYRxt}$, $\overline{RIBxt}$, $\overline{SUCCxt}$, $\overline{ETHxt}$, $\overline{ArcA}^*$, $\overline{Cra}^*$, $\overline{CraCondNeg}^\dagger$, $\overline{ftktA}^\dagger$, $\overline{ftktB}^\dagger$, $\overline{ftalA}^\dagger$, $\overline{ftalB}^\dagger$, $\overline{fpgi}^\dagger$, $\overline{Fnr}^*$, $\overline{Food}^*$, $\overline{LacI}^*$, $\overline{Mlc}^*$, $\overline{PdhR}$, $\overline{PdhRCondNeg}^\dagger$, $\overline{fdld}^\dagger$, $\overline{fdctA}^\dagger$, $\overline{fdcuA}^\dagger$, $\overline{RbsR}$, $\overline{RpiR}$

2.4 Spontaneously Created Species

ADP, ATP, AckA, Adk, AtpABCDEFGHI, COA, DcuA, Dld, Eno*, FAD, FADH, Fba*, Fbp*, FdoIHG, GalU*, GapA*, Glk*, GltA, Gnd, GpmA*, GpmB*, GpsA*, HEXT*, IcdA, MaeB, NAD*, NADH, NADP*, NADPH, NuoABEFGHIJKLMN, PI, PckA, PfkA*, PfkB*, Pgi*, Pgl, Pgm*, PitAB, PntAB, Ppa, Ppc, Pta, PykA*, Q*, QH2, Rpe, RpiA, SfcA, SucCD, TalA, TalB, TktA, TktB, TpiA*, UTP*, Zwf*

Species used as input species (in the complete network), respectively self-replicators (in the core network): *GLCxt*, *LCTSxt*, *GLxt*, *O2xt*

2.5 Spontaneously Decaying Species

13PDG, 2PG, 3PG, AC, ACCOA, ACTP, ACxt, ADP*, AKG, AMP*, ATP*, AceA, AceB, AceEF, AckA, AcnA, AcnB, Acs, AdhE, Adk, ArcA*, AtpABCDEFGHI, BDGLAC, BDGLC, Biomass, CIT, CO2, CO2xt, COA, Cra*, CraCondNeg, Crr*, CydAB, CyoABCD, D6PGC, D6PGL, DctA, DcuA, DcuB, DcuC, DcuR, DcuS, Dld, E4P, ETH, ETHxt, Eno*, F6P, FAD, FADH, FDP, FOR, FORxt, FUM, FadR*, Fba*, Fbp*, FdnGHI, FdoIHG, Fnr*, FocA, Food*, FrdABCD, FumA, FumB, FumC, G1P, G6P*, GAL1P*, GL*, GL3P, GLAC, GLC, GLCxt*, GLX, GLxt, GalE*, GalK*, GalM*, GalP*, GalR*, GalS*, GalT*, GalU*, GapA*, Glk*, GlpABC*, GlpD*, GlpF*, GlpK*, GlpR*, GltA, Gnd, GpmA*, GpmB*, GpsA*, HEXT*, ICIT, IcdA, IclR, LAC, LACxt, LCTS, LCTSxt*, LacI*, LacY*, LacZ*, LactateUP, LpdA, MAL, MaeB, Mdh, Mlc*, NAD*, NADH*, NADP*, NADPH*, Ndh, NuoABEFGHIJKLMN, O2*, O2xt*, OA, PEP*, PI*, PIxt, PPI*, PYR*, PYRxt, PckA, PdhR, PdhRCondNeg, PfkA*, PfkB*, PflAB, PflCD, Pgi, Pgk*, Pgl, Pgm*, PitAB, PntAB, Ppa, Ppc, PpsA*, Pta, PtsGHI*, PykA*, PykF*, Q*, QH2*, R5P, RIB, RIBxt, RL5P, RbsABCD, RbsK, RbsR, Rpe, RpiA, RpiB, RpiR, S7P, SOxidRadicals, SUCC, SUCCOA, SUCCxt, SdhABCD, SfcA, SucAB, SucCD, T3P1*, T3P2, TalA, TalB, TktA, TktB, TpiA*, UDPG, UDPGAL, UTP*, X5P, Zwf*

2.6 Transport Reactions

$$HEXT{+}ACxt \rightarrow AC$$
$$AC \rightarrow HEXT{+}ACxt$$
$$CO2xt \rightarrow CO2$$
$$CO2 \rightarrow CO2xt$$
$$HEXT{+}ETHxt \rightarrow ETH$$
$$ETH \rightarrow HEXT{+}ETHxt$$
$$FocA{+}FORxt \rightarrow FOR{+}FocA$$
$$FOR{+}FocA \rightarrow FocA{+}FORxt$$
$$GLCxt{+}PEP{+}Crr{+}PtsGHI \rightarrow PYR{+}G6P{+}Crr{+}PtsGHI{+}PdhRCondNeg^{\dagger *}$$
$$GLCxt{+}HEXT{+}GalP \rightarrow GLC{+}GalP^*$$
$$GLxt{+}GlpF \rightarrow GL{+}GlpF^*$$
$$GL{+}GlpF \rightarrow GLxt{+}GlpF^*$$
$$HEXT{+}LACxt{+}LactateUP \rightarrow LAC{+}LactateUP$$
$$LAC \rightarrow HEXT{+}LACxt$$
$$HEXT{+}LCTSxt{+}LacY \rightarrow LCTS{+}LacY^*$$
$$LCTS{+}LacY \rightarrow HEXT{+}LCTSxt{+}LacY^*$$
$$O2xt \rightarrow O2^*$$
$$O2 \rightarrow O2xt^*$$
$$HEXT{+}PIxt{+}PitAB \rightarrow PI{+}PitAB$$
$$PI{+}PitAB \rightarrow HEXT{+}PIxt{+}PitAB$$
$$HEXT{+}PYRxt \rightarrow PYR$$
$$PYR \rightarrow HEXT{+}PYRxt$$
$$ATP{+}RbsABCD{+}RIBxt \rightarrow ADP{+}PI{+}RIB{+}RbsABCD$$

$$\overline{FdctA}^{\dagger} + SUCCxt + HEXT + DctA \rightarrow \overline{FdctA}^{\dagger} + PdhRCondNeg^{\dagger} + SUCC + DctA$$
$$FdctA^{\dagger} + SUCC + DctA \rightarrow FdctA^{\dagger} + SUCCxt + HEXT + DctA$$
$$\overline{FdcuA}^{\dagger} + SUCCxt + HEXT + DcuA \rightarrow \overline{FdcuA}^{\dagger} + PdhRCondNeg^{\dagger} + SUCC + DcuA$$
$$FdcuA^{\dagger} + SUCC + DcuA \rightarrow FdcuA^{\dagger} + SUCCxt + HEXT + DcuA$$
$$SUCCxt + HEXT + DcuB \rightarrow SUCC + DcuB$$
$$SUCC + DcuB \rightarrow SUCCxt + HEXT + DcuB$$
$$SUCC + DcuC \rightarrow SUCCxt + HEXT + DcuC$$

2.7 Metabolic reactions

$$ACCOA + PYR + PEP + T3P1 + F6P + G6P + OA + 3PG + AKG + R5P + SUCCOA + E4P \rightarrow Biomass$$
$$AceA + ICIT \rightarrow AceA + SUCC + GLX$$
$$GLX + ACCOA + AceB \rightarrow COA + AceB + MAL$$
$$NAD + COA + AceEF + PYR + LpdA \rightarrow NADH + ACCOA + AceEF + LpdA + CO2$$
$$ADP + AckA + ACTP \rightarrow ATP + AckA + AC$$
$$ATP + AckA + AC \rightarrow ADP + AckA + ACTP$$
$$CIT + AcnA \rightarrow ICIT + AcnA$$
$$ICIT + AcnA \rightarrow CIT + AcnA$$
$$CIT + AcnB \rightarrow ICIT + AcnB$$
$$ICIT + AcnB \rightarrow CIT + AcnB$$
$$ATP + COA + AC + Acs \rightarrow ACCOA + Acs + PPI + AMP$$
$$ACCOA + AdhE + 2.0\ NADH \rightarrow COA + AdhE + ETH + 2.0\ NAD$$
$$COA + AdhE + ETH + 2.0\ NAD \rightarrow ACCOA + AdhE + 2.0\ NADH$$
$$ATP + AMP + Adk \rightarrow Adk + 2.0\ ADP$$
$$Adk + 2.0\ ADP \rightarrow ATP + AMP + Adk$$
$$ATP + AtpABCDEFGHI \rightarrow ADP + AtpABCDEFGHI + PI + 4.0\ HEXT$$
$$ADP + AtpABCDEFGHI + PI + 4.0\ HEXT \rightarrow ATP + AtpABCDEFGHI$$
$$CydAB + O2 + 2.0\ QH2 \rightarrow CydAB + 2.0\ Q + 4.0\ HEXT$$
$$O2 + CyoABCD + 2.0\ QH2 \rightarrow CyoABCD + 2.0\ Q + 5.0\ HEXT$$
$$Fdld^{\dagger} + NADH + PYR + Dld \rightarrow Fdld^{\dagger} + NAD + Dld + LAC$$
$$\overline{Fdld}^{\dagger} + NAD + Dld + LAC \rightarrow \overline{Fdld}^{\dagger} + PdhRCondNeg^{\dagger} + NADH + PYR + Dld$$
$$Q + Dld + LAC \rightarrow PdhRCondNeg^{\dagger} + QH2 + PYR + Dld$$
$$2PG + Eno \rightarrow Eno + PEP^{*}$$
$$Eno + PEP \rightarrow 2PG + Eno^{*}$$
$$Fba + FDP \rightarrow Fba + T3P2 + T3P1^{*}$$
$$Fba + T3P2 + T3P1 \rightarrow Fba + FDP^{*}$$
$$FDP + Fbp \rightarrow PI + Fbp + F6P^{*}$$
$$Q + FdnGHI + FOR \rightarrow QH2 + CO2 + FdnGHI + 2.0\ HEXT$$
$$Q + FOR + FdoIHG \rightarrow QH2 + CO2 + FdoIHG + 2.0\ HEXT$$
$$FADH + FrdABCD + FUM \rightarrow FAD + SUCC + FrdABCD$$
$$FUM + FumA \rightarrow MAL + FumA$$
$$MAL + FumA \rightarrow FUM + FumA$$
$$FUM + FumB \rightarrow MAL + FumB$$
$$MAL + FumB \rightarrow FUM + FumB$$
$$FUM + FumC \rightarrow MAL + FumC$$
$$MAL + FumC \rightarrow FUM + FumC$$
$$GalE + UDPGAL \rightarrow GalE + UDPG^{*}$$
$$GalE + UDPG \rightarrow GalE + UDPGAL^{*}$$
$$ATP + GLAC + GalK \rightarrow ADP + GalK + GAL1P^{*}$$
$$ADP + GalK + GAL1P \rightarrow ATP + GLAC + GalK^{*}$$
$$BDGLAC + GalM \rightarrow GLAC + GalM^{*}$$
$$GLAC + GalM \rightarrow BDGLAC + GalM^{*}$$
$$GalM + BDGLC \rightarrow GalM + GLC^{*}$$
$$GalM + GLC \rightarrow GalM + BDGLC^{*}$$
$$UTP + GAL1P + GalT \rightarrow PPI + UDPGAL + GalT^{*}$$
$$PPI + UDPGAL + GalT \rightarrow UTP + GAL1P + GalT^{*}$$

$$
\begin{aligned}
UTP+GalU+G1P &\rightarrow PPI+UDPG+GalU^{*}\\
PPI+UDPG+GalU &\rightarrow UTP+GalU+G1P^{*}\\
NAD+PI+T3P1+GapA &\rightarrow NADH+GapA+13PDG^{*}\\
NADH+GapA+13PDG &\rightarrow NAD+PI+T3P1+GapA^{*}\\
ATP+GLC+Glk &\rightarrow ADP+Glk+G6P^{*}\\
Q+GlpABC+GL3P &\rightarrow QH2+T3P2+GlpABC^{*}\\
Q+GL3P+GlpD &\rightarrow QH2+T3P2+GlpD^{*}\\
ATP+GL+GlpK &\rightarrow ADP+GL3P+GlpK^{*}\\
ACCOA+OA+GltA &\rightarrow COA+CIT+GltA\\
NADP+Gnd+D6PGC &\rightarrow NADPH+CO2+Gnd+RL5P\\
3PG+GpmA &\rightarrow 2PG+GpmA^{*}\\
2PG+GpmA &\rightarrow 3PG+GpmA^{*}\\
3PG+GpmB &\rightarrow 2PG+GpmB^{*}\\
2PG+GpmB &\rightarrow 3PG+GpmB^{*}\\
NADP+GL3P+GpsA &\rightarrow NADPH+T3P2+GpsA^{*}\\
NADPH+T3P2+GpsA &\rightarrow NADP+GL3P+GpsA^{*}\\
NADP+ICIT+IcdA &\rightarrow NADPH+CO2+IcdA+AKG\\
NADPH+CO2+IcdA+AKG &\rightarrow NADP+ICIT+IcdA\\
LCTS+LacZ &\rightarrow bDGLAC+GLC+LacZ^{*}\\
NADP+MAL+MaeB &\rightarrow NADPH+PYR+CO2+MaeB\\
NAD+MAL+Mdh &\rightarrow NADH+OA+Mdh\\
NADH+OA+Mdh &\rightarrow NAD+MAL+Mdh\\
NADH+Q+Ndh &\rightarrow NAD+QH2+Ndh\\
NuoABEFGHIJKLMN+2.0\ NADH+2.0\ Q &\rightarrow NuoABEFGHIJKLMN+2.0\ NAD+2.0\ QH2+\\
&\qquad 7.0\ HEXT\\
ATP+OA+PckA &\rightarrow ADP+CO2+PEP+PckA\\
ATP+F6P+PfkA &\rightarrow ADP+FDP+PfkA^{*}\\
ATP+F6P+PfkB &\rightarrow ADP+FDP+PfkB^{*}\\
COA+PYR+PflAB &\rightarrow ACCOA+FOR+PflAB\\
COA+PYR+PflCD &\rightarrow ACCOA+FOR+PflCD\\
\overline{fpgi}^{\dagger}+G6P+Pgi &\rightarrow \overline{fpgi}^{\dagger}+CraCondNeg^{\dagger}+F6P+Pgi^{*}\\
fpgi^{\dagger}+F6P+Pgi &\rightarrow fpgi^{\dagger}+G6P+Pgi^{*}\\
ADP+13PDG+Pgk &\rightarrow ATP+3PG+Pgk^{*}\\
ATP+3PG+Pgk &\rightarrow ADP+13PDG+Pgk^{*}\\
Pgl+D6PGL &\rightarrow D6PGC+Pgl\\
G1P+Pgm &\rightarrow G6P+Pgm^{*}\\
G6P+Pgm &\rightarrow G1P+Pgm^{*}\\
NAD+NADPH+PntAB &\rightarrow NADH+NADP+PntAB\\
NADH+NADP+PntAB+2.0\ HEXT &\rightarrow NAD+NADPH+PntAB\\
PPI+Ppa &\rightarrow Ppa+2.0\ PI\\
CO2+PEP+Ppc &\rightarrow PI+OA+Ppc\\
ATP+PYR+PpsA &\rightarrow AMP+PI+PEP+PpsA^{*}\\
ACCOA+PI+Pta &\rightarrow COA+ACTP+Pta\\
COA+ACTP+Pta &\rightarrow ACCOA+PI+Pta\\
ADP+PEP+PykA &\rightarrow ATP+PYR+PykA+PdhRCondNeg^{\dagger *}\\
ADP+PEP+PykF &\rightarrow ATP+PYR+PykF+PdhRCondNeg^{\dagger *}\\
ATP+RIB+RbsK &\rightarrow ADP+RbsK+R5P\\
RL5P+Rpe &\rightarrow Rpe+X5P\\
Rpe+X5P &\rightarrow RL5P+Rpe\\
RL5P+RpiA &\rightarrow R5P+RpiA\\
R5P+RpiA &\rightarrow RL5P+RpiA\\
RL5P+RpiB &\rightarrow R5P+RpiB\\
R5P+RpiB &\rightarrow RL5P+RpiB\\
FAD+SUCC+SdhABCD &\rightarrow FADH+FUM+SdhABCD\\
Q+FADH+SdhABCD &\rightarrow QH2+FAD+SdhABCD\\
QH2+FAD+SdhABCD &\rightarrow Q+FADH+SdhABCD\\
NAD+MAL+SfcA &\rightarrow NADH+PYR+CO2+SfcA\\
NAD+COA+LpdA+AKG+SucAB &\rightarrow NADH+CO2+SUCCOA+LpdA+SucAB\\
ADP+PI+SUCCOA+SucCD &\rightarrow ATP+COA+SUCC+SucCD\\
ATP+COA+SUCC+SucCD &\rightarrow ADP+PI+SUCCOA+SucCD
\end{aligned}
$$

$$\overline{ftalA}^{\dagger} + T3P1 + S7P + TalA \rightarrow \overline{ftalA}^{\dagger} + CraCondNeg^{\dagger} + F6P + TalA + E4P$$
$$ftalA^{\dagger} + F6P + TalA + E4P \rightarrow ftalA^{\dagger} + T3P1 + S7P + TalA$$
$$\overline{ftalB}^{\dagger} + T3P1 + S7P + TalB \rightarrow \overline{ftalB}^{\dagger} + CraCondNeg^{\dagger} + F6P + E4P + TalB$$
$$ftalB^{\dagger} + F6P + E4P + TalB \rightarrow ftalB^{\dagger} + T3P1 + S7P + TalB$$
$$R5P + X5P + TktA \rightarrow T3P1 + S7P + TktA$$
$$T3P1 + S7P + TktA \rightarrow R5P + X5P + TktA$$
$$\overline{ftktA}^{\dagger} + X5P + E4P + TktA \rightarrow \overline{ftktA}^{\dagger} + CraCondNeg^{\dagger} + T3P1 + F6P + TktA$$
$$ftktA^{\dagger} + T3P1 + F6P + TktA \rightarrow ftktA^{\dagger} + X5P + E4P + TktA$$
$$R5P + X5P + TktB \rightarrow T3P1 + S7P + TktB$$
$$T3P1 + S7P + TktB \rightarrow R5P + X5P + TktB$$
$$\overline{ftktB}^{\dagger} + X5P + E4P + TktB \rightarrow \overline{ftktB}^{\dagger} + CraCondNeg^{\dagger} + T3P1 + F6P + TktB$$
$$ftktB^{\dagger} + T3P1 + F6P + TktB \rightarrow ftktB^{\dagger} + X5P + E4P + TktB$$
$$T3P1 + TpiA \rightarrow T3P2 + TpiA^{*}$$
$$T3P2 + TpiA \rightarrow T3P1 + TpiA^{*}$$
$$NADP + G6P + Zwf \rightarrow NADPH + D6PGL + Zwf$$
$$NADPH + D6PGL + Zwf \rightarrow NADP + G6P + Zwf$$
$$ATP \rightarrow ADP + PI^{*}$$

