Towards structure and dynamics of metabolic networks

Dissertation

zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

eingereicht von

SERGIO GRIMBS

Arbeitsgruppe Bioinformatik
Max-Planck-Institut für molekulare Pflanzenphysiologie

Potsdam, im Januar 2009

Abstract

This work proposes solutions for several issues pertaining to metabolic network modelling, ranging from network reconstruction to multistability analysis to new modelling strategies coping with unreliable kinetic parameters. The emphasis is on the close connection between structure and dynamical behaviour of metabolic networks.

High-throughput data from various "omics" and sequencing techniques have rendered the automated metabolic network reconstruction a highly relevant problem. It is provably hard to find a suitable and fully automated algorithm to solve a mathematical abstraction of a reconstruction problem, that accounts for the uncertain, ambiguous and hence inherently probabilistic relations between genes, enzymes, reactions and metabolites.

The biosynthetic capabilities of given genome-scale metabolic networks, i.e. the metabolites that can be produced after providing some seed compounds, reflect prominent aspects of their functionality. The reverse problem of determining a minimal set of metabolites that has to be provided in order to obtain some desired target compounds, is also of importance, especially with respect to identification of drug targets and biotechnological applications. This problem is shown to be computationally hard, even after relaxation for approximation results.

A relevant property of metabolic networks viewed as dynamic systems is their capability to support multistability, as it enables switching between different modes of operation as a response to changing conditions. Chemical reaction network theory (CRNT) and its extensions provide a powerful and mathematically sound framework to obtain multistability results derived directly from the structure of a given network. CRNT is applied to compare and discriminate against several models of the Calvin cycle.

The development of detailed kinetic models is often hampered by the lack of knowledge about the kinetic properties of the involved enzymes and membrane transporters. This can be partly overcome by reformulating the Jacobian matrix in terms of *saturation parameters*, which describe the normalized influence of each metabolite on every reaction at steady state. Subsequent sampling of saturation parameters is used to evaluate the functional role of allosteric feedback regulation in the stabilization of the metabolic network. Furthermore, statistical measures for the relative impact of enzymatic reactions on local stability of the steady state are derived.

Several modelling approaches derived from assuming different simplistic kinetic mechanisms (mass-action, Michaelis-Menten, power-law, LinLog) are compared to a well established reference model of the human red blood cell. The quality of such simple models can be increased significantly by choosing a small subset of reactions, for which detailed rate equations, including allosterical effects, are established consecutively. The appropriate reactions are found by ranking the reactions according to the above-mentioned measure for their respective influence on stability.

Acknowledgement

First of all I would like to thank my supervisor, Joachim Selbig, for putting faith in me and supporting my work in all possible ways. He provided me with many helpful suggestions, important advice and constant encouragement during the course of this work.

I thank Hermann-Georg Holzhütter for cordially inviting me to visit his lab. This resulted in fruitful collaboration, for which I especially have to thank Sascha Bulik.

Furthermore, I like to thank Ralf Steuer, who introduced me to metabolic modelling and helped me with my project, especially in the beginning.

Special thanks go to Zoran Nikoloski for inspiring discussions and raising new and interesting questions. His critical comments on this project were most valuable.

This work could not have been completed without the help of Marco Ende, who took care of all technical problems and backed me up whenever necessary.

Many thanks to all present and former members of the bioinformatics group, which gave me a wonderful time at the Max-Planck-Institute and were always willing to answer my questions: Georg Basler, Liam Childs, Jan-Ole Christian, Pawel Durek, Tanja Gärtner, Detlef Groth, Stefanie Hartmann, Manuela Hische, Jan Hummel, Kathrin Jürchott, Sabrina Kleeßen, Sebastian Klie, Peter Krüger, Abdelhalim Larhlimi, Jan Lisec, Patrick May, Ram Narang, Alexander Platzer, Henning Redestig, Dirk Repsilber, Christian Schudoma, Wolfram Stacklies, Matthias Steinfath, Xiao Liang Sun, Hiroki Takahashi, Dirk Walther und Daniel Weicht.

I sincerely thank Katharina Rach for all her support and her patience.

Contents

Chapter 1

Introduction

1.1 Systems biology

Recent years have seen a shift from reductionist approaches towards a holistic and systemic view of biological processes. This shift is entailed by the emergence of *systems biology*, although it is not an easy task to find a widely accepted and precise definition or description for this term. The elements of biological systems are complex, bearing an intrinsic complexity on the system-level (Kitano, 2002a). Therefore, it is necessary to analyse the interactions of all parts of the system and the implied dynamics in order to elicit conclusions for the whole (Kitano, 2002b).

On one hand, systems biology has its origins in molecular biology (Westerhoff and Palsson, 2004) and aims at understanding the interplay of cellular components such as metabolites, proteins and genes. The recent success of genomics, proteomics and other "omics"-techniques renders genome-scale experiments possible and stimulates the increasingly important role of data integration techniques (Hwang et al., 2005). On the other hand, computational and mathematical modelling approaches are required for gaining in-depth insights into the organisation and functioning of biological systems (Kahlem and Birney, 2006). Systems biology can therefore be regarded as an integrative and interdisciplinary approach that tries to combine genome-scale data with mathematical methods in order to model and simulate complex biological systems (Klipp et al., 2005).

The work presented here takes such an interdisciplinary approach to study models of metabolic networks. In particular, the connection between structure and dynamical behaviour of metabolic networks is analyzed. As both structure and dynamics cannot be fully understood by inspection of the network elements in isolation, the analysis considers the whole system.

1.2 Biological networks

Biological networks are abstract descriptions, which can capture many essential properties of various biological systems (Alon, 2003). In general, a network is defined by a set of elements and a set of interactions between these elements. Following the notation from graph theory, they are called the *nodes* and the *edges* of the network, respectively. If the interaction bears an intrinsic directionality, the network is called *directed*, otherwise the network is *undirected*. Furthermore, if the strength or capacity of an interaction should be described explicitly, a *weight* is assigned to the edges.

The study of metabolic networks is of high relevance, because of their implications for the basic understanding of living cells and organisms and for medical applications, especially with respect to discovering drug targets (Guimerà et al., 2007b).

As networks provide a natural means for modelling the interaction of elements, network-based approaches have already been applied to many different fields of biology, from ecology to molecular biology (Calvano et al., 2005; Hood et al., 2004; Li et al., 2004). In the next section, prominent examples of networks arising in various biological subfields are reviewed, without intending to be exhaustive. This work focuses on metabolic networks, introduced in detail in Section 1.3, below.

1.2.1 Gene regulatory networks

Genes are regulated upstream by binding of specific transcription factors, which enable or prohibit the expression (transcription) of the gene. A transcription factor is itself a product of an expressed gene. Hence, the regulation of genes forms a network, where nodes represent genes and edges describe regulations. The interaction between two genes is clearly directed from the gene that encodes the transcription factor towards the regulated gene. Therefore, the gene regulatory network is directed. If a transcription factor is known to act as an activator or inhibitor on a particular gene, the corresponding edge is assigned a positive or negative sign, respectively.

Chromatin Immunoprecipitation (ChIP) is used to scan the genome for DNA binding sites of a particular protein of interest *in vivo*. Combined with microarray technology (ChIP-chip), this method detects binding events of transcription factors to genes on a genome-scale level and allows for direct readout of regulatory networks (Buck and Lieb, 2004). Limiting factors are the availability of antibodies to the transcription factors of interest and the resolution of the analyzed DNA fragments.

DNA microarrays measure the expression level of thousands of genes simultaneously. They can be used to trace changes in gene expression under various conditions and at different time points. Regulatory networks can be inferred from the obtained gene expression profiles by applying similarity measures on the gene expression patterns. Two genes are connected (coexpressed), if the similarity of the expression patterns is above a significance threshold. For instance, a widely used similarity measure is the correlation coefficient, employed by Stuart et al. (2003) on a combined set of over 3000 microarrays from multiple model organisms (*Homo sapiens*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*) to successfully predict gene functions for evolutionary conserved coexpressed genes. Partial correlation (de la Fuente et al., 2004) is another similarity measure, which can reduce the number of edges describing indirect correlation. These edges might emerge between two genes, which are uncorrelated among themselves but strongly correlated to a common third one. As similarity measures are symmetric, the inferred networks do not obtain directionality and hence lack to reproduce an important property of gene regulatory networks. One approach to overcome this problem is to analyze correlation between expression patterns stemming from time series experiments. If two genes exhibit the same expression profile except for a time-shift, then the 'early' gene might be the cause for the response of the 'late' gene. In this case, a putative direction can be assigned to the corresponding edge of the inferred regulatory network (Redestig et al., 2007).

Assuming coregulation of genes belonging to the same functional group, e.g. the encoded proteins belong to a common protein complex, gene regulatory networks can also be used to predict gene functions. This process is known as "guilt-by-association". Comparison between clusters in an inferred gene regulatory network and gene function according to the Gene Ontology (GO) annotation showed reasonable accordance (Lee et al., 2004). Using an integrated approach, DNA binding data can be combined with gene expression data to define groups of genes, which are regulated by the same transcription factors and follow the same gene expression patterns (Bar-Joseph et al., 2003). Such grouping identifies genes with similar biological function.

1.2.2 Protein-protein interaction networks

Interactions between proteins are fundamental to the functioning and survival of every living cell. They constitute the *protein-protein interaction network*, in which nodes represent proteins, and edges connect nodes if the associated proteins bind to each other. As protein binding is symmetric, the interaction network is undirected.

The *two hybrid* approach is used to experimentally detect proteins capable of interacting with a given protein of interest *in vivo* (Chien et al., 1991). In this approach, a reporter gene such as GFP, which is activated by binding of a transcription factor, is included into the genome using genetic engineering. The transcription factor is split into two fragments, one of which is fused to the protein of interest and the other one to a target protein. If the target protein interacts with the protein of interest, the transcription factor will become functional. Hence, the reporter gene is activated and produces a measurable signal.

Several model organisms such as *S. cerevisiae* (Schwikowski et al., 2000; Uetz et al., 2000) or *D. melanogaster* (Giot et al., 2003) have already been analyzed systematically by testing thousands of proteins for pairwise interaction using two hybrid approaches. However, these methods are error-prone, so each derived interaction serves as a candidate and should be confirmed or discarded by further experiments. To get a full understanding of the binding processes itself, it is necessary to reveal the three-dimensional structure of the proteins, especially of the binding sites. The structural information can then be mapped on metabolic pathways to estimate binding affinities and kinetic parameters. Unfortunately, until now it is very time-consuming to obtain detailed structural information (Aloy and Russell, 2006).

1.2.3 Further examples

Evolutionary relationships among different biological species or taxa compose a *phylogenetic tree*, which is a connected and acyclic network. Each leaf node is labeled by a species and each internal node stands for the most recent common ancestor of its decedent nodes. The estimated period of time that passed since the differentiation of species can then be used as an edge-weight. Recent work on phylogenetic trees tries to incorporate different kinds of “omics” data to gain further insights into biological ancestry of species. Structural information of metabolic pathways is used by Heymans and Singh (2003) to define a distance between species, from which a phylogenetic tree can be derived by hierarchical clustering. The necessity to cope with evolutionary events such as horizontal gene transfer, gene duplication and recombination as well as ambiguous sequence data leads to phylogenetic networks (Huson and Bryant, 2006), which are not acyclic.

Food webs reflect the predator-pray relationships between different species within a given ecosystem. High quality, comprehensive food webs can become very large and complex. Williams et al. (2002) evaluated several aquatic and terrestrial ecosystems and showed by means of graph-theory that species extinction or invasion affects species in a broader neighborhood than previously thought. Rooney et al. (2006) addressed the question of stability of food webs, i.e. the ability to respond to changes in the population sizes as well as in the carbon influx. Their results point out the critical role of top predators and asymmetrical carbon influx to achieve stability, which was observed across a range of different ecosystems.

In *social networks*, individuals are connected to each other if there exists a social relationship between them, such as friendship or kinship. These type of networks are not yet in the focus of systems biology, but can serve as an example of interdisciplinary research between sociology, mathematics, and biology. For instance, Andre et al. (2007) proposed a method that helps to contain a tuberculosis outbreak. They used the graph-theoretical concept of “betweenness” to prioritize contact persons, that have to be screened and examined first in order to obtain maximum efficiency. A statistical approach is taken by Christakis and Fowler (2007) to show that the risk of becoming obese is significantly connected to the number of obese persons in your social neighborhood, but independent from the geographical neighborhood.

1.3 Metabolic networks

Metabolism of living cells describes the complex and highly intertwined processes of subsequent chemical conversion of compounds by enzymatic or spontaneous reactions. Metabolic networks are comprised of metabolites as nodes and reactions as edges between them. For irreversible reactions, the associated edges are directed according to the reaction's direction. Reversible reactions lead to a pair of edges pointing in opposite directions.

If metabolic networks are considered as describing a dynamic system, the edges are weighted by fluxes, *i.e.*, turnover rates of molecules. The fluxes depend on the concentrations of the substrates and products of the reaction as well as on the concentration and thermodynamic properties of the participating enzyme.

Traditionally, the analysis of metabolic networks is focused on *metabolic pathways*, which are thought of, conceptually, as sets of reactions and metabolites that are functionally connected, such as the Calvin cycle or the TCA cycle. However, a precise definition of a pathway is difficult to formulate, especially with respect to delineation of pathways from one another. Nevertheless, to obtain a global view on the functionality of the entire metabolism, pathways should not be analyzed in isolation but rather on a genome-scale level (Nicholson and Wilson, 2003). This poses the general problem of metabolic modelling: the larger the network under consideration, the less experimentally validated biological knowledge on its constituting metabolites is available. Although even for large networks the structure is often known to a large extent, the detailed dynamical behaviour remains widely unknown.

1.3.1 Experimental methods and data generation

Numerous methods exist to measure kinetic parameters of enzymatic reactions (Bisswanger, 2002) *in vitro*. For instance, enzyme assays are used to follow the course of the reaction, either directly or indirectly, using coupled reactions, where the product of one reaction serves as a substrate for another reaction. As enzymes are sensitive to many environmental conditions such as pH and temperature, it is difficult to deduce *in vivo* kinetic parameters from these measurements.

Several techniques for high-throughput measurements of metabolites are now available, rendering data integration even more important to gain further insights into metabolic networks (Kell, 2004). The most widely used approach is composed of a gas or liquid chromatographic separation step combined with mass spectroscopy (GC/MS and LC/MS, respectively). While GC/MS is more accurate, LC/MS allows for analyzing a broader range of compounds, including secondary metabolites (Fernie et al., 2004). However, in a typical GC/MS or LC/MS measurement, not all detected compounds can be assigned a unique chemical structure. Another highly successful technique is nuclear magnetic resonance (NMR). In this technique, the fact that nuclear spin can lead to different energy levels of the nucleus if exposed to a strong magnetic field, is exploited in a nondestructive and noninvasive manner (Beckonert et al., 2007). The sensitivity of NMR is poor compared to MS techniques, but the information obtained by these methods is complementary and hence useful in getting a metabolic snapshot of the system at hand.

Besides metabolite concentrations, fluxes, i.e. rates of metabolite conversion, are crucial to understand network behaviour. Unlike metabolic concentrations, reaction rates are difficult to measure directly. Isotopic tracer techniques using ^{13}C are well established and provide insights into the flow of metabolites inside a network (Sauer, 2006). The basic idea is that after feeding labeled isotopes for a limited time period to the system under study, the resulting isotope distributions throughout the network allows for reconstruction of the internal fluxes. However, this approach is limited to steady state conditions and usually focuses on central metabolism. A nondestructive method to detect fluxes is given by fluorescence resonance energy transfer (FRET). FRET sensors are coupled with reporter proteins like GFP (green fluorescence protein) to monitor flux changes *in vivo* (Wiechert et al., 2007). Because FRET sensors are genetically encoded, they allow for subcellular resolution. However, a single FRET sensor can only monitor a few compounds at a time, rendering this approach time- and labour-intensive.

1.3.2 Modelling and analysis of metabolic networks

The modelling of metabolic networks ultimately aims at describing the metabolism for whole cells or even entire organisms in detail, including the kinetic parameters for all involved reactions. Such a detailed description would allow for precise predictions of biological processes. However, the available data might not be sufficient to model a metabolic network in such detail. Hence, a suitable strategy for modelling the desired level of abstraction has to be chosen to answer the biological questions at hand. The available methods range from pure structural approaches, which only incorporate basic topological information about the network, to precise description of the networks dynamics, including gene regulation (Wiechert, 2002). At present, detailed information is only available for relatively small networks, while genome-scale models of cell metabolism are still coarse.

Structural modelling is the simplest way to analyze metabolic networks that still yields valuable insights into the organisation and underlying building principle of the network topology. Only the set of participating metabolites and the set of reactions, describing the conversion of metabolites, are needed to build a graph representation of the metabolic network. Several variants for constructing the graph are possible, each of which can be directed according to reversibility and irreversibility of reactions. In a *compound graph*, each node corresponds to a metabolite and an edge is drawn between two nodes if the corresponding metabolites are substrate and product of a shared reaction. An almost dual description is given by the *reaction graph*. Here, each node represents a reaction and two nodes are connected by an edge if the product of one reaction serves as a substrate for the other reaction. Alternatively, the metabolic network can be modelled as bipartite graph, i.e. the set of nodes is split into two parts, representing the pool of metabolites and reactions, respectively. Here, edges only occur between nodes of different parts. A metabolite node is connected to a reaction node if the metabolite is either substrate or product of that reaction. Finally, the compound graphs can be extended to hypergraphs by allowing edges to not only connect two single nodes but two sets of nodes and hence allows for a consistent description of multi-substrate reactions. Both compound and reaction graphs are ambiguous in the sense that different metabolic networks can have the same compound or reaction graph. This problem is overcome by bipartite or hypergraph representation. Furthermore, every bipartite graph can be uniquely converted into a hypergraph, and vice versa (Deville et al., 2003). Well-established methods from graph theory can be used to analyze the graph representation of a metabolic network with respect to degree distribution, clustering coefficients, shortest paths and occurrence of motifs and modules, as will be described in Section 1.4.

Structural models can be enriched by incorporating the *stoichiometry* of the metabolic network, i.e. the quantitative relationship between substrates and products of each balanced reaction. For many metabolic networks, even at a genome-scale level, the stoichiometry is known. It can be summarized in a compact form as a *stoichiometric matrix* N. Each row and each column of N corresponds to a metabolite and a reaction, respectively. The entries of N describe how many molecules of each metabolite are produced or consumed by each reaction. As reaction rates take place on a much faster time scale than for instance gene regulation, a pseudosteady state assumption is justified (Llaneras and Picó, 2008). Hence, the metabolite concentrations are constant.

Describing the fluxes through all reactions by ν, the steady state condition can be formulated as $N \cdot \nu = 0$. Consequently, the null space of N defines the set of all flux vectors which satisfy the steady state condition. Using thermodynamic information, i.e. knowledge about the irreversibility of reactions, the set of feasible flux vectors can be further restricted to the so-called flux cone (Schilling et al., 1999). It is of high interest to describe the flux cone in biologically-relevant mathematical terms, as this might identify structural and functional building blocks of the metabolic network. One prominent approach uses *elementary flux modes*, or EM for short (Schuster et al., 1999). An EM is defined as a flux vector satisfying the steady state condition as well as additional constraints on irreversible reactions. Furthermore, an EM is elementary in the sense that it cannot be decomposed into smaller EMs. Hence, each EM describes a minimal set of reactions that can operate at steady state. For a given metabolic network, the set of EMs is unique, but increases drastically with network size. A different approach, called *extreme pathways*, was presented by Schilling et al. (2000). Here, by splitting every reversible reaction into a forward and a backward reaction, the flux cone is guaranteed to be pointed. The extreme pathways correspond to the extreme rays of the pointed flux cone. It can be shown that the set of extreme pathways is a subset of the set of EMs and that both sets are identical if the metabolic network does not contain reversible reactions (Klamt and Stelling, 2003; Papin et al., 2004). Recently, a new approach using an outer description of the flux cone – in contrast to EMs and extreme pathways which are both inner descriptions – was introduced by Larhlimi and Bockmayr (2008). This description is minimal and unique and significantly smaller compared to those obtained by EMs or extreme pathways.

Flux balance analysis (FBA) incorporates further constraints on the flux cone, such as maximal flux rates for every reaction (Varma and Palsson, 1994). FBA predicts the actual flux within a metabolic network under the assumption that every metabolic network evolved towards optimality with respect to an objective function. Although the true objective function is not known, good predictions can be obtained from optimizing for biomass production using standard techniques from linear optimization (Edwards et al., 2001). The response of a metabolic network after mutation, especially after gene knockouts, can be analyzed by two extensions of FBA. One approach, *minimization of metabolic adjustment* (MOMA), uses quadratic optimization to minimize the Euclidean distance of flux vectors for the initial and the mutated network (Segrè et al., 2002). The other approach, *regulatory on/off minimization* (ROOM), tries to minimize the number of significant flux changes by using a mixed integer optimization problem (Shlomi et al., 2005). While MOMA is more appropriate to describe the transient behaviour directly after gene knockout, ROOM yields better predictions for the final behaviour after metabolic adjustment.

As reaction rates vary over time, metabolic networks constitute complex dynamic systems, which is accounted for by *kinetic modelling* of metabolic networks. Here, the change of metabolite concentrations over time is described as $\frac{\delta S}{\delta t} = N \cdot \nu(S, k)$, where S denotes the vector of concentrations, N the stoichiometric matrix, ν the reaction rates and k the vector of kinetic parameters for all reactions. Several kinetic properties are of interest, especially the stability of steady states, which determines the robustness of the system under perturbations. The stability of a steady state can be determined by analyzing the eigenvalues of the Jacobian matrix (Heinrich and Schuster, 1996). If a metabolic network operates at steady state, it must be stable, as the system is constantly perturbed. Furthermore, the number of steady states a network can obtain gives insights about functionality under changing external conditions. Bifurcation analysis elucidates the influence of the kinetic parameters k on the system and identifies critical parameter values at which the behaviour changes drastically, e.g. switching from a stable to an unstable steady state or allowing oscillations.

Metabolic control analysis (MCA) is a successful and widely used approach to describe the distribution of the control of fluxes and metabolite concentrations throughout a given metabolic network (Kacser and Burns, 1973; Heinrich and Rapoport, 1974). One of the main results of

MCA demonstrates that control can be shared between different reactions, and, therefore, a rate limiting step cannot be found in every metabolic network. MCA defines *control coefficients*, which describe the relative change of fluxes and metabolite concentrations at steady state. It can be shown that the control coefficients for fluxes and for concentrations sum up to 1 and 0, respectively, stressing the analytical power and rigidity of MCA. Furthermore, local response of enzymes to changing conditions like varying substrate or product concentrations are quantified in so called *elasticities*. The elasticities are directly linked to the control coefficients by mathematically proven connectivity theorems. More details and extensions to MCA are given in Heinrich and Schuster (1996).

Explicit rate equations for every reaction are necessary if not only steady state behaviour but also time courses are to be analyzed. These rate equations constitute a system of ordinary differential equation, but can be extended for instance by algebraic equations to describe fixed relations between metabolites or stochastic differential equations to account for effects caused by molecules which only occur at very low numbers. Numerical integration can then be used to elucidate transient behaviour of metabolic networks. Several standard reaction rates derived from molecular enzyme mechanisms are available, mass-action kinetics and Michaelis-Menten kinetics being the most prominent ones. The kinetic parameters of each reaction are either obtained from literature or calculated by fitting to experimental data. Where available, knowledge about allosteric regulation or influence of external parameters (such as pH) is incorporated into the rate equations by careful modifications. However, such detailed knowledge can only be obtained for very few reactions, rendering detailed kinetic modelling at a genome-scale level a challenging endeavor.

1.3.3 Databases

The amount of available data either from biological experiments or from theoretical predictions related to biological systems in general and metabolic networks in particular is ever increasing. Sophisticated databases are inevitable to organize, sort and manage this huge amount of data and provide useful information. Several database projects dealing with various aspects of metabolic networks are already successfully used as tools for network analysis. Results obtained from such databases, especially genome-scale reconstruction of metabolic networks, have to be evaluated carefully due to non-unique metabolite identifiers and unbalanced reactions found in many databases (Poolman et al., 2006). KEGG (Kanehisa et al., 2004) is a well-established collection of databases and presents, among other things, information about many metabolic pathways, defined as known network of functional significance. The MetaCyc database (Caspi et al., 2006) provides more than 1100 curated metabolic pathways from primary and secondary metabolism for over 1500 different organisms. Information on reactions and their regulation, metabolites and enzymes together with their encoding genes is provided. Reactome (Joshi-Tope et al., 2005) is a curated database focusing on pathways from *Homo sapiens* and provides advanced graphical user interfaces. Biochemical and molecular information on all classified enzymes can be found in the BRENDA database (Chang et al., 2009). The BioModels database (Novère et al., 2006) accounts for the steadily increasing number of mathematical models for metabolic pathways. Here, the models mainly consist of systems of differential equations together with the corresponding set of kinetic parameters. The JWS-online repository for metabolic network models additionally provides an interface to simulate and analyze models easily (Olivier and Snoep, 2004).

1.4 Network analysis

Networks of various size and composition arise in various fields of biology. This raises the question of what biological conclusions can be inferred from the structure of these networks (Alm and

Arkin, 2003). Surprisingly, many biological networks, especially those originated in cell biology, exhibit characteristics, which are also found in other complex networks, arising in technological or social domains.

Terms and definitions from graph theory are used to identify and describe these recurrent characteristics. The *degree* of a node within a network is defined as the number of edges to other nodes. A *path* between two nodes a and b is given by a sequence of nodes $(x_1, \dots, x_n)$, such that $x_1 = a, x_n = b$ and there exists an edge between x_{i-1} and x_i for all $i = 2, \dots, n$. *Shortest paths* are defined straightforward. The *neighbourhood* of a node a is defined as the set of nodes connected to a. The ratio between the number of edges connecting the nodes within the neighbourhood divided by the number of edges that could possibly exist between them is called the *clustering coefficient* of a.

One recurrent characteristic is the *small-world* property (Watts and Strogatz, 1998; Wagner and Fell, 2001). A network exhibits this property if the average shortest path length is shorter and the average clustering coefficient is larger than expected from random graphs. As a consequence, any kind of information or signal can traverse a small-world network relatively fast. Another frequent characteristic is the shape of the degree distribution $P(k)$, which describes the fraction of nodes with degree k, i.e. the probability of a randomly chosen node to have degree k. If the degree distribution follows a power law, i.e. $P(k) \sim k^{-\gamma}$ for a positive exponent γ, then the network is said to be *scale-free*. Various biological networks have been shown to be scale-free (Barabási and Oltvai, 2004; Jeong et al., 2000). Furthermore, due to the power law degree distribution, scale-free networks exhibit a small number of nodes of high degree, referred to as *hubs* (Albert, 2005). These hubs are thought to serve specific purposes depending on the network type. Scale-free networks are very susceptible to failure of hubs, but highly robust to untargeted loss of randomly chosen nodes (Jeong et al., 2001). Furthermore, there is practical evidence that certain combinatorial optimization problems become tractable on networks of power law degree distribution (Ferrante et al., 2008). However, some authors raise concern about biological networks truly being scale-free, because the experimental data used to obtain these networks is highly biased (Hakes et al., 2008; Khanin and Wit, 2006).

Besides being scale-free, many biological networks are found to be modular. Here, a *module* – also named *community* – is not defined as a strictly separated subnetwork, but rather as a region where nodes are highly connected to other nodes of the very same region and sparsely connected to the rest of the network (Girvan and Newman, 2002). Modularity also allows for characterising the role of each node, which might vary from peripheral nodes of low degree with all links to the same module, to nodes which connect different modules, through to hubs connected homogeneously to all modules (Guimerà and Amaral, 2005). The occurrence of each node type can in turn be used to compare different complex networks (Guimerà et al., 2007a). In addition, nodes do not have to belong to a single module alone, but can be part of several modules, which leads to a hierarchical organisation of biological networks (Ravasz et al., 2002).

An effective method to analyse complex networks is to determine building blocks or so called *motifs* of the network. Motifs are subnetworks that occur more often in a given network than expected at random. Due to computational limitations, the networks are only scanned for very small motifs. Nevertheless, motifs with a well-defined function can be found in gene regulatory networks. For instance, small scale repetitive patterns associated with reduced response time of autoregulated genes or bistability for feedback control (Lee et al., 2002), or feed-forward loops, which account for processing external signals (Shen-Orr et al., 2002).

The work of Barabási and Albert (1999) sheds light on the underlying principles of network structure and identifies two simple laws, which govern the generation of complex networks. First, these networks result from a growth process, i.e. the number of participating nodes is not constant but increases over time. For instance, as an organism evolves over an evolutionary time scale, the number of genes and hence the size of the gene regulatory network increases. Second, new nodes tend to link to already highly-connected nodes, a phenomenon known as *preferential attachment*. Based on gene duplication (Teichmann and Babu, 2004), a more biological explanation for the structural properties of biological networks is given by Bhan et al. (2002).

1.5 Thesis outline

This work studies several problems concerning metabolic network modelling and provides means to bridge the gap between different levels of abstraction. First, problems regarding the automatic reconstruction of genome-scale network structures are addressed. Second, the network structure is used to analyze biosynthetic and dynamic capabilities of metabolic networks. Finally, a strategy is presented, which helps to develop kinetic models even for large networks by identifying crucial reactions.

Chapter 2 covers general aspects of automated reconstruction of genome-scale metabolic networks and analyses the computational complexity of this problem (published in the present form as Nikoloski et al. (2008a)). Chapter 3 examines structural methods to infer biosynthetic capabilities of metabolic networks (Borenstein et al., 2008). Especially the problem of finding a minimal set of substrates that have to be supplied to a metabolic network in order to produce certain predefined target compounds is elucidated (published as Nikoloski et al. (2008b)). In Chapter 4, chemical reaction network theory is used to compare different models of the Calvin cycle regarding their capability to obtain multiple steady states. Although this method is based on the assumption of mass-action kinetics, valuable insights about multistability can be obtained already from the structure of metabolic networks. Chapter 5 introduces structural kinetic modelling, a sampling technique to infer kinetic properties without knowing the explicit rate equations (published as Grimbs et al. (2007a)). Furthermore, this method provides a ranking of reactions according to their impact on stability of a given steady state. This method is applied to and evaluated on a model of human erythrocytes. Chapter 6 compares strategies for kinetic modelling and uses the method presented in the previous chapter to obtain models of suitable accuracy with moderate effort. This approach is exemplified by hybrid models for hepatocytes and erythrocytes (published as Bulik et al. (2009a)). Finally, a general conclusion is given in Chapter 7.

Chapter 2

Metabolic networks are NP-hard to reconstruct

High-throughput data from various omics and sequencing techniques have rendered the automated metabolic network reconstruction a highly relevant problem. Our approach reflects the inherent probabilistic nature of the steps involved in metabolic network reconstruction. Here, the goal is to arrive at networks which combine probabilistic information with the possibility to obtain a small number of disconnected network constituents by reduction of a given preliminary probabilistic metabolic network. We define *automated metabolic network reconstruction* as an optimization problem on four-partite graphs (nodes representing genes, enzymes, reactions, and metabolites) which integrates: (1) probabilistic information obtained from the existing process for metabolic reconstruction from a given genome, (2) connectedness of the raw metabolic network, and (3) clustering of components in the reconstructed metabolic network. The practical implications of our theoretical analysis refer to the quality of reconstructed metabolic networks and shed light on the problem of finding more efficient and effective methods for automated reconstruction. Our main contributions include: a completeness result for the defined problem, polynomial-time approximation algorithm, and an optimal polynomial-time algorithm for trees. Moreover, we exemplify our approach by the reconstruction of the sucrose biosynthesis pathway in *Chlamydomonas reinhardtii*.

2.1 Introduction

The availability of fully sequenced genomes, coupled with the development of effective bioinformatics methods for gene annotation, offers the possibility for reconstructing entire metabolic networks. The problem of metabolic network reconstruction is clearly related to the precise understanding of the genetic basis for metabolic organization and regulation. While preliminary metabolic networks have already been reconstructed solely based on gene annotation (Ma and Zeng, 2003; Romero et al., 2005; Reed et al., 2006), this process may discard some available information: It is often the case that the function of genes is determined by the highest similarity obtained through comparison to other already annotated organisms. However, in such practice, alternative gene functions may result in smaller but still significant similarity (Green and Karp, 2004).

The following steps are crucial for reconstructing a metabolic network based on the genome of a given organism: (I) establishing gene models, (II) sequence similarity search (*e.g.*, BLAST), (III) gene product annotation, with the help of available enzyme databases (*e.g.*, KEGG, Expasy, Brenda), (IV) enzyme-reaction association, with the help of reaction databases (*e.g.*, KEGG LIGAND (Goto et al., 2002)), and (V) pathway mapping. The outcome from steps (I)-(IV) results in a preliminary metabolic network given by sets of: enzyme-gene relationships, reaction-enzyme relationships, reactions, and metabolites, which make up the metabolic network. Finally, in step (V), the identified reactions are mapped onto a collection of pathways (*e.g.*, from KEGG (Kanehisa et al., 2004) or MetaCyc (Caspi et al., 2006)) to obtain a raw metabolic network. Currently, this network is taken to be the reconstructed metabolic network for the organism whose genome is considered as input to the process.

The preliminary metabolic network is furthermore carefully calibrated by the experimental results reported in literature. This iterative manual process can be labor-intensive and time-consuming. Even for fairly simple microorganisms such as *Escherichia coli* (Reed et al., 2003) and *Saccharomyces cerevisiae* (Duarte et al., 2004), the metabolic networks reconstructed with high-quality have taken years to assemble. There are ongoing research efforts to use the same reconstruction methodology on the human genome, with variable success directly related to the complexity of this task (Romero et al., 2005; Ma et al., 2007).

Assembling the preliminary metabolic network often employs prediction-based bioinformatics methods and is, therefore, *probabilistic*. For instance, gene annotation is based on prediction, yielding enzyme-gene relationships explicitly weighted with the accuracy of prediction (*e.g.*, in the range $(0, 1]$). Moreover, there may be an ambiguous relationship between enzymes and reactions in the reaction databases, in the sense that an enzyme in a given organism may not catalyze a reaction which is catalyzed by the same enzyme in another organism (Wu et al., 2006; Wang et al., 2006). Hence, in the absence of precise human-curated knowledge, the enzyme-reaction relationships for a given organism are also weighted with the accuracies of their computational predictions. A threshold can be imposed to include the most relevant relationships. However, it is often the case that only the highest-value predictions are included in the reconstructed metabolic network. Therefore, the possibility that, for instance, a given gene codes for more than one enzyme or that an enzyme catalyzes more than one reaction is often neglected. Finally, in step (V), only a portion of a given pathway may be included, resulting in a disconnected metabolic network. This shortcoming of the reconstruction process points at necessary clustering of connected reactions to show their functional relationships.

Therefore, we can conclude that preliminary metabolic networks, taken as the reconstructed counterparts, are often incomplete, since a large portion of available information is ignored by overlooking its probabilistic nature. As a result, much manual validation and correction is needed. To allow for inclusion of information with varying accuracy of prediction, here, we address the problem of automated reconstruction of metabolic networks. We believe that there is a need for formal definition of metabolic network reconstruction, whose analysis may result in new insights of how to approach and resolve the problem at hand.

The existing approaches for reconstructing metabolic networks include (constraint-based) elementary modes (Stelling et al., 2002) and flux balance analysis (FBA) (Edwards and Palsson, 2000a; Price et al., 2003). Elementary modes correspond to the smallest subnetworks that can operate in steady state. FBA uses linear programming to obtain a single (not necessarily unique) solution to an optimization problem (*e.g.*, with growth per substrate uptake as a function to be maximized) and can be used in the analysis of specific cell behaviors. On the other hand, elementary modes allow for investigation of the space of all meaningful physiological states, and can be used to define control-effective fluxes via their respective efficiencies (relating a mode's output to the cost for establishing the mode). In addition, elementary modes can address cellular regulation and can characterize some aspects of cellular behavior from metabolic network topology. We point out that both approaches are structural in the sense that they require the topology of a putative metabolic network together with its stoichiometry in order to elucidate mutant phenotypes, analyze network robustness, and to quantitatively predict functional features of genetic regulation. The approach described here aims at metabolic network reconstruction which satisfies the biochemical balance constraints and relies solely on graph-theoretic concepts.