2.8 Regulatory Reactions

$$\overline{IclR} \rightarrow \overline{IclR} + AceA$$
$$\overline{IclR} + \overline{ArcA} \rightarrow \overline{IclR} + \overline{ArcA} + AceB$$
$$\overline{PdhR} \rightarrow \overline{PdhR} + AceEF + LpdA$$
$$Food \rightarrow AcnA + Food$$
$$Food \rightarrow AcnB + Food$$
$$\overline{IclR} + \overline{GLCxt} + \overline{LCTSxt} + \overline{RIBxt} + \overline{GLxt} + \overline{LACxt} + \overline{PYRxt} + \overline{SUCCxt} + \overline{ETHxt} \rightarrow \overline{IclR} + \overline{GLCxt} + \overline{LCTSxt} + \overline{RIBxt} + \overline{GLxt} + \overline{LACxt} + \overline{PYRxt} + \overline{SUCCxt} + \overline{ETHxt} + Acs$$
$$\overline{O2xt} \rightarrow \overline{O2xt} + AdhE$$
$$\overline{Fnr} \rightarrow \overline{Fnr} + CydAB$$
$$ArcA \rightarrow CydAB + ArcA$$
$$\overline{ArcA} + \overline{Fnr} \rightarrow \overline{ArcA} + \overline{Fnr} + CyoABCD$$
$$Fnr \rightarrow FdnGHI + Fnr$$
$$Fnr \rightarrow FrdABCD + Fnr$$
$$FDP + Fbp + \overline{craCondNeg}^{\dagger} \rightarrow FDP + Fbp + \overline{craCondNeg}^{\dagger} + Cra$$
$$DcuR \rightarrow FrdABCD + DcuR$$
$$\overline{ArcA} + \overline{Fnr} \rightarrow \overline{ArcA} + \overline{Fnr} + FumA$$
$$Fnr \rightarrow FumB + Fnr$$
$$SOxidRadicals \rightarrow FumC + SOxidRadicals$$
$$\overline{GLCxt} + \overline{GalR} + \overline{GalS} \rightarrow \overline{GLCxt} + \overline{GalR} + \overline{GalS} + GalE^{*}$$
$$\overline{GLCxt} + \overline{GalR} + \overline{GalS} \rightarrow \overline{GLCxt} + \overline{GalR} + \overline{GalS} + GalK^{*}$$
$$\overline{GLCxt} + \overline{GalR} + \overline{GalS} \rightarrow \overline{GLCxt} + \overline{GalR} + \overline{GalS} + GalM^{*}$$
$$\overline{GLCxt} + \overline{GalR} + \overline{GalS} \rightarrow \overline{GLCxt} + \overline{GalR} + \overline{GalS} + GalT^{*}$$
$$\overline{GLCxt} + \overline{GlpR} + \overline{LCTSxt} + \overline{RIBxt} + Fnr \rightarrow \overline{GLCxt} + \overline{GlpR} + \overline{LCTSxt} + \overline{RIBxt} + GlpABC + Fnr^{*1}$$
$$\overline{GLCxt} + \overline{GlpR} + \overline{LCTSxt} + \overline{RIBxt} + \overline{ArcA} \rightarrow \overline{GLCxt} + \overline{GlpR} + \overline{LCTSxt} + \overline{RIBxt} + \overline{ArcA} + GlpD^{*1}$$
$$\overline{GLCxt} + \overline{GlpR} + \overline{LCTSxt} + \overline{RIBxt} \rightarrow \overline{GLCxt} + \overline{GlpR} + \overline{LCTSxt} + \overline{RIBxt} + GlpK^{*1}$$
$$\overline{GLCxt} + \overline{LacI} \rightarrow \overline{GLCxt} + \overline{LacI} + LacZ^{*}$$
$$\overline{ArcA} \rightarrow \overline{ArcA} + Mdh$$
$$\overline{Fnr} \rightarrow \overline{Fnr} + Ndh$$
$$ArcA \rightarrow PflAB + ArcA$$
$$Fnr \rightarrow PflAB + Fnr$$
$$ArcA \rightarrow PflCD + ArcA$$
$$Fnr \rightarrow PflCD + Fnr$$
$$Food \rightarrow Pgk + Food^{*}$$
$$Cra \rightarrow PpsA + Cra^{*}$$
$$\overline{Cra} \rightarrow \overline{Cra} + PykF^{*}$$

$$
\begin{aligned}
MAL+MaeB+\overline{PdhRCondNeg}^{\dagger} &\rightarrow MAL+MaeB+\overline{PdhRCondNeg}^{\dagger}+PdhR\\
MAL+SfcA+\overline{PdhRCondNeg}^{\dagger} &\rightarrow MAL+SfcA+\overline{PdhRCondNeg}^{\dagger}+PdhR\\
\overline{GLCxt}+\overline{LCTSxt}+\overline{RbsR} &\rightarrow \overline{GLCxt}+\overline{LCTSxt}+\overline{RbsR}+RbsK\\
\overline{RpiR} &\rightarrow \overline{RpiR}+RpiB\\
\overline{ArcA}+\overline{Fnr} &\rightarrow \overline{ArcA}+\overline{Fnr}+SdhABCD\\
\overline{PdhR} &\rightarrow \overline{PdhR}+LpdA+SucAB\\
ArcA &\rightarrow FocA+ArcA\\
Fnr &\rightarrow FocA+Fnr\\
\overline{Mlc}+Food &\rightarrow \overline{Mlc}+Crr+PtsGHI+Food^{*}\\
\overline{Cra}+Food &\rightarrow \overline{Cra}+Crr+PtsGHI+Food^{*}\\
Food &\rightarrow GalP+Food^{*}\\
\overline{GLCxt}+\overline{GlpR}+\overline{LCTSxt}+\overline{RIBxt} &\rightarrow \overline{GLCxt}+\overline{GlpR}+\overline{LCTSxt}+\overline{RIBxt}+GlpF^{*\,1}\\
\overline{GLCxt}+\overline{GLxt}+\overline{LCTSxt}+\overline{RIBxt} &\rightarrow \overline{GLCxt}+\overline{GLxt}+\overline{LCTSxt}+\overline{RIBxt}+\\
&\qquad\qquad LactateUP\\
\overline{GLCxt}+\overline{LacI} &\rightarrow \overline{GLCxt}+\overline{LacI}+LacY^{*}\\
\overline{GLCxt}+\overline{LCTSxt}+\overline{RbsR} &\rightarrow \overline{GLCxt}+\overline{LCTSxt}+\overline{RbsR}+RbsABCD\\
\overline{GLCxt}+\overline{GLxt}+\overline{LACxt}+\overline{LCTSxt}+&\\
\overline{PYRxt}+\overline{RIBxt}+\overline{ArcA}+DcuR &\rightarrow \overline{GLCxt}+\overline{GLxt}+\overline{LACxt}+\overline{LCTSxt}\ +\\
&\qquad +\ \overline{PYRxt}+\overline{RIBxt}+\overline{ArcA}+DctA+DcuR\\
\overline{GLCxt}+\overline{GLxt}+\overline{LACxt}+\overline{LCTSxt}+&\\
\overline{PYRxt}+\overline{RIBxt}+Fnr+DcuR &\rightarrow \overline{GLCxt}+\overline{GLxt}+\overline{LACxt}+\overline{LCTSxt}\\
&\qquad +\ \overline{PYRxt}+\overline{RIBxt}+DcuB+Fnr+DcuR\\
\overline{ArcA}+\overline{Fnr} &\rightarrow \overline{ArcA}+\overline{Fnr}+DcuC\\
\overline{O2xt} &\rightarrow \overline{O2xt}+ArcA^{*}\\
\overline{G6P}+FDP+fbp &\rightarrow \overline{G6P}+FDP+fbp+cra\\
DcuS &\rightarrow DcuR+DcuS\\
SUCCxt &\rightarrow SUCCxt+DcuS\\
GLCxt &\rightarrow GLCxt+FadR^{*}\\
\overline{ACxt} &\rightarrow \overline{ACxt}+FadR\\
\overline{O2xt} &\rightarrow \overline{O2xt}+Fnr^{*}\\
\overline{LCTSxt} &\rightarrow \overline{LCTSxt}+GalR^{*}\\
\overline{LCTSxt} &\rightarrow \overline{LCTSxt}+GalS^{*}\\
\overline{GLxt} &\rightarrow \overline{GLxt}+GlpR^{*}\\
FadR &\rightarrow FadR+IclR\\
\overline{LCTSxt} &\rightarrow \overline{LCTSxt}+LacI^{*}\\
\overline{GLCxt} &\rightarrow \overline{GLCxt}+Mlc^{*}\\
\overline{PYR} &\rightarrow \overline{PYR}+pdhR\\
\overline{RIBxt} &\rightarrow \overline{RIBxt}+RbsR\\
\overline{RIBxt} &\rightarrow \overline{RIBxt}+RpiR\\
GLCxt &\rightarrow GLCxt+Food^{*}\\
LCTSxt &\rightarrow LCTSxt+Food^{*}\\
RIBxt &\rightarrow RIBxt+Food\\
GLxt &\rightarrow GLxt+Food^{*}\\
LACxt &\rightarrow LACxt+Food\\
PYRxt &\rightarrow PYRxt+Food\\
SUCCxt &\rightarrow SUCCxt+Food\\
ETHxt &\rightarrow ETHxt+Food\\
ACxt &\rightarrow ACxt+Food\\
FORxt &\rightarrow FORxt+Food
\end{aligned}
$$

In reactions marked [1], $\overline{RIBxt}$ was removed from the reaction in the core network model.

3 Analysis of a Genome-Scale Metabolic Model

In order to present cases in which OT leads to more accurate predictions compared to FBA, we take a genome-scale metabolic model by Reed et al. [2]. In order to consider cellular growth, we add a decay reaction for each metabolite. The growth medium is represented by a set of influx reactions according to Table 7.

Table 7: Growth medium for the model of Reed et al. [2].

Growth medium
CO_2, Fe^{2+}, H+, water, potassium, natrium, ammonium, anorganic phosphate, sulfate, oxygen, D-glucose, acetate, glycerol, D-lactate, succinate

3.1 Role of Co-factors

One reason for wrong predictions by FBA is that FBA takes co-factors insufficiently into account. Co-factors are molecules that are necessary for some reactions to proceed. They can interact through various means with the substrates and products of a reaction.

Some of the reactions producing metabolites appearing in the biomass reaction can only proceed if certain co-factors are present. These can be molecules which participate in a reaction but are released and reconverted to their initial form in a later step. Most of these co-factors are included in the biomass reaction, but some of them are not. This can for example be the case, if these co-factors are in reality proteins or RNAs which have just been included if an intermediate metabolite is covalently bound to them.

3.1.1 Problem of Non-Metabolic Species

Whether FBA classifies a knockout as viable is not changed by adding a decay for each metabolite. However, OT reveals that this "prediction" is wrong, because in the model by Reed et al. under growth-conditions, there is no organization containing biomass. Consequently, there is no long-term behavior at which biomass is produced and thus there cannot be growth.

An analysis of this problem revealed, that this model includes metabolites that account for tRNA, the acyl carrier protein (Acp), and thioredoxin. Acp and thioredoxin participate in reactions that are essential for the production of biomass metabolites. They are only used as "co-factors" for these reactions. However, there are no production pathways for them.

The addition of an inflow for each of these metabolites then yields a biomass producing organization. We take now the resulting model in order to predict knock experiments.

As can be expected, the knockout of these production pathways is correctly predicted as lethal by OT. FBA however does neither take into account the decay, nor the production reaction, and hence predicts that the organism is viable, independently of whether the production reaction is present or not.

3.1.2 Problem of Redundant Co-factors

Another case arises, when co-factors can be substituted by other co-factors, which is for example the case for menaquinone and ubiquinone in *E. coli* (Figure 1). There is also a third metabolite, 2-demethylmenaquinone, which can be used if both, menaquinone and ubiquinone, cannot be produced by the cell [3]. However, the model we investigated cannot produce biomass when the fluxes of all reactions in which menaquinone and ubiquinone appear are constrained to zero (neither in FBA nor in OT). Thus, here, we only consider menaquinone and ubiquinone.

The presence of either species is sufficient to produce biomass (Figure 1). Thus, including them both in the biomass producing reaction would yield false results when considering knockouts in pathway for the production of one of them. This might be the reason, why both metabolites have not been included in the biomass reaction. The synthesis of both metabolites uses chorismate and octaprenyl as intermediate metabolites, which is also necessary for the production of 2-demethylmenaquinone. The knockouts of each gene essential in the octaprenyl biosynthesis is predicted as lethal by OT, while FBA predicts a viable organism. *In vivo* each of this knockout is lethal [4]. The corresponding genes are listed in Table 8

Another case arises in the knockout of *uppS*, which is correctly predicted as lethal by OT, but not by FBA. This gene synthesizes undecaprenyl diphosphate, a metabolite which acts as co-factor in the lipid metabolism. A more recent model of the metabolism of *E. coli* by Feist et al. [5] includes this metabolite in the biomass function, while neither menaquinone nor ubiquinone are included.

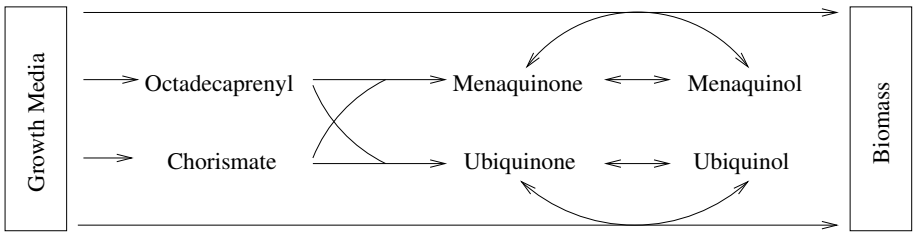


Figure 1: Schematic drawing of the role of menaquinone and ubiquinone in the production of biomass metabolites. The presence of either metabolite gives rise to an organization producing biomass. Knockouts in the essential reactions of the pathway leading to the production of octadecaprenyl are correctly predicted as lethal by OT and falsely predicted as viable by FBA. Please note that both the oxidized and the reduced versions of menaquinone and ubiquinone can be involved in the synthesis of biomass metabolites, thus the corresponding arrows are drawn in both directions.

Table 8: List of 10 lethal knockouts [4] that are correctly predicted by OT, in contrast to FBA.

Pathway	**Essential genes**
menaquinone, ubiquinone, 2-demethylmenaquinone production	*dxr, dxs, gcpE, ispA, ispB, ispD, ispE, ispF, lytB*
undecaprenyl diphosphate production	*uppS*

4 A Flux Based Algorithm for Computing Organizations

We briefly review our algorithm that computes all organizations of a given reaction network (a detailed presentation and analysis will be published elsewhere). For large networks, the computation of all organizations does not finish in reasonable time. We describe a heuristic approach that can at least compute a subset of all organizations in this case.