Contributions and organization. We define the automated reconstruction of metabolic networks as an optimization problem in Section 2.2. Our approach is exemplified in Section 2.3 by the reconstruction of the sucrose biosynthesis pathway for *Chlamydomonas reinhardtii*. The results regarding the hardness of the problem are presented in Section 2.4. The practical implications of our theoretical analysis refer to the quality of reconstructed metabolic networks and shed light on the problem of finding more efficient and effective methods for automated reconstruction. Such methods can result in biologically relevant networks that may speed up the computational analysis, but still require considerable effort for experimental validation. An optimal polynomial-time algorithm for the problem restricted to trees is described and analyzed in Section 2.5, while approximation results are shown in Section 2.6.

2.2 Problem definition

For the purpose of defining the formalism for metabolic network reconstruction, we require the assembly of a preliminary metabolic network, which we call *raw metabolic network*. One technique for obtaining the raw metabolic network includes the steps described in Section 2.1: after genes have been determined in step (I) and their similarity to genes from other organisms has been established in step (II), the function of genes can be assigned in step (III). Step (III), in fact, results in a set of enzymes that can catalyze a set of reactions. By using existing pathway databases, one can then identify to which pathway(s) the found reactions belong. Often, the existing gene annotation may cover a portion of the pathways, *i.e.*, only few of the pathways' reactions are initially included in the raw metabolic network. Other reactions may be included based on different approaches: usage of experts' knowledge or taxonomic distance between enzymes on pathways from the used database (Peregrin-Alvarez et al., 2003). In order to allow for stoichiometrically balanced reconstructed network, the raw metabolic network should not include stoichiometrically unbalanced reactions proceeding from public databases. Moreover, based on metabolomic studies, the raw metabolic network can be extended to include previously not present metabolites. In the latter case, the raw metabolic network can include reactions that use these metabolites together with the corresponding enzymes and known genes.

Here, the raw metabolic network is represented by a graph G, irrespective of the methods used in its assembly. The node set of G is a union of pairwise disjoint node sets (partitions) representing: genes, enzymes, reactions, and metabolites. The edge set of G is a union of pairwise disjoint edge sets describing gene-enzyme, enzyme-reaction, and reaction-metabolite relationships. Each edge has a weight, representing the accuracy of prediction for a particular relationship (or, its certainty). We assume that the accuracy is given by a real number from the interval $(0, 1]$. Some possible methods to obtain the edge-weights include transformation of the E-value or the BLAST score on the interval $(0, 1]$ (Green and Karp, 2004) or usage of recent databases for biochemical substructures and prediction of reaction-metabolite relationships (Kotera et al., 2008). However, the formulation of our problem and the proposed approximations are independent of the employed methods for the edge-weights. Edges of weight 0 are not included in the graph G.

In addition, if a reaction is spontaneous or is included without gene evidence, the raw metabolic network is extended to include dummy gene and enzyme nodes corresponding to the reaction. We point out that reactions included from public databases may not be chemically balanced (Poolman et al., 2006). In this case, the raw metabolic network may still include some of the chemically unbalanced reactions upon an expert's opinion and in accordance with biochemical knowledge. However, the formalism presented here does not aim at resolving this known issue of the publically available human-curated databases.

We have chosen the four-partite graph representation as it provides the minimum number of different entities sufficient for metabolic network reconstruction. The included entities are sufficient for our task since the measurement of their respective quantities (*e.g.*, gene expression, fluxes, or metabolite concentrations) yields the minimum effort for validation of the reconstructed network. The graph-theoretic representation employed here can be easily extended to include other biologically relevant entities, such as mRNA, by providing an additional node-partition for the new entity. However, we point out that our goal does not include elucidation of gene-regulatory relationships and, thus, mRNA is presently excluded.

We assume that the biosynthetic capabilities of an organism are determined by the connectedness of its metabolic pathways; therefore, the reconstructed metabolic network we extract from the raw network is based on the criterion of connectedness. In this respect, our approach to reconstruction can be regarded as a *reduction* of a given raw metabolic network to obtain the topology (structure) of a metabolic network for an investigated organism.

Automated metabolic network reconstruction is then the problem of extracting a subgraph, H, from a given raw metabolic network, G. Clearly, there are different types of subgraphs that can be extracted depending on the optimization criterion used. Here, we employ connectedness, clustering, and high accuracy of the included relationships among metabolites, enzymes, and reactions as biologically relevant criteria that subgraph H should satisfy.

We use the weakest definition of clustering via connectedness, already found as relevant for metabolic network reconstruction from pathway mapping (Duarte et al., 2007). By this definition, a node set represents a cluster if and only if it is connected. The weight of a cluster is simply the sum of edge-weights included in the cluster. Due to sparse gene annotation with the increasing complexity of an organism, one would also like to impose a bound (threshold) on the weight of the clusters as the number of reactions that may not be connected to the rest of the network (via a path from a gene to a metabolite) becomes larger. Maximization based on the weight of the cluster would be of no use for the purpose of metabolic network reconstruction, as the entire raw metabolic network is the solution to such optimization (analogous to the current practice in metabolic network reconstruction). The idea that accuracy should be maximized could be included in the formulation of our problems through a bound that must be satisfied by each of the clusters included in a solution. On the other hand, minimization leads to a smaller number of clusters and, therefore, implies a network of higher overall connectedness. We point out that the choice of the bound can result in a grouping of small clusters which must be connected (to signify functional relations), thus rendering the reconstruction as an iterative process starting with a small value for the bound.

To address the hardness of automated metabolic network reconstruction, we define the general problem and a biologically meaningful variant of automated metabolic network reconstruction.

In the *general problem*, we require that H: (1) contains all nodes (genes, enzymes, reactions, and metabolites) and (2) has a small number of clusters of weight at least as large as some imposed threshold. The problem of automated metabolic network reconstruction is then that of finding a weight-constrained generalized edge cover, since all genes and reactions (with their enzymes and metabolites) should be present in the reconstructed network, *i.e.*, subgraph H.

It is often the case that a subset of metabolites present in an organism can be identified by metabolomic techniques (Bölling and Fiehn, 2005). Knowledge about genes involved in metabolism is readily available and may imply the subset of genes that should be reflected in the reconstructed metabolic network. Therefore, in the *biologically meaningful problem* we require that H: (1) contains this (possibly empty) subset of genes, to account for metabolic functions, (2) includes all identified metabolites, and (3) contains those reactions and, consequently, the corresponding activating enzymes, that will ensure connectedness and clustering of the network in the sense of the general framework described above. As a result, the biologically meaningful problem is a version of the general problem restricted to a four-partite graph with some additional constraints, explained in the formulation, shown below.

Given a graph G, an edge cover of G is a subset of edges S such that each node in G is incident on an edge in S. In combinatorial terms, our problem has the following formulation: Given a weighted graph G, we seek to find an edge cover S, such that each connected component of the induced subgraph $G[S]$ has weight at least B, where B is a given weight bound. An edge-induced subgraph is a subset of edges of G together with any nodes that are incident on the given edges. The reconstructed metabolic network is given by the induced subgraph $G[S]$.

Formally, we address the following problem: Let $G = (V, E)$ be a connected graph and let $w : E(G) \to \mathbb{R}^+$, so that $w(e)$ denotes the weight of an edge $e \in E(G)$. For a set of edges S, let $V(G[S])$ denote the set of nodes in the graph induced by S in G.

GENERAL AUTOMATED METABOLIC NETWORK RECONSTRUCTION (GAMNR)

INSTANCE : Given a weighted graph $G = (V, E)$, with edge-weights in the range $(0, 1]$, and a positive bound B.

PROBLEM : Find an edge cover $S \subseteq E(G)$ such that the weight of each connected component H_i of $H = G[S]$ is at least B.

MEASURE : Weight of S, denoted by $\alpha^w_{1,B}$.
(min)

For the biologically meaningful problem, we have the following formulation:

BIOLOGICALLY MEANINGFUL GAMNR (BMAMNR)

INSTANCE : Given a weighted four-partite graph $G = (V, E)$, with edge-weights in the range $(0, 1]$, $V(G) = M \cup R \cup Z \cup N$, ($M$, set of metabolites, R, set of reactions, Z, set of enzymes, and N, set of genes), a subset of genes $N' \subset N$, a subset of metabolites $M' \subset M$, and a positive bound B.

PROBLEM : Find a subset of edges $S \subseteq E(G)$ such that the following five conditions are satisfied:

1. For every $u \in N'$, there exists an edge $e \in S$, such that e is incident on u, *i.e.*, every gene in the given subset N' is in the reconstructed network;

2. For every $u \in M'$, there exists an edge $e \in S$, such that e is incident on u, *i.e.*, every metabolite in the given subset M' is in the reconstructed network;

3. For every $e \in S$, there exists a path P_{xy} in S on four nodes that passes through e, such that $x \in N$ and $y \in M$, *i.e.*, each reaction included in the solution consumes or produces at least one metabolite and is catalyzed by at least one enzyme produced by at least one gene;

4. For every $w \in R \cap V(G[S])$ and any $x \in M$, x, a neighbor of w, the edge incident on x and w is in S, *i.e.*, if a reaction is included in the solution, then all of its metabolites are also in the solution;

5. The weight of each connected components H_i of $H = G[S]$ is at least B.

MEASURE : Weight of S, denoted by $\gamma^w_{1,B}$, such that $\gamma^w_{1,B} > 0$.
(min)

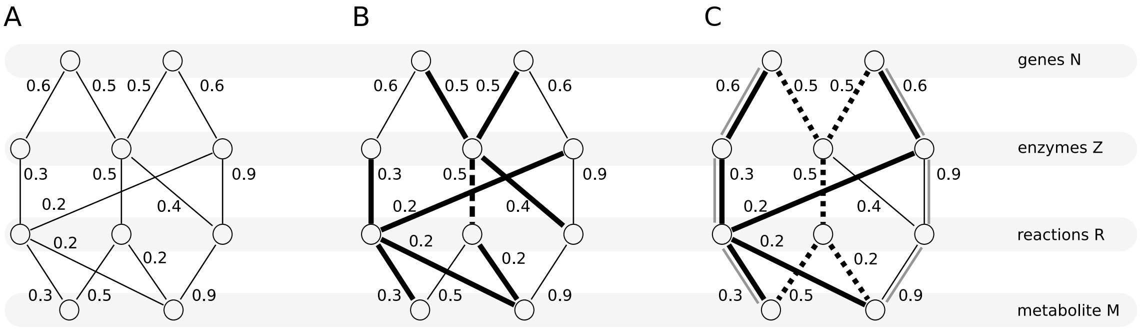


Figure 2.1: Illustration of GAMNR and BMAMNR. Contrived raw metabolic network with edge weights between 0 and 1 is shown in **A**. The nodes represent genes, enzymes, reactions and metabolites (from top to bottom). The optimal solution for GAMNR (with bound $B = 1.2$) on this example graph is highlighted in **B** by bold edges. The optimal solution consists of two separate components, which can be extended to the minimum spanning tree by adding the dotted edge. For BMAMNR, with $N' = N$ and $M' = M$ and a bound $B = 1.2$, two distinct optimal solutions can be found (**C**). Both solutions (bold edges and dotted edges) consist of a single component each. If just the highest value predictions for gene-enzyme relations are included, the solution would be given by the gray edges, which lead to a disconnected network.

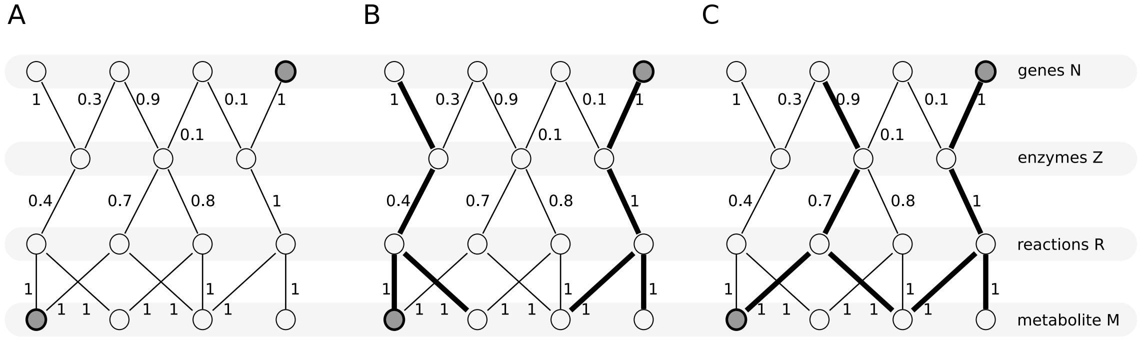


Figure 2.2: Choosing a bound in BMAMNR. The notation is the same as in Figure 2.1. Contrived raw metabolic network with edge weights between 0 and 1 is shown in **A**. For the instance of BMAMNR, N' and M' consist of the nodes marked in gray, which are to be necessarily included in the solution. Note that a solution may as well contain other gene and metabolite nodes. The optimal solution of weight $\gamma^w_{1,B} = 7.4$ for bound $B = 3$ is disconnected, depicted with bold edges in (B). Increasing the bound to $B = 7.5$ leads to a connected optimal solution of weight $\gamma^w_{1,B} = 7.6$, presented in (C).

Inclusion of condition 4 in the definition of BMAMNR together with the stoichiometrically balanced construction of the raw metabolic network guarantees that any reaction that is in the solution always appears with all of its metabolites (*i.e.*, educts and products). As a result, a network reconstructed according to the conditions of BMAMNR is guaranteed to be stoichiometrically balanced.

Figure 2.1 illustrates our definitions of GAMNR and BMAMNR by depicting the optimal solutions on an artificial raw metabolic network. Figure 2.2 demonstrates the influence of the value for the bound B on a desired property of the reconstructed network (*e.g.*, connectedness). The connectedness arises as a result of the increase in the value of B from 3 to 7.5. We envision that a practically applicable heuristic would start from a small value for B followed by an iterative increase of B based on expertise. Note that an edge-weight of 1 does not necessarily imply inclusion

	CHLRE_18029	CHLRE_78737	CHLRE_81483	CHLRE_119219	CHLRE_149366	CHLRE_176209
5.4.2.2	1	0.33	1	1	0.34	0.984
2.7.7.9	0.41	1	0.41	0.04	0	0.83
2.4.1.14	0.08	0	0.31	0	0	1
2.4.1.13	0.11	0	0.38	0	0	0.87
3.1.3.24	0	0.04	0	0	1	0.2

Table 2.1: Transformed E-values on the interval from 0 to 1 for six genes from *Chlamydomonas reinhardtii* and five possible enzymes.

of the edge in the solution. Moreover, the weights of edges incident on a given node do not sum up to 1 as the weight represents accuracy of a relationship.

In Section 2.4 we establish that GAMNR and BMAMNR are NP-hard for any bound B, while in Section 2.6 we give a polynomial-time approximation algorithm with a factor of 2 for the unweighted case and a factor that depends on the bound B and the maximum edge-weight in the weighted case for the GAMNR problem. Complexity results pertinent to BMAMNR are related to the complexity of the GAMNR problem. We see the design of approximation algorithms for GAMNR as a first step towards obtaining efficient approximation algorithms for BMAMNR.

2.3 Reconstruction of sucrose biosynthesis pathway in *Chlamydomonas reinhardtii*

To illustrate the usefulness of the formalism presented in Section 2.2, we show the process of reconstructing the sucrose biosynthesis pathway in *Chlamydomonas reinhardtii* as previously described in May et al. (2008). The raw metabolic network for the sucrose biosynthesis pathway in *Chlamydomonas reinhardtii* was obtained by first conducting a sequence similarity search for six genes: CHLRE_18029, CHLRE_78737, CHLRE_81483, CHLRE_119219, CHLRE_149366, and CHLRE_176209. Table 2.1 shows the transformed E-values on the interval from 0 to 1. Five candidates for the reaction partition were identified, including: enzymatic reaction 5.4.2.2 catalyzed by phosphoglucomutase, 2.7.7.9 catalyzed by glucose-1-phosphate uridylyltransferase, 2.4.1.14 catalyzed by sucrose-phosphatase synthase, 3.1.3.24 catalyzed by sucrose-phosphatase, and 2.4.1.13 catalyzed by sucrose synthase. For the purpose of illustrating the approach, we also include the spontaneous reaction 5.1.3.15 to the raw metabolic network. Since a spontaneous reaction should appear in any solution, we add a dummy gene node (CHLRE_dum) connected via an edge to 5.1.3.15. The raw metabolic network is presented in Figure 2.3. All weights of edges between the enzyme partition (Z) and the reaction partition (R) are 1. The same holds for the weights of edges between the reaction partition (R) and the metabolite partition (M). For the instance of BMAMNR, we set $N' = N$, *i.e.*, all genes, due to expertise, are to be included in the solution (Figure 2.4).

For the bound $B = 22$, the optimal solution of weight $\gamma^w_{1,B} = 22.83$, found by exhaustive search, is connected and includes all edges incident on reactions r_1 to r_5 and the edges of minimum weight incident on each of the seven genes. Note that r_6 is not in the optimal solution. This solution represents exactly the sucrose synthesis pathway proposed by May et al. (2008). The optimal solution of weight $\gamma^w_{1,B} = 28$ for the instance of BMAMNR with bound $B = 28$ is given by the bold edges as presented in Figure 2.3. For the same bound, a solution for BMAMNR of weight $\gamma^w_{1,B} = 28.06$ that includes all enzymes is highlighted by dashed edges, also shown in Figure 2.3.

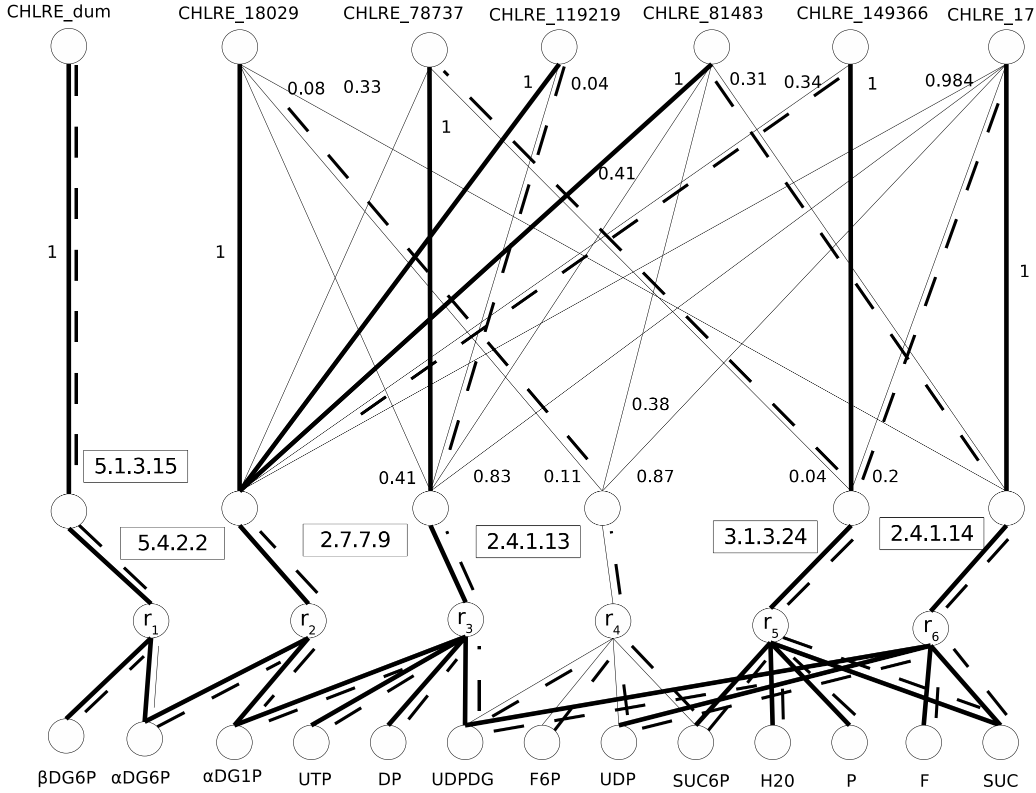


Figure 2.3: Reconstruction of sucrose biosynthesis pathway in *Chlamydomonas reinhardtii*. The raw metabolic network, G, with edge weights between 0 and 1 and nodes representing genes, enzymes, reactions and metabolites (from top to bottom). All weights of edges between the enzyme partition (Z) and the reaction (R) partitions are 1. The same holds for the weights of edges between the reaction partition (R) and the metabolite partition (M). For the sake of clarity, the weight of these edges are not shown here. The reaction associated with enzyme 5.1.3.15 is spontaneous; its inclusion in any solution requires adding a dummy gene node CHLRE_dum. Here, $N' = N$, *i.e.*, all genes should be included in the solution. The optimal solution, found by exhaustive search, of weight $\gamma^w_{1,B} = 28$ for the instance of BMAMNR with bound $B = 28$ is given by the bold edges. For the same bound, a solution for BMAMNR of weight $\gamma^w_{1,B} = 28.06$ that includes all enzymes is highlighted by dashed edges. Both solutions (bold edges and dashed edges) consist of a single component each.

2.4 Complexity of automated metabolic network reconstruction

There are several problems closely related to the GAMNR problem: The most general problem–set cover–is that of computing a minimum weighted subfamily $\mathcal{F}'$, given a family $\mathcal{F}$ of weighted subsets of a base set U, such that every element of U is covered by some subset in $\mathcal{F}'$. The version with set sizes bounded above by a constant k is known to be NP-hard (Karp, 1972) as well as MAX SNP-hard (Papadimitriou and Yannakakis, 1991). The specializations of the set cover problem for (weighted) graphs take the form of (weighted): edge cover, node cover, node dominating set and edge dominating set problems and their restrictions in regards to connectedness. Of these, only the weighted edge cover problem is known to be solvable in polynomial time (Edmonds and Johnson, 1970; Murty and Perin, 1982; Pulleyblank, 1996). Moreover, there is a clear connection between edge cover and matchings (Gallai, 1959) as well as generalized matching and the corresponding edge cover (Kirkpatrick and Hell, 1978) — a result we will repeatedly use in the next section.

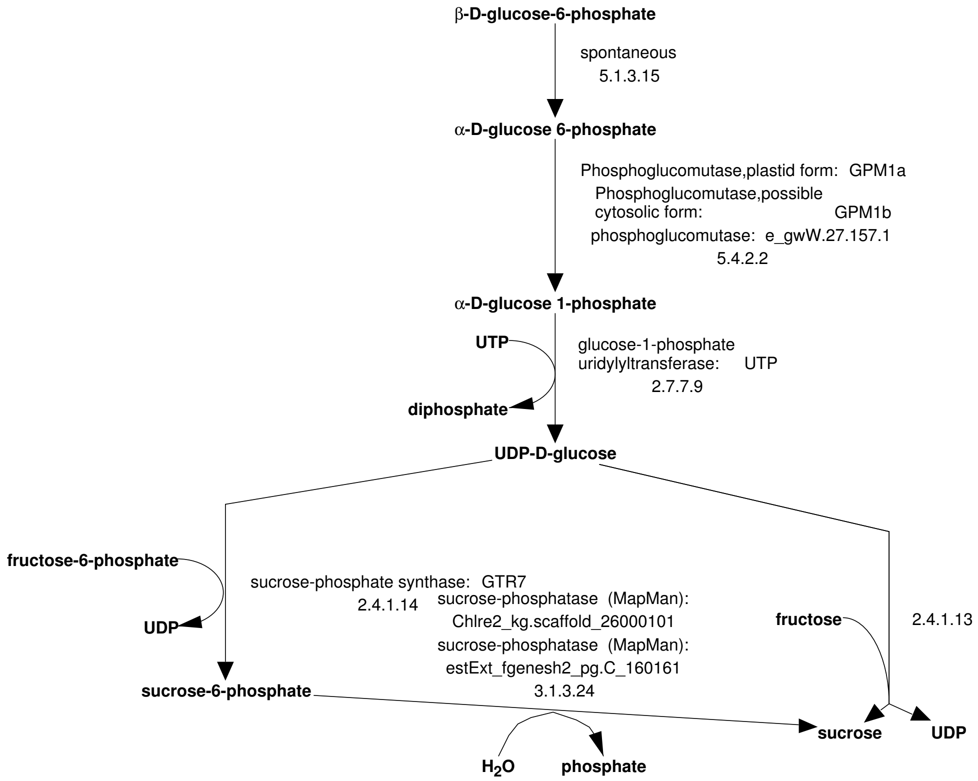

β-D-glucose-6-phosphate
spontaneous
5.1.3.15
α-D-glucose 6-phosphate
Phosphoglucomutase,plastid form: GPM1a
Phosphoglucomutase,possible cytosolic form: GPM1b
phosphoglucomutase: e_gwW.27.157.1
5.4.2.2
α-D-glucose 1-phosphate
UTP
glucose-1-phosphate uridylyltransferase: UTP
2.7.7.9
diphosphate
UDP-D-glucose
fructose-6-phosphate
sucrose-phosphate synthase: GTR7
2.4.1.14
sucrose-phosphatase (MapMan): Chlre2_kg.scaffold_26000101
sucrose-phosphatase (MapMan): estExt_fgenesh2_pg.C_160161
3.1.3.24
UDP
fructose
2.4.1.13
sucrose-6-phosphate
sucrose
UDP
H_2O
phosphate

Figure 2.4: Sucrose biosynthesis pathway in *Chlamydomonas reinhardtii*. General sucrose biosynthesis pathways. In *Chlamydomonas reinhardtii*, the sucrose biosynthesis is more likely to operate via the reactions catalyzed by sucrose-phosphate synthase 2.4.1.14 and sucrose-phosphatase 3.1.3.24 as proposed by May et al. (2008). The diagram is extracted from MetaCyc (Caspi et al., 2006). It shows all metabolites, reactions with their corresponding EC numbers, and the annotated genes.

The generalized matching problem can be cast as a weighted tree packing problem, called B-tree packing, first discussed in Kirkpatrick and Hell (1978):

WEIGHTED TREE PACKING
INSTANCE : Given a weighted graph $G = (V, E)$, with edge-weights in the range $(0, 1]$, and a positive bound B.
PROBLEM : Find a partition of $V(G)$, $P = \{P_1, \ldots, P_k\}$, such that each $G[P_i]$, $1 \leq i \leq k$, contains a tree of weight at least B.
MEASURE : Number of subsets in the partition, denoted by $\beta^w_{1,B}$.
(max)

Given an unweighted graph G, the problem of finding a B-tree packing of maximum size is already known to be NP-hard for $B \geq 2$ (Kirkpatrick and Hell, 1978). For $B = 1$, the generalized matching problem is equivalent to finding a maximum matching in G. According to the notational convention from complexity theory, we will use B whenever we argue the complexity of GAMNR and BMGAMNR. In the analysis of our approximation algorithms, B is substituted with t, *i.e.*, $t = B$.

We review the following known Gallai type result connecting the maximum size of t-tree packing to finding a minimum edge cover which induces connected components of size at least t

from Fernau and Manlove (2006):

Theorem 2.4.1. Let $G = (V, E)$ be a connected graph, where $n = |V|$, and let $1 \leq t \leq n - 1$. Then, for the minimum edge cover which induces a graph with connected components of size at least t, $\alpha_{1,t}$ and the maximum size of a t-tree packing $\beta_{1,t}$ the following holds:

$$\alpha_{1,t} + \beta_{1,t} = n.$$

From the previous theorem and the result by Kirkpatrick and Hell, we have the following corollary:

Corollary 2.4.2. The problem GAMNR is NP-hard for any positive B.

Proof. The restriction $B = 2$ and G a connected unweighted graph is NP-hard. GAMNR with all weights 1 and bound $B = 2$ is then also NP-hard by restricting it to the unweighted case. By rescaling of weights in the restriction, GAMNR in which all edge-weights are in the interval $(0, 1]$ and $B > 0$ is NP-hard. □

Corollary 2.4.3. The problem GAMNR is NP-hard for any positive B even for bipartite graphs of maximum degree 3.

Proof. The K-PATH PARTITION problem is NP-hard for bipartite graphs of maximum degree 3 for any $k \geq 3$ (Monnot and Toulouse, 2007). Therefore, the problem of packing two-trees as an instance of WEIGHTED TREE PACKING is also NP-hard on bipartite graphs of maximum degree 3. □

We can now establish the following connection between BMAMNR and GAMNR, stated in:

Theorem 2.4.4. The problem BMAMNR is NP-hard for any positive B.

Proof. We establish a restriction from GAMNR to BMAMNR, for the unweighted case, in the following way. Given an instance of GAMNR on a bipartite graph G of maximum degree 3, we name one of the partitions R and the other Z. We extend graph G to get G' by adding two more subsets of nodes M and N, such that $|M| = |R|$ and $|N| = |Z|$. Let $M' = M$ and $N' = N$. Furthermore, let the nodes in M and R (respectively, N and Z) be ordered $m_1, \ldots, m_l$ and $r_1, \ldots, r_l$, where $l = |M| = |R|$. We then add l edges e_i, $1 \leq i \leq l$, to G' such that e_i is incident on m_i and r_i. In a similar fashion, we add $n - l$, $p = |Z| = |N|$ between the ordered nodes of Z and N. Finally, we set the bound of BMAMNR to $B' = 2B + 1$. The construction of G' and B' can be carried out in polynomial time. Now, we can find a tree packing in G of size k with the given bound B if and only if we can find an optimum tree packing in G' of same size with the given bound $B' = 2B + 1$. Moreover, the edge cover in the first case will be of size $n - k$, while in the second will be $2n - k$ (we add n nodes/edges). It is trivial to check that the edge cover of G' satisfies the five conditions of the BMAMNR problem and that all nodes from M' and N' are in the solution. By the construction, graph G' is bipartite and of maximum degree 3. □

If N' and M' were not included in the definition of BMAMNR, the problem would remain NP-hard. The proof would have to be modified to add only one gene node connected to all enzymes, include only one metabolite node connected to all reactions, and to assign appropriate weights for the added edges. Nevertheless, applications of the BMAMNR problem in real-world reconstruction of metabolic network necessitate the addition of previous knowledge and experimental data from omics techniques, reflected in the present formulation.

Corollary 2.4.5. The problem BMAMNR is NP-hard for any positive B even on graph of maximum degree 3.

Proof. It follows from Corollary 2.4.3 and the reduction used in Theorem 2.4.4. □

2.5 Polynomial-time algorithm for (edge-weighted) trees

For a t-tree packing problem on a given unweighted tree T, let k be the optimal (maximum) size of the collection P. Moreover, let T' be the tree obtained by contraction of edges that belong to any T_i in P. By counting the number of nodes and the number of edges, we have:

$$k = \frac{n-x}{t+1}, \tag{2.1}$$

$$k = \frac{n-1-x'}{t}, \tag{2.2}$$

where x is the number of nodes not included in any $T_i \in P$ and x' is the number of edges in T'. An upper bound on k can then be determined by finding the smallest pair of positive integers (smallest in the lexicographic ordering of the ordered pairs (x, x')). We then have the theorem:

Theorem 2.5.1. The smallest pair in the lexicographic ordering of the solutions (x, x') of the Diophantine equation:

$$(t+1)x' - tx = n - t - 1,$$

determines an upper bound to the optimum k.

Proof. The proof follows directly by equating the two expressions for k and observing that for smallest x and x' as solutions of the Diophantine equation k is maximized. The linear Diophantine equation does not have solutions if and only if the number $(n-t-1)$ is not a multiple of the greatest common divisor of $t+1$ and t, as coefficients of the equation. Since t and $t+1$ are two consecutive integers, $n-t-1$ is always a multiple of the gcd(t, t+1) = 1. Therefore, the two numbers x and x' can be determined by the extended Euclidean algorithm. □

The idea of iterative contraction of leaves and careful bookkeeping can also be extended to devise an optimal algorithm for (weighted) trees. Algorithm 1 determines an optimal t-tree packing of a (weighted) tree. Given a (weighted) tree T, let $l(u)$ denote the label of node u. The algorithm takes T and a number t, $\phi(T) \leq t \leq w(T)$, as input and returns k, the maximum size of a t-tree packing in T. The initialization phase consists of lines 1 – 4, where k is set to 0 (*i.e.*, the t-tree packing is initially empty) and each node $u \in T$ is given an initial label $l(u) = 0$. The main idea of this dynamic algorithm is to minimize, at every step, the weight of edges not yet included in the t-tree packing. In lines 6 – 11, we check whether there is a leaf u with label smaller than t in order to decide whether or not to increase the number of trees packed in T. The label denotes the weight of edges in the subtree rooted at (and contracted onto) u. If there is such a node, one can do the necessary alteration of the parent's label (line 9) and contract one more edge (line 10), since each leaf has one parent to which it is connected via an edge. If all leaves are with labels greater than or equal to t, we choose a leaf (together with its subtree) that minimizes the weight of unused edges for inclusion in the t-tree packing (lines 12 – 22). Note that after each iteration of the while loop (lines 5 – 23), in the unweighted case, the result is a tree with at least $n-t-1$ nodes. We abuse the notation, and use "$-$" to denote contraction of an edge incident on a leaf x and removal of the resulting loop-edge.

Algorithm 1: Size of optimum t-tree packing of a tree

```
Input: $T$, tree on $n$ nodes
$t$, $\phi(T) \leq t \leq w(T)$
Output: $k$, optimum size of $t$-tree packing
$k \leftarrow 0$
foreach node $u \in T$ do
    $l(u) \leftarrow 0$
end
while $T$ is not the empty graph do
    if there is a leaf $u$, $l(u) < t$ then
        $p \leftarrow$ parent of $u$
        foreach leaf-child $v$ of $p$ do
            $l(p) \leftarrow l(p) + l(v) + w(e_{pv})$
            $T \leftarrow T - \{v\}$
            /* contracting the edge $e_{pv}$ */
        end
    else
        $m \leftarrow$ minimum of all leaf-labels
        $x \leftarrow$ leaf of label $m$
        if $w(T - \{x\}) \geq t$ then
            $T \leftarrow T - \{x\}$
            $k \leftarrow k + 1$
        else
            $T \leftarrow \emptyset$
            $k \leftarrow k + 1$
        end
    end
end
```

Proof of correctness for Algorithm 1: A *packing ordering of node-disjoint subtrees* (PODS) of a tree T for a given integer t is a sequence $C = (C_1, C_2, ..., C_k)$, such that: (1) $\forall i$, $w(G[C_i]) = \sum_{e \in E(G[C_i])} w(e) \geq t$ and exactly $(k-1)$ edges from T not included in C, *i.e.*, $|E(T) - T[\cup_{l=1}^{k} C_l]| = k - 1$; (2) for each edge e in the tree $\tilde{T}[C]$, obtained by contraction of the elements of C, there exist i and j so that the two endpoints of each edge correspond to C_i and C_j, and (3) among all leaves in tree obtained from the induced subtree $\tilde{T}[\bigcup_{l=i}^{n} C_l]$, C_i minimizes the difference $w(G[C_i]) - t$. For each i, $1 \leq i \leq k$, the difference $w(G[C_i]) - t$ is defined as the cost of the subtree C_i. A packing ordering of node-disjoint subtrees is called maximal if no element of the sequence C can be partitioned to obtain a sequence C' with more elements. To show that a maximal packing ordering of node-disjoint subtrees (MPODS) for a given T and an integer t determines an optimal (maximum) solution, we use a proof by contradiction.

First, we need to establish some supporting arguments. Note that any sequence that satisfies conditions (1) – (3) is output of the algorithm. Moreover, for a given sequence C of size k, T can be rooted by planting $\tilde{T}[C]$ at a root that corresponds to C_k.

Suppose that k is the optimum given by the algorithm, and let C be its corresponding sequence. For the same tree T, let C' be an optimal sequence of node-disjoint subtrees that satisfies condition (1) and (2), but does not satisfy condition (3) and let C' be also of size k.

We claim that the sequence C' can be reordered to guarantee that $\bigcup_{l=i}^{n} C_l$ always induces a tree. One such ordering is given by the postorder traversal of the tree $\tilde{T}[C']$ rooted at any node.

We can obtain C from C' by the following steps: (a) Reorder C' by the postorder traversal of $\tilde{T}[C']$ rooted at any node; (b) For each edge $e \in \tilde{T}[C]$, determine C'_i to which e belongs; (c) For each C_i, determine the subtree C'_j that maximizes the weight of the intersection $C_i \cap C'_j$, $1 \leq j \leq k$. (d) Starting in a bottom-up fashion–from the leaves to the root, for each C'_j that contains an edge e from $\tilde{T}[C]$: (d.i) update C'_j to be the component obtained after the removal of e that maximizes the intersection with the corresponding C_i (found in step 3), (d.ii) add the other component to the parent of C'_j in $\tilde{T}[C']$, (d.iii) remove edge e from C'_j, (d.iv) add the edge connecting C'_j and its parent in $\tilde{T}[C']$ to the subtree corresponding to the parent. If no updating is possible, re-root the tree at a child of the current root that minimizes the intersection with the corresponding subtree from C. Since every $e \in \tilde{T}[C]$ in the algorithm above becomes an edge in $\tilde{T}[C']$ (with the updated C') and the weight of such edges is not changed, $\tilde{T}[C]$ and $\tilde{T}[C']$ are isomorphic and, hence, C and C' are equivalent.

This argument guarantees that, if the algorithm produces a sequence C of size k then any other sequence of size k can be transformed to C. Let us now assume that k is not the optimal solution, and there is another sequence C_1 of size $k + 1$. By our previous argument, C_1 can be transformed to satisfy conditions (1) – (3). This implies that the sequence would have been determined by the algorithm, which is a contradiction.

For a given tree T and a number t, $\phi(T) \leq t \leq w(T)$, Algorithm 1 can easily be extended to include finding the collection P. Moreover, this algorithm in conjunction with existing efficient algorithms for minimum edge cover for weighted bipartite graphs (Edmonds and Johnson, 1970; Murty and Perin, 1982; Pulleyblank, 1996) can be used to extract the edge cover–solution of the GAMNR, as shown in Algorithm 2. This algorithm is based on the idea that any t-edge cover is at least as big as a minimum edge-cover and employs the optimality of a tree-packing produced by Algorithm 1.

Algorithm 2: Optimum t-edge cover and t-tree packing of a tree

```
Input: $T$, tree on $n$ nodes
       $t$, $\phi(T) \leq t \leq w(T)$
       $C$, sequence, output of Algorithm 1
Output: $\alpha^w_{1,t}$, optimum size of $t$-edge cover
        $P_i$, $1 \leq i \leq k$, optimum $t$-tree packing
        $S$, $t$-edge cover
$S \leftarrow \emptyset$
foreach $C_i \in C$ do
    unmark all edges
    $P \leftarrow$ minimum edge cover of $C_i$
    $P_i \leftarrow \emptyset$
    while $w(P_i) < t$ do
        $m \leftarrow$ minimum weight of unmarked edge in $C_i$ adjacent to $T[P_i]$
        $x \leftarrow$ edge of weight $m$ in $C_i$ adjacent to $T[P_i]$
        mark $x$
        $P_i \leftarrow P_i \cup \{x\}$
        $S \leftarrow S \cup \{x\}$
    end
    $S \leftarrow S \cup (P - P_i)$
end
$\alpha^w_{1,t} \leftarrow w(S)$
```

Proof of correctness for Algorithm 2 : Given a sequence C, output of Algorithm 1, let P be a (weighted) minimum edge cover of C_i. Since all nodes of C_i are covered by P, a t-tree P_i (of weight at least t) together with the edges of $P - P_i$ gives a t-edge cover for C_i. We claim that the minimality of $\alpha^w_{1,t}$ is guaranteed by: the minimality of the weighted edge cover, the fact that each C_i contains a set of edges that induces a tree of weight at least t, and the greedy spanning (sub)tree algorithm which at each step includes the edge of minimum weight. The cost of the tree P_i is minimal, since lines 5–10 build a subtree by the Prim's algorithm. If a node is not covered by an edge in the tree, it is guaranteed to be covered by an edge in $P - P_i$. Note that this argument leads to the result of Theorem 2.4.1 in the unweighted case.

2.6 Approximation results

For a given edge-weighted graph G, let $\phi(G) = \min\{w(e) \mid e \in E(G)\}$, $\Phi(G) = \max\{w(e) \mid e \in E(G)\}$, and $w(G) = \sum_{e \in E(G)} w(e)$. Furthermore, let t be a parameter such that $\phi(G) \leq t \leq w(G)$. For given t and edge subset $S \subseteq E(G)$ for which any induced connected component $H_i \subset G[S]$ satisfies $w(H_i) \geq t$, we can say that S is also a $(t - \phi(G))$-edge cover. Let $\alpha^w_{1,t} = w(S)$ denote the optimum solution of GAMNR on an edge-weighted graph G with a given bound t. Clearly, $\alpha^w_{1,t} \geq t$, since the edge cover S resulting in the optimum $\alpha^w_{1,t}$ induces at least one connected component in G of weight at least t.

Let ψ denote the weight of a minimum spanning tree T in G. Motivated by a result in (Fernau and Manlove, 2006), we observe that for any $\frac{\psi}{2} < t \leq \psi$, $\alpha^w_{1,t} = \psi$. To obtain this claim, first note that any minimum spanning tree T is a t-edge cover, which is an upper bound to the weight of an optimal S; hence, $w(S) \leq \psi$. If $G[S]$ has two connected components, then $w(S) \geq 2t > \psi$, which is a contradiction. Therefore, $G[S]$ is connected so that $w(S) \geq \psi$, yielding the claim. Moreover, for any $t \geq \psi$, we have that $\alpha^w_{1,t} = t$.

From these claims, we can establish the following theorem:

Theorem 2.6.1. GAMNR can be approximated within factor

- $\rho = 2$ for the unweighted case;
- $\rho = \left(2 + \frac{\Phi(G)}{t} \max\{i, j\}\right)$ in the weighted case.

Proof. Let each node v of G be weighted with the minimum weight over all edges incident on v, *i.e.*, $w'(v) = \min\{w(e) \mid e = \{u, v\}, u \in V(G), e \in E(G)\}$. For a minimum-weight edge cover S' we have that $w(S') \geq \frac{\sum_v w'(v)}{2}$, since each edge from S' covers exactly two nodes of G. Therefore, $\sum_v w'(v) \leq 2w(S') \leq 2w(S) = 2\alpha^w_{1,t}$.

The dynamic programming algorithm that we use for approximating the constrained edge cover in G with a bound t is the same as the optimal Algorithms 1 and 2 presented in Section 2.5 taking as input the minimum spanning tree of G with the bound t. For the analysis of the performance of this algorithm, one may show that the weight of the minimum spanning tree, ψ, is at most

$$\psi \leq \sum_v w'(v) + \sum_i w(e_i^*).$$

where e_i^* are added to the set of edges from which nodes get their weight assignment (by w') in order to get a connected induced subgraph.

Moreover, by running the optimal algorithms we get an edge cover, $S^{\#}$, satisfying the bound t. Since $S^{\#}$ is obtained from the minimum spanning tree by removing some edges, we get that:

$$w(S^{\#}) = \psi - \sum_j w(\tilde{e}_j).$$

Therefore, we have:

$$\begin{aligned}
w(S^{\#}) &= \psi - \sum_j w(\tilde{e}_j) \\
&\leq \sum_v w'(v) + \sum_i w(e_i^*) - \sum_j w(\tilde{e}_j) \\
&\leq 2\alpha_{1,t}^w + \sum_i w(e_i^*) - \sum_j w(\tilde{e}_j) \\
&\leq 2\alpha_{1,t}^w + \max\{i,j\}\,(\Phi(G) - \phi(G))
\end{aligned}$$

For unweighted graphs $\Phi(G) = \phi(G)$, and we have a 2-approximation algorithm. For weighted graphs we can find a number s, such that $\Phi(G) - \phi(G) = st$. As $\alpha_{1,t}^w > t$, we have a factor $(2 + s\max\{i,j\})$-approximation algorithm for the weighted case. More precisely, since $\phi(G) \geq 0$ and $t \geq \phi$, we have a factor $\rho = \left(2 + \frac{\Phi(G)}{t}\max\{i,j\}\right)$. □

2.7 Conclusion

We have defined the automated metabolic network reconstruction as an optimization problem that couples the probabilistic outcome of various bioinformatics methods with existing knowledge and experimental data. By reducing the raw metabolic network, the formulation mimics criteria satisfied by reconstructed metabolic networks of high quality, namely, clustering in connected subgraphs of specified accuracy. The idea that accuracy should be maximized is included in the problem's formulation via the bound (threshold) that must be satisfied by each of the clusters included in a solution. Minimization leads to a smaller number of clusters and, therefore, implies a network of higher overall connectedness.

Choosing a value for the bound imposes two, at times, opposing principles: (1) inclusion of high-weight edges and (2) ensuring high network connectedness, while meeting the requirement for the bound. Therefore, the choice for the value of the bound may affect the number of clusters, and so the reconstruction process should be *iterative*, starting from a small value. Each iteration should be evaluated by an expert in order to stop or resume the process at a higher value for the bound.

The chosen criteria of clustering and connectedness ensure that the reconstructed metabolic network would require little human validation. There could also be more than one possible result from the automated reconstruction, depending on the threshold imposed on the weight of the identified clusters (weight of a cluster is the sum of edge-weights in the cluster). Since the outcome of the reconstruction is a union of connected components, each component can be analyzed separately to speed up the computational process, while the necessary experimental effort for validation would remain unchanged.

The general problem, GAMNR, defined in Section 2.2, is closely related to finding a minimum spanning tree and its extension to a constrained weighted edge cover. The biologically meaningful problem, BMAMNR, also presented in Section 2.2, allows for integration of various types of experimental data and biochemical knowledge to carry out the reconstruction of a metabolic network. This is realized via the inclusion of a set of genes and metabolites, experimentally confirmed to be involved in the metabolism of a given organism, in the definition of BMAMNR.

We have established that both, the general and the biologically meaningful, automated reconstructions of metabolic networks, although closely connected to the abovementioned simple concepts, remain NP-hard for any given bound on the weight of the connected components induced by a constrained weighted edge cover. In terms of approximating GAMNR on a uniformly weighted graph G, we have provided a polynomial-time 2-approximation algorithm, based on its connection to a minimum spanning tree in G. For the general weighted GAMNR, we have devised a polynomial-time algorithm with approximation factor which depends on the bound t and the maximum edge-weight in the raw metabolic network. Due to the established connection between the two problems, the latter could be considered a first step towards developing an approximation algorithm for the biologically meaningful problem. Moreover, the relation among the nestedness of constrained edge covers and weighted tree packing has been employed to devise a polynomial-time algorithms that determine: (1) the optimal edge-cover—a solution to GAMNR and (2) optimal weighted tree packing. This is a first polynomial-time algorithm for the t-edge cover and t-tree packing for (weighted) trees.

We point out that reconstruction based on the graph representation of the raw metabolic network does not require stoichiometric information. Since our approach aims at structural reconstruction of metabolic networks, the stoichiometry can later be incorporated to support other studies (*e.g.* FBA). If a reaction in the raw metabolic network is chemically balanced, it remains so in the reconstructed network (provided it is included in the solution). A reaction incorporated in the raw metabolic network from public databases may not be chemically balanced. Our approach also considers such reactions as it does not aim at resolving this known issue of the publically available human-curated databases.

In Section 2.3, we exemplify the formalism on the reconstruction of the sucrose biosynthesis pathway in *Chlamydomonas reinhardtii*. The practical implications of our theoretical analysis refer to the quality of reconstructed metabolic networks: We believe that automated metabolic network reconstruction, as formulated here, can be used to elicit metabolic networks of higher accuracy. Functional annotation for alternatively spliced genes provides the possibility for inclusion of more than only one enzyme per gene, which in turn enables enlargement of the sets of identified reactions. Any dummy gene nodes included in the raw metabolic network may give hints for new investigations on the genome sequences regarding the search for alternative gene models. On the other hand, such nodes may direct further in-depth similarity inspections to annotate a yet undiscovered function of a given gene. Our results also have implications on planning knock-out experiments: For instance, accurately identified relationships from genes to enzymes which allow for two genes to produce the same enzyme would require simultaneous knock-outs of both genes in order to study its effect on the metabolism.

Our theoretical analysis leaves space for finding more efficient and effective methods for automated reconstruction: The formulation of GAMNR and BMAMNR is only a first attempt to formally address the problem of automated reconstruction. We are aware there could be other criteria, besides clustering and connectedness, which can be employed in the study of metabolic networks. This remains as one of the open problems to pursue.

Other open problems include the design of an polynomial-time approximation algorithm for the biologically meaningful reconstruction. Applications of the approximation algorithm and the very definition of GAMNR are directly dependent on the bound (threshold) B. We believe that, for a given weighted graph G, a bound which combines the smallest weight, $\phi(G)$, number of nodes, and the number of edges in G is one possible direction for further research. It would be interesting to furthermore establish which of the invariants of G should be used to arrive at a biologically meaningful bound.

Chapter 3

Hardness and approximability of the inverse scope problem

For a given metabolic network, we address the problem of determining the minimum cardinality set of substrate compounds necessary for synthesizing a set of target metabolites, called the *inverse scope problem*. We define three variants of the inverse scope problem whose solutions may indicate minimal nutritional requirements that must be met to ensure sustenance of an organism, with or without some side products. Here, we show that the inverse scope problems are NP-hard on general graphs and directed acyclic graphs (DAGs). Moreover, we show that the general inverse scope problem cannot be approximated within $n^{1/2-\epsilon}$ for any constant $\epsilon > 0$ unless P = NP. Our results have direct implications for identifying the biosynthetic capabilities of a given organism and for designing biochemical experiments.

3.1 Introduction

Availability of fully sequenced genomes for several organisms has rendered it possible to reconstruct their metabolic networks and further characterize their biosynthetic capabilities. Identifying the biosynthetic capabilities of a given organism is crucial for the development of cost-efficient energy sources, as they are directly related to plant biomass (Tsantili et al., 2007; Nissen et al., 2000; Burchhardt and Ingram, 1992). On the other hand, knowing the compounds necessary for obtaining a desired product can be employed in designing optimal environmental conditions, in the sense of minimizing the nutrients for biosynthesis, and for effective altering of bioprocesses to assist the industrial manufacture of chemicals (Burton et al., 2002).

Several mathematical methods have been developed to study the biosynthetic capabilities of metabolic networks, including: metabolic control analysis (Wildermuth, 2000), flux balance analysis (Bonarius et al., 1997), metabolic pathway analysis (Schilling et al., 1999), cybernetic modeling (Kompala et al., 1984), biochemical systems theory (Savageau, 1969), to name just a few. Many of these methods require detailed kinetic information to carry out the analysis — a condition which is often impossible to satisfy.

A method which relies only on an available metabolic network and limited knowledge about the stoichiometry of the included biochemical reactions has been recently developed and applied to study the biosynthetic capabilities of various organisms (Ebenhöh et al., 2004; Handorf et al., 2005). This method is based on the concept of a *scope*: The basic principle is that a reaction can only operate if and only if all of its substrates are available as nutrients or can be provided by other reactions in the network. Starting from the nutrients, called *seed compounds*, operable reactions and their products are added to an expanding subnetwork of a given metabolic network. This iterative process ends when no further reaction fulfills the aforementioned condition. The set of metabolites in the expanded subnetwork is called the scope of the seed compounds and represents all metabolites that can be in principle synthesized from the seed by the analyzed metabolic network (Handorf et al., 2005).

The scope concept has been applied to a variety of problems, such as: hierarchical structuring of metabolic networks (Handorf et al., 2006), comparison of metabolic capabilities of organism specific networks (Ebenhöh et al., 2005), metabolic evolution (Ebenhöh et al., 2006), and changes of metabolic capacities in response to environmental perturbations (Ebenhöh and Liebermeister, 2006).

In Handorf et al. (2008), the inverse problem was addressed as that of determining *minimal* sets of seed compounds from which metabolites that are essential for cellular maintenance and growth can be produced by a given metabolic network. There, a greedy algorithm was applied and heuristics inspired by biological knowledge were introduced to determine biologically relevant minimal nutrient requirements. Whereas by this approach a large number of minimal solutions may be obtained, the minimum cardinality set of seed compounds remains unknown and moreover, it is unclear how well this minimum was approximated by the proposed heuristic.

For a given metabolic network, we investigate the general inverse scope problem of determining the *minimum* cardinality set of seed compounds necessary for the synthesis of a specific compound or a set of compounds. In particular, the latter set may comprise metabolic precursors that an organism requires for maintenance or growth. Therefore, solving this inverse problem may indicate minimal nutritional requirements that must be met to ensure sustenance of the organism. The nutrients which can be provided in synthesis are often restricted to a specific set, in which case we address the inverse problem with a forbidden set. In addition, we address the problem of finding the minimum cardinality set of seed compounds that are necessary for synthesis of a given set of compounds and, at the same time, guarantee that a specific set of compounds are not created as side products. This is the inverse problem with two forbidden sets.

The problems addressed here have applications that span various fields: In a sensor network with directed communications, one is interested in finding the minimum number of nodes that can be used for fast delivery of information. In the field of computational geometry, one may formulate the problem of determining the minimum number of flood-lights that can illuminate a given polygon (Bagga et al., 1996), while in automated reasoning one may seek automated deduction with minimum number of axioms (Duffy, 1991). We note that none of the related variants has the constraint that all precursors must be present for an action to take place.

Contributions. Here, we show that the inverse scope problems are NP-hard on general graphs and directed acyclic graphs (DAGs). We also demonstrate that the inverse scope problem with two forbidden sets on general graphs cannot be approximated within $n^{1/2-\epsilon}$ for any constant $\epsilon > 0$ unless P = NP. In addition, we discuss the practical implications of the hardness of approximation results.

3.2 Problem definition

A *metabolic network* is typically represented by a directed bipartite graph $G = (V, E)$. The vertex set of G can be partitioned into two subsets: V_r, containing *reaction nodes*, and V_m, comprised of *metabolite nodes*, such that $V_r \cup V_m = V(G)$. The edges in $E(G)$ are directed either from a node $u \in V_m$ to a node $v \in V_r$, in which case the metabolite u is called a *substrate* of the reaction v, or from a node $v \in V_r$ to a node $u \in V_m$, when u is called a *product* of the reaction v.

The scope concept is related to reachability in the metabolic network graph G: A reaction node $v \in V_r$ is reachable if all of its substrates are reachable. Given a subset S of metabolite nodes, a node $u \in V_m$ is reachable either if $u \in S$ or if u is a product of a reachable reaction. With these clarifications, we can present a precise mathematical formulation for the scope of a set of seed compounds:

Definition 3.2.1. Given a metabolic network $G = (V, E)$ and a set $S \subseteq V_m$, the scope of S, denoted by $R(S)$, is the set of all metabolite nodes reachable from S.

For a given metabolic network $G = (V, E)$ and a set $S \subseteq V_m$, the scope $R(S)$ can be determined in polynomial time of the order $O(|E| \cdot |V|)$, as can be established by analyzing the following algorithm:

Algorithm 3: Scope for a set of seed metabolites S in a metabolic network G

```
Input: $G = (V_m \cup V_r, E)$, metabolic network
$S$, set of seed metabolites, $S \subseteq V_m$
Output: $R(S)$, scope of $S$
mark all nodes in $V(G)$ unreachable
mark all nodes in $V_r$ unvisited
mark all nodes in $S$ reachable
repeat
    foreach node $v \in V_r$ do
        if $v$ is reachable then
            mark $v$ as reachable
        end
    end
    if there is a reachable unvisited node $v \in V_r$ then
        mark $v$ visited
        mark successors of $v$ reachable
    end
until no reachable unvisited nodes in $V_r$
$R(S) \leftarrow$ all reachable nodes in $V_m$
```

We define the inverse scope problem as follows:

INVERSE SCOPE (IS)
INSTANCE : Given a metabolic network $G = (V, E)$ and a subset of metabolites $P \subseteq V_m$.
PROBLEM : Find a subset of metabolites $S \subseteq V_m$ such that $P \subseteq R(S)$.
MEASURE : Cardinality of S.
(min)

Often, there is a restriction to the subset of metabolites from which we would like to identify S, the seed compounds synthesizing P. In that case, we address the inverse scope problem with a forbidden set $V(G) - S'$, such that $S \subseteq S'$, defined below:

INVERSE SCOPE WITH A FORBIDDEN SET (ISFS)
INSTANCE : Given a metabolic network $G = (V, E)$ and two subsets of metabolites $V(G) - S', P \subseteq V_m$, where $V(G) - S'$ is the forbidden set and P is the set of products.
PROBLEM : Find a subset of metabolites $S \subseteq S'$ such that $P \subseteq R(S)$.
MEASURE : Cardinality of S.
(min)

It is interesting to also consider the problem of determining the set of nutrients to be provided in the synthesis of a given set of products while not yielding a pre-specified set of side products.

The study of this problem may, for instance, indicate how to design a biochemical experiment to minimize the effect of some undesirable compounds. To ensure that a specific set of metabolites is not synthesized, we modify ISFS as follows:

INVERSE SCOPE WITH TWO FORBIDDEN SETS (IS2FS)
INSTANCE : Given a metabolic network $G = (V, E)$ and three subsets of metabolites $V(G) - S', F, P \subseteq V_m$, where $V(G) - S'$ and F are the forbidden sets and P is the set of products.
PROBLEM : Find a subset of metabolites $S \subseteq S'$ such that $P \subseteq R(S)$ and $F \cap R(S) = \emptyset$.
MEASURE : Cardinality of S.
(min)

Remark 3.2.1. Note that by taking $S' = V(G)$ in ISFS, every instance of IS becomes an instance of ISFS, and ISFS can be restricted to IS. In addition, every instance of IS2FS with $F = \emptyset$ is an instance of ISFS. Therefore, IS2FS is the most general of the three problems.

For completeness, we show the decision versions of the three problems defined above:

INVERSE SCOPE DECISION (ISD)
INSTANCE : Given a metabolic network $G = (V, E)$, subset of metabolites $P \subseteq V_m$, and an integer K.
PROBLEM : Does there exist a subset of metabolites $S \subseteq V_m$ such that $P \subseteq R(S)$ and $|S| \leq K$.

INVERSE SCOPE WITH A FORBIDDEN SET DECISION (ISFSD)
INSTANCE : Given a metabolic network $G = (V, E)$, two subsets of metabolites $V(G) - S', P \subseteq V_m$, where $V(G) - S'$ is the forbidden set and P is the set of products, and an integer K.
PROBLEM : Does there exist a subset of metabolites $S \subseteq S'$ such that $P \subseteq R(S)$ and $|S| \leq K$.

INVERSE SCOPE WITH TWO FORBIDDEN SETS DECISION (IS2FSD)
INSTANCE : Given a metabolic network $G = (V, E)$, three subsets of metabolites $V(G) - S', F, P \subseteq V_m$, where $V(G) - S'$ and F are the forbidden sets and P is the set of products, and an integer K.
PROBLEM : Does there exist a subset of metabolites $S \subseteq S'$ such that $P \subseteq R(S)$, $F \cap R(S) = \emptyset$, and $|S| \leq K$.

In the next section we present the results regarding the NP-hardness of the three inverse scope problems.

3.3 Hardness results

An optimization problem Π is shown to be NP-hard by establishing a polynomial time reduction from a problem known to be NP-complete to the decision version of Π. First, we show that IS2FS is NP-hard on a general graph by providing a reduction from MINIMUM DISTINGUISHED ONES (MIN-DONES). We also show that ISFS is NP-hard even on DAGs by providing a reduction from the SET COVER (SC) problem. In a similar way, we show that IS, too, is NP-hard when restricted to DAGs. These results will later be used for obtaining the approximation results for the three problems.

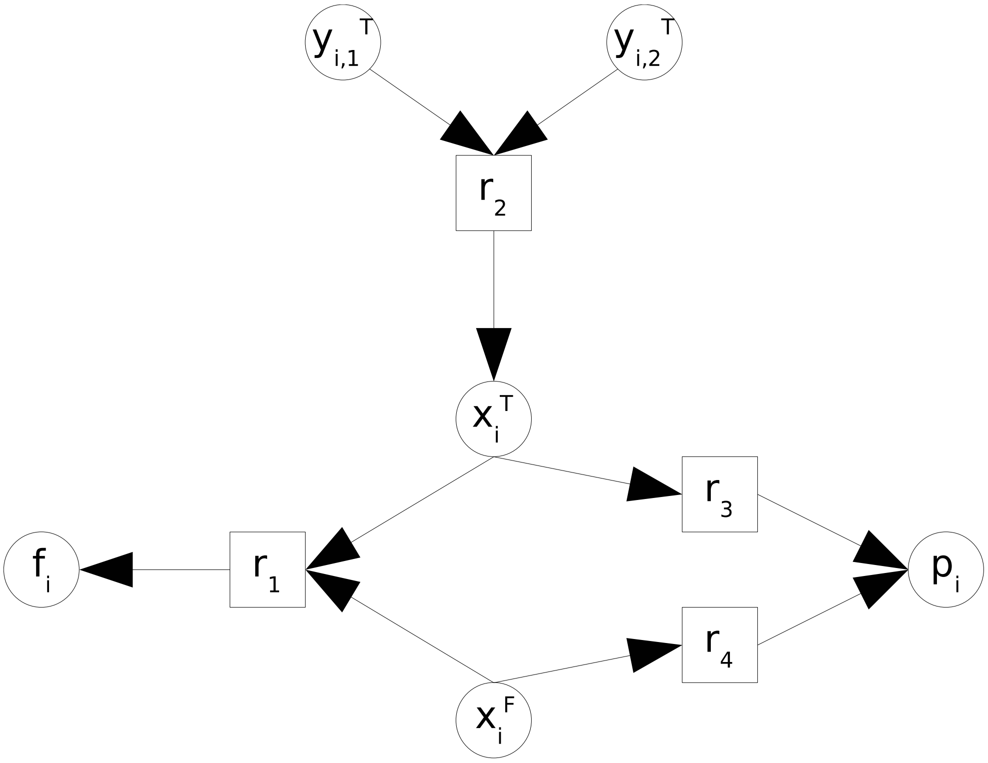


Figure 3.1: Gadget for the construction of IS2FSD instance.

Theorem 3.3.1. INVERSE SCOPE WITH TWO FORBIDDEN SETS DECISION problem is NP-complete.

Proof. First, we need to show that IS2FSD is in NP. Given an instance of IS2FSD with three subsets of nodes $S, F, P \subseteq V_m(G)$, one can find $R(S)$, by employing Algorithm 3, and check whether $P \subseteq R(S)$, $F \cap R(S) = \emptyset$ and $|S| \leq K$ in polynomial time.

Next, we provide a reduction from the MIN-ONESD problem. An instance of the decision version of MIN-ONES is given by a set of n variables Z, collection C of disjunctive clauses of 3 literals, and an integer K' (a literal is a variable or a negated variable in Z). The problem is then to find a truth assignment for Z that satisfies every clause in C such that the number of variables in Z that are set to true in the assignment is at most K'.

Given an instance of MIN-ONESD, we can construct an instance of IS2FSD, a bipartite directed graph $G = (V, E)$ with $V(G) = V_m \cup V_r$, three subsets of nodes $S', F, P \subseteq V_m(G)$, and an integer K as follows: For each variable $x_i \in Z$, we use the gadget shown in Figure 3.1. The gadget is composed of six nodes $y_{i,1}^T$, $y_{i,2}^T$, x_i^T, x_i^F, p_i, and f_i connected through four reactions—r_i^1 with x_i^T and x_i^F as substrates and f_i as a product, r_i^2 with $y_{i,1}^T$ and $y_{i,2}^T$ as substrates and x_i^T as a product, r_i^3 with x_i^T as substrate and p_i as a product, and r_i^4 with x_i^F as substrate and p_i as a product. Moreover, for each clause in C we add a node c_j. A node x_i^T is connected to c_j via a reaction if variable x_i appears non-negated in c_j; similarly, a node x_i^F is connected to c_j via a reaction if variable x_i appears negated in c_j. Finally, we let S' be composed of all $y_{i,1}^T$, $y_{i,2}^T$, and x_i^F nodes, P be composed of all c_j and p_i nodes, while F be comprised of all f_i nodes.

Note that x_i^T is reached if and only if its two corresponding nodes $y_{i,1}^T$ and $y_{i,2}^T$ are included as substrates. Moreover, the inclusion of p_i nodes in P and f_i nodes in F ensures that exactly one of the x_i^T and x_i^F is chosen. Therefore, to complete the construction, we set $K = n + K'$.

A solution to MIN-ONESD can be transformed to a solution of IS2FSD of cardinality $K =$

$n+K'$ by taking those nodes $y^T_{i,1}$ and $y^T_{i,2}$ in S' whose corresponding variable x_i is set to TRUE in the solution of MIN-ONESD. Moreover, the solution also includes the x^F_i whose value can be set to FALSE. All c_j nodes can be reached, since a solution to MIN-ONESD guarantees that at least one literal in the clause c_j has value TRUE. Similarly, all p_i nodes are also reached. Moreover, a valid truth assignment guarantees that no pair x^T_i and x^F_i is in the solution of MIN-ONESD; thus, the nodes in F cannot be accessed.

Given a solution S to IS2FSD on G, the solution to MIN-ONESD can be obtained by assigning value TRUE to that variable x_i in Z whose corresponding nodes $y^T_{i,1}$ and $y^T_{i,2}$ are in S; the remaining variables are assigned value FALSE. Since all c_j nodes are reachable, then each of them has at least one directed path from a node in S and thus the value of the corresponding clause is TRUE. Moreover, since no node in F is in the scope of S, the reconstructed truth assignment is valid.

Since IS2FSD can be solved if and only if there is a solution to MIN-ONESD, we have the NP-completeness of the problem in the theorem. □

Corollary 3.3.2. The INVERSE SCOPE WITH TWO FORBIDDEN SETS problem is NP-hard on general graphs.

For the inverse scope with a forbidden set we have:

Theorem 3.3.3. INVERSE SCOPE WITH A FORBIDDEN SET DECISION is NP-complete even on DAGs.

Proof. The decision version of ISFS is in NP since for any given set of metabolites S, we can find $R(S)$, by using Algorithm 3, and check whether $P \subseteq R(S)$ in polynomial time.

We provide a polynomial time reduction from the SCD problem: An instance of the SCD problem is given by a collection C of subsets from a finite set U and an integer K'. The problem then is to determine whether there is a set cover for U of cardinality at most K', *i.e.* a subset $C' \subseteq C$ such that every element of U belongs to at least one member of C' and $|C'| \leq K'$.

Given an instance of the SCD problem we design an instance of the ISFSD problem as follows: First, we build the metabolic network G which must be bipartite. Let the number of subsets in the collection C be denoted by p. For every subset $C_i \in C$ we create a reaction node r_i, $1 \leq i \leq p$; thus, we have p reaction nodes. Let the number of elements in U be denoted by n. For every element $x_j \in U$ we create a metabolite node, denoted by x_j, $1 \leq j \leq n$. Furthermore, we create p additional metabolite nodes, denoted by y_i, $1 \leq i \leq p$. Therefore, $V_r = \{r_i \mid 1 \leq i \leq p\}$ and $V_m = \{x_j \mid 1 \leq j \leq n\} \cup \{y_i \mid 1 \leq i \leq p\}$. Finally, we set $P = \{x_j | 1 \leq j \leq n\}$

A reaction r_i is connected via a directed edge to a metabolite x_j if and only if the subset C_i corresponding to the node r_i contains the element x represent by the node x_j. Additionally, we include a directed edge from y_i to r_i, $1 \leq i \leq n$.

Finally, we set $S' = \{y_i \mid 1 \leq i \leq p\}$ and let the integer K of the decision version of the ISFS problem be equal to K'. This construction can be completed in time polynomial in the size of the SC instance. The construction is illustrated in Figure 3.2.

If we have a solution to the SC problem and it is given by a subset C', $|C'| \leq K'$ then the solution to the ISFS problem, the subset $S \subseteq S'$, is comprised of the nodes in $\{y_i \mid 1 \leq i \leq p\}$ that are connected via a directed edge to reaction nodes r_i representing the subsets in C', since $R(S) = P$.

Conversely, if we have a solution to the ISFS problem, *i.e.*, a subset $S \subset S'$, $|S| \leq K$ then R(S) = P. The solution to the SC problem, a subset $C' \subseteq C$, can be obtained by including those elements $C_i \in C$ corresponding to the reaction nodes r_i to which there exists an edge from $y_i \in S$ in G.

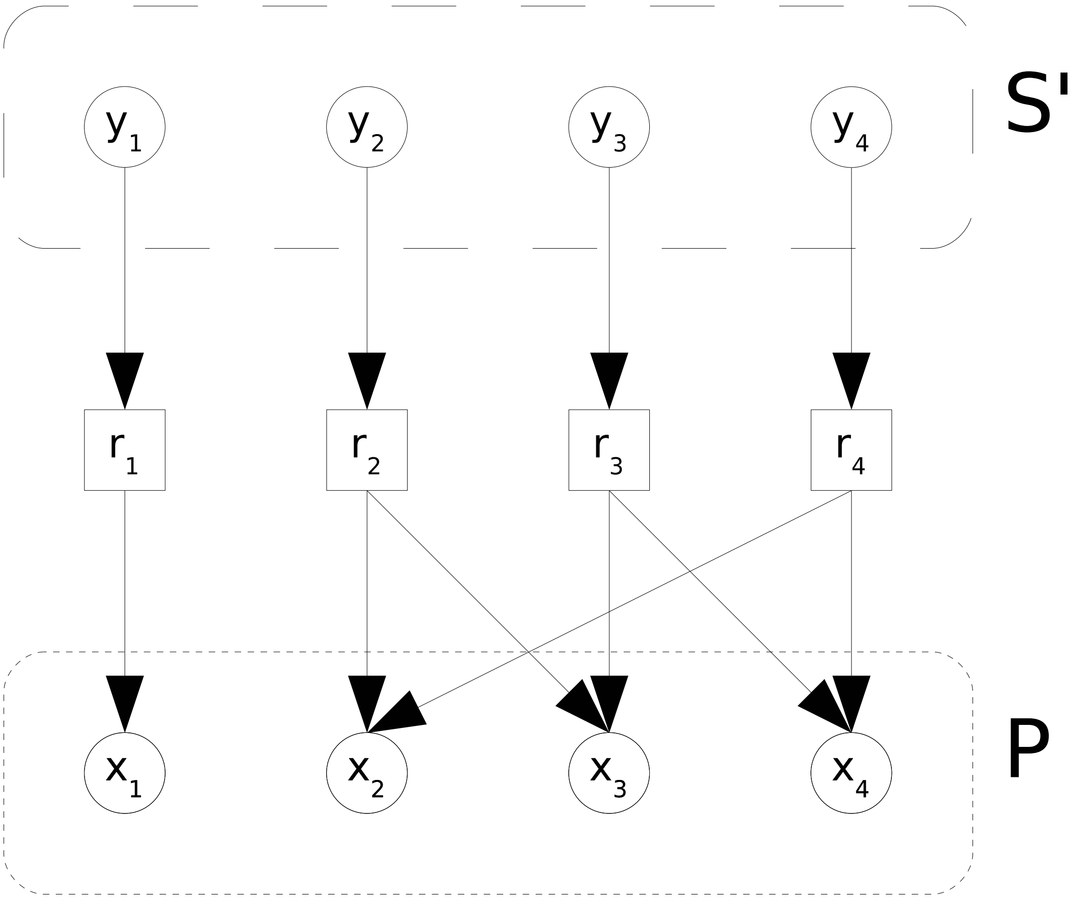


Figure 3.2: Instance of ISFS obtained from the instance of SC with $U = \{1, 2, 3, 4\}$ and $C = \{\{1\}, \{2, 3\}, \{3, 4\}, \{2, 4\}\}$.

Since the ISFS problem can be solved if and only if there is a solution to the SC problem, we have the NP-completeness of the problem in the theorem. Furthermore, the polynomial time construction results in an acyclic directed graphs (DAGs), so ISFSD is NP-complete on DAGs. □

We then have the following corollary:

Corollary 3.3.4. The INVERSE SCOPE WITH A FORBIDDEN SET problem is NP-hard even on DAGs.

We can use a similar approach as in the proof of Theorem 3.3.3 to obtain the following result:

Theorem 3.3.5. The INVERSE SCOPE problem is NP-hard even when restricted to DAGs.

Proof. The ISD problem is in NP even when restricted to DAGs: For any given set of metabolites S and an integer K, we can find $R(S)$, by using Algorithm 3, and check whether $P \subseteq R(S)$ and $|S| \leq K$ in polynomial time.

Given an instance of the SCD problem we design an instance of the ISD problem, a graph G and an integer K, on DAGs as follows: First, we build the metabolic network G which must be bipartite. Let the number of subsets in the collection C be denoted by p. For every subset $C_i \in C$ we create a metabolite node y_i; thus, we have p metabolite nodes from this step of the construction. Let the number of elements in U be denoted by n. For every element $x_j \in U$ we create a metabolite node, denoted by x_j^1, $1 \leq j \leq n$.