4.1 Using the self-maintenance cone in flux-space

To be an organization, a set of species has to fulfill two properties: closure and self-maintenance. Starting with the latter condition, methods from convex analysis can be employed to compute organizations. Given a reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$ and its $m \times n$ stoichiometric matrix $\mathbf{S}$, a flux vector $v \in \mathbb{R}^n$ fulfilling the self-maintenance condition must be found to show that a species set is self-maintaining. All such flux vectors lie in a convex polyhedral cone $\mathcal{P}$ in the n-dimensional flux space $\mathbb{R}^n$, originating in the point of origin. The cone is defined by the $n + m$ inequalities:

$$v \geq \mathbf{0} \text{ and } \mathbf{S} \cdot v \geq \mathbf{0}.$$

The constraints can be transformed into a matrix $\mathcal{A}$ representing the extreme rays or spanning vectors of $\mathcal{P}$ [6]. Each point within $\mathcal{P}$ can be written as a linear combination of these extreme rays. Thus, we can compute organizations by searching for combinations of extreme rays whose corresponding set of species fulfills the closure condition.

To compute the extreme rays for a given cone $\mathcal{P}$, we implemented the well-known Schuster algorithm [7] to compute elementary modes. This algorithm computes the extreme rays of a convex cone $\mathcal{P}'$ defined by $v \geq \mathbf{0}$ and $\mathbf{S} \cdot v = \mathbf{0}$. By adding an outflux for each metabolite to the stoichiometric matrix used for the elementary mode computation, the algorithm can also compute the extreme rays of the cone $\mathcal{P}$.

4.2 The extreme ray algorithm

The extreme ray algorithm takes as input a reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$ and a set of extreme rays V_B defining the convex polyhedral cone $\mathcal{P}$ containing all self-maintenance flux vectors. The algorithm delivers all organizations of the network. If instead a convex polyhedral cone $\mathcal{P}' \subset \mathcal{P}$ is supplied as input, only those organizations are computed, whose self-maintaining flux vectors lie within $\mathcal{P}'$. This is useful when, for example, only balanced organizations are to be computed. The cone $\mathcal{P}'$ containing all steady state flux distribution then serves as input for the algorithm. Given a flux vector v, the algorithm only relies on the set of reactions that have positive fluxes in v and not on the specific flux values. Hence, we define v^{set} as the set of reaction indices containing all reactions that have positive fluxes in v. Considering the set of extreme rays V_B defining $\mathcal{P}$, V_B^{set} describes the set of reaction sets v_B^{set} corresponding to the extreme

rays $v_B \in V_B$. Note that generally $|V^{set}| \leq |V|$, as a reaction set v^{set} can correspond to more than one flux vector v. The species set that corresponds to a reaction set v^{set} is denoted by $M(v^{set})$. It contains all reactants and products of the reactions contained in v^{set}.

The computation of the organizations is split into three parts. In the first part, all elementary organizations are computed. An organization O is elementary if no other organizations exist such that the union of their reactions equals the reactions of O. The elementary organizations are then used in the second part of the algorithm to find all organizations that have different sets of reactions. In the final step, organizations are computed that have the same set of reactions like the already computed organizations, but contain different sets of species. In this step organizations are discovered that contain isolated species that do not participate in any reaction of the organization.

4.2.1 Step 1: computing elementary organizations

The central function in the computation of elementary organizations is *organizationsAbove()*. Given a self-maintenance flux vector $v \in \mathcal{P}$ and its corresponding reaction set v^{set}, it computes all organizations O that contain $M(v^{set})$ and for which there exists no other organization being a subset of O and a superset of $M(v^{set})$. (More precisely, *at least* those organizations are computed. Under certain circumstances, organizations are also in the result set for which a subset O_s is also an organization and contains $M(v^{set})$. However, in such a case O_s is also contained in the resulting set of organizations.) Hence, the smallest organizations containing $M(v^{set})$ are computed.

First, the closure of the reaction set v^{set}, respectively $M(v^{set})$, is computed. This is done by taking the species set $M(v^{set})$ and iteratively adding all species to the set that can be created by reactions of the network from the species set. The reaction set $v^{set}_{Closure}$ contains all reactions that can take place in the generated closed set of species. If this reaction set is identical to v^{set}, the species set $M(v^{set})$ is an organization. The reaction set $v^{set}_{Closure}$ contains more reactions than v^{set} when either species were added, or $M(v^{set})$ is closed but v^{set} does not contain all reactions that are possible in this set. One such reaction is taken, and all reaction sets $v^{set}_B \in V^{set}_B$ that contain this reaction are consecutively combined with the original reaction set v^{set} and the function is called again recursively. As the initial reaction set v^{set} and the extreme ray reaction sets v^{set}_B correspond to flux vectors fulfilling the self-maintenance condition, also a flux vector v_u fulfilling the self-maintenance property exists for the union $v^{set}_u = v^{set} \cup v^{set}_B$. Hence, all reaction sets that are considered in the recursive function calls are associated with self-maintaining flux vectors. To obtain all elementary organizations, the function *organizationsAbove()* is called for each reaction set $v^{set}_B \in V^{set}_B$ corresponding to an extreme ray defining $\mathcal{P}$. The central function as pseudo code:

Function `organizationsAbove`

Input: reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$, set V^{set}_B of reaction sets corresponding to the extreme rays spanning the self-maintenance cone $\mathcal{P}$, reaction set v^{set} corresponding to a self-maintenance flux vector $v \in \mathcal{P}$

Output: the set of the smallest organizations containing the reactions of v^{set}

```
result ← {};
v^set_Closure ← closure(v^set );
if v^set_Closure == v^set then
    result ← { M(v^set )};
else
    select one reaction r with r ∈ (v^set_Closure \ v^set);
    foreach v^set_B ∈ V^set_B with r ∈ v^set_B do
        result = result ∪ organizationsAbove(v^set ∪ v^set_B );
    end
end
return(result ) ;
```

4.2.2 Step 2 & 3: computing all organizations

In the first step of the algorithm, the function *organizationsAbove()* was called for each of the reaction sets corresponding to the extreme rays defining $\mathcal{P}$.

In the second part, organizations are determined that are combinations of elementary organizations. This is done by taking all possible combinations of two elementary organizations and calling *organizationsAbove()* for the union of their reaction sets. For every newly discovered organization, this organization must be again combined with each of the elementary organizations and *organizationsAbove()* must be called again for the reaction set unions.

The organizations we have obtained so far all possess a different set of reactions. Consequently, the third step consists of searching for organizations having the same set of reactions as already discovered ones. Hence, we need to determine for all discovered organizations all species sets, that can be added to the organization without changing its set of reactions.

4.3 An heuristic approach to organization computation

The presented algorithm requires the extreme rays of the cone $\mathcal{P}$ as input. For larger networks, the time needed for their computation exceeds practical limits. In order to compute at least a subset of all organizations in such cases, a heuristic approach can be employed. Instead of starting with the set of all extreme rays V_B, the first step of the algorithm can be skipped by directly starting with a set of elementary organizations. To obtain such a set, a simple heuristic approach is used. A random walk through the reaction network delivers a set of species. After computing the closure of this species set, it is tested whether the closure also fulfills the self-maintenance condition. This is done by solving the linear programming problem defined by the self-maintenance constraints and a dummy objective function. After having determined a sufficient large set of organizations, the associated set of elementary organizations O_{el} is determined. The reaction sets of the organizations in O_{el} are then used as input to the second step of the extreme ray algorithm to compute the complete set of organizations that can be found by combining organizations from O_{el}. The heuristic approach was able to correctly determine the whole set of organizations for all tested networks to which we could also apply the exact method.

To verify the results, each knockout experiment has been run two times independently for two hours. Than it was checked whether the predictions for both runs were equal. Since we found no differences, we consider the results reliable.

References

[1] M. W. Covert and B. Palsson. Transcriptional regulation in constraints-based metabolic models of escherichia coli. *J Biol Chem*, 277(31):28058–28064, Aug 2002.

[2] Jennifer L Reed, Thuy D Vo, Christophe H Schilling, and Bernhard O Palsson. An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). *Genome Biol*, 4(9):R54, 2003. doi: 10.1186/gb-2003-4-9-r54. URL `http://dx.doi.org/10.1186/gb-2003-4-9-r54`.

[3] P. T. Lee, A. Y. Hsu, H. T. Ha, and C. F. Clarke. A c-methyltransferase involved in both ubiquinone and menaquinone biosynthesis: isolation and identification of the escherichia coli ubie gene. *J Bacteriol*, 179(5):1748–1754, Mar 1997.

[4] Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A Datsenko, Masaru Tomita, Barry L Wanner, and Hirotada Mori. Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. *Mol Syst Biol*, 2:2006.0008, 2006. doi: 10.1038/msb4100050. URL `http://dx.doi.org/10.1038/msb4100050`.

[5] Adam M Feist, Christopher S Henry, Jennifer L Reed, Markus Krummenacker, Andrew R Joyce, Peter D Karp, Linda J Broadbelt, Vassily Hatzimanikatis, and Bernhard Palsson. A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Mol Syst Biol*, 3:121, 2007. doi: 10.1038/msb4100155. URL `http://dx.doi.org/10.1038/msb4100155`.

[6] Julien Gagneur and Steffen Klamt. Computation of elementary modes: a unifying framework an the new binary approach. *BMC Bioinformatics*, 5:175, 2004.

[7] S. Schuster, T. Dandekar, and D. A. Fell. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*, 17(2):53–60, Feb 1999.

BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–11

Using Chemical Organization Theory for Model-Checking (Supplementary Material)

Christoph Kaleta*[,1,2], Stephan Richter*[,1], and Peter Dittrich†[,1]

[1]Bio Systems Analysis Group, Jena Centre for Bioinformatics (JCB) and Department of Mathematics and Computer Science, and [2]Department of Bioinformatics, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

1 EMULATING FLUX-BASED NETWORK ANALYSIS METHODS

Flux-based methods allow to predict whether a reaction can appear in any steady state. These methods have additional requirements on the network, like definition of upper bounds of inflow reactions for FBA [Varma and Palsson, 1994] or network size in elementary mode analysis [Schuster et al., 1999] and extreme pathway analysis [Schilling et al., 2000]. Since we only want to test whether a reaction can appear in a steady-state flux a simplified approach that yields the same results like elementary mode analysis as well as extreme pathway analysis can be used. This approach is similar to FBA and allows to determine whether a reaction can appear in any steady state of the model using linear programming.

Given the stoichiometric matrix $\mathbf{N}$ of a model of n reactions for which we want to determine whether reaction i can appear in a steady-state flux, the constraints of the linear program are

1. $\mathbf{N} \cdot \mathbf{v} = 0$
2. $\mathbf{v} \geq 0$
3. $v_i \geq 1$

with $\mathbf{v}$ being the vector of variables. A dummy objective function can be used if required. Since we have not defined any upper bound, constraint 3 is equal to searching a steady-state flux $\mathbf{v}$ with a positive flux in v_i. Since we are also interested in states where some species can accumulate over time, the steady-state condition is relaxed to a constraint similar to the self-maintenance condition in chemical organization theory (OT). Thus, when checking if reaction i can be part of a flux of the reactions such that the concentration of no molecule declines, the constraints read

1. $\mathbf{N} \cdot \mathbf{v} \geq 0$
2. $\mathbf{v} \geq 0$
3. $v_i \geq 1$

If the linear program is feasible, flux-based methods would predict i to be present in a flux of the system where the concentration of no molecule declines.

2 EXTENDING THE STEADY-STATE LEMMA TO GROWTH STATES

In Dittrich and Speroni di Fenizio [2007] it is shown that each steady state of a reaction network can be mapped to an organization of the system, if the reaction network obeys the condition that the kinetic law of a reaction implicates a non-zero flux if and only if all educts have a positive concentration [Feinberg and Horn, 1974]. Additionally we assume that during simulation no species has a negative concentration and each flux is positive, i.e., reversible reactions are split into irreversible forward and backward reactions. Here we will demonstrate that, if the network is simulated using ordinary differential equations, the steady-state lemma can be extended to every phase of the simulation in which there is a non-negative concentration change for each species.

Given two points $\mathbf{x}(t_1)$ and $\mathbf{x}(t_2)$ with $0 < t_1 < t_2$ and $\mathbf{x}(t_1), \mathbf{x}(t_2) \in \mathbb{R}^n$ in the trajectory $\mathbf{x}(t)$ of the concentration of the species of a reaction network during simulation, we call the time span $[t_1, t_2]$ a *growth phase* if $\mathbf{x}(t_1) \geq \mathbf{x}(t_2)$. As the species set s_t present at a time-point t we identify each species having a positive concentration, i.e., $s_t = \{i \mid x_i(t) > 0\}$. If the kinetic laws of the network fulfill the Feinberg conditions s_t necessarily fulfills the closure-condition if $t > 0$. If we determine s_t at a time-point t_g during a growth phase, $t_g \in [t_1, t_2]$, there additionally exists a flux vector d fulfilling the self-maintenance condition for s_{t_g}) and hence s_{t_g} constitutes an organization. The existence of d will be demonstrated in the following. In passing we note that a non-negative concentration change for each species in the interval $[t_1, t_2]$ implies that $s_{t_g} = s_{t_1} = s_{t_2}$.

Given the stoichiometric matrix $\mathbf{N}$ of a reaction network and the kinetic laws of the reactions as $\mathbf{v}(t)$, the ordinary differential equation

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{N}\mathbf{v}(t) \tag{1}$$

describes the trajectory of the concentrations of the species $\mathbf{x}(t)$ given the starting point $\mathbf{x}(0)$. Usually $\mathbf{v}(t)$ is written as being dependent on $\mathbf{x}(t)$. Here we assume that we have already solved the differential equation in (1) since the following proof relies only on the existence of the solution $\mathbf{v}(t)$. Thus, we can compute the concentration change of the species of the network in the interval $[t_1, t_2]$ as

$$\mathbf{x}(t_2) - \mathbf{x}(t_1) = \int_{t_1}^{t_2} \frac{d\mathbf{x}(t)}{dt} dt = \int_{t_1}^{t_2} \mathbf{N}\mathbf{v}(t) dt = \mathbf{N} \int_{t_1}^{t_2} \mathbf{v}(t) dt \tag{2}$$

*Authors contributed equally

†to whom correspondence should be addressed (dittrich@minet.uni-jena.de)

Now, if we assume that the system is in a growth phase during t_1 and t_2 we can choose $\mathbf{d}$ as

$$\mathbf{d} = \int_{t_1}^{t_2} \mathbf{v}(t)dt \tag{3}$$

Since we assume $\mathbf{v}(t)$ to be non-negative, $\mathbf{d}$ has only non-negative entries. From the condition, that every reaction has a non-zero flux, if and only if all its substrates are present, i.e., have a positive concentration, we can derive that $\mathbf{d}$ has positive values for each reaction implied by the species set s_{t_g}. Additionally we can see from (2) that $\mathbf{Nd} \geq 0$, hence, $\mathbf{d}$ fulfills the self-maintenance condition for s_{t_g}. This implies, that s_{t_g} is an organization.

In consequence, each growth phase of the simulation of a reaction network corresponds to an organization of the system. Moreover, if there exists a *growth state* of the network, i.e., $t_2 \to \infty$ this state also corresponds to an organization. Please note that a growth state contains the steady-state condition, i.e., $\mathbf{Nv} = 0$ and $\mathbf{v} \geq 0$, as special case. Hence, the steady-state lemma can be generalized to a growth-state lemma.

3 DETAILED DESCRIPTION OF THE SBML PROCESSING ALGORITHM

3.1 Libraries used

We implemented our analysis tool in Java and used the JigCell SBML parser, available under the DARPA BioCOMP Open Source License on `http://jigcell.biol.vt.edu`. This code is used to open, modify, and save the analyzed SBML models.

3.2 Overview on the processing steps

In order to perform the OT analysis, the model is passed through several analysis and adjustment steps:

1. reading and testing the SBML code
2. searching for defined meta-ids
3. building a look-up table for used functions
4. building a look-up table for predefined parameters

For each reaction in a model, we perform:

5. analysis of the structure of the kinetic laws
6. adaptation of the reaction structure to fulfill the Feinberg condition
7. adaptation of the kinetic laws to preserve the dynamics

3.3 Description of the steps

In the subsequent sections we will use the following terms:

- *support*: a modifier set is supporting a reaction if a non-zero concentration of the modifiers allows a non-zero flux of the reaction
- *absent*, *absence*, *deletion*: a modifier's concentration is set to zero

3.3.1 Reading and testing the SBML code Prior to all analysis steps the models have to be read in. This is done using the JigCell SBML parser, which also checks the syntactic structure while loading the document. As a result of this syntax checking, we found *MODEL8262229752* to contain syntactical errors.

3.3.2 Searching for defined meta-ids All reaction and species entities in a model have a unique id. For the later creation of new reactions it is necessary to create new ids, which is only possible, if we know the existing ids in the model. New ids are given names like `metaid_xxxxxxx`, where `xxxxxxx` stands for the first free number including leading zeros.

3.3.3 Building function and parameter look-up tables For several subsequent steps, we need a table of all used function names, their respective parameters and the assigned function. Therefore, a data structure mapping each function name to this information is created by analyzing the model's function definitions. In a similar way, all kinetic parameters defined in the SBML model are stored in a mapping structure, which allows a replacement of parameter occurrences by their values when resolving kinetic laws.

3.3.4 Analysis of the structure of the kinetic laws The most important and complex step in the process is the examination of the kinetic law of each reaction. For this purpose, every rate law in the SBML document is parsed into a tree structure using the JigCell library. The tree structure of reaction 3 of the example network is shown in Figure 1. The main goal of this step is to gather the set of

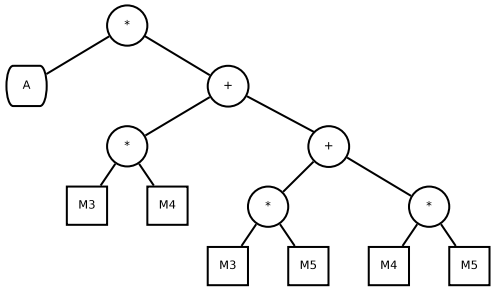


Fig. 1. The tree corresponding to the rate law of reaction 3 from the example network given in the main paper. Circles indicate operator nodes, round boxes denote species and squared boxes correspond to parameters.

all modifiers involved in a reaction and, moreover, to determine the set of supporting sets. This is achieved by checking which modifiers are omittable, i.e., can be absent without reducing the flux of the reaction to zero. Therefore, a list of all modifiers is obtained from the reaction definition. Out of this set, the power set of all involved modifiers is calculated and passed to a data structure, which we will refer to as *untested modifiers* in the following. Please note that if a set of modifiers is replaced by zero values, this means the concentrations of the modifiers in the complement set are left positive. Therefore, we check the support of a single modifier by testing the effect of deleting all other modifiers. The following steps are performed iteratively over all sets in the untested set:
The largest untested set is obtained from the data structure. The

modifier species contained within this set are assumed to be absent. Thus, all their occurrences are replaced by zero values in the kinetic law. Literally spoken, we test support of the smallest set by deleting the largest. In our example, the first untested set would be $\{M3, M4, M5\}$. Hence, these modifiers are replaced by zero values, as can be seen in Figure 2. Then, beginning from the leaves,

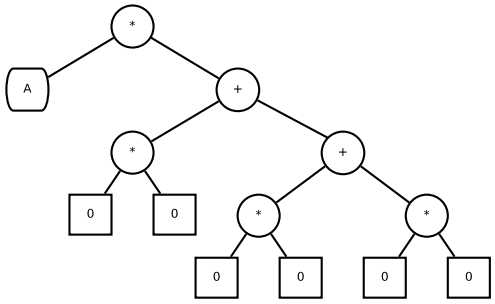


Fig. 2. The same tree after application of zero values to the (absent) modifiers.

the tree structure (related to the current kinetic law) is resolved by application of mathematical rules. Simultaneously, all occurrences of parameters are replaced by their values from the look-up table. In the example, the right subtree of the root node will be solved to zero, while the left leaf node will be replaced by the value of A, if given, or stay a symbol otherwise (Figure 3). As one can

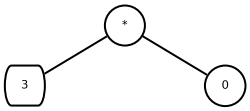


Fig. 3. The right subtree solved to zero, the left subtree was replaced by A's value, say 3

see, if all modifiers are absent, this reaction will have a zero flux. Consequently, the set $\{M3, M4, M5\}$ is marked as not omittable and removed from the untested sets. Due to the ordering by size, the next sets to be tested are $\{M3, M4\}$, $\{M3, M5\}$ and $\{M4, M5\}$. In each case we find that the flux in the reaction is constrained to zero if the modifiers are deleted (Figure 4). Informally spoken, this means: if you delete any pair of modifiers in $R3$, the reaction will stop, or, no single modifier supports the reaction. Therefore those sets are marked as not omittable, too.

The next untested sets are $\{M3\}$, $\{M4\}$ and $\{M5\}$. As can be seen by application of a zero concentration to modifier $\{M3\}$, we no longer obtain a zero value as result, but a function depending on the concentration of $\{M4\}$ and $\{M5\}$. As we assume their concentrations to be positive, we find that reaction R3 can have a positive flux if $M3$ has a zero concentration (Figure 5). The single-modifier sets $\{M3\}$, $\{M4\}$ and $\{M5\}$ are marked as omittable and hence their complements $\{M4, M5\}$, $\{M3, M5\}$ and $\{M3, M4\}$ are the supporting sets of the reaction. Since those sets already enable a positive flux in the reaction, there is no need to additionally test whether the entire modifier set is a supporting set. In our

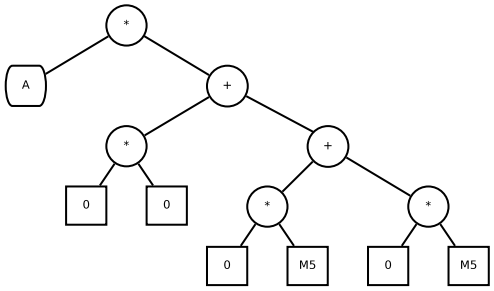


Fig. 4. The right subtree is solved to zero, if the two modifiers $M3$ and $M4$ are absent

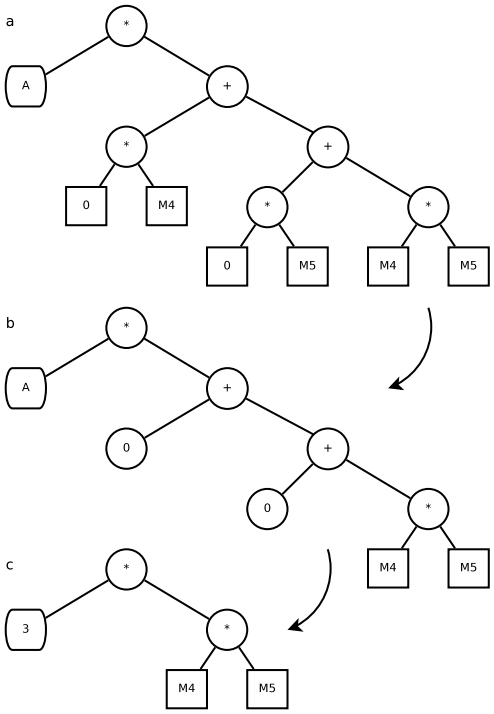


Fig. 5. In (a), we applied a zero concentration to $\{M3\}$, solved the multiplications by zero (b) and obtained the rate law $v_{R3} = 3 \cdot [M4] \cdot [M5]$ after the application of A's value 3 (c).

algorithm, this is achieved by marking all proper subsets of the sets $\{M3\}$, $\{M4\}$ and $\{M5\}$ as omittable, which accordingly means, that all supersets of $\{M4, M5\}$, $\{M3, M5\}$ and $\{M3, M4\}$ are recognized to support the reaction.

3.3.5 Adapting the reaction structure and the kinetic laws After determining the minimal supporting modifier sets, we adapt the reactions in order to correctly apply OT. Our algorithm distinguishes 3 cases:

1. None of the modifiers is omittable, i.e., the (only) supporting set is the entire modifier set. In this case all modifiers are removed from the reaction and added as educts and products to

C. Kaleta, S. Richter, P. Dittrich

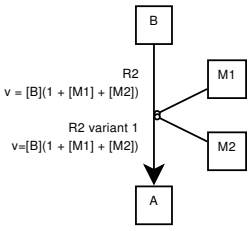


Fig. 6. Since the supporting set is the empty set, no modifier has to be moved to the educt or product side, hence also the kinetic laws are left unchanged.

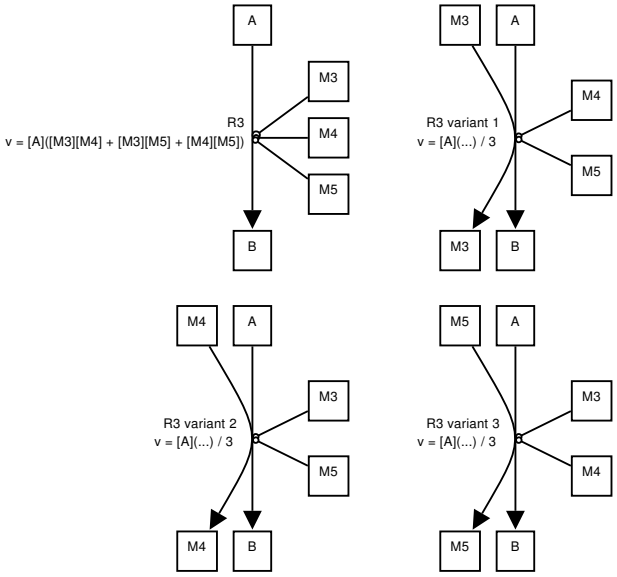


Fig. 7. We obtain three supporting modifier sets, leading to three reaction variants. Note the division by 3 in the kinetic laws.

the reaction structure. Since the number of reactions does not change along with this modification, we do not need to change the kinetic law of the reaction. But due to the alteration of the reaction structure, we rename the reaction to `<oldname> variant 0`.

2. We have exactly one supporting set, as it is the case in reaction $R2$, for example (Figure 6, here the empty set is the only supporting set). Thus, every modifier apart from this set can be set to zero without constraining the flux of the reaction to zero. Hence, we have only to move the modifiers of the single supporting set to educt and product side. Again, we do not need to modify the kinetic law and the modified reaction will be renamed to `<oldname> variant 1`.
3. If more than one supporting set is found, a new reaction for each supporting set is derived from the original one. For each derived reaction, the respective set of modifiers of the supporting set is moved to the educt and product side, while all other stay modifiers. While the original reaction is removed, the new reactions get names like `<oldname> variant 1...<oldname> variant <number of variants>`. Since we change the number of reactions in this case, we have to adapt their kinetic laws by dividing them by the number of derived reactions. For reaction $R3$ we obtain the three reactions depicted in Figure 7.

4 ABBREVIATIONS

Table 1. Abbreviations.

Abbreviation	Protein
APC	adenomatous polyposis coli
Dsh	dishevelled
ERK	extracellular signal related-kinase
GSK-3β	glycogen synthase kinase 3β
RKIP	Raf-1 kinase inhibitor protein
TCF	T-cell factor.

DETAILED LIST OF NETWORKS OF THE BIOMODELS DATABASE

In the following details about the analysis of all models of the BioModels Database are given. Models in which inconsistencies have been identified are shaded in light gray. The first three rows give general details about the models. Numbers in brackets indicate the number of reactions of the original network that can increase through the processing of the kinetic laws. The number of species remains constant. The third column gives the number of reactive organization in the modified and, in brackets, of the original network. In the forth and fifth column species and reactions that can persist in a long-term simulation of the processed network are given. OT denotes the prediction by chemical organization theory and FBM the predictions by flux-based methods. In some cases FBM identifies more reactions to be present in a long-term simulation than OT. Those cases are shaded in dark-gray. The sixth and seventh column give the same numbers when inflow reactions for species with an event setting their concentration to a positive value at a certain time-point are added. In cases where the original network already contained all species, those numbers are omitted. In cases marked with $^{(*)}$ the heuristical approach for organization computation [Centler et al., 2008] had to be applied. Since we searched only for organizations containing the complete system, computations were aborted as soon as such an organizations was found.

Curated branch

Model	Species/ Reactions	Organizations	First Step		Second Step	
			OT (species/reactions)	FBM (reactions)	OT (species/reactions)	FBM (reactions)
BIOMD0000000001	12/34(34)	2(2)	12/34	34	-	-
BIOMD0000000002	13/34(34)	3(3)	13/34	34	-	-
BIOMD0000000003	3/7(7)	1(1)	3/7	7	-	-
BIOMD0000000004	5/7(7)	4(4)	5/7	7	-	-
BIOMD0000000005	9/15(15)	2(2)	9/15	15	-	-
BIOMD0000000006	3/5(5)	1(1)	3/5	5	-	-
BIOMD0000000007	22/35(35)	2(1)	22/35	35	-	-
BIOMD0000000008	5/13(13)	1(1)	5/13	13	-	-
BIOMD0000000009	22/30(30)	8(8)	22/30	30	-	-
BIOMD0000000010	8/10(10)	4(8)	8/10	10	-	-
BIOMD0000000011	22/30(30)	8(8)	22/30	30	-	-
BIOMD0000000012	6/12(12)	1(1)	6/12	12	-	-
BIOMD0000000013	27/48(48)	4(4)	27/48	48	-	-
BIOMD0000000014$^{(*)}$	86/300(300)	53(72)	86/300	300	-	-
BIOMD0000000015	18/41(41)	1(1)	18/41	41	-	-
BIOMD0000000016	7/12(12)	1(1)	7/12	12	-	-
BIOMD0000000017	19/28(28)	3(3)	19/28	28	-	-
BIOMD0000000018	33/73(73)	1(1)	33/73	73	-	-
BIOMD0000000019$^{(*)}$	100/256(256)	1036(1036)	100/256	256	-	-
BIOMD0000000021	10/26(26)	1(1)	10/26	26	-	-
BIOMD0000000022	13/40(40)	1(1)	13/40	40	-	-
BIOMD0000000023	13/38(38)	1(1)	13/38	38	-	-
BIOMD0000000024	3/6(6)	1(1)	3/6	6	-	-
BIOMD0000000025	4/10(10)	2(1)	4/10	10	-	-
BIOMD0000000026	11/16(16)	3(3)	11/16	16	-	-
BIOMD0000000027	5/8(4)	2(2)	5/8	8	-	-
BIOMD0000000028	16/27(27)	3(3)	16/27	27	-	-
BIOMD0000000029	6/11(7)	2(2)	6/11	11	-	-
BIOMD0000000030	18/32(32)	3(3)	18/32	32	-	-
BIOMD0000000031	6/10(4)	2(2)	6/10	10	-	-
BIOMD0000000032	37/96(96)	390(260)	37/96	96	-	-
BIOMD0000000033	32/56(48)	4976(8192)	32/56	56	-	-
BIOMD0000000034	9/22(22)	1(1)	9/22	22	-	-
BIOMD0000000035	10/18(18)	4(4)	10/18	18	-	-
BIOMD0000000036	3/7(7)	1(1)	3/7	7	-	-
BIOMD0000000037	12/14(14)	1(2)	3/6	10	12/14	14
BIOMD0000000038	17/28(28)	12(12)	17/28	28	-	-
BIOMD0000000039	5/7(7)	3(3)	5/7	7	-	-
BIOMD0000000040	5/9(9)	2(2)	5/9	9	-	-
BIOMD0000000041	10/17(17)	12(12)	9/17	17	9/17	17
BIOMD0000000042	15/34(34)	1(1)	15/34	34	-	-
BIOMD0000000043	5/7(7)	3(4)	5/7	7	-	-
BIOMD0000000044	7/8(8)	2(2)	3/4	5	6/7	7
BIOMD0000000045	4/8(8)	2(2)	4/8	8	-	-
BIOMD0000000046	16/23(23)	2(2)	16/23	23	-	-