A metabolite y_i is connected via a directed path of length 2 with a middle reaction node r_i to a metabolite x_j^1 if and only if the subset C_i, corresponding to the node y_i, contains the element x represented by the node x_j. Additionally, we include $p-1$ copies of each node x_j^1, denoted by x_j^i, $2 \leq i \leq p$, and connect them to the in-neighbors of x_j^1. Finally, we let P contain all x_j^i nodes and $K' = K$. We increase the number of elements per set to ensure that only nodes among y_i are chosen as a solution to ISD, so that it can be transformed to a solution of SCD.

If we have a solution to the SCD problem and it is given by a subset C', $|C'| \leq K'$ then the solution to the ISD problem, the subset $S \subseteq V_m$, is comprised of the nodes in $\{y_i \mid 1 \leq i \leq p\}$ representing the subsets in C', since $R(S) = P$.

Conversely, if we have a solution to the ISD problem, *i.e.*, a subset $S \subseteq V_m$, $|S| \leq K$, then $R(S) = P$. The set C' can be obtained in the following way: Let S contains a node u from $\{x_j^i \mid 1 \leq i \leq p, 1 \leq j \leq n\}$. There are two cases: If u can be reached by some y_i, then a solution to SCD excludes this element from S. If u cannot be reached by some y_i, then none of its remaining $p-1$ copies can be reached ($p \geq 2$). Including one y_i representing a set that contains u can always decrease the cardinality of S by at least two. The solution to the SCD problem, a subset $C' \subseteq C$, therefore includes those elements $C_i \in C$ corresponding to the elements in S as well as the nodes added by the algorithm to cover nodes from $\{x_j^i \mid 1 \leq i \leq p, 1 \leq j \leq n\}$ which are initially included in S but are removed by the previous algorithm. It follows that $|C''| \leq K'$.

Since the ISD problem can be solved if and only if there is a solution to the SCD problem, we have the NP-completeness of the problem in the theorem. Furthermore, the polynomial time construction results in an acyclic directed graphs (DAGs), so ISD is NP-complete on DAGs. □

3.4 Approximation results

Let us recall a few definitions about approximability. Given an instance x of an optimization problem A and a feasible solution y of x, we denote by $m(x, y)$ the value of the solution y, and by $opt_A(x)$ the value of an optimum solution of x. Here, we consider minimization problems. The performance ratio of the solution y for an instance x of a minimization problem A is

$$R(x, y) = \frac{m(x, y)}{opt_A(x)}.$$

For a constant $\rho > 1$, an algorithm is a ρ-approximation if for any instance x of the problem it returns a solution y such that $R(x, y) \leq \rho$. We say that an optimization problem is constant approximable if, for some $\rho > 1$, there exists a polynomial-time ρ-approximation for it. APX is the class of optimization problems that are constant approximable. An optimization problem has a polynomial-time approximation scheme (a PTAS, for short) if, for every constant $\epsilon > 0$, there exists a polynomial-time $(1 + \epsilon)$-approximation for it.

L-reduction was introduced as a transformation of optimization problems which keeps the approximability features (Papadimitriou and Yannakakis, 1991). L-reductions in studies of approximability of optimization problems play a similar role to that of polynomial reductions in the studies of computational complexity of decision problems. For completeness we include the following definition:

Definition 3.4.1. Let A and B be two optimization problems. Then A is said to be *L-reducible* to B if there are two constants $\alpha, \beta > 0$ such that:

1. there exists a function, computable in polynomial time, which transforms each instance x of A into an instance x' of B such that $opt_B(x') \leq \alpha \cdot opt_A(x)$,

2. there exists a function, computable in polynomial time, which transforms each solution y' of x' into a solution y of x such that $|m(x,y) - opt_A(x)| \leq \beta \cdot |m(x',y') - opt_B(x')|$.

Remark 3.4.1. This reduction preserves PTAS, *i.e.*, if A is L-reducible to B and B has a PTAS then A has a PTAS as well.

Remark 3.4.2. From the above, if δ is a lower bound of the worst-case approximation factor of A, then $\rho = \frac{\delta}{\alpha \cdot \beta}$ is a lower bound of the worst-case relative error of B.

We employ L-reduction to obtain results about the lower bound of the worst-case approximation factor for IS2FS.

Theorem 3.4.2. Inverse Scope with Two Forbidden Sets on a graph G with n nodes cannot be approximated to within a factor of $n^{1/2-\epsilon}$ in polynomial time for any constant $\epsilon > 0$, unless P = NP.

Proof. To construct an L-reduction, we first choose A to be MIN-ONES and B, IS2FS. Jonsson (1998) has shown that MIN-ONES is NPO PB-complete, and is not approximable within $|Z|^{1/2-\epsilon}$ for any $\epsilon > 0$. Given an instance x of MIN-ONES, we construct an instance x' of IS2FS the same as in Theorem 3.3.1. From the proof of Theorem 3.3.1, we have $opt_B(x') = n + opt_A(x)$, so $opt_B(x) \leq opt_A(x')$ and $\alpha = 1$. Moreover:

$$|m(x,y) - opt_A(x)| \leq |m(x',y') - opt_B(x')|,$$

so $\beta = 1$. Thus, we have $n^{1/2-\epsilon}$ for the lower bound of the worst-case approximation factor of IS2FS by Remark 3.4.2. □

Handorf et al. (2008) developed a heuristic for finding minimal sets of seed compounds from which metabolites that are essential for cellular maintenance can be produced. As every minimum set of seed compounds is also minimal, the heuristic can approximate the IS problem. The heuristic takes as input an ordered list of all metabolites in a given metabolic network and a set of target metabolites. It then continually removes a metabolite from the beginning of the list, while recalculating the scope of the remainder of the list. If the resulting scope does not contain the full target set, the metabolite is inserted back in the list; otherwise, it remains permanently removed. Clearly, the set of metabolites contained in the list after the exhaustive search represents a minimal seed, as the removal of any metabolite would result in a scope that does not contain all target metabolites. Since different orderings of the list may result in a different minimal set of seed metabolites, it is not known how well this heuristic approximates the IS problem. It remains as an open problem to develop provably good approximation algorithms for all of the addressed inverse scope problems.

3.5 Instances of IS and ISFS in P

IS and ISFS are solvable in polynomial time on trees. In a tree metabolic network, each metabolite, other than the root, is a product of a reaction with only one substrate. In other words, a tree metabolic network is a tree rooted in a metabolite node. Given a tree metabolic network T and a node u, let S_u be the set of predecessors of u. Given two sets $S', P \subseteq V_m(T)$, the solution to ISFS on T is given by the following greedy algorithm:

Algorithm 4: Algorithm for ISFS on a tree

```
   Input: $T = (V_m \cup V_r, E)$, metabolic tree network
   $P \subseteq V_m$, set of metabolites
   $S' \subseteq V_m$, the set to choose seeds
   Output: $S$, $S \subseteq S'$, such that $P \subseteq R(S)$
1  foreach node $u \in P$ do
2      find the set $S_u$
3  end
4  $S \leftarrow \emptyset$
5  $L \leftarrow$ order list of nodes from $P$
6  while there is a node $u \in L$ do
7      while there is node $v \neq u$, $v \in L$ with $S_v \cap S_u \cap S' \neq \emptyset$ do
8          $S_u \leftarrow S_u \cap S_v \cap S'$ remove $v$ from $L$
9      end
10     $S \leftarrow S \cup$ last common ancestor in $S_u$
11     remove $u$ from $L$
12 end
13 output $S$
```

The algorithm works by finding the last common ancestor for a subset of P as large as possible (lines 6 – 12). Since this subset is reachable from one node only, its cardinality cannot be decreased and the algorithm is optimal.

Given two directed graphs G and H, the Cartesian product $G \Box H$ is a graph with node set $V(G) \times V(H)$ such that there is an edge $\{(u_1, v_1), (u_2, v_2)\} \in E(G \Box H)$ if and only if: (1) $u_1 = u_2$ and $\{v_1, v_2\} \in E(H)$ or (2) $v_1 = v_2$ and $\{u_1, u_2\} \in E(G)$.

Given a tree metabolic network T in which each reaction has precisely one substrate and one product, let $\tilde{T}$ be the tree obtained by the following steps: (1) for each reaction node, connect the substrate with all the products, (2) remove reaction nodes. Note that $\tilde{T}$ includes only the metabolite nodes of T. These tree will be called *reduced*.

Given a directed graph G, let G_s be the graph in which each edge of G is subdivided (while keeping the direction of the edge). The nodes used in the subdivision can be treated as reaction nodes, and all nodes in $V(G)$ as metabolite nodes.

If there is a graph G which can be represented as subdivision of the directed product of $\tilde{T_1}$ and $\tilde{T_2}$ in the instance of a ISFS with a set P, such that ISFS has a non-empty solution on T_1 and T_2, then:

$$S((\tilde{T_1} \Box \tilde{T_2})_s) = \min\{S(T_1), S(T_2)\},$$

where $S(T_1)$ is a solution to ISFS with P mapped onto $V_m(T_1)$, and $S(T_2)$ is a solution to ISFS with P mapped onto $V_m(T_2)$. Therefore, the problem is polynomially solvable if G can be obtained by subdividing the Cartesian product of two reduced tree metabolic networks.

We anticipate that similar constructions may lead to ways of decomposing metabolic networks into smaller parts on which the inverse scope problems may be polynomially solvable. However, we leave this as an open problem and a direction for future research.

3.6 Discussion

The inverse scope problem discussed here is of great importance for biological research since its solution allows to computationally predict minimal nutrient requirements for the cultivation of organisms or to identify cost efficient combinations of substrates for biotechnological applications.

The hardness of approximation results obtained in Section 3.4 bear some important implications to the application of the inverse scope problems. For a general graph, we show that IS2FS cannot be approximated within $n^{1/2-\epsilon}$ for any constant $\epsilon > 0$ unless P = NP. The hardness of approximation and parameterized complexity of IS and ISFS on general graphs remain as open problems. Our conjecture is that their complexity on general graphs strongly depends on the existence of directed cycles in the given metabolic network.

Our results imply that divising an efficient approximation algorithm will depend on finding biologically meaningful metabolic networks representations with no or a small number of cycles. Analyses of real metabolic networks have demonstrated the abundance of directed cycles, which result in high clustering and small average path length (Jeong et al., 2000). Furthermore, one may observe that the directed cycles are predominantly induced on the ubiquitous compounds, such as ATP and NADH. Under physiological conditions, a cell maintains such substances at rather constant levels guaranteeing their availability to the many processes in which they are required. It is therefore unrealistic to assume that these compounds have to be produced a priori. This allows for an alteration of the network structure reflecting that ubiquitous compounds are always available, while still describing the biochemical capabilities of the considered organism. Such a reduction will considerably reduce the number of cycles. Another type of cylces results from the representation of a metabolic network as bipartite graph, in which a reversible reaction is represented by two reaction nodes which are connected to an identical set of reactants with directions of all corresponding edges reversed. This results in cycles of length four for each reversible reaction. To remove such cycles without altering the biochemical capabilities of the network is more challenging. A possible approach is to study networks with flux balance analysis to identify those reactions which under physiological conditions always proceed in one direction. We are currently working on applying our findings to metabolic networks obtained from the KEGG database.

In addition, our analysis demonstrates that the concerted interrelation of biochemical processes responsible for efficient systematic adjustment of an organism to changing environmental conditions are indeed complex and not yet well-understood.

Chapter 4

Structure and bistability

The number of organisms for which there exist preliminary genome-scale reconstructions of their metabolic network, together with their coverage and accuracy, is rapidly growing (Reed et al., 2006). Metabolic networks represent dynamic systems whose behaviour is difficult to predict by visual inspection alone. Hence, computational approaches are inevitable in the analysis of metabolic networks.

One relevant property of metabolic networks is their capability to exhibit multistability and, consequently, a hysteresis behaviour, since it enables switching between different modes of operation as a response to changing conditions. This also holds true for other biological networks, especially signalling pathways (Angeli et al., 2004). Given a metabolic network consisting of parameter-dependent reactions converting a set of metabolites, it is extremely difficult to determine for which regions of the parameter space, *i.e.* for which assignment to the present kinetic parameters, multiple positive steady states can occur. This information might in turn prove useful when discriminating and comparing several model candidates with respect to experimental data. Chemical reaction network theory (CRNT), explained in more detail in this chapter, is a powerful and mathematically sound framework to obtain results about bistability. The main benefit of CRNT is that it derives all results directly from the underlying structure of a metabolic network and thus avoids tedious and largely incomplete numerical exploration of the parameter space.

4.1 Chemical Reaction Network Theory

The development of CRNT started in the early 70s with the work of Horn and Jackson (1972) and ever since has been refined and extended by Feinberg (Feinberg, 1995a,b). Current extensions to this theory are briefly presented in Section 4.1.3.

The key idea is as follows: although the system of ordinary differential equations describing the dynamics of a metabolic network in terms of metabolite concentrations is in general non-linear, assuming mass-action kinetics for every reaction introduces enough linearity to establish fundamental results about multistability. More technically, the differential equations describe a non-linear function in terms of metabolite concentrations. This function can also be expressed with respect to *complexes* which are constituted of the left- and right-hand sides of each reaction. CRNT precisely describes the conditions for this mapping to be linear and the consequences for the occurrence of multiple positive steady states. The beauty of CRNT lies in the fact that these conditions are solely dependent on the structure of the network, and most of them can easily be calculated even for large networks.

4.1.1 Introduction

To understand CRNT a few necessary concepts and definitions are briefly presented. For simplicity, the notation will be slightly changed compared to Feinberg (1995a). A very good introduction to CRNT appears in Gunawardena (2003).

Consider the following example reaction network with three species (metabolites) A, B, and C and two reversible reactions:

$$\begin{aligned} A + B &\rightleftarrows 2C \\ C &\rightleftarrows 0 \end{aligned} \tag{4.1}$$

Under the assumption that the reaction network follows mass-action kinetics, each reaction v_i is dependent on one parameter k_i and on the substrate concentrations:

$$\begin{aligned} v_1 &= k_1 \cdot [A] \cdot [B], \\ v_2 &= k_2 \cdot [C]^2, \\ v_3 &= k_3 \cdot [C], \\ v_4 &= k_4. \end{aligned} \tag{4.2}$$

The changes of species concentration over time can then be expressed as a system of ordinary differential equations:

$$\begin{aligned} \dot{[A]} &= -v_1 + v_2, \\ \dot{[B]} &= -v_1 + v_2, \\ \dot{[C]} &= 2v_1 - 2v_2 - v_3 + v_4. \end{aligned} \tag{4.3}$$

In matrix notation, system 4.3 can be rewritten as

$$\begin{pmatrix} \dot{[A]} \\ \dot{[B]} \\ \dot{[C]} \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 2 & -2 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}. \tag{4.4}$$

Consequently, every reaction network with n species and r reactions can be written as

$$\dot{x} = N \cdot v(k, x),$$

where x is a vector of size n containing the metabolite concentrations, v is a vector of size r describing the reaction rates, and N is the stoichiometric $n \times r$ matrix. In general, the rank of N is not maximal. Let $s = \text{rank}(N)$, then the number of conservation relations within the reaction network is $n - s$. In example 4.1, the rank is 2 and the only conservation relation is $[A] = [B]$.

Furthermore, the combination of species which form together either the educts or the products of a reaction are called *complexes*. In example 4.1, there are four complexes, namely $A + B$, $2C$, C, and the zero complex 0, which accounts for having an open system with mass influx and efflux. Note that C and $2C$ form different complexes because their stoichiometric coefficients are not equal. Formally, let the set of species be denoted as $\mathcal{S}$ and the set of complexes as $\mathcal{C}$. Each complex $y \in \mathcal{C}$ is then a multiset of $\mathcal{S}$. Moreover, each reaction can be written as $y \longrightarrow y'$ with $y, y' \in \mathcal{C}$. A reaction network is defined in the following way:

Definition 4.1.1. A *chemical reaction network* is a 3-tupel $(\mathcal{S}, \mathcal{C}, \mathcal{R})$, where $\mathcal{S}$ is a finite set of species, $\mathcal{C}$ is a finite set of multisets of species, called complexes and $\mathcal{R}$ is a relation on $\mathcal{C}$, denoted as $y \longrightarrow y'$ for $(y, y') \in \mathcal{R}$ and $y, y' \in \mathcal{C}$, which represents a reaction converting y into y'.

This definition does allow for reaction networks in which mass conservation is violated, for instance $A \rightleftarrows 2A$. However, this does not affect the strength of the results.

The set of complexes can then be unambiguously partitioned into a set of *linkage classes*, defined as maximal sets of complexes, which are only connected by reactions to complexes within the same linkage class, but disconnected from other complexes.

Definition 4.1.2. Let $\mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R})$ be a chemical reaction network and let $y, y' \in \mathcal{C}$. The complex y is *directly linked* to y', denoted as $y \leftrightarrow y'$, if either $y \rightarrow y' \in \mathcal{R}$ or $y' \rightarrow y \in \mathcal{R}$. Furthermore, y is *linked* to y', denoted as $y \sim y'$, if either $y = y'$ or there exist $y_1, \ldots, y_m \in \mathcal{C}$ with $y = y_1 \leftrightarrow y_2 \leftrightarrow \cdots \leftrightarrow y_m = y'$. Finally, y and y' belong to the same linkage class, if and only if $y \sim y'$.

So the linkage classes are the equivalence classes under the equivalence relation $\sim$. The two linkage classes of example 4.1 are $\{A + B, 2C\}$ and $\{C, 0\}$. The total number of linkage classes in a reaction network is denoted by l.

Finally, the *deficiency* of a reaction network can be defined as:

Definition 4.1.3. Let $\mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R})$ be a chemical reaction network and let N be the stoichiometric matrix associated to it. Let $m = |\mathcal{C}|$ be the number of complexes, $s = rank(N)$ the rank of the stoichiometric matrix and l the number of linkage classes. The *deficiency* δ of $\mathcal{N}$ is then defined as

$$\delta = m - l - s.$$

The *deficiency of a linkage class* in defined straightforward as the deficiency obtained from a reaction network only comprised of this linkage class. Note that the deficiency is dependent on the number of complexes, the number of linkage classes and the rank of the stoichiometric matrix, all of which are fully described by the structure of the network alone. Hence, the deficiency is independent of parameter values for k. To see that the deficiency is non-negative, suppose that the number of linkage classes and the number of complexes are fixed. If two complexes within a linkage class are connected by a new edge, *i.e.* if a new reaction is introduced, the rank of the stoichiometric matrix is unaltered because the additional reaction is linearly dependent on the already existing reactions forming the linkage class. Hence, two networks with the same complexes and the same linkage classes have the same rank for their stoichiometric matrices. Furthermore, one can construct a network which preserves a given set of compounds and linkage classes and only consists of $m - l$ reactions. Therefore, the rank of the stoichiometric matrix of such a network is at most $m - l$, leading to the general inequality $s \leq m - l$ (Feinberg, 1995a).

Characterization of the deficiency

To get an intuition of what aspects of the structure are described by the deficiency, two different classes of reaction networks are analyzed in more detail. The first class provides a partial characterization of reaction networks of deficiency 0 and the second one allows to create reaction networks of arbitrary large deficiency.

Definition 4.1.4. Let $\mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R})$ be a chemical reaction network and let $A, B \in \mathcal{C}$. A and B are called *disjunct* from each other if supp$(A) \cap$ supp$(B) = \emptyset$, i.e. if the sets of species constituting the two complexes A and B are disjunct. A reaction network $\mathcal{N}$ is called disjunct, if all of its complexes are pairwise disjunct.

Theorem 4.1.5. If a reaction network is disjunct, then the deficiency is zero.

Several intermediate results are needed to prove this theorem.

Lemma 4.1.6. Let $\mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R})$ be a disjunct reaction network with m complexes $C_1, \ldots, C_m$. Let $\mathcal{N}^*$ be another reaction network with species $\mathcal{C}^* = \{S_1, \ldots, S_m\}$, complexes $\mathcal{C}^* = \{\{S_1\}, \ldots, \{S_m\}\}$ and reactions $\{S_i\} \to \{S_j\}$ if and only if $C_i \to C_j$ in $\mathcal{N}$. Then $\mathcal{N}^*$ is disjunct and $\delta(G) = \delta(G^*)$.

Proof. The network $\mathcal{N}^*$ is disjunct by construction. Furthermore, the number of complexes and linkage classes is the same in both $\mathcal{N}$ and $\mathcal{N}^*$, so $m = m^*$ and $l = l^*$. Since $\mathcal{N}$ is disjunct, all rows in the associated stoichiometric matrix belonging to the same complex are pairwise linearly dependent and linearly independent of all other rows. Hence, each complex can be replaced by a complex containing just one species, while the reactions between the complexes are preserved. Therefore, the rank of the stoichiometric matrices associated with $\mathcal{N}$ and $\mathcal{N}^*$ respectively, is the same. Hence, $\delta(\mathcal{N}) = \delta(\mathcal{N}^*)$. □

Lemma 4.1.7. For every disjunct reaction network $\mathcal{N}$, the deficiency of the entire network is the sum of deficiencies of the linkage classes.

Proof. The number of linkage classes is l and the number of complexes in the linkage classes sum up to m. Since $\mathcal{N}$ is disjunct, the complexes can be reordered, such that the stoichiometric matrix is in block diagonal form, where each block corresponds to one linkage class. The rank of the stoichiometric matrix is the sum of the ranks of each block. Hence, the total rank s is the sum of ranks obtained from the restriction of the stoichiometric matrix to each linkage class. □

Lemma 4.1.8. Let $\mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R})$ be a disjunct reaction network with only one linkage class. Furthermore, let $|\mathcal{R}| = |\mathcal{C}| - 1$, *i.e.* the complexes are linked in a path. Then the deficiency of $\mathcal{N}$ is 0.

Proof. Since G is disjunct, the stoichiometric matrix has full rank, *i.e.* $m - 1$. Therefore, $\delta(\mathcal{N}) = m - 1 - (m - 1) = 0$. □

The deficiency is dependent on the connectivity between complexes and not on the reactions which establish the connections.

Lemma 4.1.9. Two reaction networks with the same complexes and the same linkage classes have the same deficiency.

Proof. See Feinberg (1995a) Remark 2.9. □

Finally, theorem 4.1.5 can be proved.

Proof of theorem 4.1.5. For a given reaction network $\mathcal{N}$, a network $\mathcal{N}^*$ with $\delta(\mathcal{N}) = \delta(\mathcal{N}^*)$ can be created according to Lemma 4.1.6. The network $\mathcal{N}^*$ is also disjunct, so the deficiency of $\mathcal{N}^*$ is the sum of the deficiencies of each linkage class (Lemma 4.1.7). Each linkage class of $\mathcal{N}^*$ can be simplified to a path while preserving the deficiency (Lemma 4.1.9). Finally, Lemma 4.1.8 ensures that the deficiency of each linkage class and hence the deficiency of the network $\mathcal{N}$ itself is 0. □

The following example shows how to construct a reaction network of arbitrary deficiency d (Guberman, 2003).

Example 4.1.10.

$$\begin{array}{rcl} A & \to & B \\ 2A & \to & 2B \\ & \vdots & \\ (d+1)A & \to & (d+1)B \end{array}$$

Clearly, the deficiency d is achieved by interrelating complexes through common species. Taking this example and theorem 4.1.5, the deficiency of a network can be seen as a characteristic which describes the internal dependencies of complexes that do not arise from reactions but from shared species that constitute the complexes.

4.1.2 Overview of existing theorems and findings

The deficiency of a reaction network allows to draw rigorous conclusions about multistability of the network. However, the following overview of theorems does not cover all possible networks and therefore some networks remain inconclusive. Extending CRNT to also cover the remaining is an ongoing work.

Deficiency Zero Theorem

If the deficiency of a reaction network is zero, then, assuming mass-action kinetics for all reactions, no set of positive parameter values for k exists that leads to multiple steady states. An example of such a network was already given by 4.1.

Theorem 4.1.11 (Deficiency Zero Theorem)**.** Let $\mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R})$ be a chemical reaction network. If the deficiency of $\mathcal{N}$ is zero, then, under the assumption of mass-action kinetics, $\mathcal{N}$ cannot obtain multiple positive steady states.

Proof. A proof of a slightly stricter version of this theorem can be found in Feinberg (1995a). □

Deficiency One Theorem

The Deficiency Zero Theorem can be extended to networks of higher deficiency which satisfy some additional mild conditions, that can be expressed by some further definitions. Each linkage class can be further decomposed into *strong linkage classes*. These are defined as sets of complexes, such that there exists a directed path between any two complexes within the same strong linkage class. Furthermore, a strong linkage class is called *terminal strong linkage class*, if it contains no complex that reacts to a complex in a different strong linkage class. See the following two examples for clarification:

$$B \leftarrow A \rightarrow C, \tag{4.5}$$

$$B \rightleftarrows A \rightarrow C. \tag{4.6}$$

In both examples, the reaction network consists of three complexes A, B, and C and one linkage class. In example 4.5 there are three strong linkage classes {A},{B}, and {C} of which two are terminal ({B} and {C}). In contrast, example 4.6 consists of only two strong linkage classes ({A, B} and {C}) and one single terminal strong linkage class ({C}).

The Deficiency One Theorem can be applied to reaction networks for which the following conditions are fulfilled:

(1) The deficiency of each linkage class is less or equal to one.

(2) The deficiencies of all linkage classes sum up to the deficiency of the entire network.

(3) Each linkage class contains precisely one terminal strong linkage class.

Again, for such a network, no positive parameter values for k exist that allow multistability.

Theorem 4.1.12 (Deficiency One Theorem)**.** Let $\mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R})$ be a chemical reaction network of deficiency δ and let $\mathcal{L}$ be the decomposition of $\mathcal{C}$ into linkage classes. Furthermore, let δ_L be the deficiency of the network induced by $L \in \mathcal{L}$. If

1. $\delta_L \leq 1 \quad \forall L \in \mathcal{L}$,
2. $\sum\limits_{L \in \mathcal{L}} \delta_L = \delta$,
3. each linkage class contains precisely one terminal strong linkage class,

the reaction network $\mathcal{N}$, taken with mass-action kinetics, can not obtain multiple positive steady states.

Proof. A proof is given in Feinberg (1995a). □

The usefulness of this theorem can be seen in the following example.

$$\begin{tikzcd}
B + C \arrow[r] & A \arrow[r] & 0 \arrow[r, shift left] \arrow[d, leftrightarrow] & B \arrow[l, shift left] \\
 & & C & \\
 & A + C \arrow[r, shift left] & D \arrow[l, shift left] &
\end{tikzcd} \tag{4.7}$$

The deficiency of this example reaction network is 1 ($m = 7, l = 2, s = 4$), hence the Deficiency Zero Theorem is not applicable. However, the deficiency of the upper and the lower linkage class is 1 and 0, respectively, satisfying the first two conditions of the Deficiency One Theorem. Furthermore, each linkage class contains only one terminal strong linkage class ($\{0, B, C\}$ for the upper linkage class and $\{A + C, D\}$ for the lower one). Hence, the deficiency one theorem can be applied and guarantees that network 4.7 cannot admit multiple steady states. See Feinberg (1995a) for more examples showing the necessity of each of the three conditions.

Deficiency One Algorithm

The Deficiency One Theorem can be applied to reaction networks of any deficiency, as long as the preconditions are met. However, there are even networks of deficiency 1, which do not fall into this class. The following example was taken from Ellison (1998).

Example 4.1.13.

$$0 \rightleftarrows A \rightleftarrows B$$
$$2A + B \rightleftarrows 3A$$

The deficiency is 1, but condition 2 of the deficiency one theorem is violated. For some of these networks, the *Deficiency One Algorithm* can be applied to analyze the capacity for supporting multiple positive steady states. For each network the algorithm constructs a set of linear inequality systems. If any of these systems has a solution which satisfies some further mild conditions, the network can exhibit multistability. Furthermore, the algorithm also provides the precise values for the kinetic parameters k and the metabolite concentrations for the steady states.

Again, some more definitions are needed (Ellison, 1998).

Definition 4.1.14. Let n be the number of species in a reaction network $\mathcal{N}$ and M the corresponding stoichiometric matrix. A vector $v \in \mathbb{R}^n$ is called *sign compatible* if there exists a vector $\sigma \in \text{im}(M)$ and some positive numbers p_i such that $v_i = p_i \sigma_i$ for all $i = 1, \ldots, n$.

Definition 4.1.15. Let $\mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R})$. A pair of complexes $\{y, y'\}$ is called a *cut pair* if the following two conditions are fulfilled:

1. $y \to y' \in \mathcal{R}$ or $y' \to y \in \mathcal{R}$
2. Removing the reactions between y and y' disconnects the linkage class containing both y and y'.

Definition 4.1.16. A reaction network $\mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R})$ is *regular* if it satisfies the following three conditions:

1. For every $y \to y' \in \mathcal{R}$, there exists a positive number $\alpha_{y \to y'}$ such that
$$\sum_{y \to y' \in \mathcal{R}} \alpha_{y \to y'}(y' - y) = 0$$
2. Each linkage class contains only one terminal strong linkage class.
3. If $y \to y' \in \mathcal{R}$ or $y' \to y \in \mathcal{R}$, and y and y' belong to the same terminal strong linkage class, then $\{y, y'\}$ is a cut pair.

For each regular reaction network of deficiency 1 the algorithm decides whether the network has the capacity to support multiple steady states or not. In brief, the algorithm divides the set of complexes into three different groups according to their membership to terminal strong linkage classes. There are only finitely many possible decompositions into such groups. For each decomposition a system of linear equalities and inequalities between the complexes is established, depending on which of the three groups contains the complexes. The reaction network can obtain multiple steady states if and only if at least one of the systems has a sign compatible solution. For a full description of the Deficiency One Algorithm see Ellison (1998) or Feinberg (1988). The application of this algorithm to example 4.1.13, for which the Deficiency One Theorem does not apply, shows that this network cannot support multiple steady states.

Advanced Deficiency Algorithm

The Deficiency One Theorem can be generalized to be applicable to reaction networks of any deficiency without further constraints (Ellison, 1998). However, the drawback of this *Advanced Deficiency Algorithm* is that for some reaction networks a huge number of non-linear inequality systems have to be solved. Even for networks of moderate size this can become computationally intractable.

Similar to the Deficiency One Algorithm, the Advanced Deficiency Algorithm creates systems of inequalities which have to be checked for falsifiability. The major difference is that here the reactions and not the complexes are grouped together. The grouping is based on the concept of colinearity classes, which are corresponding to linkage classes. In short, two reactions belong to the same colinearity class, if the flux ratio between them is the same for all steady states.

Definition 4.1.17. An *orientation* $\mathcal{O}$ of a reaction network $\mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R})$ is a subset of $\mathcal{R}$, such that $\mathcal{O}$ contains each irreversible reaction and precisely one direction of each reversible reaction from $\mathcal{R}$. Let $M_{\mathcal{O}}$ be the stoichiometric matrix restricted to reactions in $\mathcal{O}$ and let $(v^1, \ldots, v^d)$ be a basis of $\ker(M_{\mathcal{O}})$. For each reaction $y \to y'$ let $v^i_{y \to y'}$ denote the component of v^i which corresponds to $y \to y'$. Then a vector $w_{y \to y'}$ can be defined as $[v^1_{y \to y'}, \ldots, v^d_{y \to y'}]$. Two reactions $y \to y'$ and $p \to p'$ are in the same *colinearity class* if and only if there exists a non-zero number c such that $w_{y \to y'} = c w_{p \to p'}$.

There exist many different possibilities for grouping the reactions, leading to a huge number of inequality systems which are in general non-linear. However, a sufficient but not necessary condition is given in Ellison (1998) to guarantee linearity of all constructed inequality systems. Again, the reaction network can obtain multiple steady states if and only if at least one of these systems has a sign compatible solution. Therefore, even in case of non-linear inequalities it makes sense to look at the linear part of the system alone, as unsatisfiability of the linear inequalities implies unsatisfiability of the entire system. A clear characterization of reaction networks which leads to purely linear systems as well as finding concise pruning rules for non-linear systems remain as open problems.

Both the Deficiency One Algorithm and the Advanced Deficiency Algorithm are implemented in the *chemical reaction network toolbox* (Feinberg and Ellison, 2000). However, the current version is restricted to reaction networks of at most 20 complexes due to computational limitations, which is already too small for most biological networks.

4.1.3 Advancements and extensions

Several approaches exist which either extend CRNT or are closely related to it. Two will be explained briefly. Also noteworthy are the algebraic approach presented by Gatermann and Wolfrum (2005) and the work on consequences of multistability in subnetworks by Li (1997), although somewhat restricted to reaction networks of very special structure.

Subnetwork analysis

Conradi et al. (2007) proposed a method to draw conclusions about multistationarity of a reaction network by analyzing special subnetworks. In particular, they investigated subnetworks defined by elementary flux modes (see Section 1.3.2) called *stoichiometric generators*. If the reaction network consists of m complexes and r reactions, then I_c is defined as the $m \times r$ incidence matrix of the complex network. Hence, each reaction is represented by a column of I_c with exactly one entry -1 for the educt and one entry 1 for the product complex. All other entries are zero. An elementary flux mode E is a stoichiometric generator if $I_c E \neq 0$. For a stoichiometric generator E it can be shown that if every linkage class of the subnetwork induced by E contains only one terminal strong linkage class, then the deficiency one algorithm can be applied to the subnetwork. If the subnetwork is capable of supporting two steady states, then these steady states might be extended to the initial network. Therefore, some further conditions arising from the implicit function theorem have to be met, which can be tested by solving a system of linear equations. However, if no multistability is found for any of the subnetworks, nothing can be concluded for the entire network. Altogether, this approach allows to analyze reaction networks of previously intractable sizes by decomposing them into sophisticated subnetworks. Still, even the calculation of all elementary flux modes can be computational demanding (Klamt and Stelling, 2002; Acuña et al., 2008).

SR-graphs

An important class of reaction networks, especially for biochemical engineers, is obtained from the context of *continuous flow stirred tank reactors* (CFSTR). For a CFSTR, constant temperature and pressure is assumed as well as constant import and export of all species. Although these assumptions are far from biological settings, some of the results for CFSTR networks translate to general reaction networks and, therefore, can be applied to metabolic networks.

Using CFSTR networks, Craciun and Feinberg (2005) introduce an algebraic approach by defining a polynomial function $p(c, k)$ associated with the reaction network, where c is the vector of species concentration and k is the vector of kinetic parameters for the rate equations. This function consists of the right-hand side of the system of differential equations under mass action kinetics. It is shown that if $p(c,k)$ is injective for all positive choices of k, the reaction network cannot support multiple positive steady states. However, the opposite is not true unless the nonnegativity condition for the steady states is dropped. For a given reaction network, $p(c,k)$ is proved to be injective if and only if all the coefficients in the expansion of $\det(\frac{\partial p}{\partial c}(c,k))$ are nonnegative. Furthermore, a criterion is introduced which decides if a reaction network, for which $p(c,k)$ is not injective, can admit multiple positive steady states or not. This criterion is restricted to reaction networks in the CFSTR context (all species can be exported) and can be calculated with the help of standard computer algebra systems. However, one might additionally have to solve some time-consuming polynomial optimization problems, rendering the calculation practically impossible.

A more graph-based approach was presented by Craciun and Feinberg (2006b), where multiple steady states could be ruled out for reaction networks based on properties of the associated *species-reaction graph* (SR-graph).

Definition 4.1.18. For a reaction network $\mathcal{N} = (\mathcal{S}, \mathcal{C}, \mathcal{R})$ the SR-graph $G_{\mathcal{N}}$ is a bipartite, undirected graph with one *species node* for every species in $\mathcal{S}$ and precisely one *reaction node* for every reversible or irreversible reaction in $\mathcal{R}$. A species node s is connected to a reaction node r given by $y \to y'$ or $y \rightleftarrows y'$ if and only if $s \in y$ or $s \in y'$ (so s is part of at least one of the two complexes that constitute the reaction r). For $s \in y$ the edge is assigned the *complex label* y and the *stoichiometric coefficient*, which is the multiplicity of s in y. For $s \in y'$ the labeling is done accordingly.

The structure of a SR-graph is of high interest, especially cycles within such a graph need further classification.