Model	Species/ Reactions	Organizations	First Step		Second Step	
			OT (species/reactions)	FBM (reactions)	OT (species/reactions)	FBM (reactions)
BIOMD0000000047	2/3(3)	2(2)	2/3	3	-	-
BIOMD0000000048	23/47(47)	14(14)	23/47	47	-	-
BIOMD0000000049(*)	99/248(226)	3439(4513)	99/248	248	-	-
BIOMD0000000050	14/16(16)	1(1)	0/0	0	14/16	16
BIOMD0000000051	18/96(96)	1(1)	18/96	96	-	-
BIOMD0000000052	11/11(11)	1(1)	0/0	2	11/11	11
BIOMD0000000053	6/12(12)	1(1)	0/0	8	6/12	12
BIOMD0000000054	3/5(5)	1(1)	3/5	5	-	-
BIOMD0000000055	13/32(32)	1(1)	13/32	32	-	-
BIOMD0000000056	54/100(94)	392(16)	54/100	100	-	-
BIOMD0000000057	6/10(10)	2(2)	6/10	10	-	-
BIOMD0000000058	4/11(11)	1(1)	4/11	11	-	-
BIOMD0000000059	6/18(18)	2(1)	6/18	18	-	-
BIOMD0000000060	4/6(6)	2(2)	4/6	6	-	-
BIOMD0000000061	25/54(54)	8(8)	25/54	54	-	-
BIOMD0000000062	3/10(10)	1(1)	3/10	10	-	-
BIOMD0000000063	9/24(24)	1(1)	9/24	24	-	-
BIOMD0000000064	26/48(42)	16(2)	26/48	48	-	-
BIOMD0000000065	9/18(18)	1(1)	9/18	18	-	-
BIOMD0000000066	11/14(14)	13(13)	11/14	14	-	-
BIOMD0000000067	8/18(18)	1(1)	8/18	18	-	-
BIOMD0000000068	9/20(16)	1(1)	9/20	20	-	-
BIOMD0000000069	10/12(12)	8(8)	9/11	11	10/12	12
BIOMD0000000070	45/86(86)	168(168)	45/86	86	-	-
BIOMD0000000071	17/34(34)	3(3)	17/34	34	-	-
BIOMD0000000072	7/8(8)	1(2)	1/2	7	7/8	8
BIOMD0000000073	16/52(52)	1(1)	16/52	52	-	-
BIOMD0000000074	19/62(62)	1(1)	19/62	62	-	-
BIOMD0000000075	12/22(22)	8(8)	12/22	22	-	-
BIOMD0000000076	3/8(8)	1(1)	3/8	8	-	-
BIOMD0000000077	8/8(8)	3(3)	8/8	8	-	-
BIOMD0000000078	16/52(52)	1(1)	16/52	52	-	-
BIOMD0000000079	3/6(6)	1(1)	3/6	6	-	-
BIOMD0000000080	10/10(10)	3(3)	7/6	8	10/10	10
BIOMD0000000081	23/32(32)	18(18)	23/32	32	-	-
BIOMD0000000082	10/10(10)	3(3)	7/6	8	10/10	10
BIOMD0000000083	19/62(62)	1(1)	19/62	62	-	-
BIOMD0000000084	8/16(16)	16(16)	8/16	16	-	-
BIOMD0000000085	17/34(34)	10(10)	17/34	34	-	-
BIOMD0000000086	17/48(48)	12(12)	17/48	48	-	-
BIOMD0000000087	55/45(45)	1952(1952)	29/29	32	55/45	45
BIOMD0000000088	105/182(178)	936(1584)	68/122	167	105/182	182
BIOMD0000000089	16/41(41)	1(1)	16/41	41	-	-
BIOMD0000000090	26/47(47)	1(1)	26/47	47	-	-
BIOMD0000000091	16/25(25)	8(8)	16/25	25	-	-
BIOMD0000000092	4/6(6)	2(2)	4/6	6	-	-
BIOMD0000000093	34/48(48)	5(3)	11/16	30	31/43	43
BIOMD0000000094	34/47(47)	2(3)	5/5	27	24/24	40
BIOMD0000000095	19/46(46)	1(1)	19/46	46	-	-
BIOMD0000000096	19/46(46)	1(1)	19/46	46	-	-
BIOMD0000000097	19/46(46)	1(1)	19/46	46	-	-
BIOMD0000000098	2/6(6)	1(1)	2/6	6	-	-
BIOMD0000000099	7/14(14)	1(1)	7/14	14	-	-
BIOMD0000000100	5/12(12)	1(1)	5/12	12	-	-
BIOMD0000000101	6/13(13)	1(1)	6/13	13	-	-
BIOMD0000000102	13/37(37)	1(1)	13/37	37	-	-
BIOMD0000000103	17/61(61)	1(1)	17/61	61	-	-
BIOMD0000000104	6/2(2)	1(1)	1/0	0	6/2	2
BIOMD0000000105	39/102(102)	3(3)	13/16	49	39/102	102
BIOMD0000000106	25/44(44)	26(1)	14/18	40	25/44	44
BIOMD0000000107	14/23(23)	2(1)	14/23	23	-	-

Model	Species/ Reactions	Organizations	First Step		Second Step	
			OT (species/reactions)	FBM (reactions)	OT (species/reactions)	FBM (reactions)
BIOMD0000000108	9/20(18)	1(1)	9/20	20	-	-
BIOMD0000000109	61/138(138)	269(28)	48/91	128	61/138	138
BIOMD0000000110	15/30(30)	1(1)	15/30	30	-	-
BIOMD0000000111	10/20(19)	2(1)	10/20	20	-	-
BIOMD0000000112	10/12(12)	4(3)	4/4	11	10/12	12
BIOMD0000000113	4/8(8)	1(1)	4/8	8	-	-
BIOMD0000000114	2/5(5)	1(1)	2/5	5	-	-
BIOMD0000000115	2/5(5)	1(1)	2/5	5	-	-
BIOMD0000000116	6/10(10)	1(1)	6/10	10	-	-
BIOMD0000000117	2/6(6)	1(1)	2/6	6	-	-
BIOMD0000000119	1/1(1)	1(1)	1/1	1	-	-
BIOMD0000000120	5/10(10)	1(1)	3/5	9	5/10	10
BIOMD0000000121	6/10(10)	2(2)	6/10	10	-	-
BIOMD0000000122	14/38(38)	6(6)	14/38	38	-	-
BIOMD0000000123	14/34(34)	14(14)	14/34	34	-	-
BIOMD0000000124	2/2(2)	1(1)	2/2	2	-	-
BIOMD0000000125	5/7(7)	4(1)	5/7	7	-	-
BIOMD0000000126	9/22(22)	2(2)	9/22	22	-	-
BIOMD0000000128	3/3(3)	1(1)	3/3	3	-	-
BIOMD0000000137	21/32(32)	24(24)	21/32	32	-	-
BIOMD0000000138	1/1(1)	1(1)	1/1	1	-	-
BIOMD0000000139	24/64(64)	2(2)	17/39	63	24/64	64
BIOMD0000000140	24/64(64)	2(2)	17/39	63	24/64	64
BIOMD0000000143	20/20(20)	1(1)	4/4	4	7/5	5
BIOMD0000000144	18/56(56)	2(1)	17/54	55	18/56	56
BIOMD0000000145	7/12(12)	2(1)	7/12	12	-	-
BIOMD0000000146	36/54(47)	29(336)	36/54	54	-	-
BIOMD0000000147	24/70(70)	2(2)	17/45	69	24/70	70
BIOMD0000000148	7/16(12)	4(1)	7/16	16	-	-
BIOMD0000000149	28/39(39)	150(384)	28/39	39	-	-
BIOMD0000000150	4/4(4)	2(2)	4/4	4	-	-
BIOMD0000000151	68/114(114)	80(96)	49/71	111	19/9	112
BIOMD0000000152	64/122(122)	143(143)	61/116	120	64/122	122
BIOMD0000000153	75/154(154)	147(147)	72/148	152	75/154	154
BIOMD0000000154	2/5(5)	1(1)	2/5	5	-	-
BIOMD0000000155	2/4(4)	1(1)	2/4	4	-	-
BIOMD0000000156	3/6(6)	1(1)	3/6	6	-	-
BIOMD0000000157	3/7(7)	1(1)	3/7	7	-	-
BIOMD0000000158	3/7(7)	1(1)	3/7	7	-	-
BIOMD0000000159	3/7(7)	1(1)	3/7	7	-	-
BIOMD0000000160	25/43(43)	32(1)	25/43	43	-	-
BIOMD0000000161	46/92(92)	4160(2112)	46/92	92	-	-
BIOMD0000000162	32/106(106)	1(1)	32/106	106	-	-
BIOMD0000000163	16/26(26)	4(4)	16/26	26	-	-
BIOMD0000000164	26/58(58)	24(12)	26/58	58	-	-
BIOMD0000000165	37/62(62)	624(624)	37/62	62	-	-
BIOMD0000000166	3/18(18)	1(1)	3/18	18	-	-
BIOMD0000000167	9/16(16)	4(2)	9/16	16	-	-
BIOMD0000000168	7/11(10)	4(1)	7/11	11	-	-
BIOMD0000000169	11/29(27)	1(1)	11/29	29	-	-
BIOMD0000000170	7/17(17)	1(1)	7/17	17	-	-
BIOMD0000000171	12/31(31)	1(1)	12/31	31	-	-
BIOMD0000000172	25/47(47)	9(9)	25/47	47	-	-
BIOMD0000000173	26/48(48)	16(16)	25/47	47	26/48	48
BIOMD0000000174	4/10(10)	2(1)	4/10	10	-	-
BIOMD0000000175	120/214(198)	319248(319248)	86/128	208	120/214	214
BIOMD0000000176	25/48(48)	9(9)	25/48	48	-	-
BIOMD0000000177	28/55(55)	9(9)	28/55	55	-	-
BIOMD0000000178	6/6(6)	1(1)	1/2	2	6/6	6
BIOMD0000000179	7/17(17)	1(1)	7/17	17	-	-
BIOMD0000000180	8/23(23)	1(1)	8/23	23	-	-

Model	Species/ Reactions	Organizations	First Step		Second Step	
			OT (species/reactions)	FBM (reactions)	OT (species/reactions)	FBM (reactions)
BIOMD0000000181	6/18(18)	1(1)	6/18	18	-	-
BIOMD0000000182	37/64(64)	1920(1920)	37/64	64	-	-
BIOMD0000000183	67/352(352)	16(16)	67/352	352	-	-
BIOMD0000000184	3/7(7)	1(1)	3/7	7	-	-
BIOMD0000000185	8/20(20)	1(1)	8/20	20	-	-

Non-Curated branch

The non-curated branch of the BioModels Database contains 74 models. Of these, we could validate 55 with the same methods as for the curated branch. For 18 networks we were not able to compute the entire set of organizations. This was caused by too large a set of organizations in 9 cases ($> 10^6$ organizations) in which we had to abort computation due to constraints in memory. However, in these cases we already found an organization encompassing the whole species set. The remaining 9 networks were too large for a detailed analysis since they contained more than 500 reactions. Centler et al. [2008] analyzed such a network by applying a heuristic whose results need to be carefully checked. This is due to the nature of the heuristic algorithm which only approximates the set of organizations. Hence, some organizations might not be found if the heuristic is aborted too early. Since these models are contained in the non-curated branch of the BioModels Database, which is not the central focus of this work, and error-checking is time-consuming we applied only flux-based methods in these cases. Finally, we could not open *MODEL8262229752* with the JigCell SBML parser. The parser indicated that the model contains a syntactic error. The cases where we could compute chemical organizations are listed next. Computations were performed as described in the previous section.

Model	Species/ Reactions	Organizations	First Step		Second Step	
			OT (species/reactions)	FBM (reactions)	OT (species/reactions)	FBM (reactions)
MODEL0212154960	5/10(10)	1(1)	0/0	6	3/5	7
MODEL0995500644	12/13(13)	1(1)	11/12	12	12/13	13
MODEL1502077979	7/13(13)	1(1)	7/13	13	-	-
MODEL2463576061	330/222(222)	48(48)	33/14	22	33/14	22
MODEL2463683119	680/470(470)	256(256)	71/26	40	71/26	40
MODEL2504064544	19/52(51)	1(1)	19/52	52	-	-
MODEL4665428627	61/146(146)	6(6)	9/12	96	61/146	146
MODEL4734733125	14/24(24)	4(4)	14/24	24	-	-
MODEL4779732381	14/17(17)	4(4)	11/14	16	14/17	17
MODEL4780441670	8/11(11)	2(2)	5/8	10	8/11	11
MODEL4780784080	14/24(24)	2(2)	10/21	23	14/24	24
MODEL4816599063	12/18(18)	4(4)	12/18	18	-	-
MODEL4821294342	12/26(26)	2(2)	12/26	26	-	-
MODEL4968912141	8/10(10)	3(3)	8/10	10	-	-
MODEL4969417017	18/28(28)	11(11)	15/22	26	18/28	28
MODEL5073396359	22/70(70)	2(2)	7/16	59	9/17	60
MODEL5662324959	628/2212(2212)	1(1)	628/2212	2212	-	-
MODEL5662377562	628/2212(2212)	1(1)	628/2212	2212	-	-
MODEL5662398146	628/2212(2212)	1(1)	628/2212	2212	-	-
MODEL5662425708	628/2212(2212)	1(1)	628/2212	2212	-	-
MODEL5974712823	10/6(6)	2(2)	8/5	5	8/6	6
MODEL6623617994	22/36(36)	22(22)	22/36	36	-	-
MODEL6623628741	10/8(8)	2(2)	10/8	8	-	-
MODEL6624091635	34/80(80)	14(14)	34/80	80	-	-
MODEL6624199343	5/10(10)	2(2)	5/10	10	-	-
MODEL6762427183	0/0(0)	1(1)	0/0	0	-	-
MODEL8568434338	225/219(219)	1(1)	28/0	0	28/0	0
MODEL8583955822	12/30(30)	1(1)	12/30	30	-	-
MODEL8584137422	12/30(30)	1(1)	12/30	30	-	-
MODEL8584292730	13/38(38)	1(1)	13/38	38	-	-
MODEL8584468482	13/36(36)	1(1)	13/36	36	-	-
MODEL8938094216	15/18(18)	19(24)	15/18	18	-	-
MODEL9070467164	94/179(179)	4656(4656)	94/179	179	-	-
MODEL9071122126	64/116(116)	280(280)	64/116	116	-	-
MODEL9071773985	73/147(147)	380(380)	73/147	147	-	-
MODEL9077438479	29/48(48)	78(90)	29/48	48	-	-
MODEL9079179924	81/146(146)	4512(4512)	81/146	146	-	-
MODEL9079740062	29/48(48)	78(90)	29/48	48	-	-
MODEL9080388197	15/26(26)	2(2)	15/26	26	-	-
MODEL9080747936	50/90(90)	288(288)	50/90	90	-	-
MODEL9081220742	188/350(350)	$> 10^6 (> 10^6)$	188/350	350	-	-
MODEL9085850385	59/104(104)	640(640)	59/104	104	-	-
MODEL9086207764	284/580(580)	$> 10^6 (> 10^6)$	284/580	580	-	-
MODEL9086518048	286/594(594)	$> 10^6 (> 10^6)$	286/594	594	-	-
MODEL9086628127	16/32(32)	3(3)	16/32	32	-	-
MODEL9086926384	85/156(156)	1112(1112)	85/156	156	-	-
MODEL9086953089	114/206(206)	43245(46416)	114/206	206	-	-

Model	Species/ Reactions	Organizations	First Step		Second Step	
			OT (species/reactions)	FBM (reactions)	OT (species/reactions)	FBM (reactions)
MODEL9087255381	289/602(602)	$> 10^6 (> 10^6)$	289/602	602	-	-
MODEL9087474843	290/602(602)	$> 10^6 (> 10^6)$	290/602	602	-	-
MODEL9087766308	5/4(4)	2(2)	5/4	4	-	-
MODEL9087988095	5/4(4)	2(2)	5/4	4	-	-
MODEL9088169066	5/4(4)	2(2)	5/4	4	-	-
MODEL9088294310	5/4(4)	2(2)	5/4	4	-	-
MODEL9089491423	196/364(364)	$> 10^6 (> 10^6)$	196/364	364	-	-
MODEL9089538076	200/374(374)	$> 10^6 (> 10^6)$	200/374	374	-	-
MODEL9089914876	192/358(358)	$> 10^6 (> 10^6)$	192/358	358	-	-
MODEL9147091146	77/142(142)	752(752)	77/142	142	-	-
MODEL9147232940	64/112(112)	996(996)	64/112	112	-	-
MODEL9147975215	37/49(49)	72(72)	35/46	48	37/49	49
MODEL9200487367	5/9(9)	2(2)	5/9	9	-	-
MODEL9852292468	73/66(66)	1(1)	0/0	0	61/52	56

For the remaining 9 cases we adapted a different approach by just checking whether the entire species set is an organization. If we did not find this set to be an organization, we added, similar to the original approach, an inflow for every species whose concentration is set to a non-zero value at a certain time-point. In contrast, we were able to test whether each reaction could be present in a steady-state or growth state flux using the emulation method for flux-based methods described in Section 1.

Model	Species/ Reactions	Organizations	First Step		Second Step	
			OT (species/reactions)	FBM (reactions)	OT (species/reactions)	FBM (reactions)
MODEL0403888565	377/805(805)	n.a.	n.a.	445	377/805	805
MODEL0403928902	377/805(805)	n.a.	n.a.	445	377/805	805
MODEL0403954746	377/805(805)	n.a.	n.a.	445	377/805	805
MODEL0403988150	377/806(806)	n.a.	n.a.	446	377/805	805
MODEL0404023805	377/806(806)	n.a.	n.a.	446	377/805	805
MODEL2021729243	2715/4370(4370)	n.a.	n.a.	3733	n.a.	3733
MODEL2021747594	2715/4370(4370)	n.a.	n.a.	3592	n.a.	3592
MODEL3023609334	1972/3842(3842)	n.a.	n.a.	3752	n.a.	3752
MODEL3023641273	1972/3842(3842)	n.a.	n.a.	3752	n.a.	3752
MODEL4132046015	408/534(534)	n.a.	n.a.	32	n.a.	32

Discussion

Next, we will briefly discuss the models of the non-curated branch of the BioModels Database in which we detected inconsistencies. Overall we found 11 models containing inconsistencies. However, a first examination reveals that in 6 of these models, *MODEL2021729243*, *MODEL2021747594*, *MODEL2463576061*, *MODEL2463683119*, *MODEL4132046015* and *MODEL8568434338*, neither an input nor an initial concentration for any species is given. Thus, it can be assumed that these models have been constructed for the purpose of a structural and not a dynamic analysis.

Another two models, *MODEL3023609334* and *MODEL3023641273*, represent a genome-scale reconstruction of the metabolism of *E. coli*. Since the metabolism of *E. coli* is not yet entirely understood, it contains species which are only consumed and thus also the reactions using them as educts cannot have a positive flux. Most of the reactions and species that cannot be present in the limit behavior are due to this missing knowledge. Additionally, some species and reactions are not present because they belong to uptake and utilization pathways for metabolites that are not contained in the growth media.

The remaining models are *MODEL0212154960*, modeling the vectorial transport of bromosulfophthalein over epithelial cells [Bartholomé et al., 2007], *MODEL9852292468*, modeling lipid-mediated thrombin generation [Bungay et al., 2003], and *MODEL5073396359*, modeling apoptosome-dependent caspase activation [Rehm et al., 2006]. Comparing *MODEL0212154960* to the model presented in Bartholomé et al. [2007], we found that a reaction was missing in the SBML model, while it was described in the supplementary material of the publication. Adding this reaction, we found all species present in an organization. During simulation, the two species absent from any organization had indeed a zero concentration. In *MODEL9852292468* we found reactions that had a negative flux during simulation while they were set to irreversible in the model. Relaxing the irreversibility constraint in these cases we found all species present in an organization. Finally, analyzing *MODEL5073396359* and comparing to the supplementary material of Rehm et al. [2006] we found that several species were not supplied as input in the SBML model, while the description of the model in the publication contained such an inflow. In this case, even the dynamic behavior of the SBML model did not match the behavior of the model described in the original publication. Adding an inflow or an initial concentration resolved the inconsistencies. In all three inconsistent models we found that chemical organization theory predicted the reactions that can have a non-zero flux during simulation more accurately than flux-based methods. The reasons for these differences follow the scheme outlined in the main article. Thus, flux-based methods find a steady state flux through a set of reactions encompassing species than

can not be produced at positive rate. However, due to a drain of one or several of these species by interconversion to other species or decay, they cannot persist in a long-term simulation. Since OT takes this drain into account, the species are found absent from any organization.

REFERENCES

K. Bartholomé, M. Rius, K. Letschert, D. Keller, J. Timmer, and D. Keppler. Data-based mathematical modeling of vectorial transport across double-transfected polarized cells. *Drug Metab Dispos*, 35(9):1476–1481, Sep 2007. doi: 10.1124/dmd.107.015636.

S. D. Bungay, P. A. Gentry, and R. D. Gentry. A mathematical model of lipid-mediated thrombin generation. *Math Med Biol*, 20(1):105–129, Mar 2003.

F. Centler, C. Kaleta, P. S. di Fenizio, and P. Dittrich. Computing chemical organizations in biological networks. *Bioinformatics*, 24(14):1611–1618, 2008.

P. Dittrich and P. Speroni di Fenizio. Chemical organization theory. *Bull Math Biol*, 69(4):1199–1231, 2007.

M. Feinberg and F. J. M. Horn. Dynamics of open chemical systems and the algebraic structure of the underlying reaction network. *Chem. Eng. Sci.*, 29(3):775–787, 1974.

M. Rehm, H. J. Huber, H. Dussmann, and J. H. M. Prehn. Systems analysis of effector caspase activation and its control by X-linked inhibitor of apoptosis protein. *EMBO J*, 25(18): 4338–4349, Sep 2006. doi: 10.1038/sj.emboj.7601295.

C. H. Schilling, D. Letscher, and B. Ø. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol*, 203(3):229–248, Apr 2000. doi: 10.1006/jtbi.2000.1073.

S. Schuster, T. Dandekar, and D. A. Fell. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*, 17(2):53–60, Feb 1999.

A. Varma and B. O. Palsson. Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/Technology*, 12(10):994–998, October 1994. doi: 10.1038/nbt1094-994.

Can the Whole Be Less Than the Sum of its Parts? Pathway Analysis in Genome-Scale Metabolic Networks Using Elementary Flux Patterns - Supplementary Material -

Christoph Kaleta[1,2], Luís Filipe de Figueiredo[1], and Stefan Schuster[1]

[1] Department of Bioinformatics, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany
[2] Bio Systems Analysis Group, Jena Centre for Bioinformatics (JCB) and School of Mathematics and Computer Science, Friedrich Schiller University Jena, D-07743 Jena, Germany

Corresponding Author:
Christoph Kaleta
Department of Bioinformatics, Friedrich Schiller University Jena
Ernst-Abbe-Platz 2, D-07743 Jena, Germany
Phone: +49-3641-949583
Fax: +49-3641-946452
E-Mail: Christoph.Kaleta@uni-jena.de

S1 Input-Species

The following species are provided as input to the model. They are necessary to produce biomass.

D-glucose, ammonium, nitrate, sulfate, Fe^{2+}, Fe^{3+}, CO_2, H^+, potassium, calcium, cobalt, molybdate, sodium, phosphate, oxygen, water, chloride, Cu^{2+}, Mg^{2+}, Mn^{2+} and Zn^{2+}

S2 Abbreviations

Table 1: List of abbreviated species names.

Abbreviation	**Species**
2-H3OP	2-hydroxy 3-oxopropanoate
2-PG	D-glycerate 2-phosphate
2,3-G6P	2-dehydro 3-deoxy-D-gluconate 6-phosphate
3PG	3-phospho-D-glycerate
5-Aizc	5-amino 1,5-phospho-D-ribosyl-imidazole 4-carboxylate
6-Pgc	6-phospho-D-gluconate
Ac	acetate
AcCoA	acetyl-CoA
Aicar	5-amino 1,5-Phospho-D-ribosyl-imidazole 4-carboxamide
Aics	S-2,5-amino 1,5-phospho-D-ribosyl-imidazole 4-carboxamidosuccinate
Ala	L-alanine
Arg	L-arginine
Asp	L-aspartate
Cit	citrate
CoA	coenzyme A
Cys	cysteine
DHAP	dihydroxyacetone phosphate
FDP	D-fructose 1,6-bisphosphate
Fum	fumarate
G3P	glyceraldehyde 3-phosphate
G6P	D-glucose 6-phosphate
GABA	4-Aminobutanoate
Gl	glycerol
Glc	D-glucose
Glcn	gluconate
Glu	L-glutamate
Glx	glyoxylate
Glyc	R-glycerate

Abbreviation	Species
H20	water
His	histidine
Hpyr	hydroxypyruvate
IMP	inosine monophosphate
ICit	isocitrate
Mal	L-malate
MmCoA	R-methylmalonyl-CoA
NH4	ammonium
OAA	oxaloacetate
OG	2-oxoglutarate
PEP	phosphoenolpyruvate
Prpp	5-phospho-alpha-D-ribose 1-diphosphate
Pyr	pyruvate
R5P	α-D-ribose 5-phosphate
Ru5P	D-ribulose 5-phosphate
Sl2a6o	N-succinyl 2-L-amino 6-oxoheptanedioate
SO4	sulfate
SucArg	N2-succinyl-L-arginine
Succ	succinate
SucCoa	succinyl-CoA
Suchms	O-succinyl-L-homoserine
Sucsal	succinic-semialdehyde
Thdp	2,3,4,5-tetrahydrodipicolinate
Urdglyc	ureidoglycolate
X5P	D-xylulose 5-phosphate

S3 Enzyme names

Table 2: List of abbreviated enzyme names.

Abbreviation	Species
AceB/GlcB	malate synthase
AceEF	pyruvate dehdrogenase
Acn	aconitase
AllA	ureidoglycolate hydrolase
AspA	aspartate ammonia-lyase
AspC	aspartate aminotransferase

[illegible]	[illegible]
Eda	2-keto-3-deoxy-6-phosphogluconate aldolase
Edd	6-phosphogluconate dehydratase
Eno	enolase
Fum	fumarase
GabD	succinate-semialdehyde dehydrogenase
GarK	glycerate kinase I
GarR/GlxR	tartronate semialdehyde reductase
Gcd/YliI	glucose dehydrogenase / aldose sugar dehydrogenase
Gcl	glyoxylate carboligase
Gdh	glutamate dehydrogenase
GltA	citrate synthase
GlxK	glycerate kinase II
Hyi	hydroxypyruvate isomerase
Icd	isocitrate dehydrogenase
Icl	isocitrate lyase
IdnK/GntK	D-gluconate kinase
Mdh	malate dehydrogenase
Mqo	malate dehydrogenase
Pck	phosphoenolpyruvate carboxykinase
Ppc	phosphoenolpyruvate carboxylase
Pps	phosphoenolpyruvate synthase
PurB	adenylosuccinate lyase
PurC	phosphoribosyaminoimidazole-succinocarboxamide synthetase
PuuE/GabT	4-aminobutyrate aminotransferase
Pyk	pyruvate kinase
Sdh/Frd	succinate dehydrogenase / fumarate reductase
SucAB/LpdA	a-ketoglutarate dehydrogenase / dihydrolipoamide dehydrogenase
SucCD	succinyl-CoA synthetase

S4 Runtime complexity of the computation of elementary flux patterns

For the analysis of the runtime complexity of the algorithm for the computation of elementary flux patterns it is necessary to take a closer look on the mixed-integer linear program (MILP) used for the computation of elementary flux patterns. The MILP contains twice the number of reactions as integer variables. It has been shown by Lenstra (1983) that the runtime of solving a MILP is exponential in the number of integer variables and polynomial in the number of real variables. We have to solve the MILP as many times as there are elementary flux patterns. A trivial upper boundary of the number of flux patterns is 2^k which is also an upper boundary

for the number of elementary flux patterns. Multiplying two exponential functions yields another exponential function. Hence, the upper boundary for the running time of the presented algorithm is exponential in the number of reactions in the subsystem and polynomial in the number of reactions of the entire system.

S5 Relationship between elementary modes and elementary flux patterns

Here, we want to examine the properties of a flux vector $\mathbf{v} \in \mathrm{R}^n$ fulfilling the flux pattern condition for a certain elementary flux pattern s_v in more detail. Thus, we first examine the special case in which a subsystem encompasses the entire system. Second, we show that each elementary flux pattern is part of at least one elementary mode in the entire system. Third, we demonstrate how such an elementary mode can be found. Forth, we outline a linear program that allows to determine whether an elementary mode of the subsystem is part of a steady-state flux of the entire system. We will use the same notation as in the main document. For a metabolic network with n reactions among m metabolites or species $\mathbf{M}$ defines the $m \times n$ stoichiometric matrix of which the first k columns, *i.e.,* reactions, are assumed to belong to the subsystem.

S5.1 A subsystem encompassing the entire system

If the subnetwork encompasses the entire network ($k = n$), each elementary flux pattern corresponds to an elementary mode of the network. This can be shown through the definition of a flux pattern. Given a flux $\mathbf{v}$ in such a system that obeys the condition $\mathbf{Mv} \geq 0$ and $\mathbf{v} \geq 0$ (*i.e.*, it is a steady-state flux), we identify r as the set of reactions having a non-zero flux in $\mathbf{v}$. As it can be easily seen, r fulfills the flux pattern condition (conditions 3-6 of the main document) since the subsystem encompasses the whole system. Hence, we can find a flux pattern to each steady state flux of the network. Each steady state can be written as a positive linear combination of elementary modes and each flux pattern as a set-union of elementary flux patterns. Thus, each elementary flux pattern s uniquely corresponds to the non-zero indices of an elementary mode $\mathbf{e}$ and $\mathbf{e}$ is a flux-vector fulfilling the flux pattern condition for s.

S5.2 Elementary modes associated to elementary flux patterns

Since $\mathbf{v}$ fulfills the steady-state condition, it can be written as a positive linear combination of a set of h elementary modes e_1,...,e_h. Each of these elementary modes can be assigned to one corresponding flux pattern in the subsystem s_{e_1},..., s_{e_h} by identifying those reactions of the elementary mode having a non-zero flux in the subsystem. If some elementary mode does not use any reaction of the subsystem, the corresponding flux pattern might be equal to the empty set. Since $s_{e_1} \cup \cdots \cup s_{e_h} = s_v$ and s_v fulfills the elementarity condition, at least one $s_{e^*} \in \{s_{e_1} \cdots s_{e_h}\}$ needs to be equal to s_v. Consequently, for each elementary flux pattern there is at least one elementary mode e^* in the complete system, being a solution vector of the flux pattern condition. Thus, each elementary flux pattern is part of at least one global flux corresponding to an elementary mode of the complete system. This elementary mode can be computed using an approach outlined in Section S5.3. Therefore, elementary flux patterns can also be used to study elementary modes in genome-scale metabolic networks using routes through the predefined subsystem.

S5.3 Finding elementary modes containing an elementary flux pattern

Next, we will outline how an elementary mode fulfilling the flux pattern condition for an elementary flux pattern s can be found.

As outlined in Section 2.3 we can start with a flux vector $\mathbf{v}$ that fulfills the flux pattern condition for s, *i.e.*, a vector $\mathbf{v}$ that is a feasible solution of the linear program outlined in the Appendix. It is necessary to decompose $\mathbf{v}$ into elementary modes in order to obtain an elementary mode $\mathbf{e}*$ containing exactly the reactions of s in the subsystem. Thus, we want $\mathbf{v}$ already to use as few reactions as possible and hence to be the linear combination of only few elementary modes. This can be achieved by adding an objective function to the linear program used to check whether a reaction set is a flux pattern (see the Appendix for more details).
We obtain an initial $\mathbf{v}$ by solving the linear program

$$min \sum_{i=1}^{n} v_i$$

subject to

(1) $\mathbf{M} \cdot \mathbf{v} = 0$

(2) $\mathbf{v} \geq 0$

(3) $\forall i \in s : v_i \geq 1$

(4) $\forall j \in \{1, .., k\} \setminus s : \ v_j = 0$

Hence, we find a $\mathbf{v}$ that fulfills the flux pattern condition for s and has the least overall flux, if a minimal flux of 1 for the reactions of the flux pattern is required. This approach is similar to flux-minimization as described in Holzhütter (2004). Practical experience shows that the solution of this linear program often returns an elementary mode immediately. However, this is not necessarily the case. Thus, we need to decompose $\mathbf{v}$ into the set of elementary modes of which it is a positive linear combination. This can be done by computing the elementary modes of the system just made up by the reactions having a non-zero flux in $\mathbf{v}$. Note that this approach might also find additional elementary modes that are not necessary to decompose $\mathbf{v}$. In this case we just need to identify an elementary mode $\mathbf{e}^*$ corresponding to a flux vector fulfilling the flux pattern condition for s. This flux pattern is necessarily contained within the set of elementary modes we obtain (*cf.* Supplementary Material S5.2). Doing this we can even analyze elementary modes of the complete system, *i.e.*, "genome-scale" elementary modes using reactions from the subsystem.

S5.4 Checking the feasibility of elementary modes

In order to check whether an elementary mode $\mathbf{e}$ of the subsystem is a part of a steady-state flux in the entire system we need to check if there exists a steady-state flux $\mathbf{v}$ in the complete system such that the flux ratios of $\mathbf{v}$ in the subsystem are a multiple of $\mathbf{e}$.

First, we have to derive $\mathbf{e}'$ from $\mathbf{e}$ by removing indices corresponding to exchange reactions that have been added to the subsystem for elementary mode analysis. Next, we check whether the following condition is fulfilled

$$\exists \mathbf{v} \in \mathbb{R}^n, \ c > 0 : \ \mathbf{v} \geq 0 \text{ and } \mathbf{M} \cdot \mathbf{v} = 0 \text{ and } \forall i \in \{1..k\} : \ v_i = c \cdot e'_i$$

This can be easily transformed into a linear program almost identical to the one used to determine whether a set of reactions fulfills the flux pattern condition (see the Appendix for more details).

S6 Complete MILP and proof of elementarity of the computed flux pattern

Given the stoichiometric matrix $\mathbf{M}$ of which the first k columns correspond to the subsystem and the set of previously found elementary flux patterns S, the MILP used for the computation of elementary flux patterns is

$$min \sum_{i=1}^{k} b_i$$

subject to

(LP 1) $\mathbf{M} \cdot \mathbf{v} = 0$

(MILP 1) $\forall i \in \{1,..,k\} : \; b_i \leq v_i \leq c \cdot b_i$

(MILP 2) $\forall i \in \{1,..,k\} : \; b_i - h_i \geq 0$

(MILP 3) $\forall s' \in S : \; \sum_{i \in s'} b_i + h_i \leq |s'|$

(MILP 4) $\sum_{i=1}^{k} h_i \geq 1$

(LP 2) $\mathbf{v} \geq 0$

(MILP 5) $\mathbf{b}, \mathbf{h} \in \{0,1\}^k$

with the real variables $\mathbf{v}$ and the binary variables $\mathbf{b}$ and $\mathbf{h}$.