Definition 4.1.19. Let $\mathcal{N}$ be a reaction network and $G_{\mathcal{N}}$ its associated SR-graph. A pair of edges adjacent to the same reaction node is called a *c-pair* if they bear the same complex labeling. If a cycle contains an even number of c-pairs it is called an *even-cycle*; otherwise, it is an *odd-cycle*. If alternating multiplication and division of stoichiometric coefficients of edges along a cycle gives 1, the cycle is called *one-cycle*. Two cycles are said to have an *S-R intersection* if their common edges generate a simple path from a species to a reaction node or are a disjoint union of such paths.

Clearly, a cycle can be both an even-cycle and a one-cycle. Only odd-cycle and even-cycle are mutually exclusive. For a given reaction network the SR-graph can be constructed easily and once all the cycles are identified, they can be characterized according to the definition above. Finally, with the help of the SR-graph it can be analyzed whether or not the reaction network can support multiple positive steady states.

Theorem 4.1.20. Let $\mathcal{N}$ be a reaction network (in the context of CFSTR) and $G_{\mathcal{N}}$ its associated SR-graph. If all cycles of $G_{\mathcal{N}}$ are odd-cycles or one-cycles and no two even-cycles have an S-R intersection, the reaction network $\mathcal{N}$ has not the capacity for multiple positive steady states.

The proof of this theorem and a slightly stronger version of it can be found in Craciun and Feinberg (2006b). Here, the results are still restricted to reaction networks in which all species are imported and exported with a constant positive rate. This assumption can be dropped by using *entrapped species models* (Craciun and Feinberg, 2006a), where the set of species is subdivided into two disjunct sets, one containing the species that are imported and exported as in the CFSTR context and the other one containing all other, *entrapped*, species without such transport reactions. In that sense, every reaction network can be considered as an entrapped species model. Theorem 4.1.20 can be extended to such networks under the mild modification that multiple positive steady states are not ruled out generally but *degenerated* steady states are still possible. Here, degenerated means that there are infinitely many other positive steady states in close proximity which are consistent with possible conservation constraints on the entrapped species. Degenerated steady states are pathological and do not reflect the intuition behind multistationarity in metabolic networks as they arise from fragile mathematical artifacts (Craciun and Feinberg, 2006a; Feinberg, 1987).

So far, CRNT as well as SR-graphs are based on the assumption that all reactions within the reaction network are governed by mass-action kinetics. One way to overcome this strong assumption is to model each reaction on the level of enzyme catalysis. For instance, if an enzymatic reaction $A \to B$ follows Michaelis-Menten kinetics, it can be replaced by the small network $A + E \leftrightarrows AE \to B + E$, where E is the enzyme and AE is the intermediate complex build by physical interaction between the enzyme and the substrate A. For all three reactions within the new network mass-action kinetics can be assumed and hence CRNT or SR-graphs can be applied to this new network. Several basic mechanisms for enzyme catalysis such as competitive or uncompetitive inhibition have been evaluated systematically by means of SR-graphs (Craciun et al., 2006). As one might expect, the capacity for multiple positive steady states heavily depends on the structural details of the analyzed mechanisms.

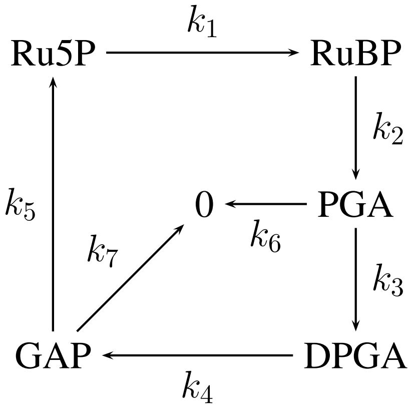
4.2 Bistability in the Calvin cycle

One of the most important metabolic pathways is the process of carbon fixation in chloroplast stroma, which is known as the *Calvin cycle*. Under consumption of ATP and NADPH, CO_2 is inserted into the metabolism to produce new carbon-rich molecules. The cycle consists of three phases: (*i*) *carboxylation*, during which the enzyme *RuBisCO* adds CO_2 to Ribulose-1,5-bisphosphate (RuBP) to get two molecules of phosphoglycerate (PGA), (*ii*) *reduction*, converting the obtained PGA into glyceraldehyde-3-phosphate (GAP) and dihydroxyacetone phosphate (DHAP), and (*iii*) *regeneration*, which recovers RuBP after several intermediate steps (Berg et al., 2002).

Due to nonlinear interplay of the participating reactions and metabolites, the Calvin cycle is a complex dynamic system. Like for any complex system, the question whether multiple steady states can be achieved is of high interest. Although multistationarity was found in comparable dynamic systems like cell cycle (Zwolak et al., 2004; Tyson et al., 2001), analysis of the Calvin cycle with respect to multiple steady states is still fragmentary, not least because of difficulties in obtaining experimental data. Pettersson and Ryde-Pettersson (1988) found two steady states for their model, of which one was shown to be unstable and therefore considered to be of no biological relevance. Nevertheless, the remaining stable steady state was in good accordance with previous experiments (Flügge et al., 1980; Heldt et al., 1977). Poolman et al. (2000) also found two steady states, which furthermore were confirmed experimentally (Poolman et al., 2001). However, the two steady states occur in leafs of different age and therefore have different capacities of utilizing the produced carbonhydrates (Olçer et al., 2001). It is still unclear to which extent these results hold within one single leaf. A systematic approach was taken by Zhu et al. (2008), using a sophisticated algorithm to find all roots of a system of polynomials. The application of this approach to a very simplistic model of the Calvin cycle revealed 40 steady states, of which 39 were biological infeasible due to extremely small or even negative metabolite concentrations. However, their analysis was limited to a given set of kinetic parameters and their close vicinity.

4.3 Hierarchy of Calvin cycle models

The methods described in Sections 4.1.2 and 4.1.3 will be used to analyze the capacity for multiple positive steady states over the entire parameter space of various Calvin cycle models. With respect to their level of abstractions these models form a hierarchy, ranging from a very simple model



$$\begin{aligned}
Ru5P &\xrightarrow{k_1} RuBP\\
RuBP &\xrightarrow{k_2} 2PGA\\
PGA &\xrightarrow{k_3} DPGA\\
DPGA &\xrightarrow{k_4} GAP\\
GAP &\xrightarrow{k_5} 0.6Ru5P\\
PGA &\xrightarrow{k_6} 0\\
GAP &\xrightarrow{k_7} 0
\end{aligned} \tag{4.8}$$

Figure 4.1: Simple model of the Calvin cycle as presented in Zhu et al. (2008). RuBP: Ribulose 1,5-bisphosphate; PGA: 3-Phosphoglycerate; DPGA: 1,3-Bisphosphoglycerate; GAP: Glyceraldehyde 3-phosphate; Ru5P: Ribulose 5-phosphate.

of only five metabolites to an elaborate one including compartmentalization of metabolites and additional pathways like sucrose synthesis.

4.3.1 Model of Zhu

The model of Zhu et al. (2008) contains only five internal metabolites, arranged as a cycle with two additional transport reactions (Figure 4.1). The simplification of the regeneration phase to the overall reaction k_5 leads to non-integer stoichiometric coefficients, which therefore do not describe the number of molecules participating in a single reaction.

Reaction network 4.8 has a deficiency of 1 and two linkage classes of deficiency 0 each. Hence, neither the Deficiency Zero Theorem (Theorem 4.1.11) nor the Deficiency One Theorem (Theorem 4.1.12) is applicable. One of the linkage classes ($\{PGA, DPGA, GAP, 0.6Ru5P, 0\}$) consists of two terminal strong linkage classes ($\{0.6Ru5P\}$ and $\{0\}$), so the network is not regular and hence violating the preconditions of the Deficiency One Algorithm. Applying the Advanced Deficiency Algorithm to network 4.8 revealed multiple steady states. However, reaction networks like 4.8 that fulfill conditions 1 and 3, but not condition 2 of definition 4.1.16 are to some extent considered *pathological*, as they allow for an infinite number of steady states (Feinberg, 1995b, 1987, Appendix 4). More importantly, the capability of obtaining multiple steady states vanishes even under very subtle changes in the network structure, which is why such networks do not represent good model candidates.

A slightly different network is obtained by changing the stoichiometry of reaction k_5 in network 4.8 into:

$$5GAP \xrightarrow{k_5} 3Ru5P \tag{4.9}$$

This modified network is of deficiency 1 and composed of two linkage classes of deficiency 0 each. But in contrast to network 4.8 the Deficiency One Algorithm is now applicable. It guarantees that no multiple positive steady states are possible, no matter what values of the mass-action kinetic parameters are chosen. But even the existence of a single steady state is not ensured and depends on some of the kinetic parameters. This can be seen by analysing the set of differential equations

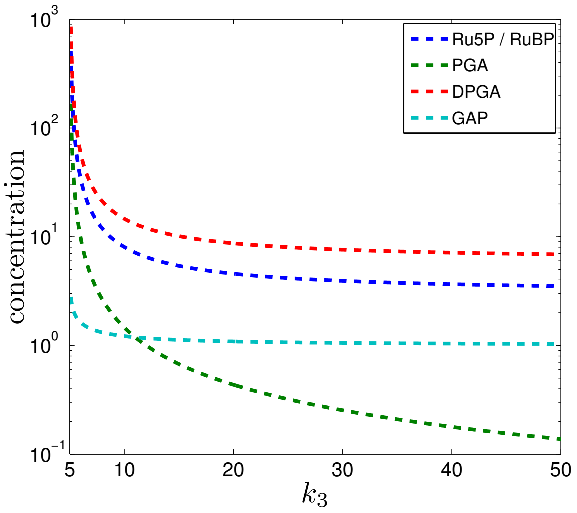


Figure 4.2: Steady state concentrations for network 4.9. The parameter k_3 is varied and all other parameters are fixed to 1. As k_3 approaches $5 \cdot k_6$, the concentrations go to infinity. For even smaller values of k_3, no steady state exists at all. All steady states are unstable, as indicated by the dotted lines. $Ru5P$ and $RuBP$ always have the same concentration.

associated with the modified network.

$$\begin{aligned}
\frac{dRu5P}{dt} &= -k_1 \cdot Ru5P + 3 \cdot k_5 \cdot GAP^5 \\
\frac{dRuBP}{dt} &= k_1 \cdot Ru5P - k_2 \cdot RuBP \\
\frac{dPGA}{dt} &= 2 \cdot k_2 \cdot RuBP - k_3 \cdot PGA - k_6 \cdot PGA \\
\frac{dDPGA}{dt} &= k_3 \cdot PGA - k_4 \cdot DPGA \\
\frac{dGAP}{dt} &= k_4 \cdot DPGA - 5 \cdot k_5 \cdot GAP^5 - k_7 \cdot GAP
\end{aligned} \tag{4.10}$$

To obtain the steady state solutions, the left-hand sides of 4.10 are set to zero. Expressing every equation in terms of $RuBP$ leads to $Ru5P = \frac{k_2}{k_1} \cdot RuBP$ and $PGA = \frac{2 \cdot k_2}{k_3 + k_6} \cdot RuBP$. Subsequent substitutions give $DPGA = \frac{k_3}{k_4} \cdot PGA = \frac{2 \cdot k_2 \cdot k_3}{k_4 \cdot (k_3 + k_6)} \cdot RuBP$ and $GAP = \sqrt[5]{\frac{k_2 \cdot}{3 \cdot k_5} \cdot RuBP}$. Finally, this leads to:

$$\begin{aligned}
0 &= k_4 \cdot DPGA - k_7 \cdot GAP - 5 \cdot k_5 \cdot GAP^5 \\
&= \frac{2 \cdot k_2 \cdot k_3}{(k_3 + k_6)} - k_7 \cdot \sqrt[5]{\frac{k_2}{3 \cdot k_5} \cdot RuBP} - \frac{5}{3} \cdot k_2 \cdot RuBP \\
&= \left(\frac{2 \cdot k_2 \cdot k_3}{(k_3 + k_6)} - k_7 \cdot \sqrt[5]{\frac{k_2}{3 \cdot k_5 \cdot RuBP^4}} - \frac{5}{3} \cdot k_2 \right) \cdot RuBP
\end{aligned}$$

The latter equation has five distinct solutions, of which only one is positive as well as real, given by

$$RuBP = \sqrt[4]{\frac{k_2 \cdot k_7^5}{3 \cdot k_5 \left(\frac{2 \cdot k_2 \cdot k_3}{k_3 + k_6} - \frac{5}{3} \cdot k_2 \right)^5}}$$

for $\frac{2 \cdot k_2 \cdot k_3}{k_3 + k_6} - \frac{5}{3} \cdot k_2 > 0$ or equivalently $k_3 > 5 \cdot k_6$. This imposes a lower threshold for k_3 in terms of k_6. More precisely, if k_3 is below this threshold, not even a single steady state can occur, no matter what values are obtained for all remaining parameters k. Figure 4.2 shows the change of steady state concentration for varying k_3, while all other k's are fixed to one.

To analyze the stability of the steady state solution, one has to calculate the eigenvalues of the Jacobian matrix J of system 4.10.

$$J = \begin{bmatrix} -k_1 & 0 & 0 & 0 & 15 \cdot k_5 \cdot GAP^4 \\ k_1 & -k_2 & 0 & 0 & 0 \\ 0 & 2 \cdot k_2 & -k_3 - k_6 & 0 & 0 \\ 0 & 0 & k_3 & -k_4 & 0 \\ 0 & 0 & 0 & k_4 & -25 \cdot k_5 \cdot GAP^4 - k_7 \end{bmatrix}$$

The roots of the characteristic polynomial $\chi_J(\lambda) = \det(J - \lambda \cdot E)$, where E stands for the identity matrix, determine the eigenvalues of J. The characteristic polynomial can be calculated by a subsequent minor expansion across the first row, leading to

$$\chi_J(\lambda) = (k_1 - \lambda)(-k_2 - \lambda)(-k_3 - k_6 - \lambda)(-k_4 - \lambda)(-25k_5GAP^4 - k_7 - \lambda) + 30k_1k_2k_3k_4k_5GAP^4$$

The expansion of χ_J to the form $\chi_J(\lambda) = \alpha_0\lambda^0 + \alpha_1\lambda^1 + \alpha_2\lambda^2 + \alpha_3\lambda^3 + \alpha_4\lambda^4 + \alpha_5\lambda^5$ shows that α_1 to α_5 are negative. The remaining coefficient α_0 can be expressed as $\alpha_0 = k_1k_2k_4\left(k_7\left(-k_3 - k_6\right) + 5k_5GAP^4\left(k_3 - 5k_6\right)\right)$. Substituting $GAP^4 = \frac{k_7}{k_5(\frac{6k_3}{k_3+k_6}-5)}$, obtained from the steady state relation between $RuBP$ and GAP, one finally gets $\alpha_0 = 4k_1k_2k_4k_7(k_3 + k_6)$. Hence, α_0 is always positive. From Descartes' rule of sign it follows that $\chi_J(\lambda)$ has exactly one positive root and therefore one positive eigenvalue. Consequently, the entire parameter space of network 4.9 does not contain any stable steady states, which clearly makes this network an extremely poor model.

So far, the analysis was based on assuming simple mass-action kinetics for all participating reactions. This limitation can be overcome by breaking down more complex kinetic rate laws into elementary enzymatic mechanisms, which in turn follow mass-action kinetics. Applying this approach, network 4.8 can be rewritten as:

$$\begin{array}{rcccl} Ru5P + E_1 & \overset{k_1}{\underset{k_2}{\leftrightarrows}} & Ru5PE_1 & \xrightarrow{k_3} & RuBP + E_1 \\ RuBP + E_2 & \overset{k_4}{\underset{k_5}{\leftrightarrows}} & RuBPE_2 & \xrightarrow{k_6} & 2PGA + E_2 \\ PGA + E_3 & \overset{k_7}{\underset{k_8}{\leftrightarrows}} & PGAE_3 & \xrightarrow{k_9} & DPGA + E_3 \\ DPGA + E_4 & \overset{k_{10}}{\underset{k_{11}}{\leftrightarrows}} & DPGAE_4 & \xrightarrow{k_{12}} & GAP + E_4 \\ 5GAP + E_5 & \overset{k_{13}}{\underset{k_{14}}{\leftrightarrows}} & GAPE5 & \xrightarrow{k_{15}} & 3Ru5P + E_5 \\ PGA + E_6 & \overset{k_{16}}{\underset{k_{17}}{\leftrightarrows}} & PGAE_6 & \xrightarrow{k_{18}} & E_6 \\ GAP + E_7 & \overset{k_{19}}{\underset{k_{20}}{\leftrightarrows}} & GAPE_7 & \xrightarrow{k_{21}} & E_7 \end{array} \qquad (4.11)$$

Network 4.11 has a deficiency of 2 and is composed of seven linkage classes, each of deficiency 0. Therefore, neither Deficiency Zero Theorem nor Deficiency One Theorem nor Deficiency One Algorithm are applicable. Furthermore, since network 4.11 consists of 21 complexes, it already exceeds the computational capabilities of the CRNT Toolbox (Feinberg and Ellison, 2000) to run the Advanced Deficiency Algorithm.

The SR-graph of network 4.11 is shown in Figure 4.3. It contains one large even-cycle, comprising ten consecutive c-pairs. Furthermore, alternate multiplication and division of the stoichiometric coefficients along the cycle does not give the result 1, so this cycle is not a one-cycle. Thus

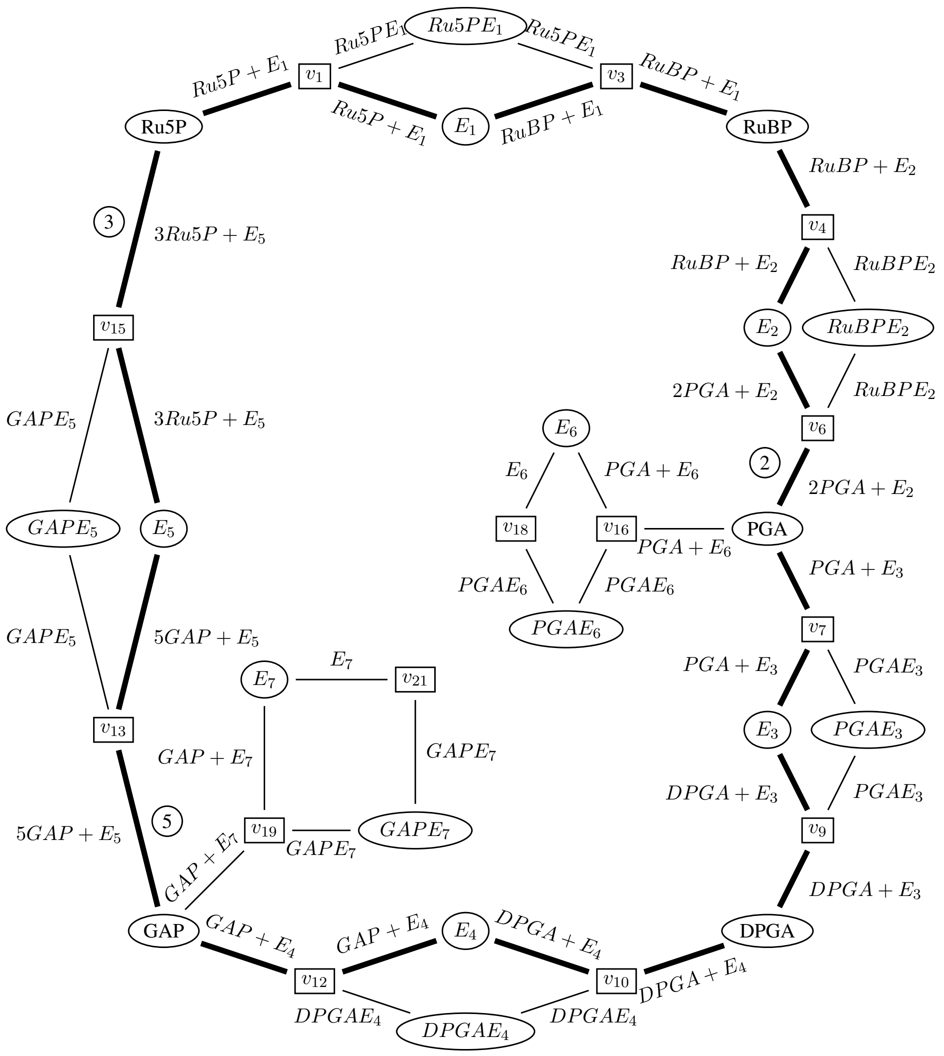


Figure 4.3: SR-graph of network 4.11. If not explicitly stated otherwise, the stoichiometric coefficients are always 1. The bold lines form an even-cycle which is not a one-cycle. Hence Theorem 4.1.20 cannot be applied and therefore no conclusions about multistability can be drawn from the SR-graph.

the preconditions of Theorem 4.1.20 are violated, leaving the question of multiple positive steady states open.

Subnetwork analysis revealed only two elementary modes for network 4.11, which arise from shutting down one of the two transporter reactions [1]. Both of them are capable of obtaining multiple steady states, which can be calculated by the CRNT Toolbox. Furthermore, by means presented by Conradi et al. (2007) these steady states of the subnetworks induced by the elementary modes can be extended to the full network. To see this, consider the following system of differential

[1] $\nu_1^{EM} = \{3, 0, 3, 3, 0, 3, 5, 0, 5, 5, 0, 5, 1, 0, 1, 1, 0, 1, 0, 0, 0\}$
$\nu_2^{EM} = \{3, 0, 3, 3, 0, 3, 6, 0, 6, 6, 0, 6, 1, 0, 1, 0, 0, 0, 1, 0, 1\}$

k_1	=	1.3666169	k_8	=	1	k_{15}	=	0.26920841
k_2	=	1	k_9	=	10.30969	k_{16}	=	1
k_3	=	2.4579323	k_{10}	=	6.1626543	k_{17}	=	1
k_4	=	0.99119923	k_{11}	=	1	k_{18}	=	1
k_5	=	1	k_{12}	=	6.6747969	k_{19}	=	1.5770407
k_6	=	2.0237445	k_{13}	=	15.141035	k_{20}	=	1
k_7	=	9.9649223	k_{14}	=	1	k_{21}	=	1.7182818

Table 4.1: Parameter assignment for the system of differential equations 4.12 which allow for multiple positive steady states. Parameters set to 1 are precisely those associated with reactions that are not present in the first elementary mode which was used to construct a subnetwork. For simplicity, all of these parameters are chosen to have the same value.

equation obtained from reaction network 4.11.

$$
\begin{aligned}
\tfrac{dRu5P}{dt} &= -k_1 \cdot Ru5P \cdot E_1 + k_2 \cdot Ru5PE_1 + 3 \cdot k_{15} \cdot GAPE_5 \\
\tfrac{dE_1}{dt} &= -k_1 \cdot Ru5P \cdot E_1 + k_2 \cdot Ru5PE_1 + k_3 \cdot Ru5PE_1 \\
\tfrac{dRu5PE_1}{dt} &= k_1 \cdot Ru5P \cdot E_1 - k_2 \cdot Ru5PE_1 - k_3 \cdot Ru5PE_1 \\
\tfrac{dRuBP}{dt} &= k_3 \cdot Ru5PE_1 - k_4 \cdot RuBP \cdot E_2 + k_5 \cdot RuBPE_2 \\
\tfrac{dE_2}{dt} &= -k_4 \cdot RuBP \cdot E_2 + k_5 \cdot RuBPE_2 + k_6 \cdot RuBPE_2 \\
\tfrac{dRuBPE_2}{dt} &= k_4 \cdot RuBP \cdot E_2 - k_5 \cdot RuBPE_2 - k_6 \cdot RuBPE_2 \\
\tfrac{dPGA}{dt} &= 2 \cdot k_6 \cdot RuBPE_2 - k_7 \cdot PGA \cdot E_3 + k_8 \cdot PGAE_3 - k_{16} \cdot PGA \cdot E_6 + k_{17} \cdot PGAE_6 \\
\tfrac{dE_3}{dt} &= -k_7 \cdot PGA \cdot E_3 + k_8 \cdot PGAE_3 + k_9 \cdot PGAE3 \\
\tfrac{dPGAE_3}{dt} &= k_7 \cdot PGA \cdot E_3 - k_8 \cdot PGAE_3 - k_9 \cdot PGAE3 \\
\tfrac{dDPGA}{dt} &= k_9 \cdot PGAE_3 - k_{10} \cdot DPGA \cdot E_4 + k_{11} \cdot DPGAE_4 \\
\tfrac{dE_4}{dt} &= -k_{10} \cdot DPGA \cdot E_4 + k_{11} \cdot DPGAE_4 + k_{12} \cdot DPGAE_4 \\
\tfrac{dDPGAE_4}{dt} &= k_{10} \cdot DPGA \cdot E_4 - k_{11} \cdot DPGAE_4 - k_{12} \cdot DPGAE_4 \\
\tfrac{dGAP}{dt} &= k_{12} \cdot DPGAE_4 - 5 \cdot k_{13} \cdot GAP^5 \cdot E_5 + 5 \cdot k_{14} \cdot GAPE_5 - k_{19} \cdot GAP \cdot E_7 \\
&\quad + k_{20} \cdot GAPE_7 \\
\tfrac{dE_5}{dt} &= -k_{13} \cdot GAP^5 \cdot E_5 + k_{14} \cdot GAPE_5 + k_{15} \cdot GAPE_5 \\
\tfrac{dGAPE_5}{dt} &= k_{13} \cdot GAP^5 \cdot E_5 - k_{14} \cdot GAPE_5 - k_{15} \cdot GAPE_5 \\
\tfrac{dE_6}{dt} &= -k_{16} \cdot PGA \cdot E_6 + k_{17} \cdot PGAE_6 + k_{18} \cdot PGAE_6 \\
\tfrac{dPGAE_6}{dt} &= k_{16} \cdot PGA \cdot E_6 - k_{17} \cdot PGAE_6 - k_{18} \cdot PGAE_6 \\
\tfrac{dE_7}{dt} &= -k_{19} \cdot GAP \cdot E_7 + k_{20} \cdot GAPE_7 + k_{21} \cdot GAPE_7 \\
\tfrac{dGAPE_7}{dt} &= k_{19} \cdot GAP \cdot E_7 - k_{20} \cdot GAPE_7 - k_{21} \cdot GAPE_7
\end{aligned}
\tag{4.12}
$$

Using parameters shown in Table 4.1, this system does have the capability to obtain multiple positive steady states as can be seen by the two steady states presented in Table 4.2. Furthermore, Figure 4.4 shows the corresponding bifurcation diagram for some of the metabolites, using the sum of the concentration of E_3 and $PGAE_3$ as bifurcation parameter.

Besides the simplicity of this model and the coarse modelling of the Michaelis-Menten kinetics, one further concern arises from the parameter values and the steady state concentrations of the metabolites. Since the applied method only aims at answering whether multiple positive steady states might occur, the resulting values for the parameters and metabolite concentrations are clearly outside any biological meaningful range. Furthermore, one of the two observed steady states is unstable and hence of no biological relevance. However, there is some freedom in choosing those parameters associated with reactions not included in the subnetwork induced by the elementary mode under consideration. This can be further exploited to test whether multistability also occurs for biological feasible parameter values and metabolite concentrations. Interestingly, an isolated

metabolite	steady state 1	steady state 2	metabolite	steady state 1	steady state 2
$Ru5P$	1.8031031	6.3319974	E_4	2.0319109	1.1136257
E_1	2.6499371	1.2957393	$DPGAE_4$	1.310466	2.2287516
$Ru5PE_1$	1.8883671	3.242564	GAP	0.6439413	1.2921397
$RuBP$	2.1738771	7.6340526	E_5	4.3510314	0.2296564
E_2	3.2184725	1.573736	$GAPE_5$	5.7470699	9.8684449
$RuBPE_2$	2.2935102	3.9382467	E_6	1.4641405	0.9364545
PGA	0.7319781	2.2714297	$PGAE_6$	0.5358594	1.0635454
E_3	1.3155187	0.7209939	E_7	1.5754002	1.2367929
$PGAE_3$	0.8484346	1.4429594	$GAPE_7$	0.5885531	0.9271604
$DPGA$	0.8031938	2.4924215			

Table 4.2: Two different positive steady states obtained from system 4.12 using the parameters shown in Table 4.1. The first steady state is unstable while the second one is stable.

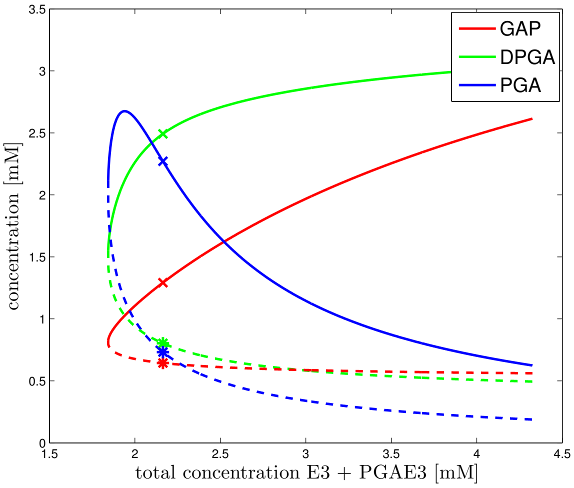


Figure 4.4: Bifurcation diagram for the system 4.12 using the parameters from Table 4.1. Stable steady states are depicted by a solid line, unstable steady states by a dashed line. The stars and crosses mark the concentrations at steady state 1 and 2, respectively (Table 4.2). The sum of concentrations of E_3 and $PGAE_3$ is chosen as bifurcation parameter.

reaction network of the form of $A + E \leftrightarrows AE \longrightarrow B$, which is exactly the set of reactions that were included into network 4.8 to simulate Michaelis-Menten kinetics, does not support multiple steady states on its own (Craciun et al., 2006). Therefore, the fact that multiple steady states exist for network 4.11 and not for network 4.8 does not arise from local structural properties but rather from the overall structure of the entire network.

4.3.2 Models of Pettersson and Poolman

A more detailed model of the Calvin cycle was introduced by Pettersson and Ryde-Pettersson (1988), consisting of 18 metabolites and 20 reactions, out of which 9 are irreversible. Here, the three phases of the Calvin cycle as well as starch synthesis are described explicitly, whereas the light reaction is modelled as an overall and simplified reaction converting ADP and P into ATP. This model was extended by Poolman et al. (2001) who introduced, among other things, a starch degradation reaction. The reaction network looks as follows, where reaction k_{32} only occurs in Poolman's model (Figure 4.5).

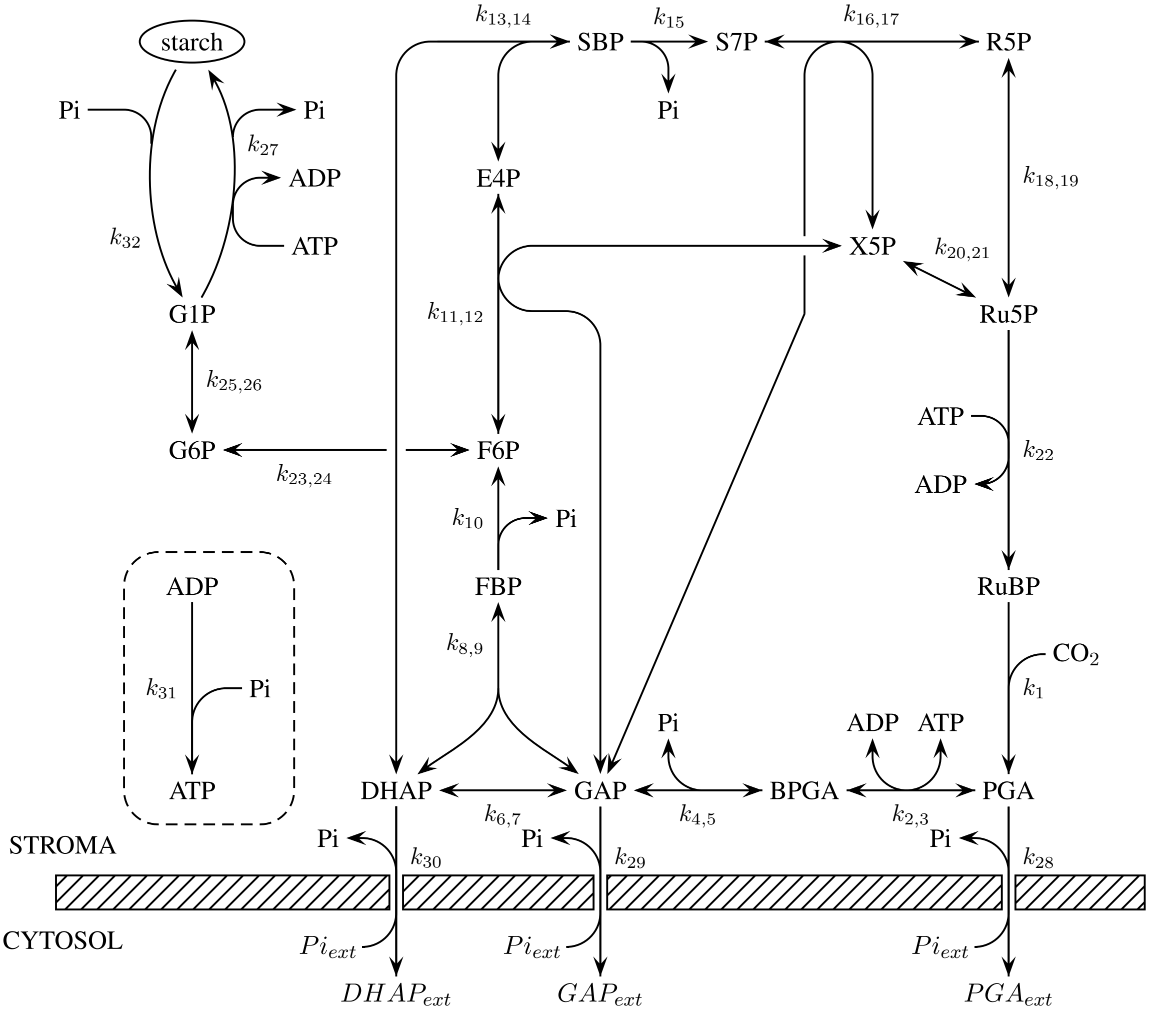


Figure 4.5: Graphical representation of models of Pettersson and Ryde-Pettersson (1988) and Poolman et al. (2001). The concentrations of $CO_2, Pi_{ext}, PGA_{ext}, GAP_{ext}, DHAP_{ext}$ and starch are kept constant. PGA, 3-phosphoglyceric acid; BPGA, 2,3-bisphosphoglyceric acid; GAP, glyceraldehyde 3-phosphate; DHAP, dihydroxyacetone phosphate; FBP, fructose 1,6-bisphosphate; F6P, fructose 6-phosphate; G6P, glucose 6-phosphate; G1P, glucose 1-phosphate; E4P, erythrose 4-phosphate; SBP, sedoheptulose 1,7-bisphosphate; S7P, sedoheptulose 7-phosphate; R5P, ribose 5-phosphate; X5P, xylulose 5-phosphate; RuSP, ribulose 5-phosphate; RuBP, ribulose 1,5-bisphosphate.

$$
\begin{array}{rcl rcl}
RuBP & \xrightarrow{k_1} & 2PGA & R5P & \underset{k_{19}}{\overset{k_{18}}{\leftrightarrows}} & Ru5P \\
PGA+ATP & \underset{k_3}{\overset{k_2}{\leftrightarrows}} & BPGA+ADP & X5P & \underset{k_{21}}{\overset{k_{20}}{\leftrightarrows}} & Ru5P \\
BPGA & \underset{k_5}{\overset{k_4}{\leftrightarrows}} & Pi+GAP & Ru5P+ATP & \xrightarrow{k_{22}} & RuBP+ADP \\
GAP & \underset{k_7}{\overset{k_6}{\leftrightarrows}} & DHAP & F6P & \underset{k_{24}}{\overset{k_{23}}{\leftrightarrows}} & G6P \\
GAP+DHAP & \underset{k_9}{\overset{k_8}{\leftrightarrows}} & FBP & G6P & \underset{k_{26}}{\overset{k_{25}}{\leftrightarrows}} & G1P \\
FBP & \xrightarrow{k_{10}} & F6P+Pi & G1P+ATP & \xrightarrow{k_{27}} & ADP+2Pi \\
F6P+GAP & \underset{k_{12}}{\overset{k_{11}}{\leftrightarrows}} & E4P+X5P & PGA & \xrightarrow{k_{28}} & Pi \\
DHAP+E4P & \underset{k_{14}}{\overset{k_{13}}{\leftrightarrows}} & SBP & GAP & \xrightarrow{k_{29}} & Pi \\
SBP & \xrightarrow{k_{15}} & S7P+Pi & DHAP & \xrightarrow{k_{30}} & Pi \\
S7P+GAP & \underset{k_{17}}{\overset{k_{16}}{\leftrightarrows}} & X5P+R5P & ADP+Pi & \xrightarrow{k_{31}} & ATP \\
 & & Pi & \xrightarrow{k_{32}} & G1P &
\end{array}
\tag{4.13}
$$

Both models consist of 32 complexes, but the model of Pettersson consists of 13 linkage classes while the model of Poolman consists of only 12 linkage classes. Hence, the deficiencies for the reaction networks are 3 and 4, respectively. In both cases, the deficiency of each linkage class is 0 and therefore neither Deficiency Zero Theorem nor Deficiency One Theorem nor Deficiency One Algorithm can be applied. Again, the size of this network already exceeds the computational capability of the CRNT Toolbox.