By minimizing the number of elements of $\Theta(\mathbf{b})$ in the objective function, we achieve that we find an elementary flux pattern. If $\Theta(\mathbf{b})$ is not elementary it can be written as a union of a set of elementary flux patterns S' ($\Theta(\mathbf{b}) \notin S'$) that contain at least one element $\Theta(\mathbf{t})$ that has not been found by the previous iterations. Else $\mathbf{b}$ could not be part of a feasible solution of the MILP and constraints (MILP 2) to (MILP 4) would be violated. Furthermore, $\Theta(\mathbf{t})$ is, by definition, a proper subset of $\Theta(\mathbf{b})$ and hence $\sum_{i=1}^{k} t_i < \sum_{i=1}^{k} b_i$. Thus, the objective function is not minimal. Hence, $\Theta(\mathbf{b})$ has to be an elementary flux pattern.

S7 Relationship between flux coupling analysis and elementary flux pattern analysis

Next, we want to outline how the coupling of reactions as determined by flux coupling analysis (Burgard et al. 2004) can be derived from calculating elementary flux patterns. Given two reactions i and j, flux coupling analysis consists of determining the relationships between the fluxes through both reaction in a steady-state flux. There are three types of coupling (Burgard et al. 2004). First, i is called directionally coupled to j if a non-zero flux in i implies a non-zero flux in j, but not necessarily the reverse. That is, reaction i requires a non-zero flux in j to appear in a steady-state flux. Second, i and j are called partially coupled if a non-zero flux in i implies a non-zero flux in j and vice versa. In such a case, i is directionally coupled to j and j is directionally coupled to i. Third, i and j are called fully coupled if a non-zero flux in i implies a specific flux in j. That is, $\frac{v_i}{v_j} = const.$ for every steady-state flux $\mathbf{v} \in \mathbb{R}^n$.

In this context we can compute the coupling between two reactions by defining the reactions i and j as the subsystem. Assuming that i and j can appear in a steady-state flux, we obtain one of the following sets of flux patterns S

1. $S = \{\{i\},\{j\}\}$: in this case both reactions can have a non-zero flux independent from each other. Hence, they are not coupled.

2. $S = \{\{i\}, \{i, j\}\}$: in this case reaction j can only have a non-zero flux if reaction i has a non-zero flux. Thus, j is directionally coupled to i.

3. $S = \{\{i, j\}, \{j\}\}$: similar to 2., Here, i is directionally coupled to j.

4. $S = \{\{i, j\}\}$: in this case, j can only have a non-zero flux if reaction i has a non-zero flux and vice versa. Thus, both reactions are partially coupled.

References

Burgard AP, Nikolaev EV, Schilling CH, Maranas CD 2004. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res* **14**(2): 301–312.

Holzhütter HG 2004. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur J Biochem* **271**(14): 2905–2922.

Lenstra H 1983. Integer programming with a fixed number of variables. *Math Oper Res* **8**: 538–548.

SYSTEMATIC DETERMINATION OF GLUCONEOGENIC PATHWAYS FROM FATTY ACIDS IN HUMANS
- Supplemental Material -

Christoph Kaleta[1,2], **Luís Filipe de Figueiredo**[1], **Reinhard Guthke**[3], **Stefan Schuster**[1]
From Department of Bioinformatics, Friedrich-Schiller-Universität Jena[1], and Bio Systems Analysis Group, Friedrich-Schiller-Universität Jena[2], and Systems Biology/Bioinformatics Group, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute[3],
D-07743 Jena, Germany
Running head: Gluconeogenesis from Fatty Acids in Humans

Address correspondence to: Christoph Kaleta Department of Bioinformatics,
Friedrich-Schiller-Universität Jena Ernst-Abbe-Platz 2, D-07743 Jena, Germany, Tel: +49-3641-949583.
Fax: +49-3641-946452;
E-Mail: Christoph.Kaleta@uni-jena.de

A List of reactions

Table 1: List of reactions and alternative catalyzing enzymes. In some cases several of the mentioned proteins are necessary to catalyze the reaction (for more information see the BiGG database, `http://bigg.ucsd.edu`).

Reaction	Catalyzing enzymes
Acat1	acetyl-Coenzyme A acetyltransferase 1 (acetoacetyl Coenzyme A thiolase)
Acat2	acetyl-Coenzyme A acetyltransferase 2 (acetoacetyl Coenzyme A thiolase)
Aco1	aconitase 1, soluble
Aco2	aconitase 2, mitochondrial
Adh	alcohol dehydrogenase IV alcohol dehydrogenase V alcohol dehydrogenase 1A (class I), alpha polypeptide alcohol dehydrogenase IB (class I), beta polypeptide alcohol dehydrogenase 1C (class I), gamma polypeptide alcohol dehydrogenase 6 (class V) alcohol dehydrogenase 7 (class IV), mu or sigma polypeptide alcohol dehydrogenase, iron containing, 1 (liver) alcohol dehydrogenase 1AB alcohol dehydrogenase 1AC alcohol dehydrogenase 1BC
Adh5	alcohol dehydrogenase V
Akr	aldo-keto reductase family 1, member A1 aldo-keto reductase family 1, member B1 (aldose reductase) aldo-keto reductase family 7, member A2 (aflatoxin aldehyde reductase)
Akr1b1	aldo-keto reductase family 1, member B1 (aldose reductase)
Akr7a2	aldo-keto reductase family 7, member A2 (aflatoxin aldehyde reductase)

Reaction	Catalyzing enzymes
Aldh	aldehyde dehydrogenase 1 family, member A1 aldehyde dehydrogenase 1 family, member A2 aldehyde dehydrogenase 1 family, member A3 aldehyde dehydrogenase 3 family, member A1 aldehyde dehydrogenase 3 family, member A2 aldehyde dehydrogenase 3 family, member B1 aldehyde dehydrogenase 3 family, member B2 aldehyde dehydrogenase 7 family, member A1 aldehyde dehydrogenase 9 family, member A1
Aqp9	aquaporin 9
Bpgm	2,3-bisphosphoglycerate mutase
Crns (Carnitine shuttle)	Solute carrier family 25, carnitine/acylcarnitine translocase, member 20, mitochondrial carnitine acetyltransferase, mitochondrial
Cyp2e1	cytochrome P450, family 2, subfamily E, polypeptide 1
Fbp	fructose-1,6-bisphosphatase 1 (liver) fructose-1,6-bisphosphatase 2 (muscle)
Flj32499	hypothetical protein FLJ32499, mitochondrial
Flj39207	C219-reactive peptide
Glo1	glyoxalase I
Grhpr	glyoxylate reductase/hydroxypyruvate reductase
Hadha	hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), alpha subunit hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), beta subunit
Hagh	hydroxyacylglutathione hydrolase hydroxyacylglutathione hydrolase-like
Hmgcl	3-hydroxylmethyl-3-methylglutaryl-Coenzyme A lyase 3-hydroxymethyl-3-methylglutaryl-Coenzyme A lyase-like 1, mitochondrial
Hmgcs1	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1 (soluble)
Hmgcs2	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (mitochondrial)
Idh2	isocitrate dehydrogenase 2 (NADP+), mitochondrial
Idh1	isocitrate dehydrogenase 1 (NADP+), soluble
Ireb2	iron-responsive element binding protein

Reaction	Catalyzing enzymes
Ldh	lactate dehydrogenase A (liver, skeletal muscle)
	lactate dehydrogenase B (heart, red blood cell)
	lactate dehydrogenase C (testis)
	lactate dehydrogenase A-like 6B (testis)
	lactate dehydrogenase A-like 6A
	lactate dehydrogenase (heart, red blood cell, brain, kidney, liver, skeletal muscle)
Ldha	lactate dehydrogenase A (liver, skeletal muscle)
Ldhd	lactate dehydrogenase D
Mdh1	malate dehydrogenase 1 (NAD), soluble
	malate dehydrogenase 1b (NAD), soluble
Me1	malic enzyme 1, NADP(+)-dependent, cytosolic
Oxct	3-oxoacid CoA transferase 1, mitochondrial
	3-oxoacid CoA transferase 2, mitochondrial
Pc	pyruvate carboxylase, mitochondrial (liver, kidney, intestine)
Pck1	phosphoenolpyruvate carboxykinase 1, soluble (liver, kidney, intestine)
Pck2	phosphoenolpyruvate carboxykinase 2, mitochondrial (liver, kidney, intestine)
Pgam	brain phosphoglycerate mutase 1
	muscle phosphoglycerate mutase 2
Pgk	phosphoglycerate kinase 1
	phosphoglycerate kinase 2 (testis)
Slc16a1	solute carrier family 16 (monocarboxylic acid transporters), member 1
Slc25a1	solute carrier family 25 (mitochondrial carrier; citrate transporter), member 1
Slc25a10	solute carrier family 25 (dicarboxylate transporter) member 10, mitochondrial
Uev	ubiquitin-conjugating enzyme E2-like
Zadh	zinc binding alcohol dehydrogenase

B List of abbreviations

Abbreviation	Compound	Abbreviation	Compound
12ppd-S	(S)-propane-1,2-diol	HmgCoA	hydroxymethylglutaryl-CoA
AaCoA	acetoacetyl-CoA	Lac-D	D-lactate
AcAc	acetoacetate	Lac-L	L-lactate
AcCoA	acetyl-CoA	Lald-D	D-lactaldehyde
CoA	coenzyme A	Lald-L	L-lactaldehyde
DHAP	dihydroxyacetone phosphate	Lgt-S	(R)-S-lactoylglutathione
FDP	fructose-1,6-bisphosphate	Mthgxl	methylglyoxal
FicytC	ferricytochrome c	Pyr	pyruvate
FocytC	ferrocytochrome c	R5P	ribose-5-phosphate
G3P	Glyceraldehyde 3-phosphate	Ru5P	ribulose-5-phosphate
G6P	glucose-6 phosphate	Succ	succinate
Glc	glucose	SuccCoA	succinyl-CoA
Gthrd	reduced glutathione	X5P	xylulose-5-phosphate

C Determining reactions essential for a pathway

In order to determine which reactions are essential on a pathway a linear program can be used. It consists in checking whether there still exists a steady-state flux through the final reaction of the pathway, if the flux through each reaction of an initial pathway is individually constrained to zero. If we find no steady-state flux, the corresponding reaction is essential for the pathway.

We use the same notation like in the main document. Thus, $\mathbf{N}$ corresponds to the $n \times m$ stoichiometric matrix of a reaction network of n reactions and m metabolites. Given an initial pathway of length l, that is, a set of indices of reactions, $i_1, ..., i_l$ with $1 \leq i_1, ..., i_l \leq n$ we assume that i_l is the index of the final reaction r_{i_l} of the pathway. In order to determine the essential reactions of the pathway we check for each i_p with $1 \leq p \leq l-1$ whether the linear program

$$\mathbf{v} \geq 0 \tag{1}$$
$$\mathbf{N} \cdot \mathbf{v} = 0 \tag{2}$$
$$v_{i_p} = 0 \tag{3}$$
$$v_{i_l} \geq 1 \tag{4}$$
$$\tag{5}$$

with the variables $\mathbf{v}$ is feasible. Thus, we check if there still exists a steady-state flux through r_{i_l} if constraining the flux through r_{i_l} to zero. In consequence, if the linear program in non-feasible for a particular i_p, r_{i_p} is an essential reaction. Note that we split reversible reactions into irreversible forward and backward direction. Hence, equation (1) corresponds to an irreversibility constraint.

D Elementary flux patterns in acetone degradation

#	Reactions	Enzyme	Rat.	#	Reactions	Enzyme	Rat.
1	NADPH + $Acetol_c$ → $NADP^+$ + 12ppd-S_c NAD^+ + 12ppd-S_c → NADH + Lald-L_c NAD^+ + Lald-L_c → NADH + Lac-L_c Lac-L_c → Lac-L_m NAD^+ + Lac-L_m → NADH + Pyr_m	Akr Add/Zadh Aldh Aqp9 Ldha	-2/+3	2	NADPH + $Acetol_c$ → $NADP^+$ + 12ppd-S_c NAD^+ + 12ppd-S_c → NADH + Lald-L_c NAD^+ + Lald-L_c → NADH + Lac-L_c Lald-L_c → Lald-L_m NAD^+ + Lac-L_m → NADH + Pyr_m	Akr Add/Zadh Aldh Slc16a1 Ldha	-2/+3
3	NADPH + $Acetol_c$ → $NADP^+$ + 12ppd-S_c NAD^+ + 12ppd-S_c → NADH + Lald-L_c NAD^+ + Lald-L_c → NADH + Lac-L_c $FicytC_m$ + Lac-L_c → $FocytC_m$ + Pyr_c Pyr_c → Pyr_m	Akr Add/Zadh Aldh Flj32499 Slc16a1	-2/+2	4	NADPH + $Acetol_c$ → $NADP^+$ + 12ppd-S_c NAD^+ + 12ppd-S_c → NADH + Lald-L_c NAD^+ + Lald-L_c → NADH + Lac-L_c NAD^+ + Lac-L_c → NADH + Pyr_c Pyr_c → Pyr_m	Akr Add/Zadh Aldh Ldh/Uev3 Slc16a1	-2/+3
5 + 6	NADPH + $Acetol_c$ → $NADP^+$ + 12ppd-S_c NAD^+ + 12ppd-S_c → NADH + Lald-L_c NAD^+ + Lald-L_c → NADH + $Mthglx_c$ NADPH + $Mthglx_c$ → $NADP^+$ + Lald-D_c NAD^+ + Lald-D_c → NADH + Lac-D_c NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Akr Add/Zadh Add/Zadh Akr/Grhpr Aldh Ldhd Slc16a1	-3/+4	7	NADPH + $Acetol_c$ → $NADP^+$ + 12ppd-S_c NAD^+ + 12ppd-S_c → NADH + Lald-L_c NAD^+ + Lald-L_c → NADH + $Mthglx_c$ $NADP^+$ + $Mthglx_c$ → NADPH + Pyr_c Pyr_c → Pyr_m	Akr Add/Zadh Add/Zadh Akr1b1 Slc16a1	-1/+2
8 + 9	NADPH + $Acetol_c$ → $NADP^+$ + 12ppd-S_c NAD^+ + 12ppd-S_c → NADH + Lald-L_c NAD^+ + Lald-L_c → NADH + $Mthglx_c$ NADPH + $Mthglx_c$ → $NADP^+$ + Lald-D_c NAD^+ + Lald-D_c + $Gthrd_c$ → NADH + Lgt-S_c Lgt-S_c → Lac-D_c + $Gthrd_c$ NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Akr Add/Zadh Add/Zadh Akr/Grhpr Adh5 Hagh Ldhd Slc16a1	-3/+4	10	NADPH + $Acetol_c$ → $NADP^+$ + 12ppd-S_c NAD^+ + 12ppd-S_c → NADH + Lald-L_c NAD^+ + Lald-L_c → NADH + $Mthglx_c$ NADH + $Mthglx_c$ → NAD^+ + Lald-D_c NAD^+ + Lald-D_c → NADH + Lac-D_c NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Akr Add/Zadh Add/Zadh Akr/Grhpr Aldh Ldhd Slc16a1	-2/+3
11	NADPH + $Acetol_c$ → $NADP^+$ + 12ppd-S_c NAD^+ + 12ppd-S_c → NADH + Lald-L_c NAD^+ + Lald-L_c → NADH + $Mthglx_c$ NADH + $Mthglx_c$ → NAD^+ + Lald-D_c NAD^+ + Lald-D_c + $Gthrd_c$ → NADH + Lgt-S_c Lgt-S_c → Lac-D_c + $Gthrd_c$ NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Akr Add/Zadh Add/Zadh Akr/Grhpr Adh5 Hagh Ldhd Slc16a1	-2/+3	12	NADPH + $Acetol_c$ → $NADP^+$ + 12ppd-S_c NAD^+ + 12ppd-S_c → NADH + Lald-L_c NAD^+ + Lald-L_c → NADH + $Mthglx_c$ $Gthrd_c$ + $Mthglx_c$ → Lgt-S_c NADH + Lgt-S_c → NAD^+ + Lald-D_c + $Gthrd_c$ NAD^+ + Lald-D_c → NADH + Lac-D_c NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Akr Add/Zadh Add/Zadh Glo1 Adh5 Aldh Ldhd Slc16a1	-2/+3
13	NADPH + $Acetol_c$ → $NADP^+$ + 12ppd-S_c NAD^+ + 12ppd-S_c → NADH + Lald-L_c NAD^+ + Lald-L_c → NADH + $Mthglx_c$ $Gthrd_c$ + $Mthglx_c$ → Lgt-S_c Lgt-S_c → Lac-D_c + $Gthrd_c$ NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Akr Add/Zadh Add/Zadh Glo1 Hagh Ldhd Slc16a1	-2/+3	14	NADPH + $Acetol_c$ → $NADP^+$ + $Mthglx_c$ $NADP^+$ + $Mthglx_c$ → NADPH + Pyr_c Pyr_c → Pyr_m	Cyp2e1 Akr1b1 Slc16a1	-1/0
15 + 16	NADPH + $Acetol_c$ → $NADP^+$ + $Mthglx_c$ NADPH + $Mthglx_c$ → $NADP^+$ + Lald-D_c NAD^+ + Lald-D_c → NADH + Lac-D_c NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Cyp2e1 Akr/Grhpr Aldh Ldhd Slc16a1	-3/+2	17	NADPH + $Acetol_c$ → $NADP^+$ + $Mthglx_c$ NADH + $Mthglx_c$ → NAD^+ + Lald-D_c NAD^+ + Lald-D_c → NADH + Lac-D_c NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Cyp2e1 Akr/Grhpr Aldh Ldhd Slc16a1	-2/+1
18	NADPH + $Acetol_c$ → $NADP^+$ + $Mthglx_c$ NADH + $Mthglx_c$ → NAD^+ + Lald-D_c NAD^+ + Lald-D_c + $Gthrd_c$ → NADH + Lgt-S_c Lgt-S_c → Lac-D_c + $Gthrd_c$ NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Cyp2e1 Akr/Grhpr Adh5 Hagh Ldhd Slc16a1	-2/+1	19 + 20	NADPH + $Acetol_c$ → $NADP^+$ + $Mthglx_c$ NADPH + $Mthglx_c$ → $NADP^+$ + Lald-D_c NAD^+ + Lald-D_c + $Gthrd_c$ → NADH + Lgt-S_c Lgt-S_c → Lac-D_c + $Gthrd_c$ NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Cyp2e1 Akr/Grhpr Adh5 Hagh Ldhd Slc16a1	-3/+2
21	NADPH + $Acetol_c$ → $NADP^+$ + $Mthglx_c$ $Gthrd_c$ + $Mthglx_c$ → Lgt-S_c NADH + Lgt-S_c → NAD^+ + Lald-D_c + $Gthrd_c$ NAD^+ + Lald-D_c → NADH + Lac-D_c NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Cyp2e1 Glo1 Adh5 Aldh Ldhd Slc16a1	-2/+1	22	NADPH + $Acetol_c$ → $NADP^+$ + $Mthglx_c$ $Gthrd_c$ + $Mthglx_c$ → Lgt-S_c Lgt-S_c → Lac-D_c + $Gthrd_c$ NAD^+ + Lac-D_c → NADH + Pyr_c Pyr_c → Pyr_m	Cyp2e1 Glo1 Hagh Ldhd Slc16a1	-2/+1

Table 2: List of possible pathways for the conversion of acetol to pyruvate. The second column indicates the reactions and the third column the catalyzing enzymes. The forth column gives the ratio of NADPH consumed to NADH produced in each pathway (including the NADPH oxidation for the reduction of acetone to acetol not shown). Pathways shaded in gray have not been reported previously in acetone metabolism (using (1) as a reference). Note that the pathways 5 and 6, 8 and 9, 15 and 16, as well as 19 and 20 are identical in reactions and NADPH/NADH-ratio but use a different enzyme for the reaction of Mthgxl to Lald-D. Except pathways 1 and 2, a direct import is assumed the principal source of mitochondrial pyruvate. Indices indicate the localization of the metabolites: c - cytosol, m - mitochondrion. A list of the abbreviations as can be found in the appendix A and B.

E ATP consumption for gluconeogenesis from fatty acids

Here we want to discuss the requirements in terms of NADPH, NADH and ATP of the presented pathways. Overall, 1 mole of glucose and 2 moles of carbon dioxide are produced from 4 mole acetyl-CoA. The following considerations will focus on the production of one glyceraldehyde 3-phosphate from two acetyl-CoA. The following steps in gluconeogenesis produce glucose from two glyceraldehyde 3-phosphate without further NADH or ATP consumption. It is assumed that a gradient of 10 protons from intermembrane space to the inner mitochondrion equals an amount of 2.5 ATP that can be produced from oxidation of 1 mitochondrial NADH (2).

The alternative pathways in ketone metabolism only yield energy in form of ATP or GTP, if acetoacetate is produced from acetoacetyl-CoA. In one case, via the 3-Oxoacid CoA transferase, one succinyl-CoA is produced from succinate and coenzyme A. Succinyl-CoA can be hydrolyzed via production of one ATP or one GTP. In the cytosol, acetoacetate can be produced directly from acetoacetyl-CoA accompanied with production of one ATP from AMP and pyrophosphate. The only transport reaction changing the mitochondrial proton gradient is the import of acetoacetate into the mitochondrion. In this case the gradient is reduced by one proton, equaling $\frac{1}{4}$ ATP. The subsequent export of acetone removes a proton from the mitochondrion and in consequence increases the membrane potential by $\frac{1}{4}$ ATP. When equaling the phosphorylation of AMP to ATP with two times the reaction of ADP to ATP, the production of acetoacetate can yield in between $\frac{1}{4}$ and 2 ATP.

In acetone metabolism there are large differences in terms of NADPH oxidized and NAD^+reduced in the different pathways. The reduction of one ferricytochrome can be equaled with the reduction of one NAD^+. Replenishing one NADPH using the described pathway via the malic enzyme costs an equivalent of 4 ATP. One ATP is consumed in the carboxylation of pyruvate to oxaloacetate. Two protons are added to the mitochondrion by import of pyruvate and reaction of water and carbon dioxide to HCO_3^-, necessary for the carboxylation step. The conversion to malate oxidizes one NADH and the subsequent transport to the cytosol proceeds via the antiport of one phosphate ion. The negatively charged phosphate increases the membrane potential by two. Consequently the import of pyruvate and carboxylation to oxaloacetate costs an equivalent of $1\frac{1}{2}$ ATP and the reduction to malate with subsequent transport into the cytosol 2 ATP.

The amount of NADH produced during acetone metabolism can be oxidized in the mitochondrion necessitating the use of the malate-aspartate shuttle for the import. Since this shuttle also involves the export of one phosphate ion, the overall ATP yield from one cytosolic NADH is 2 ATP. Thus, the energy-balance in acetone metabolism depends on the amount of NADPH consumed and NADH produced.

Pathway 21 and 22 (Table D) additionally produce one mitochondrial NADH instead of an cytosolic from oxidation of L-lactate to pyruvate. In this case the ATP-consumption is reduced by $\frac{1}{2}$ ATP, since the NADH can be oxidized directly without prior import into the mitochondrion. Furthermore, when Aquaporine 9 is used for the import of L-lactate into the cytosol, no symport of a proton is necessary. This reduces the ATP-balance by a further $\frac{1}{4}$ ATP. A more detailed analysis of energy requirements in acetone metabolism can be found in (3).

Further down in the pathway, the conversion of one cytosolic pyruvate to cytosolic phosphoenolpyruvate costs $2\frac{1}{2}$ ATP. This is the cost of the cycling pathway described above plus the hydrolization of one GTP to decarboxylate oxaloacetate. In the following, one ATP and one NADH are necessary to form glyceraldehyde 3-phosphate. This sums up to a cost of 3 ATP.

Overall the energy requirements of the pathways are mainly determined by the pathway used in ketogenesis and acetone metabolism. The most efficient pathway in acetone metabolism is pathway 7 (Table D) which has a netto-balance of one NADPH consumed and 2 NADH produced. Overall this would result in an investment of

$-$	2 ATP	(conversion of acetyl-CoA to acetone)
$+$	$(1 \times 3\frac{1}{2})$ ATP	(NADPH replenishment)
$-$	(2×2) ATP	(cytosolic NADH)
$+$	$5\frac{1}{2}$ ATP	(Pyruvate to glyceralde 3-phosphate)
$+$	**3 ATP**	(overall ATP consumption)

Thus, the production of one glucose from 4 acetyl-CoA costs at least 6 ATP.

Taking by contrast a pathway which is in favor of NADPH consumption over NADH production, the balance is (pathway 14, Table D)

$-$	2 ATP	(conversion of acetyl-CoA to acetone)
$+$	$(1 \times 3\frac{1}{2})$	(NADPH replenishment)
$-$	(0×2) ATP	(cytosolic NADH)
$+$	$5\frac{1}{2}$ ATP	(Pyruvate to glyceralde 3-phosphate)
$+$	**7 ATP**	(overall ATP consumption)

resulting in a cost of 14 ATP for each glucose produced.

The most ATP-intensive routes can be constructed from pathways 15, 16, 19 or 20 and no ATP production in ketogenesis

$-$	$\frac{1}{4}$ ATP	(conversion of acetyl-CoA to acetone)
$+$	$(3 \times 3\frac{1}{2}$ ATP)	(NADPH replenishment)
$-$	(2×2) ATP	(cytosolic NADH)
$+$	$5\frac{1}{2}$ ATP	(Pyruvate to glyceralde 3-phosphate)
$+$	$\mathbf{11\frac{3}{4}}$ **ATP**	(overall ATP consumption)

This requires $23\frac{1}{2}$ ATP to produce one glucose.

Furthermore, it needs to be mentioned that high amounts of NADH production would also induce an intensive usage of the malate-aspartate shuttle in order to import this compound for oxidation into the mitochondrion. This shuttle involves the TCA-cycle intermediates oxaloacetate, malate and oxoglutarate. Especially oxoglutarate, which is not part of the mentioned pathways would further decrease the energy-production capability of the TCA-cycle. Another point is the increased mitochondrial NADH concentration during ketogenesis (3) which would be further increased by the import of cytosolic NADH. Thus, pathway 7 might be the most favorable, despite its higher consumption in terms of ATP equivalents.

References

1. Kalapos, M. P. (2003) On the mammalian acetone metabolism: from chemistry to clinical implications. Biochim. Biophys. Acta **1621**, 122–139
2. Lehninger, A., Nelson, D. L. and Cox, M. M. (2008) Lehninger Principles of Biochemistry W. H. Freeman 5th edn.
3. Kalapos, M. P. (1999) Possible physiological roles of acetone metabolism in humans. Med. Hypotheses **53**, 236–242

Angabe zum Eigenanteil

Titel	**Literaturangabe**	**Autoren**	**Arbeits-anteil**
Analyzing molecular reaction networks: From pathways to chemical organizations.	*Mol. Biotechnol.*, 34(2):117-123, Oct 2006.	**C. Kaleta**, F. Centler, P. Dittrich.	50% 25% 25%
Computing chemical organizations in biological networks.	*Bioinformatics*, 24(14):1611-1618, Jul 2008.	F. Centler, **C. Kaleta**, P.S. di Fenizio, P. Dittrich.	40% 35% 10% 15%
Phenotype prediction in regulated metabolic networks.	*BMC Syst. Biol.*, 2:37, 2008.	**C. Kaleta**, F. Centler, P.S. di Fenizio, P. Dittrich.	50% 30% 10% 10%
Using chemical organization theory for model-checking.	*Bioinformatics*, 25(15):1915-1922, Aug 2009.	**C. Kaleta**, S. Richter, P. Dittrich.	40% 40% 20%
Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns.	*Genome Res.*, 2009, in print.	**C. Kaleta**, L.F. de Figueiredo, S. Schuster.	70% 10% 20%
Systematic determination of gluconeogenic pathways from fatty acids in humans.	*Biochem. J.*, 2009, in preparation.	**C. Kaleta**, L.F. de Figueiredo, R. Guthke, S. Schuster.	70% 10% 5% 15%

.....

bestätigt Stefan Schuster

Lebenslauf

Persönliche Daten

Name: Christoph Kaleta

Geburtsdatum: 16. Mai 1983

Geburtsort: Weimar, Deutschland

Nationalität: deutsch

Familienstand: ledig

Ausbildung

1994-2002: Friedrich-Schiller-Gymnasium Weimar,

2003-2008: Studium der Bioinformatik an der Friedrich-Schiller-Universität Jena, Abschluss als Diplom-Bioinformatiker

seit 2008: Promotion am Lehrstuhl für Bioinformatik der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena bei Prof. Dr. Stefan Schuster

Publikationen

1. C. Kaleta, F. Centler, and P. Dittrich. Analyzing molecular reaction networks: From pathways to chemical organizations. *Mol. Biotechnol.*, 34(2):117-123, Oct 2006.

2. R. Schwarz, C. Liang, C. Kaleta, M. Kühnel, E. Hoffmann, S. Kuznetsov, M. Hecker, G. Griffith, S. Schuster and T. Dandekar. Integrated network reconstruction, visualization and analysis using YANAsquare. *BMC Bioinformatics 2007*, 8, 313.

3. C. Kaleta, F. Centler, P. S. di Fenizio, and P. Dittrich. Phenotype prediction in regulated metabolic networks. *BMC Syst. Biol.*, 2:37, 2008.

4. F. Centler, C. Kaleta, P. S. di Fenizio, and P. Dittrich. Computing chemical organizations in biological networks. *Bioinformatics*, 24(14):1611-1618, Jul 2008.

5. L.F. de Figueiredo, S. Schuster, C. Kaleta and D.A. Fell. Can sugars be produced from fatty acids? A test case for pathway analysis tool. *Bioinformatics*, 25(1): 152-158, Jan 2009.

6. C. Kaleta, S. Richter, and P. Dittrich. Using chemical organization theory for model-checking. *Bioinformatics*, 25(15):1915-1922, Aug 2009.

7. C. Kaleta, L. F. de Figueiredo, and S. Schuster. Detecting metabolic conversions in genome-scale metabolic networks on the basis of elementary flux patterns in subnetworks. In: I. Troch, F. Breitenecker (eds.), *Proceedings MATHMOD 09 Vienna - Full Papers CD Volume*, 2009, pages 748-759.

8. C. Kaleta, L. F. de Figueiredo, and S. Schuster. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res.*, 2009, *in print*.

9. C. Kaleta, S. Nikolajewa, S. Schuster, and R. Guthke. Use of the concept of directed information for the inference of gene regulatory networks in *Escherichia coli*. *BMC Bioinformatics*, 2009, *submitted*.

10. C. Kaleta, L. F. de Figueiredo, R. Guthke, and S. Schuster. Systematic determination of gluconeogenic pathways from fatty acids in humans. *Biochem. J.*, 2009, *in preparation*.

11. C. Kaleta. From Artificial Chemistry to Systems Biology: Software for Chemical Organization Theory. In: A. Adamatzky and M. Komosinski (eds.), *Artificial Life Models in Software Vol. II*, Springer, London, 2009, pages 319-342.

12. L. F. de Figueiredo, A. Podhorski, A. Rubio, C. Kaleta, J. E. Beasley, S. Schuster, and F. J. Planes. Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics, 2009, *revised version submitted*.

13. C. Kaleta, L.F. de Figueiredo, J. Behre, and S. Schuster. EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. In: German Conference on Bioinformatics 2009, *accepted*.

Vorträge

1. C. Kaleta. Network inference in *Escherichia coli* using directed information. Workshop: Gene Regulatory Network Inference, Jena, September 25-26, 2008.

2. C. Kaleta, F. Centler, P. S. di Fenizio, S. Schuster, P. Dittrich. Analyzing molecular reaction networks: From elementary modes to chemical organizations. Workshop: Metabolic networks: dynamics, evolution, and topology, Potsdam-Golm, November 21, 2008.

3. C. Kaleta, L. F. de Figueiredo, and S. Schuster. Detecting metabolic conversions in genome-scale metabolic networks on the basis of elementary flux patterns in subnetworks. MATHMOD 2009, Vienna, February 11 - 14, 2009.

4. C. Kaleta, L. F. de Figueiredo, and S. Schuster. Steady-state analysis in genome-scale metabolic networks using elementary flux patterns. Genomic and Systems Biology of Molecular Networks - Workshop Zeuthen 2009, Zeuthen, March 26 - 28, 2009.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt habe. Mir ist die geltende Promotionsordnung bekannt und ich habe weder die Hilfe eines Promotionsberaters in Anspruch genommen, noch haben Dritte unmittelbare oder mittelbare geldwerte Leistungen für Arbeit erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die vorgelegte Dissertation wurde noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Weiterhin habe ich mich mit der vorliegenden Arbeit an keiner anderen Hochschule um den akademischen Grad doctor rerum naturalium (Dr. rer. nat.) beworben und weder früher noch gegenwärtig die Eröffnung eines Verfahrens zum Erwerb des o.g. akademischen Grades an einer anderen Hochschule beantragt.

Bei der Auswahl und Auswertung des Materials, sowie bei der Herstellung des Manuskripts hat mich der Lehrstuhl für Bioinformatik der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena unter der Leitung von Prof. Dr. Stefan Schuster und die Biosystemanalysegruppe unter der Leitung von PD Dr. Peter Dittrich unterstützt.

Jena, den 7. April 2010

.....

(Christoph Kaleta)