Furthermore, also the approach using SR-graphs is not applicable for these models. As can be seen in Figure 4.6, the SR-graph contains an even-cycle that is not a 1-cycle, thus violating the conditions of theorem 4.1.20. On the other hand, subnetwork analysis of the model of Pettersson reveals four elementary modes, which differ mainly in whether starch synthesis (k_{27}) or one of the three export reactions ($k_{28} - k_{30}$) is used. All of the subnetworks induced by these elementary modes contain at least 26 complexes, exceeding the capabilities of the currently available implementation of the Deficiency One Algorithm. Introducing the starch degradation step (k_{31}) in the model of Poolman extends the number of elementary modes to eight. However, except for a trivial elementary mode composed of starch synthesis and degradation together with the light reaction (k_{27}, k_{31} and k_{32}) which cannot admit multiple positive steady states, all other elementary modes contain again at least 26 complexes. Altogether, even under simplified mass-action kinetics, the question whether these models support multiple steady states remains open.

4.3.3 Extended model of the Calvin cycle

Zhu et al. (2007) extended the previous models by including the photosynthetic carbon oxygenation pathway and sucrose synthesis. Several new reactions were added which take place in the cytosol, leading to a compartmentation of the reaction network. Hence, every metabolite which may appear in the stroma as well as in the cytosol is modelled as two distinct compounds, such as 3-phosphoglycerate (PGA and PGAc respectively) which results in a total of 31 compounds (Figure 4.7). However, the models of Poolman and Pettersson cannot be perfectly embedded into this model, because here several metabolites are pooled together (e.g. GAP and DHAP). The reaction

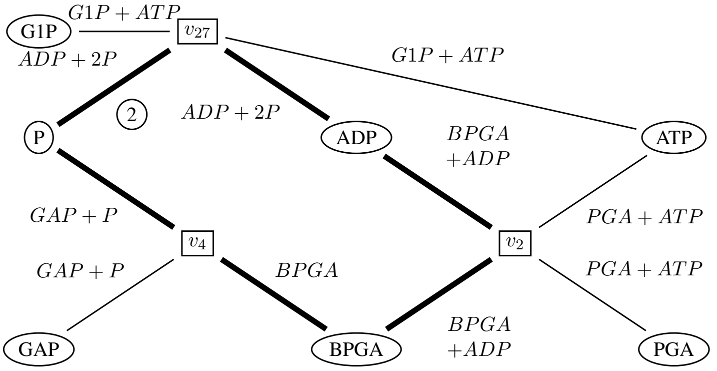


Figure 4.6: Part of the SR-graph for the models of Pettersson and Poolman. The bold lines form an even-cycle (two c-pairs labeled as *ADP + 2P* and *BPGA + ADP*), which is not a 1-cycle. Therefore, Theorem 4.1.20 cannot be applied.

network is as follows:

$$
\begin{array}{rcl rcl rcl}
RuBP & \xrightarrow{k_1} & 2\,PGA & T3P & \xrightarrow{k_{22}} & T3Pc & 2\,T3Pc & \underset{k_{41}}{\overset{k_{40}}{\leftrightarrows}} & FBPc \\
PGA+ATP & \underset{k_3}{\overset{k_2}{\leftrightarrows}} & DPGA & PGA & \xrightarrow{k_{23}} & PGAc & FBPc & \underset{k_{43}}{\overset{k_{42}}{\leftrightarrows}} & HexPc \\
DPGA & \xrightarrow{k_4} & T3P & RuBP & \xrightarrow{k_{24}} & PGA+PGCA & HexPc+UTPc & \underset{k_{45}}{\overset{k_{44}}{\leftrightarrows}} & OPOPc+UDPGc \\
2\,T3P & \underset{k_6}{\overset{k_5}{\leftrightarrows}} & FBP & PGCA & \xrightarrow{k_{25}} & GCA & HexPc+UDPGc & \underset{k_{47}}{\overset{k_{46}}{\leftrightarrows}} & SUCPc \\
FBP & \underset{k_8}{\overset{k_7}{\leftrightarrows}} & HexP & GCEA+ATP & \underset{k_{27}}{\overset{k_{26}}{\leftrightarrows}} & PGA & SUCPc & \underset{k_{49}}{\overset{k_{48}}{\leftrightarrows}} & SUCc \\
T3P+HexP & \underset{k_{10}}{\overset{k_9}{\leftrightarrows}} & E4P+PenP & GCEAc & \underset{k_{29}}{\overset{k_{28}}{\leftrightarrows}} & GCEA & F26BPc & \xrightarrow{k_{50}} & HexPc \\
E4P+T3P & \underset{k_{12}}{\overset{k_{11}}{\leftrightarrows}} & SBP & GCAc & \underset{k_{31}}{\overset{k_{30}}{\leftrightarrows}} & GCA & HexPc+ATPc & \underset{k_{52}}{\overset{k_{51}}{\leftrightarrows}} & F26BPc \\
SBP & \underset{k_{14}}{\overset{k_{13}}{\leftrightarrows}} & S7P & GCAc & \xrightarrow{k_{32}} & GOAc & ATPc & \underset{k_{54}}{\overset{k_{53}}{\leftrightarrows}} & UTPc \\
S7P+T3P & \underset{k_{16}}{\overset{k_{15}}{\leftrightarrows}} & 2\,PenP & GOAc+SERc & \underset{k_{34}}{\overset{k_{33}}{\leftrightarrows}} & GLYc+HPRc & OPOPc & \xrightarrow{k_{55}} & 0 \\
PenP+ATP & \underset{k_{18}}{\overset{k_{17}}{\leftrightarrows}} & RuBP & HPRc & \underset{k_{36}}{\overset{k_{35}}{\leftrightarrows}} & GCEAc & SUCc & \xrightarrow{k_{56}} & 0 \\
0 & \underset{k_{20}}{\overset{k_{19}}{\leftrightarrows}} & ATP & GOAc & \underset{k_{38}}{\overset{k_{37}}{\leftrightarrows}} & GLYc & PGAc & \xrightarrow{k_{57}} & 0 \\
HexP+ATP & \xrightarrow{k_{21}} & 0 & 2\,GLYc & \xrightarrow{k_{39}} & SERc & 0 & \xrightarrow{k_{58}} & ATPc
\end{array}
\tag{4.14}
$$

Network 4.14 is comprised of 49 complexes and 13 linkage classes. The rank of the stoichiometric matrix is 31, which produces a deficiency of 5 for the network. Furthermore, the deficiency of each linkage class is 0. So Deficiency Zero Theorem, Deficiency One Theorem and Deficiency One Algorithm are not applicable. The high number of complexes renders it impossible to use the Advanced Deficiency Algorithm. Also, the analysis of the SR-graph does not resolve the question of bistability as can be seen in Figure 4.8. The subnetwork analysis reveals five elementary modes. Besides one futile cycle, they all consist of the core Calvin cycle plus one of the following pathways: starch synthesis, PGA export, photosynthetic carbon oxygenation or sucrose synthesis. Three of the subnetworks induced by each elementary mode cannot admit multiple positive steady states. For the remaining two, the induced subnetworks are already too large to be handled by the CRNT Toolbox. In summary, no definite answer can be given to the question whether this extended model supports multiple steady states.

4.4 Conclusion

Multistability of a metabolic network is a very important and interesting dynamic property, as it is the cause for a switchlike behaviour. However, it is not trivial to determine regions in the parameter space in which multistability occurs. In that respect, the presented methods constitute

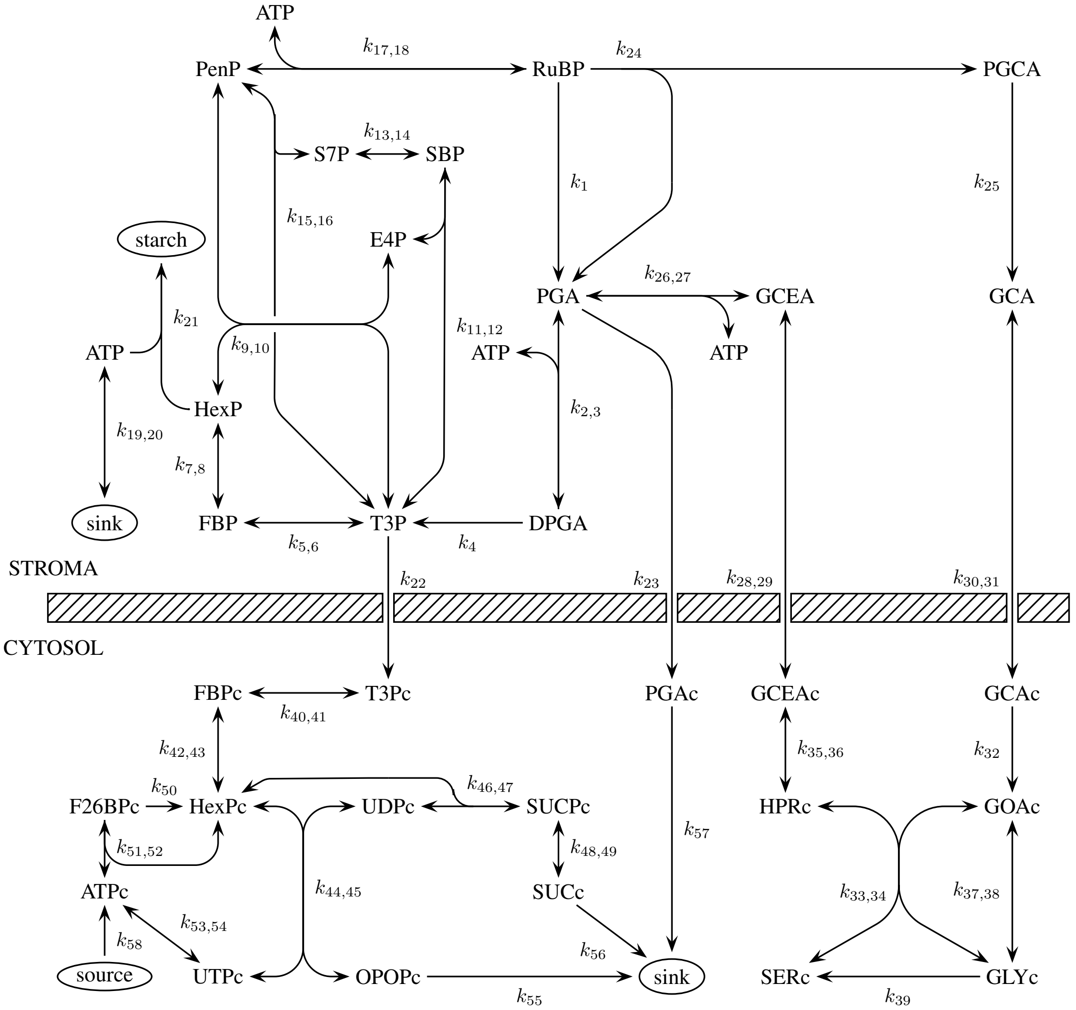


Figure 4.7: Graphical representation of the extended model (Zhu et al., 2007). The nodes denoted as 'starch', 'source' and 'sink' serve as external metabolites. DPGA, 1,3-Bisphosphoglycerate; E4P, Erythrose-4-phosphate; FBP, Fructose-1,6-bisphosphate; F26BP, Fructose-2,6-bisphosphate; GCA, Glycolate; GCEA, Glycerate; GLY, Glycine; GOA, Glyoxylate; HexP, Hexose-phosphate, includes Fructose-6-phosphate, Glucose-6-phosphate and GLucose-1-phosphate; HPR, Hydroxypyruvate; OPOP, Pyrophosphate; PenP, sum of concentrations of Ribose-5-phosphate, Ribulose-5-phosphate and Xylulose-5-phosphate; PGA, 3-phosphoglycerate; PGCA, Phosphoglycolate; RuBP, Ribulose-1,5-bisphosphate; S7P, Seduheptulose-7-phosphate SBP, Seduheptulose-1,7-bisphosphate SER, Serine; SUC, Succrose; SUCP, Sucrose phosphate; T3P, Triose-phosphate, including Dihydroxyacetone-phosphate and Glyceraldehyde-3-phosphate

powerful means to investigate the entire parameter space of metabolic networks, at least under the assumption of mass-action kinetics. Table 4.3 summarises the results obtained from applying these methods to several models of the Calvin cycle, which differ in their level of abstraction as well as in the number of considered reactions. For small networks definite results were found. Interestingly, weakening the mass-action assumption by explicitly modelling enzyme mechanisms leads to multiple positive steady states. However, no results could be found for the larger net-

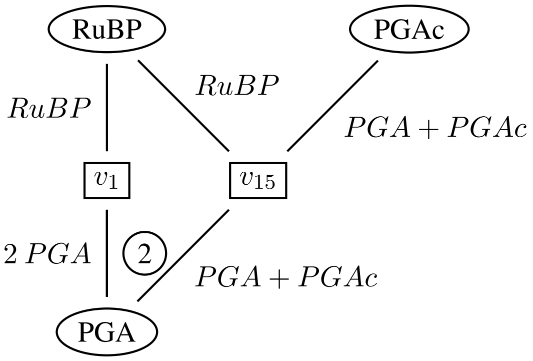


Figure 4.8: Part of the SR-graph for the extended model shown in network 4.14. The depicted cycle forms an even-cycle which is not a 1-cycle. Note that although edges are labeled with $RuBP$ they do not form a c-pair as they are not incident to a common *reaction* node. Therefore, Theorem 4.1.20 cannot be applied.

works, mainly due to the limitations imposed by the CRNT Toolbox. Therefore, it is crucial to improve the existing implementations of the Deficiency One Algorithm and the Advanced Deficiency Algorithm. Especially the latter is of high importance, since it can, in principle, be applied to all sorts of reaction networks but might require solving nonlinear systems. Furthermore, the existing methods should be extended such that also the stability of the calculated steady states is considered, since one is usually only interested in stable steady states.

	model of Zhu with mass-action kinetics (network 4.9)	model of Zhu with Michaelis-Menten kinetics (network 4.11)	model of Pettersson/ Poolman (network 4.13)	extended model (network 4.14)
species	5	19	18	31
complexes	9	21	32	49
reactions	7	21	31/32	58
Deficiency Zero Theorem	not applicable	not applicable	not applicable	not applicable
Deficiency One Theorem	not applicable	not applicable	not applicable	not applicable
Deficiency One Algorithm	no multiple positive steady states possible	not applicable	not applicable	not applicable
Advanced Deficiency Algorithm	–	exceeds computational capabilities	exceeds computational capabilities	exceeds computational capabilities
SR-graphs	–	not applicable (contains even-cycle that is not a one-cycle)	not applicable (contains even-cycle that is not a one-cycle)	not applicable (contains even-cycle that is not a one-cycle)
subnetwork analysis	–	2 elementary modes subnetworks can admit multiple positive steady states and can be extended	4/7 elementary modes all subnetworks exceed computational capabilities	5 elementary modes two subnetworks exceed computational capabilities
multiple positive steady states	NO	YES	still open	still open

Table 4.3: Summary of results regarding multiple positive steady states for several models of the Calvin cycle. Only for the small networks definite results could be obtained. The analysis of the larger networks is hampered by the limitations of the current implementation of deficiency 1 and advanced deficiency algorithm.

Chapter 5

The stability and robustness of metabolic states: Identifying stabilizing sites in metabolic networks

The dynamic behavior of metabolic networks is governed by numerous regulatory mechanisms, such as reversible phosphorylation, binding of allosteric effectors or temporal gene expression, by which the activity of the participating enzymes can be adjusted to the functional requirements of the cell. For most of the cellular enzymes, such regulatory mechanisms are at best qualitatively known, whereas detailed enzyme-kinetic models are lacking. To explore the possible dynamic behavior of metabolic networks in cases of lacking or incomplete enzyme-kinetic information, we present a computational approach based on structural kinetic modelling. We derive statistical measures for the relative impact of enzyme-kinetic parameters on dynamic properties (such as local stability) and apply our approach to the metabolism of human erythrocytes. Our findings show that allosteric enzyme regulation significantly enhances the stability of the network and extends its potential dynamic behavior. Moreover, our approach allows to differentiate quantitatively between metabolic states related to senescence and metabolic collapse of the human erythrocyte. We think that the proposed method represents an important intermediate step on the long way from topological network analysis to detailed kinetic modelling of complex metabolic networks.

5.1 Introduction

One of the most challenging goals of computational systems biology is the development of detailed kinetic models to simulate and predict the dynamic response of metabolic networks towards, for example, changes in kinetic parameters due to pharmacological interventions or variations of environmental conditions. However, for complex metabolic networks comprised of several interwoven pathways, detailed kinetic modelling is usually not possible due to the inevitable lack of knowledge about the kinetic properties of the involved enzymes and membrane transporters. In this work, we extend a recently proposed method that bridges between topology-based approaches and explicit kinetic models of metabolic networks (Steuer et al., 2006). In the face of lacking or incomplete enzyme-kinetic information, we *i)* derive and compare statistical measures for the relative impact of enzymatic reactions and parameters on the dynamic properties (such as local stability) of metabolic networks; *ii)* evaluate the functional role of allosteric feedback regulation in the stabilization of metabolic networks; and *iii)* propose measures to quantitatively evaluate the stability and robustness properties of metabolic states.

Our approach is exemplified and validated using a representation of the metabolic network of the human erythrocyte. Due to the fundamental role of erythrocytes in the oxygen supply of cells, as well as the relative simplicity of its metabolism, erythrocytes have been subject to extensive experimental and theoretical research for decades. Numerous explicit mathematical models have been developed since the late 1970s (Rapoport et al., 1976; Ataullakhanov et al., 1981; Holzhütter et al., 1985; McIntyre et al., 1989; Joshi and Palsson, 1989; Ni and Savageau, 1996a; Mulquiney and Kuchel, 1999; Nakayama et al., 2005), providing a suitable benchmark to assess the reliability of our method.

Our approach is motivated by the increasing experimental accessibility of cellular characteristics, such as metabolic fluxes and concentrations of metabolic intermediates (Fernie et al., 2004; Goodacre et al., 2004; Sauer, 2004). Each metabolic state, characterized by a flux distribution and metabolite concentrations, is associated with a unique spectrum of dynamic properties, as defined by the ensemble of all possible kinetic models consistent with the respective state. Our main focus thus lies on a quantitative characterization and comparison of the stability properties of metabolic states.

In particular, transitions to instability, occurring via a loss of a stable steady state, were previously argued to play a crucial role in senescence and metabolic collapse of erythrocytes, and may act as a primary signal for cell removal in patients with haemolytic anaemia (de Atauri et al., 2006; Schuster and Holzhütter, 1995). While usually an investigation of such transitions necessitates the construction of explicit kinetic models, our approach allows to draw quantitative conclusions about the stability of metabolic states in response to an increased ATP demand, occurring, for example, under conditions of osmotic or mechanic stress (Dariyerli et al., 2004; Kodícek, 1986). It is demonstrated that different metabolic states, each satisfying the flux balance equation and thermodynamic constraints, can nonetheless show drastic differences in the ability to ensure stability and maintain metabolic homeostasis.

As our method requires no detailed information about enzyme-kinetic rate equations and parameters, and due to its computational simplicity, it is applicable to large metabolic networks. In particular, as the construction of explicit kinetic model is usually not feasible, our method significantly extends previous approaches to metabolic robustness, often based on topological or stoichiometric considerations alone (Edwards and Palsson, 2000b; Tekir et al., 2006; Deutscher et al., 2006; Stelling et al., 2002). We argue that dynamic aspects of metabolic networks are becoming more and more important in view of modern techniques like siRNA knockdowns or genetic modifications (Bailey, 1991; Becker et al., 2005) to modify the activity of individual enzymes *in vivo*. It has to be expected that such perturbations may give rise to fundamental changes in the dynamic behavior of the underlying network.

5.2 Results

5.2.1 The parametrization of metabolic states

A metabolic network is a set of coupled chemical reactions and transports processes. Neglecting spatial variations of the metabolite concentrations within the reactions compartments the time-dependent changes of the metabolite concentrations can be described by a set of differential equations of the form $\dot{\boldsymbol{S}} = \boldsymbol{N}\boldsymbol{\nu}(\boldsymbol{S})$, where $\boldsymbol{S}$ denotes the m-dimensional vector of metabolite concentrations, $\boldsymbol{N}$ the $m \times r$-dimensional stoichiometric matrix and $\boldsymbol{\nu}(\boldsymbol{S})$, a r-dimensional vector of enzyme kinetic reaction rates. In the case of lacking or incomplete enzyme-kinetic data and assuming the existence of a stationary state $\boldsymbol{S^0}$, the differential equation can be interpreted as linear equation for the stationary reaction rates $\boldsymbol{\nu^0} = \boldsymbol{\nu}(\boldsymbol{S^0})$. The mass balance equation $\boldsymbol{N}\boldsymbol{\nu^0} = \boldsymbol{0}$ provides the conceptual foundation for current stoichiometry-based approaches to metabolic network analysis and brought forth a number of highly successful applications to determine the structure

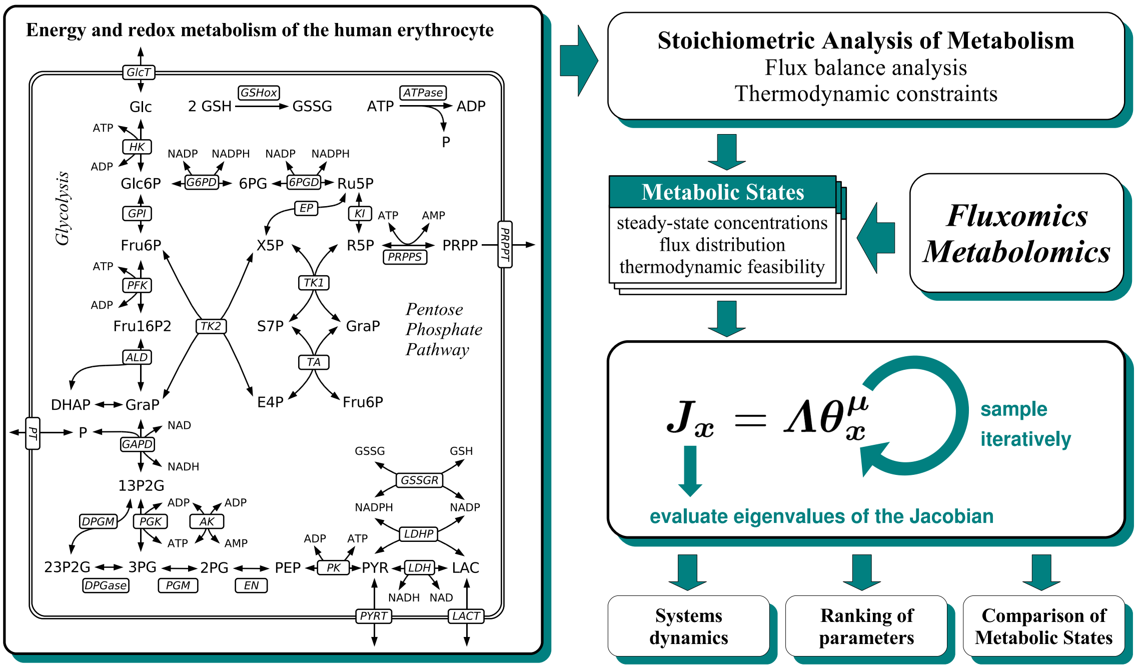


Figure 5.1: Energy and redox metabolism of the human erythrocyte and the proposed workflow: The stoichiometry and the steady state concentrations and fluxes are extracted from existing models and available experimental data. The Jacobian matrix is established and the intervals for the saturation parameters are specified according to available biological information and/or additional constraints of interest. The saturation parameters are sampled repeatedly and the eigenvalues of the Jacobian are evaluated. Abbreviations: Glc, glucose; Glc6P, glucose 6-phosphate; Fru6P, fructose 6-phosphate; Fru16P2, fructose 1,6-bisphosphate; GraP, glyceraldehyde 3-phosphate; DHAP, dihydroxyacetone phosphate; 13P2G, 1,3-bisphosphoglycerate; 23P2G, 2,3-bisphosphoglycerate; 3PG, 3-phosphoglycerate; 2PG, 2-phosphoglycerate; PEP, phosphoenolpyruvate; PYR, pyruvate; LAC, lactate; 6PG, 6-phosphoglycanate; Ru5P, ribulose 5-phosphate; Xul5P, xylulose 5-phosphate; Rib5P, ribose 5-phosphate; S7P, sedoheptulose 7-phosphate; E4P, erythrose 4-phosphate; PRPP, phosphoribosyl pyrophosphate; GSH, reduced glutathione; GSSG, oxidized glutathione; GlcT, glucose transport; HK, hexokinase; GPI, glucose-6-phosphate isomerase; PFK, phosphofructokinase; ALD, aldolase; TPI, triosephosphate isomerase; GAPD, glyceraldehyde phosphate dehydrogenase; PGK, phosphoglycerate kinase; DPGM, 2,3-bisphosphoglycerate mutase; DPGase, 2,3-bisphosphoglycerate phosphatase; PGM, 3-phosphoglycerate mutase; EN, enolase; PK, pyruvate kinase; LDH(P), lactate dehydrogenase; Lact, lactate transport; AK adenylate kinase; G6PD, glucose-6-phosphate dehydrogenase; 6PGD, 6-phosphogluconate dehydrogenase; GSSGR, glutathione reductase; EP, ribose phosphate epimerase; KI, ribose phosphate isomerase; TK, transketolase; TA, transaldolase; PRPPS, phosphoribosypyrophosphate synthetase; PRPPT, phosphoribosypyrophosphate transport

and function of metabolic networks (Schuster et al., 1999; Stelling et al., 2002; Varma and Palsson, 1994). Recently, the flux-balance equation was supplemented with thermodynamic constraints, providing a link between feasible flux distributions and metabolite concentrations (Kümmel et al., 2006; Henry et al., 2006; Holzhütter, 2004; Hoppe et al., 2007).

However, the mass balance equation itself, along with its associated thermodynamic constraints, does not allow to draw any conclusions about the possible dynamics or potential instabilities of a metabolic state. To obtain information about essential aspects of the dynamics, we thus augment the mass balance equation with the first-order expansion of the differential equation. Given a metabolic system at a (possibly unknown and not necessarily unique) metabolic state, characterized by $\boldsymbol{\nu}^{\mathbf{0}}$ and $\boldsymbol{S}^{\mathbf{0}}$, the system of differential equations can be approximated by a Taylor series expansion.

$$\frac{d\boldsymbol{S}}{dt} = \underbrace{\boldsymbol{N}\boldsymbol{\nu}(\boldsymbol{S}^{\mathbf{0}})}_{=\mathbf{0}} + \underbrace{\boldsymbol{N}\left.\frac{\partial\boldsymbol{\nu}}{\partial\boldsymbol{S}}\right|_{\boldsymbol{S}^{\mathbf{0}}}}_{=:\boldsymbol{J}} \left(\boldsymbol{S} - \boldsymbol{S}^{\mathbf{0}}\right) + \ldots \tag{5.1}$$

The first term describes the steady state properties of the system, as exploited by flux-balance analysis to constrain the stoichiometrically feasible flux distributions. Along similar lines, taking the next term of the expansion into account, the structure of the *Jacobian matrix* $\boldsymbol{J}$ determines and constraints the possible dynamics of the system at each metabolic state.

Our method builds upon a statistical evaluation of the Jacobian matrix. Based on the formalism of structural kinetic modelling (Steuer et al., 2006), we construct a parametric representation of the Jacobian matrix, such that each element covers the comprehensive parameter space at a specific metabolic state. In particular, the Jacobian matrix can be written as product of two matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\theta}_{\boldsymbol{x}}^{\boldsymbol{\mu}}$. The elements of the matrix $\boldsymbol{\Lambda}$ are fully specified by the metabolic state of the system. In addition, the (usually unknown) elements of the matrix $\boldsymbol{\theta}_{\boldsymbol{x}}^{\boldsymbol{\mu}}$, specify the relative saturation of each enzyme with respect to its ligands and can be assigned to well defined intervals even when the explicit functional form of the rate equations is not known. In the following, these matrix elements are denoted *saturation parameters* θ_m^r, where r stands for the reaction saturated by metabolite m. A brief mathematical synopsis is given in *Materials and Methods*.

Evaluating the Jacobian matrix with respect to the (unknown) elements of the matrix $\boldsymbol{\theta}_{\boldsymbol{x}}^{\boldsymbol{\mu}}$ then defines the spectrum or scope of dynamic behavior at the respective metabolic state. The proposed workflow is summarized in Figure 5.1: First, the stoichiometry and a metabolic state $\boldsymbol{\nu}^{\mathbf{0}}$ and $\boldsymbol{S}^{\mathbf{0}}$ are specified, based on available experimental data and existing mathematical models. Second, an ensemble of models (Jacobians) is generated by assigning random values to the elements of $\boldsymbol{\theta}_{\boldsymbol{x}}^{\boldsymbol{\mu}}$, obeying the defined intervals. Evaluating the eigenvalues of the Jacobian matrix repetitively, allows to investigate and compare the scope of dynamic behavior under different preconditions, e.g., such as suppressed or absent allosteric regulation. In this respect, especially the largest real part of the eigenvalues, denoted by $\lambda_{Re}^{\max}$, is of interest, as it relates to the slowest timescale of the system and, if positive, implies (local) instability of the metabolic state. The metabolic state is stable only if all eigenvalues have a negative real part (see also *Materials and Methods*).

5.2.2 The role of regulation

Allosteric regulation is one of the main mechanisms to control enzyme activity. Since allosteric regulation occurs within metabolic networks as feedback or feedforward loops, it operates network wide and affects the dynamic properties at a systems level. The presented approach is used to analyze the effects of allosteric regulation on stability in a systematic way. Two sets of models under different preconditions are created, both corresponding to the normal *in vivo* conditions of the erythrocyte (see Figure 5.1 for a schematic representation of the energy and redox metabolism of the human erythrocyte and the used abbreviations). Within the first set of models $\mathbf{C_{noreg}}$ all saturation parameters associated with allosteric effectors are fixed to zero, corresponding to absence of regulatory interactions. In addition, a second set $\mathbf{C_{reg}}$ is constructed by assigning all saturation parameters, including those for allosteric regulation, to their respective intervals (see *Materials and Methods* and the *Supplementary information* (Grimbs et al., 2007b) for details).

Each model (Jacobian) is evaluated according to its spectrum of eigenvalues, with $\lambda_{Re}^{\max} > 0$ implying instability of the metabolic state. In the case of absent allosteric regulation, corresponding to the set $\mathbf{C_{noreg}}$, the proportion of dynamically stable models is approximately 81%. Thus,

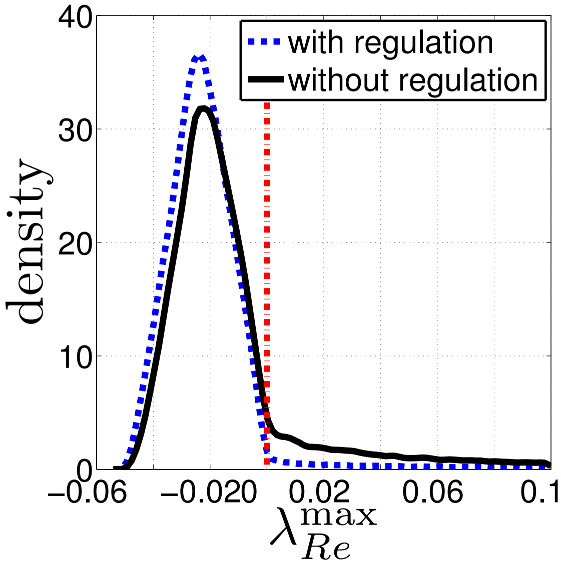


Figure 5.2: The probability density function of the largest real part $\lambda_{Re}^{\max}$ within the spectrum of eigenvalues for both sets $\mathbf{C_{noreg}}$ and $\mathbf{C_{reg}}$. A value $\lambda_{Re}^{\max} > 0$ implies instability. Allosteric regulation significantly suppresses the number of unstable models. Interestingly, the probability density function for $\mathbf{C_{reg}}$ does not show a shift towards more negative $\lambda_{Re}^{\max}$ but rather an increase at the peak.

although the vast majority of models are stable, the proportion of unstable models cannot be neglected. As dynamic stability is mandatory for the existence of the metabolic state, it indicates a substantial risk for the unregulated network to be driven out of the observed steady state when changes of the binding constants for the substrates occur for genetic or pharmacological reasons.

Within the whole set, no model (Jacobian) is found exhibiting more than one eigenvalue larger than zero, suggesting that the occurrence of a Hopf bifurcation is at least rare under the precondition of suppressed allosteric regulation. Since a Hopf bifurcation indicates the transition to sustained oscillation, such dynamical behavior seems unlikely under these conditions.

Looking at the set $\mathbf{C_{reg}}$, thus including allosteric regulation, the proportion of stable models shifts to 91%, which is significantly higher than in case of suppressed regulation (see *Materials and Methods*). Similar observations were made by Ni and Savageau (1996b), where additional regulation was introduced to a model to stabilize the steady state. Note that regulation is here only defined qualitatively, i.e, the actual strength of each regulatory interaction is chosen randomly and varies between the individual samples. Nonetheless, even without specific fine-tuning of parameters, allosteric regulation results in a higher proportion of stable networks. This is presumably evolutionary advantageous, since a larger parameter subspace corresponding to stable models increases the flexibility to optimize parameter towards additional requirements other than stability.

Within the set of unstable models, 592 out of 10^6 samples in $\mathbf{C_{reg}}$ have two eigenvalues greater than zero, in each case exhibiting complex conjugate imaginary parts. A smaller fraction (41 samples) show three eigenvalues larger than zero, pointing to bifurcations of higher co-dimension. Though restricted to a very small region in parameter space, allosteric regulation thus expands the scope of dynamical behavior by increasing the region in parameter space where oscillatory or more complex dynamics can be expected. For comparison, the estimated probability density functions of the largest real parts $\lambda_{Re}^{\max}$ within the spectrum of eigenvalues are shown in Figure 5.2.

5.2.3 The ranking of parameters

Stability of a metabolic steady state is an emergent systemic property that is brought about by the kinetic properties of all enzymes. Nevertheless, changes in the kinetic parameters of individual enzymes may have quite differential impact on the stability of a given steady state. To evaluate and compare the degree of influence of individual parameters and reactions on the stability and response to perturbations, we rank the parameters according to several objective measures. Three

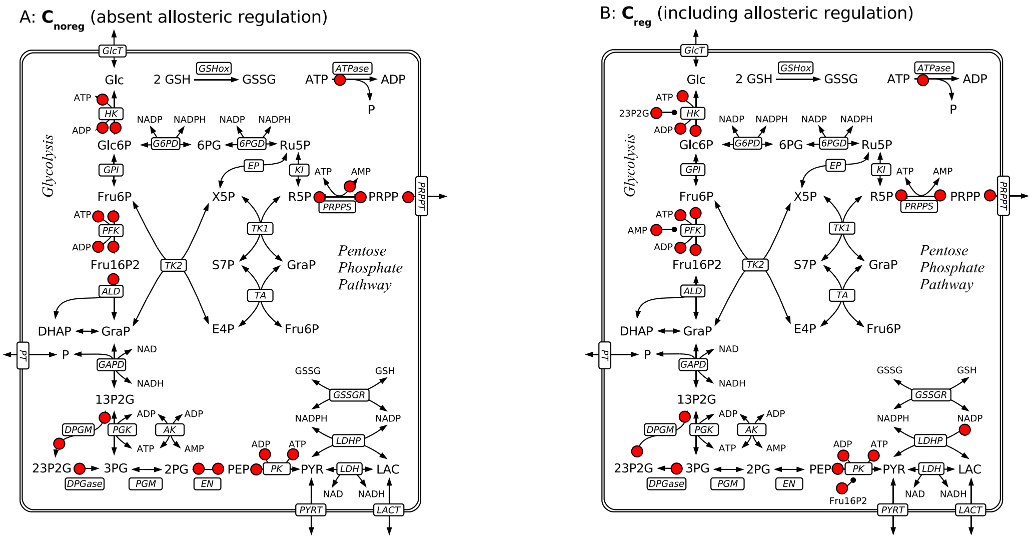


Figure 5.3: All significant saturation parameters for both sets of models $\mathbf{C_{noreg}}$ (absent allosteric regulation, left plot) and $\mathbf{C_{reg}}$ (included allosteric regulation, right plot). For brevity, only the ranking according to the correlation coefficient is considered.

distinct measures are used and compared to each other, namely the (Pearson) correlation coefficient, the mutual information and the Kolmogorov-Smirnov-test (KS-test), see *Materials and Methods* for explicit definitions.

All three measures were evaluated for all saturation parameters for both sets of models $\mathbf{C_{noreg}}$ and $\mathbf{C_{reg}}$. Although the detailed ranking of the parameters is not identical, it is still consistent with respect to all different measures. Figure 5.3 depicts the significant parameters for both sets of models $\mathbf{C_{noreg}}$ and $\mathbf{C_{reg}}$. Figure 5.4 exemplifies the influence of the most highly ranked parameters on the stability of the metabolic state. Shown is the percentage of stable models within the parameter space as a function of selected saturation parameters. For a more detailed discussion and comparison of the different rankings see also Grimbs et al. (2007b).

The ranking of parameters allows for several significant conclusions about the role of regulation within the metabolic network. First, almost all high ranked parameters are associated with reactions involved in ATP production or consumption. Especially the PFK, HK and PK play an important role in stabilizing the network. Interestingly, evaluating the set $\mathbf{C_{noreg}}$ reveals that, although no additional information about putative sites for allosteric regulation is included, mainly those parameters are ranked very high that affect reactions that are known to be allosterically regulated (see *Materials and Methods* for a statistical verification of this assertion). We point out that for the construction of $\mathbf{C_{noreg}}$ only the stoichiometry and the metabolic state under normal conditions were used. This emphasizes the usefulness of our approach to analyze metabolic networks that are not as well studied as the one of erythrocytes and where detailed information about allosteric regulation is not available.

Several more observations can be made from the ranking of the parameters. First, the high ranked parameters almost all belong to enzymes of the glycolytic pathway. Intriguingly, kinetic alterations of the allosterically regulated enzyme G6PD which is known to control the flux through the pentose phosphate pathway does not show significant impact on stability. This corresponds with results from de Atauri et al. (2006), obtained from an explicit kinetic model. The authors show that the metabolic network breaks down for low concentrations of the glycolytic enzymes, whereas such a transition to instability does not occur if the enzymes of the pentose phosphate pathway are

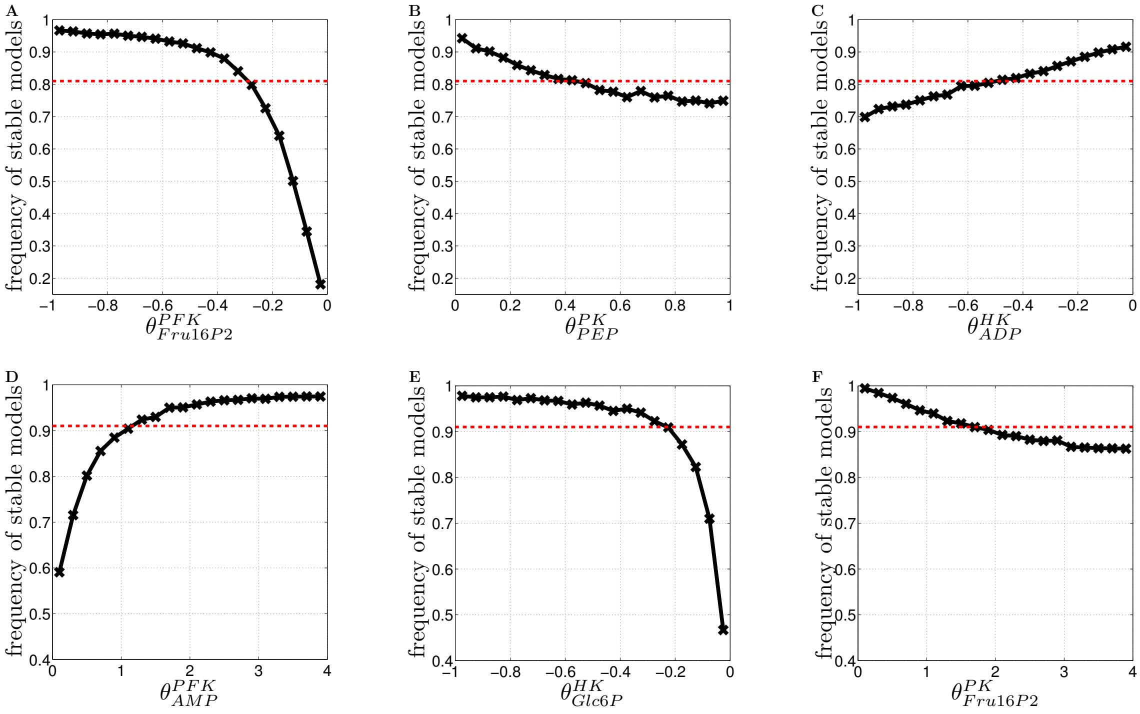


Figure 5.4: Relationships between selected saturation parameters and the probability of stable models. Shown is the case of suppressed allosteric regulation (**A - C**) and for included allosteric regulation (**D - F**). In each case, a single saturation parameter is fixed while all other parameters are chosen randomly. The dotted red line marks the overall averages, 81% in the case of $\mathbf{C_{noreg}}$ and 91% in the case of $\mathbf{C_{reg}}$. The plots **D** and **F** show saturation parameters associated with allosteric regulation.

at very low concentrations. Second, all parameters associated with 23P2G as an allosteric effector are relatively low ranked. This indicates that the main function of these regulatory mechanisms is not to maintain or achieve stability.

5.2.4 Comparison with the explicit model

The availability of a comprehensive and well-established mathematical model of the erythrocyte metabolism (Schuster and Holzhütter, 1995) allows to validate our method by comparing the ranking of saturation parameters with results of metabolic control analysis (MCA, see Heinrich and Schuster (1996)). As a metabolic instability of the erythrocyte may occur if the energy demand exceeds the glycolytic ATP production we study the impact that changes in the kinetic parameters of the various enzymes of the network have on ATP utilization. The relative change of the rate of ATP utilization (υ_{ATPase}) elicited by a (small) change of the Michaelis constant characterizing affinity of metabolite M to enzyme E is given by the flux control coefficient

$$C^{\mathrm{ATPase}}_{\mathrm{K^E_M}} = \frac{\partial \ln \nu_{ATPase}}{\partial \ln K^E_M} = \frac{K^E_M}{\nu_{ATPase}} \frac{\partial \nu_{ATPase}}{\partial K^E_M} \tag{5.2}$$

Note that negative values of the flux coefficient indicate that decreasing value of the Michaelis constant (corresponding to increasing saturation) increases the rate of ATP production and thus stabilizes the steady state. Calculating the flux control coefficient for all 85 Michaelis constants occurring in the rate laws of the mathematical model and ranking them in ascending order reveals

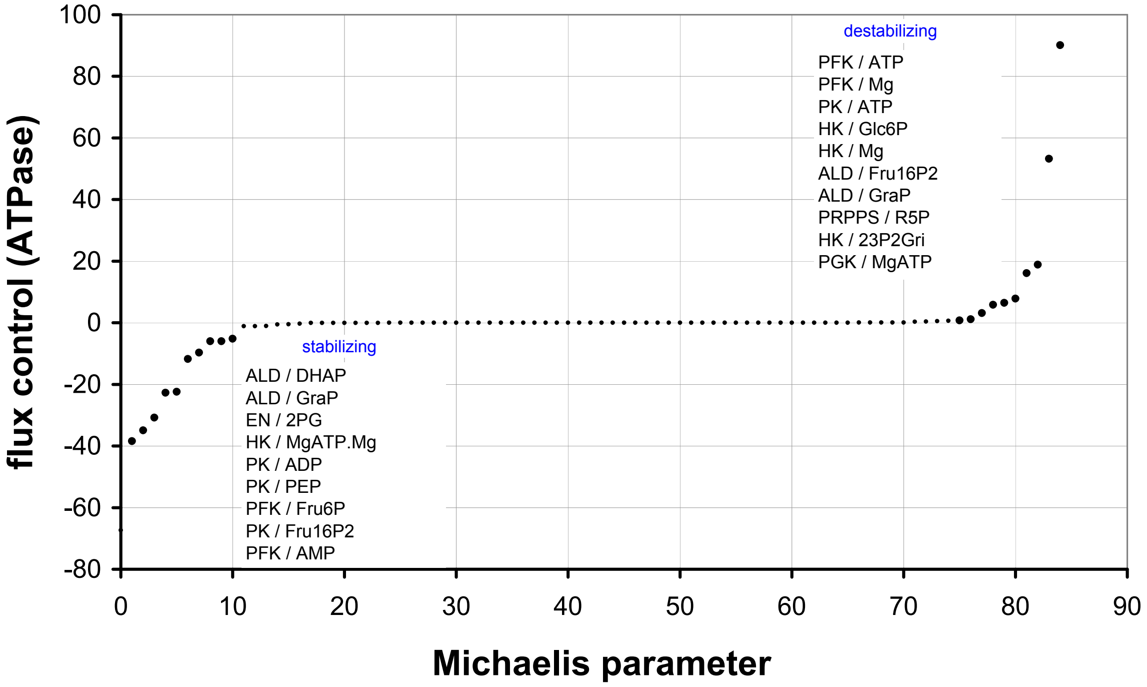


Figure 5.5: Values of the flux control coefficient defined by equation (5.2) plotted in ascending order for the 85 Michaelis constants contained in the rate laws of kinetic model. Bold points indicate the flux control coefficients for the explicitly depicted 20 Michaelis constants that upon decrease (i.e. increase of saturation) exert the most stabilizing and destabilizing influence on ATP supply, respectively.

that only 10 affinity parameters each contribute significantly to the energetic stabilization and destabilization of the network (see Figure 5.5). Changes of the Michaelis constants for binding of AMP and ATP to the phosphofructokinase (PFK) have by far the highest impact on the ATP production. This underlines the well-known central regulatory importance of this enzyme for red cell glycolysis. Remarkably, the set of 20 regulatory most relevant Michaelis constants determined by metabolic control analysis of the basis of the full mathematical model comprises all saturation parameters identified by our random sampling method.

5.2.5 Robustness of metabolic states

As yet our analysis has focused on the analysis of a single metabolic state corresponding to the normal *in vivo* conditions of the erythrocyte. However, the energy metabolism of this cell has to cope with large fluctuations of the ATP demand as the activity of the Na/K-ATPase, accounting for about 70% of the total ATP utilization, is greatly enhanced under conditions of osmotic stress (Dariyerli et al., 2004) or mechanic stress exerted during passage of the cell through thin capillaries (Kodícek, 1986). Moreover, because of lacking de novo protein synthesis the erythrocyte is extremely susceptible to enzyme deficiencies which typically result in an impairment of glycolytic ATP production and subsequent break down (hemolysis) of the cell (Jacobasch and Rapoport, 1996).

To demonstrate the discriminatory power of our approach to detect changes in the stability properties of metabolic states, we thus consider a second metabolic state of the erythrocyte characterized by an increased energy demand. To this end we use the kinetic model to calculate fluxes and metabolite concentrations at a 6-fold higher energetic load as compared to the normal reference state. Switching from the normal in vivo state (kATPase = 1.6 mM/h) to the new steady state

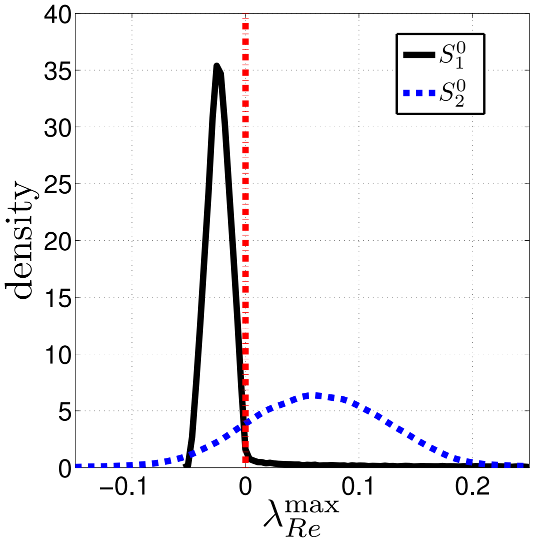


Figure 5.6: The probability density functions of $\lambda_{Re}^{\max}$ for two metabolic states. Under increased energy load ($\mathbf{S_2^0}$) a significantly higher percentage of unstable models is observed, as compared to normal conditions ($\mathbf{S_1^0}$). In both cases allosteric regulation is included.

at increased energetic load (kATPase = 10 mM/h) , the glycolytic flux increases from 1.5 mM/h to 2.33 mM/h (=155%) whereas the ATP concentration decreases from 1.6 mM to 0.56 mM (= 35%). These relative changes are in excellent agreement with experimental data (160% increase of glycolytic flux at 35% decrease of ATP) obtained by successively decoupling glycolysis from ATP consumption by means of arsenate titrations (Ataullakhanov et al., 1981).

The parametrization of the second state $\boldsymbol{S_2^0}$ is performed as described before, with all saturation parameters, including allosteric regulation, sampled randomly from their respective intervals. Figure 5.6 depicts the resulting distribution of the largest real eigenvalue $\lambda_{Re}^{\max}$ within the spectra of eigenvalues, as compared to the distribution under normal conditions (state $\boldsymbol{S_1^0}$). While for normal *in vivo* conditions, the proportion of stable models (Jacobians) was approximately 91%, this value drops drastically to only about 13.2% for the second state $\boldsymbol{S_2^0}$. This is in accordance with our earlier observation that energy related reactions are most crucial with respect to stability, as well as with the fact that in the detailed kinetic model of Schuster and Holzhütter (1995) the system is able to compensate an increased energy load only up to an upper critical value, but breaks down if the energy demand is increased any further. We emphasize that both states cannot be discriminated based on stoichiometric considerations alone: Both satisfy the flux-balance equation and are consistent with thermodynamic constraints.

The proportion of unstable models alone, evaluated over the comprehensive parameter space of a metabolic state, does not necessarily imply actual instability of the respective flux distribution. However, the proportion of unstable models has significant consequences for the ability of the system to maintain the considered metabolic state at perturbations of enzyme-kinetic parameters (Morohashi et al., 2002). To evaluate the robustness properties of both states quantitatively, we consider two distinct scenarios: First, for both metabolic states random instances of stable models are repeatedly selected from the parameter space and the set of parameters is subsequently perturbed within a given radius. The percentage of perturbations that remain stable, as a function of the magnitude or radius of the perturbations, then serves as a quantitative measure of robustness. See Figure 5.7A for a schematic representation. Second, to make the results for both metabolic states more comparable, the parameter space from which random instances of models are selected is restricted to a small interval with $\lambda_{Re}^{\max} \in [0, -0.01]$ for both states. Again each parameter set is perturbed with increasing radius and the frequency with which a given magnitude of perturbations leads to instability is recorded. The results are shown in Figure 5.7.

Clearly, for both metabolic states the probability that a perturbation results in a loss of stability increases with increasing magnitude of perturbations. However, starting (by construction) with

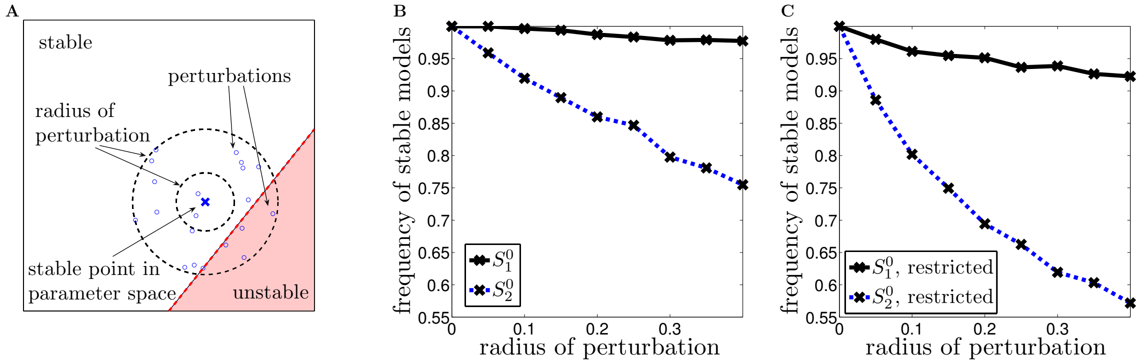


Figure 5.7: The robustness of the two metabolic states. A stable point in parameter space is chosen randomly. Within a given radius the point is perturbed several times. Each such perturbation is checked for stability (**A**). 1000 stable points were chosen randomly with 10 perturbations each. A radius of r allows the perturbation to vary within $\theta \pm r \cdot \theta$, where θ is the initially samples value (**B**). In the second scenario, denoted as 'restricted', only stable states close to the instability were considered $\lambda_{Re}^{\max} \in [-0.01, 0)$ (**C**).

initially 100% of stable models, the fraction of models that become unstable increases significantly faster for the state $\boldsymbol{S_2^0}$. This effect is even more pronounced in the second scenario. Here, the stable models for both metabolic states are initially restricted to similar real parts within the spectra of the eigenvalues, and thus to similar distances to the bifurcation. Nonetheless, a perturbation of the (initially stable) metabolic state $\boldsymbol{S_2^0}$ is much more likely to result in a transition to instability than corresponding perturbations of the normal state $\boldsymbol{S_1^0}$. In this sense, the *in vivo* metabolic state $\boldsymbol{S_1^0}$, and concomitantly also its observed flux distribution, is more robust than the second state $\boldsymbol{S_2^0}$. We emphasize again that our quantification of robustness does not involve any knowledge about the explicit functional form of the rate equations or kinetic parameters. Nonetheless, different metabolic states can be clearly differentiated, based only on information about metabolite concentrations and associated flux patterns. In this respect, our approach gives valuable insights on the qualitative and quantitative dynamic behavior of metabolic states that cannot be obtained by considering the stoichiometric balance equation alone and is also applicable to situations where detailed knowledge about the explicit rate equations is not yet available.

5.3 Materials and Methods

5.3.1 Models of the human erythrocyte

To exemplify and validate our approach, we mainly draw upon a previously published model of Schuster and Holzhütter (1995), consisting of 30 metabolites and 31 reactions (see Figure 5.1 for a schematic representation). The model was slightly modified to account for free inorganic phosphate and additional transport reactions for the educt glucose, the intermediate phosphate and the end products phosphoribosyl pyrophosphate, pyruvate and lactate. Mg-complexes were omitted. All reactions, except ATPase, GSHox and PRPPT, were treated as reversible. As ATPase and GSHox are merged overall reactions, describing energy consumption and oxidative load, product inhibition for these reactions was not included, i.e., ADP and GSSG have no influence on ATPase and GSHox, respectively.

The kinetic model was used to calculate the steady metabolic state of the human erythrocytes under normal *in vivo* conditions (state $\boldsymbol{S_1^0}$). Metabolite concentrations and flux values are given in Grimbs et al. (2007b). In addition to the normal *in vivo* state $\boldsymbol{S_1^0}$, a second steady state $\boldsymbol{S_2^0}$ was calculated, corresponding to an increased energy demand of the cell. Analogous simulations were performed previously to explore senescence and metabolic collapse of erythrocytes (Schuster and Holzhütter, 1995; de Atauri et al., 2006). See also (Tekir et al., 2006) for an analysis of red-blood cell enzymopathies based on stoichiometric analysis.

5.3.2 Structural kinetic modelling

Our analysis is based on a decomposition of the Jacobian matrix of a metabolic system at a state $\boldsymbol{S^0}$ into a product of two matrices. Given a metabolic system consisting of m metabolites and r reactions, the set of differential equations $\dot{\boldsymbol{S}} = \boldsymbol{N\nu}(\boldsymbol{S})$ which describe the time-dependent behavior of all metabolite concentrations $S_i(t)$ can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{S_i(t)}{S_i^0} = \sum_{j=1}^{r} \underbrace{\frac{\nu_j^0}{S_i^0} N_{ij}}_{:=\Lambda_{ij}} \underbrace{\frac{\nu_j(\boldsymbol{S})}{\nu_j^0}}_{:=\mu_j(\boldsymbol{S})} \tag{5.3}$$

where S_i^0 and $\nu_j^0 = \nu_j(\boldsymbol{S^0})$ denote the metabolic state at which the system is to be evaluated. Using the definitions given in (5.3) and the variable transformation $x_i(t) = S_i(t)/S_i^0$, the Jacobian with respect to the normalized variables $\boldsymbol{x}$ is

$$\boldsymbol{J_x} = \boldsymbol{\Lambda\theta_x^\mu} \qquad \text{with} \qquad \boldsymbol{\theta_x^\mu} := \left.\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x^0}=\boldsymbol{1}} \tag{5.4}$$

The scaled Jacobian $\boldsymbol{J_x}$ is related to the original Jacobian by a simple similarity transformation and it is fully specified by the parameter matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\theta_x^\mu}$. The elements of $\boldsymbol{\Lambda}$ describe the timescales of the system, as specified by the metabolic state $\boldsymbol{S^0}$ and $\boldsymbol{\nu}(\boldsymbol{S^0})$. The (usually unknown) elements of $\boldsymbol{\theta_x^\mu}$ are defined as the normalized derivatives of the reaction rates and, analogous to the scaled elasticity coefficients of Metabolic Control Analysis, denote the *effective kinetic order* or *normalized saturation* of each reaction with respect to its substrates. Each element θ_S^ν is constrained to the interval of $[0, 1]$ if the metabolite S is a substrate and $[0, -1]$ if S is a product of the reaction $\nu(S)$. Additional nonzero terms arise from allosteric regulation. Allosteric regulation is included by assigning the corresponding parameter to intervals, such that $\theta_S^\nu \in [0, -n]$ for inhibition and $\theta_S^\nu \in [0, n]$ for activation of a reaction ν by S, respectively. A detailed derivation is given elsewhere (Steuer et al., 2006).

Statistical sampling of the parameter space

The stability and possible dynamics of the metabolic network are evaluated at a given metabolic state, characterized by metabolite concentrations $\boldsymbol{S^0}$ and fluxes $\boldsymbol{\nu}(\boldsymbol{S^0})$. The vector of reaction rates satisfies the steady state condition $\boldsymbol{N\nu}(\boldsymbol{S^0}) = \boldsymbol{0}$ and is described by $r - \mathrm{rank}(\boldsymbol{N})$ free parameters. The vector of metabolite concentrations $\boldsymbol{S^0}$ is restricted by thermodynamic constraints (Kümmel et al., 2006; Henry et al., 2006) and approximated by values adapted from Schuster and Holzhütter (1995). The metabolic state fully specifies the matrix $\boldsymbol{\Lambda}$.

To evaluate the dynamic capabilities at a given metabolic state, the nonzero elements of the matrix $\boldsymbol{\theta_x^\mu}$ are sampled from their predefined intervals (Steuer et al., 2006; Wang et al., 2004), while the elements of the matrix $\boldsymbol{\Lambda}$ are restricted to the respective metabolic state. The schematic workflow is shown in Figure 5.1.

Specifically, each reversible enzyme-kinetic reaction $\nu_i(\boldsymbol{S})$ is split into a forward $\nu_i^+(\boldsymbol{S})$ and backward rate $\nu_i^-(\boldsymbol{S})$ and described by the overall steady state flux ν_i^0, the flux ratio $\gamma = \nu^-(\boldsymbol{S^0})/\nu^+(\boldsymbol{S^0})$, the steady state $\boldsymbol{S^0}$, as well as by a set of saturation parameters $\theta_{S_i}^\nu$. Though the method does not presuppose a specific functional form of the rate equations, we illustrate the parametrization using a generic form of enzyme-kinetic rate equations, such as

$$A + B \leftrightarrow P + Q \qquad \nu = \frac{v_m\,(AB - PQ/K_{\mathrm{eq}})}{f(A,B,P,Q)} \,, \tag{5.5}$$

where $f(A,B,P,Q)$ denotes a first order polynomial. The reaction is characterized by the net flux ν^0, the steady state concentrations A^0, B^0, C^0, D^0, as well as the flux ratio $\gamma = \nu^-/\nu^+$, relating to the (often accessible) equilibrium constant K_{eq}. The four unknown saturation coefficients apply to forward and backward rate separately, obeying the relationships $\theta_A^{\nu^+} \in [0,1]$ and $\theta_A^{\nu^-} = \theta_A^{\nu^+} - 1 \in [0,-1]$ for substrates and $\theta_P^{\nu^+} \in [0,-1]$ and $\theta_P^{\nu^-} = \theta_P^{\nu^+} + 1 \in [0,1]$ for products, respectively.

The model of the erythrocyte is parametrized by 87 saturation parameters for substrate and product dependencies of each reaction, as well as 10 additional parameters corresponding to allosteric regulation. See Grimbs et al. (2007b) for a detailed listing. The parameters are denoted with reactions (superscript) and substrate (subscript) respectively, i.e., $\theta^{\mathrm{PFK}}_{\mathrm{Fru6P}}$ denotes the dependence of the phosphofructokinase (PFK) on fructose 6-phosphate (Fru6P).

Stability and dynamics of metabolic states

Our method is based upon a statistical evaluation of the Jacobian matrix. In particular, a metabolic state that satisfies the steady state condition $\boldsymbol{N\nu}(\boldsymbol{S^0}) = \mathbf{0}$ must not necessarily be stable. Rather, its dynamic stability is determined by the eigenvalues of the Jacobian at the respective state. Each eigenvalue describes the behavior of the system after an (infinitesimal) perturbation of the concentrations (Heinrich and Schuster, 1996). The possible dynamics in the vicinity of a metabolic state are schematized in Figure 5.8: The metabolic state can either be *i)* a stable (attracting) steady state, characterized by a largest real part of the eigenvalues $\lambda^{\max}_{\mathrm{Re}} < 0$, *ii)* an unstable (repelling) state, characterized by a positive largest real part $\lambda^{\max}_{\mathrm{Re}} > 0$ within the spectrum of eigenvalues, or, *iii)* a stable (attracting) focus, characterized by nonzero (complex conjugate) imaginary parts $\lambda^{\max}_{\mathrm{Re}} \pm \lambda^{\max}_{\mathrm{Im}}$ with $\lambda^{\max}_{\mathrm{Re}} < 0$ and $\lambda^{\max}_{\mathrm{Im}} \neq 0$, or *iv)* an unstable focus, characterized by a positive real part $\lambda^{\max}_{\mathrm{Re}} > 0$ and complex conjugate eigenvalues $\lambda^{\max}_{\mathrm{Im}} \neq 0$.

Of particular interest are also transitions between the scenarios (bifurcations), most importantly the Hopf bifurcation, where a pair of complex conjugate eigenvalues cross the imaginary axis (stable → unstable focus) and bifurcations of the saddle-node type, where the largest real part within the eigenvalues crosses the imaginary axis (stable node → unstable saddle). Further types of bifurcations are discussed in Steuer et al. (2006). We emphasize that stability does not imply constancy of a metabolic state. Rather, local dynamic stability is mandatory for the existence of the state, but all actual states will fluctuate around their average values (Steuer et al., 2003). All reported results are robust against small deviations of metabolite concentrations and flux values, i.e., an analysis with small alterations of the metabolic state under normal conditions yields identical results.

A simple example

To illustrate our approach, we briefly consider the simple example pathway depicted in Figure 5.9. Within our approach, not assuming any further knowledge of the explicit rate equations and parameters, the system is parametrized by the matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\theta^{\mu}_{A}}$

$$\boldsymbol{\Lambda} = \begin{bmatrix} \frac{\nu_1^0}{A^0} & -\frac{\nu_2^0}{A^0} & -\frac{\nu_3^0}{A^0} \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} 0 \\ \theta_A^{\mu_2} \\ \theta_A^{\mu_3} \end{bmatrix} \,, \tag{5.6}$$

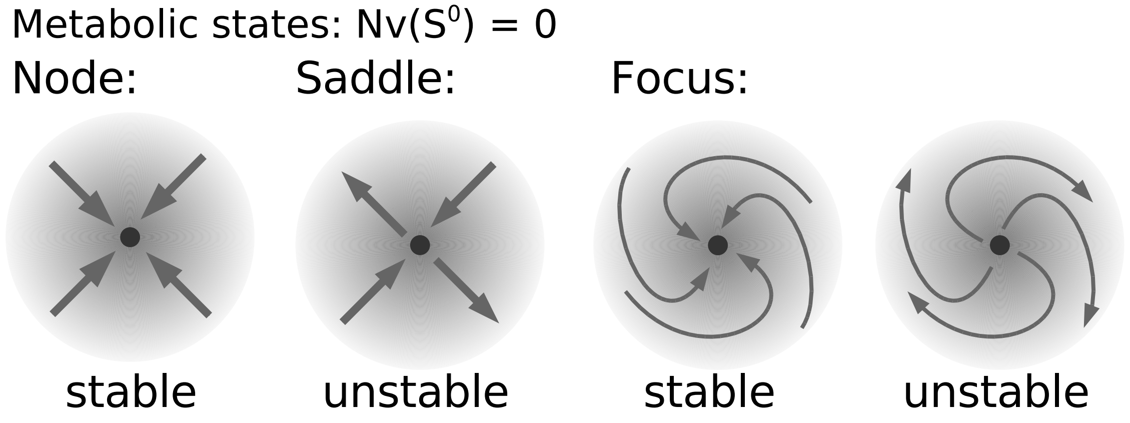


Figure 5.8: The stoichiometric balance equation $\boldsymbol{N\nu^0} = \mathbf{0}$ does not imply actual stability of a metabolic state. Though each of the depicted scenarios fulfills the steady state condition, the dynamic behavior can be classified as (Heinrich and Schuster, 1996): *i)* stable node, *ii)* unstable saddle *iii)* stable focus, corresponding to damped transient oscillations, and *iv)* unstable focus, corresponding to (undamped) transient oscillations.

where $\mathbf{\Lambda}$ defined the metabolic state of the system, constraint by $\nu_1^0 = \nu_2^0 + \nu_3^0$. For simplicity, we assume linear dependence of $\nu_2(A)$ on its substrate A, thus $\theta_A^{\mu_2} = 1$. The parameter $\theta_A^{\mu_3} \in [1-n, 1]$ includes possible nonlinear inhibition of $\nu_3(A)$ by its substrate A.

As the units of times and concentrations are arbitrary, we set $\nu_1^0 := 1$ and $A^0 = 1$ without loss of generality. The Jacobian at an observed metabolic state (specified by the matrix $\mathbf{\Lambda}$) is thus given as

$$\boldsymbol{J} = -1 + \nu_3^0(1 - \theta_A^{\mu_3}) \ . \tag{5.7}$$

Figure 5.9 shows the region of stability of the observed state ν_3^0 versus the (unknown) parameter $\theta_A^{\mu_3}$. More importantly, the observed metabolic state, here only characterized by ν_3^0, restricts the stability properties of the state. For small flux the observed state is always stable, i.e., there exists no set of parameters such that the state is unstable. However for high flux ν_3^0 the system might lose stability and is stable only in a small region of the (unknown) parameter space.

This behavior is again exemplified in Figure 5.10, using explicit differential equations. Starting in the vicinity of the metabolic state $\{A^0, \nu_3^0\}$, all other parameters are chosen randomly from the comprehensive parameter space. For small ν_3^0 (left plot), the system will always decay back to the state. Hence, the steady state remains stable, independent of the actual value of θ_A^3. However, for large ν_3^0 (right plot) the probability of instability increases. For some perturbations of θ_A^3 the steady state becomes unstable, i.e. the initial state cannot be restored. The systems transits into a new stable steady state with concentrations of **A** different from the initial value **A** = 1. In this sense, the observed metabolic state puts constraints on the possible dynamics and allows to quantify the existence size of unstable regions in parameter space. The perturbation analysis shows that the metabolic state with small ν_3^0 is more robust against changes in parameters than a metabolic state with large ν_3^0.

5.3.3 The role of regulation

The evaluation of models (Jacobians) for $\mathbf{C_{reg}}$ and $\mathbf{C_{noreg}}$ was repeated 10^3 times and the percentage of unstable instances recorded. In both cases, the variance of the values, due to finite sampling effects, was estimated numerically. A t-test was used to test the average values for both cases $\mathbf{C_{reg}}$ and $\mathbf{C_{noreg}}$ against each other. The null hypothesis that the expectation value in case

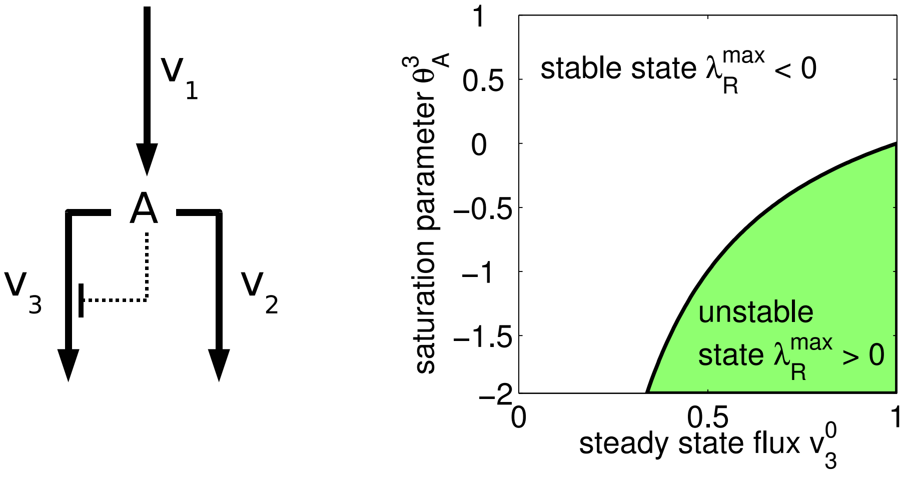


Figure 5.9: *Left:* A simple example, consisting of $r = 3$ reactions and $m = 1$ metabolite. ***Right:*** The dynamic behavior of the pathway as a function of the metabolic state ν_3^0 and the (unknown) saturation parameter.

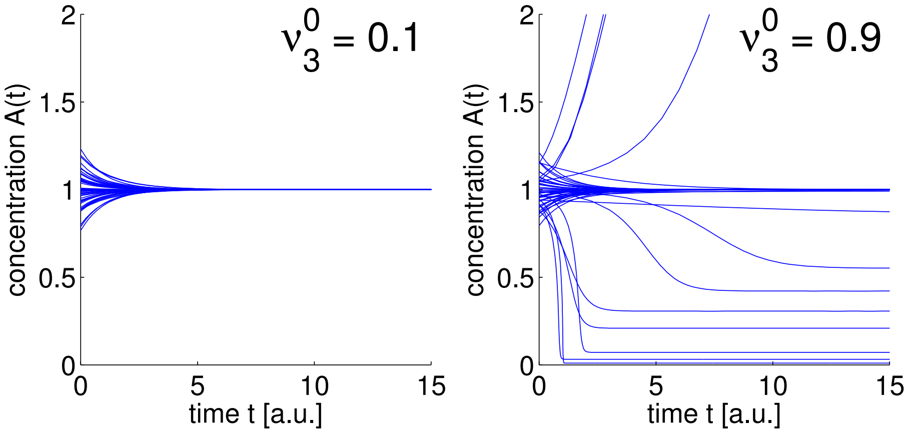


Figure 5.10: The time course of the concentration $A(t)$ using explicit kinetic simulations. Starting in the vicinity of a metabolic state, the observed flux distribution puts constraints on the possible dynamics of the pathway. For small ν_3^0 (left plot), the system will always decay back to the state. For large ν_3^0 (right plot) the probability of instability increases.

of allowed allosteric regulation is equal or lower than in case of suppressed allostric regulation is rejected with a p-value below 10^{-320}. So allosteric regulation significantly increases the frequency of stable models.

Furthermore, we tested if the observed increase is specific for the actual set of regulation parameters or if it can be achieved by any randomly chosen set of allosteric regulation parameters. To this end, instead of the actual regulation parameters, 10 putative allosteric regulations were selected randomly and the increase in the percentage of stable models was recorded. Most random sets ($\sim 85\%$) of regulation parameters lead to a decrease in the percentage of stable models, demonstrating that not every possible arbitrary allosteric regulation has a positive effect on stability.

5.3.4 Ranking of parameters

To assess the relative impact of enzyme-kinetic parameters on the stability properties of a given metabolic state, we employed and compared several measures of dependency.

The most common choice to detect dependencies between variables is the (Pearson) *correlation coefficient* r, defined as

$$r(X,Y) := \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{5.8}$$

where x_i and y_i are n realizations of the random variables X and Y. Although the Pearson correlation only detects linear dependencies, it holds the advantage that its sign specifies whether a parameter must be increased or decreased to obtain a higher percentage of stable models. Nonetheless, the Pearson correlation suffers from several drawbacks, such as sensitivity to non-gaussian and skewed distributions, making more elaborate measures necessary (Kumar and Shoukri, 2007).

A more general measure of dependency is given by the *mutual information* (Shannon, 1948), defined as

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \ , \tag{5.9}$$

where $H(X) = \sum_k p_k(x) \log p_k(x)$ denotes the entropy of the variable X, measured from a binned histogram such that each bin occurs with probability p_k. The entropy $H(X,Y)$ denotes the joint entropy of X and Y. Among its main advantages is that the mutual information is zero *if and only if* both variables are statistically independent. Evaluating the mutual information between network parameters and the resulting largest real part of the eigenvalue $\lambda_{Re}^{\max}$ thus accounts for arbitrary nonlinear dependencies and does not presuppose gaussian or uniform distribution of the parameters. A detailed account of its numerical estimation is given elsewhere (Steuer et al., 2002).

Based on a slightly different concept, the *Kolmogorov-Smirnov test* is used to test for the equality of two distributions. The null-hypothesis in the context of our analysis is as follows: If a saturation parameter has no impact on the stability of the metabolic system, then its distribution within the restricted subset of stable models equals (in a statistical sense) its initial distribution for the comprehensive set of models. On the other hand, if the distribution of a parameter within the restricted set of stable models shows a strong deviation from the initial distribution, a significant dependency can be expected. In this sense, the KS-test tests whether two random variables X and Y have the same distribution. The cumulative frequencies F_X and F_Y and the maximal difference

$$D = \sup_{z \in \mathbb{R}} |F_X(z) - F_Y(z)| \tag{5.10}$$

are calculated. If the test statistic D is greater than the critical value for the sample size, the null hypothesis that both distributions are equal is rejected. Since D is always identically distributed, the KS-test is independent of the distribution of X and Y. If the test rejects the null hypothesis, given a sufficiently small p-value, the parameter under consideration has significant impact on stability. All measures yielded consistent results, as depicted in Figure. 5.11.

Significance of ranking

To test for the significance of the correlation coefficient and the mutual information, we employed a permutation test: The values of $\lambda_{Re}^{\max}$ were randomly permuted in order to abolish any relationship between the saturation parameters and $\lambda_{Re}^{\max}$. This yielded mean values of the correlation coefficient and the mutual information close to zero ($< 5 \cdot 10^{-5}$) and standard deviation of $3.1 \cdot 10^{-3}$ and $2.5 \cdot 10^{-4}$, respectively. The correlation coefficient and the mutual information for the high ranked parameters are indeed significantly larger than those obtained in case of totally unrelated saturation parameters.

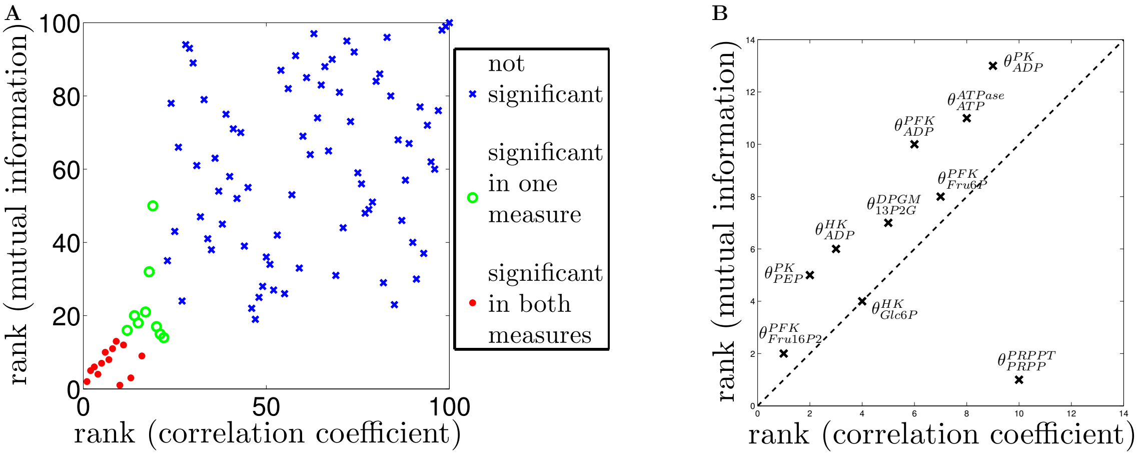


Figure 5.11: The ranking according to the correlation coefficient is plotted versus the ranking according to mutual information in case of suppressed allosteric regulation (**A**). For the significant parameters (red dots) the two measures give consistent result. The top ranking parameters from a total of 87 are shown in **B**. The outlier θ^{PRPPT}_{PRPP} is caused by nonlinear dependencies between θ^{PRPPT}_{PRPP} and $\lambda^{\max}_{Re}$. Note that θ^{PRPPT}_{PRPP} is still high ranked under both measures.

To verify the assertion that there is an enrichment of actual feedback parameters in the top ranking parameters, we conducted two statistical tests: *i)* For the ensemble of models with absent regulation ($\boldsymbol{C_{noreg}}$), we verify that high ranking parameters are primarily associated with reactions that are actually allosterically regulated (PK, PFK, PK, G6PD and 6PDG). To this end, we record the fraction of parameters associated with regulated reactions within the top k ranked parameters. This number is then compared (statistically) to the number that must be expected if the high-ranking parameters are indeed randomly distributed across all reactions. The results are depicted in Figure 5.12A (as a function of k) and show a clear significant enrichment of parameters associated with regulated reactions among the top ranking parameters. *i)* For the ensemble of models including allosteric regulation ($\boldsymbol{C_{reg}}$), we evaluate if regulation parameters are statistically overrepresented in the set of top ranking parameters. Again, we record the expected number of regulation parameters within the k top-ranking parameters, based on a purely random distribution of parameters. This value is compared to the actual number of regulation parameters within the set of top-ranked parameters. The result is again significant and depicted in Figure 5.12B.

5.4 Discussion and Conclusions

A central goal of metabolic regulation is homeostasis, i.e. the maintenance of a stable quasi-stationary state under largely varying external conditions. In this work we present a kinetic-free approach that enables the identification of those enzymes and putative allosteric regulators having the largest impact on the stability of experimentally observed metabolic steady states. Our analysis was focused on three different aspects: First, the role of allosteric regulation was elucidated by comparing the dynamic behavior of the network under suppressed and allowed allosteric regulation. The proportion of stable models is significantly increased by allosteric regulation, showing that feedback regulation has a stabilizing effect. Second, three statistical measures were introduced to quantify the influence of enzymatic reactions and saturation parameters on stability in a systematic way. The parameters were ranked according to these measures. Intriguingly, almost all high ranked parameters are involved in one of the three reactions HK, PFK or PK, corresponding

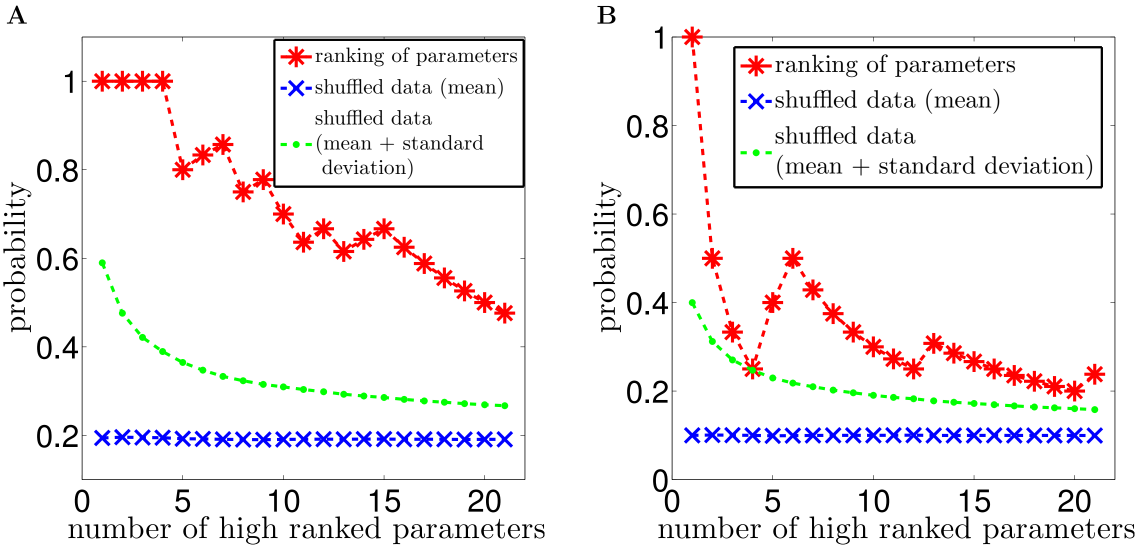


Figure 5.12: Probability of a parameter among the high ranked parameters to belong to a special subset of parameters (for $\mathbf{C_{noreg}}$: parameters that correspond to allosterically regulated reactions (**A**); for $\mathbf{C_{reg}}$: regulation parameters (**B**)). On the x-axis the number of parameters that are considered to be high ranked is shown (always beginning at the topmost). The results for the rankings given by $\mathbf{C_{noreg}}$ and $\mathbf{C_{reg}}$ are denoted by the red stars. The blue crosses show the results for shuffled rankings.

to those reactions that are indeed highly regulated and almost irreversible. We note that these results also hold when knowledge about allosteric regulation and irreversibility is not presupposed in the initial analysis. Third, we provided a quantitative measure to analyze different metabolic states with respect to their robustness towards perturbations in parameters. We compared the *in vivo* state with a second metabolic state, corresponding to an increased energy demand of the cell. With respect to robustness, the *in vivo* state is clearly superior, in accordance with the fact that high energy demand will lead to a metabolic collapse of the red blood cell.

Our approach is essentially based on knowledge of the stoichiometry and the metabolic state of the system. While detailed kinetic models are available for only very few metabolic networks, knowledge of metabolite concentrations and flux distributions becomes increasingly experimentally accessible (Fernie et al., 2004; Goodacre et al., 2004; Sauer, 2004). In this respect, we have demonstrated that different metabolic states, only characterized by a flux distribution and metabolite concentrations, are indeed associated with a unique spectrum of dynamic capabilities – and can be differentiated based on their stability properties. As our method specifically samples the parameter space associated with a given metabolic state, it can be directly related to experimental observations and thus improves methods based on a straightforward sampling of kinetic parameters within an explicit kinetic model (von Dassow et al., 2000).

In particular, we expect that recent efforts for biotechnological modifications of metabolic systems will concomitantly result in fundamental changes in the dynamic behavior of these networks. While a desired flux distribution might be stoichiometrically feasible, unanticipated changes in dynamic properties can lead to a failure of network function. The weak preconditions and the semi-automatic and straightforward manner of its implementation thus make our approach a suitable starting point to elucidate and detect changes in dynamic properties of metabolic networks for which the construction of detailed kinetic models is not yet possible.

Chapter 6

Kinetic hybrid models composed of mechanistic and simplified enzymatic rate laws - a promising method for speeding up the kinetic modelling of complex metabolic networks

Kinetic modelling of complex metabolic networks – a central goal of computational systems biology – is currently hampered by the lack of reliable rate equations for the majority of the underlying biochemical reactions and membrane transporters. On the basis of biochemically substantiated evidence that metabolic control is exerted by a narrow set of key regulatory enzymes, we propose here a hybrid modelling approach in which only the central regulatory enzymes are described by detailed mechanistic rate equations, and the majority of enzymes are approximated by simplified (nonmechanistic) rate equations (e.g. mass-action, LinLog, Michaelis-Menten and power law) capturing only a few basic kinetic features and hence containing only a small number of parameters to be experimentally determined. To check the reliability of this approach, we have applied it to two different metabolic networks, the energy and redox metabolism of red blood cells, and the purine metabolism of hepatocytes, using in both cases available comprehensive mechanistic models as reference standards. Identification of the central regulatory enzymes was performed by employing only information on network topology and the metabolic data for a single reference state of the network (Grimbs et al., 2007a). Calculations of stationary and temporary states under various physiological challenges demonstrate the good performance of the hybrid models. We propose the hybrid modelling approach as a means to speed up the development of reliable kinetic models for complex metabolic networks.

6.1 Introduction

Kinetic modelling is the only reliable computational approach to relate stationary and temporal states of reaction networks to the underlying molecular processes. The ultimate goal of computational systems biology is the kinetic modelling of complete cellular reaction networks comprising gene regulation, signalling and metabolism. Kinetic models are based on rate equations for the underlying reactions and transport processes. However, even for whole cell metabolic networks–although they have been under biochemical investigation for decades–only a low percentage of enzymes and an even lower percentage of membrane transporters have been kinetically characterized to an extent that would allow us to set up physiologically feasible rate equations. For the foreseeable future, full availability of 'true' rate equations for all enzymes is certainly an illusion, because of the lack of methods with which to efficiently gain insights into all kinetic effects controlling a given enzyme *in vivo*. Currently, there is not even systematic in vitro screening for all possible modes of regulation that a given enzyme is subjected to. In principle, such an approach would imply the testing of all cellular metabolites as potential allosteric effectors, all cellular kinases and phosphatases as potential chemical modifiers, and all cellular membranes as potential activating or inactivating scaffolds. However, the experimental effort actually required can be drastically reduced, considering that only a few metabolites exert significant regulation of enzymes, and that the signature of phosphorylation sites and membrane-binding domains is similar in most proteins studied so far. Another critical aspect regarding the use of mechanistic rate equations developed for individual enzymes under test tube conditions is the need for subsequent tuning of parameter values to take into account the influence of the cellular milieu, which is imperfectly captured in the in vitro assay (Teusink et al., 2000; Wilkinson et al., 2008).

Therefore, instead of waiting for 'everything', it has been proposed that we should start with 'something' by using simplified rate equations that can be established with modest experimental effort. At the extreme, parameters of such simplified rate equations can even be inferred from the known stoichiometry of a biochemical reaction (Smallbone et al., 2007).

The predictive capacity of the approximate modelling approaches published so far has not been critically tested for a broader range of perturbations that the considered network has to cope with under physiological conditions. One objective of our work was thus to assess the range of physiological conditions under which a kinetic model of erythrocyte metabolism based exclusively on simplified rate equations may still adequately describe the system's behaviour. This was done by replacing the full mechanistic rate equations for the 25 enzymes and five transporters involved in the model (Schuster and Holzhütter, 1995) by various types of simplified rate equations, and using these simplified models to calculate stationary load characteristics with respect to changes in the consumption of ATP and glutathione (GSH), the two cardinal metabolites that mainly determine the integrity of the cell. The goodness of these simplified models was evaluated by using the solutions of the full mechanistic model as the reference standard. In most cases that were tested, the simplified models failed to reproduce the 'exact' load characteristics even in a rather narrow vicinity around the reference *in vivo* state.

A second, and even more important, goal of our work was to test a novel modelling approach based on 'mixed' kinetic models composed of detailed and simplified enzymatic rate equations. Assuming a typical situation, where only the stoichiometry of the network and the fluxes as well as metabolite concentrations of a specific steady state are known, we identified central regulatory enzymes by using the recently proposed sampling method of structural kinetic modelling (SKM) (Grimbs et al., 2007a). For the small number of regulatory enzymes, the full mechanistic rate equations were used, whereas all other enzymes were described by simplified rate equations as before. These mixed kinetic models yielded significantly better load characteristics for almost all variants of simplified rate equations tested. Hence, the development of kinetic hybrid models composed of rate equations of different mechanistic strictness according to the regulatory importance of the respective enzymes may be a meaningful strategy to economize the experimental effort required for a mechanism-based understanding of the kinetics of complex metabolic networks.

The mathematical models described here have been submitted to the Online Cellular Systems Modelling Database and can be accessed free of charge at `http://jjj.biochem.sun.ac.za/database/bulik/index.html`.

6.2 Results

6.2.1 Test case 1 - a metabolic network of erythrocytes

To investigate the suitability of different variants of kinetic network models considered in this work, we have chosen a metabolic network of human erythrocytes for which detailed mechanistic rate laws of the participating enzymes are available (Schuster and Holzhütter, 1995). The network consists of 23 individual enzymatic reactions, five transport processes, and two overall reactions representing two cardinal physiological functions of the network, the permanent re-production of energy (ATP) and of the antioxidant GSH. The network comprises as main pathways glycolysis and the hexose monophosphate shunt, consisting of an oxidative and nonoxidative part (Figure 6.1). Setting the blood concentrations of glucose, lactate, pyruvate and phosphate to typical *in vivo* values creates a stable stationary working state of the system, which was taken as a reference state for the adjustment of the simplified rate laws and for the construction of the Jacobian matrix used for the analysis of stability. Enzymatic rate laws and other details of the full kinetic model are given in Bulik et al. (2009b).

6.2.2 Comparing simplified and mechanistic rate equations for individual reactions

We first studied the differences associated with replacing the exact rate equations of the erythrocyte network with the various types of simplified rate equations given in Table 6.1. In order to mimic the most common situation where the regulatory *in vivo* control of an enzyme by allosteric effectors, reversible phosphorylation and other mechanisms is not known, the simplified equations take into account only the influence of substrates and products on the reaction rate. The rate of metabolic enzymes determined by network perturbations of intact cells (Gombert and Nielsen, 2000; Speers and Cravatt, 2004) is inevitably influenced by changes of their allosteric effectors. To mimic this effect, fitting of the simplified rate equations to the 'true' mechanistic rate equations was done by varying the concentrations of reaction substrates and products as well as the concentrations of the respective modifier metabolites occurring in the mechanistic rate equations (see below).

The mass-action (MA) rate law represents the simplest possible rate law taking into account reversibility of the reaction and yielding a vanishing flux at thermodynamic equilibrium. It contains as parameters only the unknown forward rate constant k and the thermodynamic equilibrium constant (K), which does not depend on enzyme properties and is related to the standard Gibb's free energy ΔG_0 of the reaction by $K = \exp(\Delta G_0/RT)$. A numerical value for K or ΔG_0 can be determined from calorimetric or photometric measurements (Goldberg, 1999), or can be computed from the structure of the participating metabolites (Forsythe et al., 1997). The numerical value of the turnover rate constant k is commonly chosen such that the predicted flux rate equals the measured flux rate in a given reference state of the network. In this way, the value of k implicitly takes into account all unknown *in vivo* effects influencing the enzyme activity, such as allosteric effectors, the ionic milieu, molecular crowding, or binding to other proteins or membranes. The LinLog (LL) rate law (Delgado and Liao, 1992; Rottenberg, 1973) is inspired by the concept of linear nonequilibrium thermodynamics, which sets the reaction rate proportional to the thermodynamic driving force ΔG, the free energy change, which depends on the concentration of the reactants in a logarithmic manner. Nielsen (1997) proposed adding additional logarithmic concentration terms to include allosteric effectors. A further generalization was to neglect the stoichiometric coupling of the coefficients of the logarithmic concentration terms dictated by the free energy equation; that is, these coefficients are regarded as being independent of each other. We also included a special stoichiometric variant of the LinLog model (LLst) recently proposed by Smallbone et al. (2007), in which the coefficients of the logarithmic concentrations are simply given by the stoichiometric coefficient of the respective metabolites. The power law (PL) was

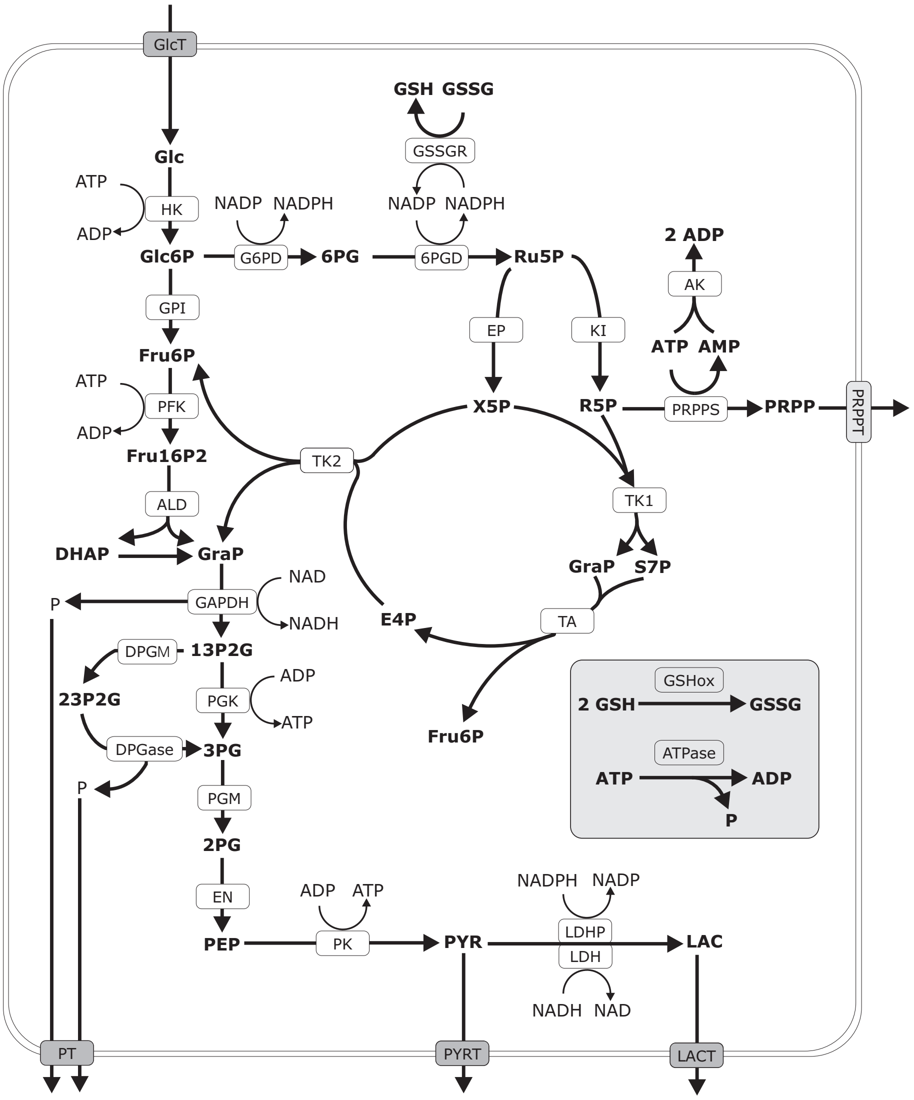


Figure 6.1: Erythrocyte energy metabolism. Reaction scheme of erythrocyte energy metabolism comprising glycolysis, the pentose phosphate shunt and provision of reduced GSH. The ATPase and GSH oxidase reactions are overall reactions representing the total ATP demand and reduced GSH consumption. 1,3PG, 1,3-bisphosphoglycerate; 2,3PG, 2,3-bisphosphoglycerate; 2PG, 2-phosphoglycerate; 3PG, 3-phosphoglycerate; 6PG, 6-phosphoglycanate; 6PGD, 6-phosphogluconate dehydrogenase; AK, adenylate kinase; ALD, aldolase; DPGase, 2,3-bisphosphoglycerate phosphatase; DPGM, 2,3-bisphosphoglycerate mutase; E4P, erythrose 4-phosphate; EN, enolase; EP, ribose phosphate epimerase; Fru1,6P2, fructose 1,6-bisphosphate; Fru6P, fructose 6-phosphate; G6PD, glucose-6-phosphate dehydrogenase; Glc6P, glucose 6-phosphate; GlcT, glucose transport; GPI, glucose-6-phosphate isomerase; GraP, glyceraldehyde 3-phosphate; DHAP, dihydroxyacetone phosphate; GSHox, glutathione oxidase; GSSG, oxidized glutathione; GSSGR, glutathione reductase; HK, hexokinase; KI, ribose phosphate isomerase; LAC, lactate; LACT, lactate transport; LDH, lactate dehydrogenase; PEP, phosphoenolpyruvate; PFK, phosphofructokinase; PGK, phosphoglycerate kinase; PGM, 3-phosphoglycerate mutase; PK, pyruvate kinase; PRPP, phosphoribosyl pyrophosphate; PRPPS, phosphoribosylpyrophosphate synthetase; PRPPT, phosphoribosylpyrophosphate transport; PYR, pyruvate; Rib5P, ribose 5-phosphate; Ru5P, ribulose 5-phosphate; S7P, sedoheptulose 7-phosphate; TA, transaldolase; TK, transketolase; TPI, triose phosphate isomerase; Xul5P, xylulose 5-phosphate.

Rate law	Formula	Comments
Linear mass action (MA)	$v = k \cdot \left(\prod_i S_i^{\mu_i} - \frac{1}{K_{Eq}} \prod_i P_i^{\nu_i}\right)$	
Power law (PL)	$v = k \prod_i \left(\frac{S_i}{S_i^0}\right)^{a_i} \prod_i \left(\frac{P_i}{P_i^0}\right)^{b_i} \left(\prod_i S_i^{\mu_i} - \frac{1}{K_{Eq}} \prod_i P_i^{\nu_i}\right)$	a_i, b_i - dimensionless constants S_i^0, P_i^0 - concentrations of substrates and products at a stationary reference state
LinLog (LL)	$v = v^0 \cdot \left(1 + \sum_i a_i \log\left(\frac{S_i}{S_i^0}\right) + \sum_i b_i \log\left(\frac{P_i}{P_i^0}\right)\right)$	a_i, b_i - empirical rate constants v^0, S_i^0, P_i^0 - flux and concentrations of substrates and products at a stationary reference state
Michaelis-Menten (MM)	$v = \frac{v_{max}\left(\prod_i S_i^{\mu_i} - \frac{1}{K_{Eq}} \prod_i P_i^{\nu_i}\right)}{\prod_i (1+a_i S_i)^{\mu_i} + \prod_i (1+b_i P_i)^{\nu_i} - 1}$	a_i, b_i - inverse half-concentrations of substrates and products

Table 6.1: Simplified rate expressions used in the kinetic model of erythrocyte metabolism. S_i and P_i denote the concentrations of the reaction substrates and products, respectively. The integer constants μ_i and ν_i are the stoichiometric coefficients with which the ith substrate and product enter the reaction. K denotes the thermodynamic equilibrium constant and k the catalytic constant of the subject enzyme, and v the flux of the reaction. The empirical parameters a_i and b_i have different meanings in the PL, LL and MM rate laws. The notation of the PL rate equation differs from the conventional form in that the rate is here decomposed into an MA term and a residual PL term. Hence, the PL exponents for substrates and products commonly used in most applications correspond to $a_i + \mu_i$ and $b_i + \nu_i$. The form of the MM equation used is based on the assumption that all μ_i substrate molecules and ν_i product molecules bind simultaneously (and not consecutively and not cooperatively) to the enzyme.

originally introduced by Savageau (1969). It has no mechanistic basis, i.e. it cannot be derived from a binding scheme of enzymeligand interactions using basic rules of chemical kinetics, but it provides a conceptual basis for the efficient numerical simulation and analysis of nonlinear kinetic systems (Voit and Radivoyevitch, 2000). The Michaelis-Menten (MM) equation was the first mechanistic rate law that took into account a fundamental property of enzyme-catalysed reactions, namely the formation of an enzyme-substrate complex explaining the saturation behaviour at increasing substrate concentrations. The form of the MM rate law given in Table 6.1 refers to a simplified reaction scheme in which the substrates and products bind to the enzyme in random order and without cooperative effects, i.e. without mutually influencing their binding constants.

The simplified rate equations were parameterized as described in Experimental procedures. For all 30 reactions of the network, the best-fit model parameters and the scatter plots of rates calculated by means of the simplified and mechanistic rate law, respectively, are given in Bulik et al. (2009b). In what follows, the distance between the paired values $\tilde{x}_i$ and x_i $(i = 1, 2, ...n)$ of any variable X computed by the exact and the approximate model, respectively, is measured by the normalized root mean square distance (NRMSD):

$$NRMSD(X) = \left[\frac{\sum_{i=1}^{n}(x_i - \tilde{x}_i)^2}{\sum_{i=1}^{n}\tilde{x}_i^2}\right]^{1/2}$$

Table 6.2 depicts the differences between the paired values of the exact and simplified rate laws. Generally, all simplified rate laws provided a poor approximation of the exact one (differences larger than 50%) for those reactions catalysed by regulatory enzymes such as HK, PFK, PK or G6PD, which have in common the fact that they are controlled by multiple effectors. For example, the rate of G6PD is allosterically controlled by Glc6P, ATP and 2,3-bisphosphoglycerate. Moreover, the enzyme uses free NADP and NADPH as substrates, whereas in the cell a large proportion of the pyridine nucleotides is protein bound. Obviously, simplified rate equations that do not explicitly take into account such regulatory effects fail to provide good approximations to the 'true' rate equations.

Averaging the NRMSD values across the 30 reactions of the network ranks the four types of simplified rate equations tested as follows: MM and PL perform best, with the PL approach resulting in slightly smaller average NRMSD values, and the MM approach describing more enzyme kinetics with the highest accuracy. The LL approach takes third place, followed by MA. This ranking is not unexpected, considering that the mathematical structure of the PL rate equations allows better fitting to complex nonlinear kinetic data than the linear or bilinear MA rate equations. Intriguingly, the LL rate law was able to reproduce the exact rates in sufficient quality for none of the reactions except the ATPase reaction. On the other hand, the quality achieved with the LL rate law fluctuated less from one reaction to the other than with the other simplified rate laws.

6.2.3 Calculation of stationary system states calculated with approximate models

To check how the inaccuracies of the simplified rate laws translate into inaccuracies of the whole network model, we calculated stationary metabolite concentrations and fluxes at varying values of four model parameters (in the following referred to as load parameters) defining the physiological conditions that the erythrocyte has typically to cope with: the energetic load (utilization of ATP), the oxidative load (consumption of GSH or, equivalently, NADPH) and the concentrations of the two external metabolites glucose and lactate in the blood. Changes of the energetic load are due to changes in the activity of the Na^+ /K^+-ATPase, accounting for about 70% of the total ATP utilization in the erythrocyte, as well as to preservation of red cell membrane deformability (Weed et al., 1969). Under conditions of osmotic stress (Dariyerli et al., 2004) or mechanical stress exerted during passage of the cell through thin capillaries (Kodícek, 1986), the ATP demand may increase by a factor of 35. The oxidative load of erythrocytes may rise by two orders of magnitude in the presence of oxidative drugs or intake of fava beans (McMillan et al., 2001). The average concentration of glucose in the blood amounts to 5.5 mM, but may vary between 3.0 mM in acute hypoglycaemia to 15 mM in severe untreated diabetes mellitus. The concentration of lactate in the blood is mainly determined by the extent of anaerobic glycolysis in skeletal muscle. It may rise from its normal value of 1 mM up to 8 mM during intensive physical exercise of long duration (Petibois and Deleris, 2004).

Stationary load characteristics for the 29 metabolites and 30 fluxes were constructed by varying the values of each of the four load parameters k_{ATPase} (rate constant for ATP utilization), k_{ox} (rate constant for GSH consumption), glucose concentration, and lactate concentration, within the following physiologically feasible ranges:

$\frac{1}{2}k^0_{ATPase}$	$\leq$	k_{ATPase}	$\leq$	$2k^0_{ATPase}$	small variation of the energentic load
$\frac{1}{5}k^0_{ATPase}$	$\leq$	k_{ATPase}	$\leq$	$5k^0_{ATPase}$	large variation of the energetic load
$\frac{1}{50}k^0_{ox}$	$\leq$	k_{ox}	$\leq$	$50k^0_{ox}$	variation of the oxidative load
3 mM	$\leq$	[Gluc]	$\leq$	15 mM	variation of blood glucose concentration
1 mM	$\leq$	[Lac]	$\leq$	8 mM	variation of blood lactate concentration

$k^0_{ATPase} = 1.6h^{-1}$ and $k^0_{ox} = 0.03h^{-1}$, respectively, denote the reference values for the chosen *in vivo* state of the cell. Differences between the load characteristics obtained by means of the exact model and the approximate models composed of the various types of simplified rate equations were evaluated by the NRMSD value defined in Experimental procedures. NRMSD

	Simplified rate law				
Reaction	MA (%)	PL (%)	LL (%)	LLst (%)	MM (%)
GlcT	16.5	1.3	10.1	**90.1**	16.0
HK	*43.5*	8.8	9.1	**62.8**	19.4
GPI	5.7	1.5	12.1	**99.0**	0.0
PFK	**83.3**	**60.5**	**58.7**	**90.8**	**79.9**
ALD	*33.6*	2.0	*22.2*	**78.3**	0.2
TPI	7.0	1.0	16.0	**99.8**	0.0
GAPD	*21.2*	1.7	*32.6*	**99.5**	0.1
PGK	**54.7**	**52.1**	*24.6*	**97.5**	**52.4**
DPGM	0.0	0.0	9.7	*33.2*	0.0
DPGase	0.0	0.0	9.5	*35.2*	0.0
PGM	0.5	0.1	17.2	**86.7**	0.0
EN	0.4	0.1	16.1	**68.2**	0.0
PK	*37.6*	*37.5*	*40.5*	**50.2**	*37.4*
LDH	0.0	0.0	*29.1*	**92.6**	0.0
LDH(P)	1.4	0.1	8.4	**62.4**	1.1
ATPase	0.7	0.1	0.3	*46.9*	0.0
AK	14.6	3.0	18.1	**100.0**	0.3
G6PD	12.3	9.4	*22.5*	*42.8*	10.6
GSSGR	3.7	1.0	15.7	**102.0**	4.7
GSHox	0.0	0.0	0.0	**89.5**	0.0
EP	0.9	0.2	17.1	**100.0**	0.0
KI	0.2	0.1	17.7	**98.9**	0.2
TK1	*28.6*	1.5	*29.7*	**50.2**	0.7
TA	*25.3*	3.6	*20.5*	**98.0**	2.5
PRPPS	10.2	0.2	8.7	*49.1*	0.8
TK2	*33.2*	3.0	*30.5*	**97.9**	0.9
PT	0.0	0.0	*25.5*	**100.0**	0.0
LacT	0.0	0.0	*25.5*	**100.0**	0.0
PyrT	0.0	0.0	*25.5*	**100.0**	0.0

Table 6.2: Differences between simplified and detailed rate laws. The differences between simplified and detailed rate laws for the individual reactions of the erythrocyte network are given as NRMSD values defined in Experimental procedures. Differences larger than 20% are in italic; differences larger than 50% are marked in bold. The scatter grams of the paired rate values for each reaction are given in Bulik et al. (2009b). 6PGD, 6-phosphogluconate dehydrogenase; AK, adenylate kinase; ALD, aldolase; DPGase, 2,3-bisphosphoglycerate phosphatase; EN, enolase; EP, ribose phosphate epimerase; GAPD, glyceraldehyde phosphate dehydrogenease; GlcT, glucose transport; GPI, glucose-6-phosphate isomerase; GSSGR, glutathione reductase; KI, ribose phosphate isomerase; LDH(P), lactate dehydrogenase (NADP dependent); PGK, phosphoglycerate kinase; PGM, 3-phosphoglycerate mutase; PRPPS, phosphoribosylpyrophosphate synthetase; PyrT, pyruvate transport; TA, transaldolase; TPI, triose phosphate isomerase; TK1, transketolase 1; TK2, transketolase 2.

values were computed across the range of the perturbed parameters for which a stationary solution was found with the approximate models. All individual load characteristics and the associated NRMSD values are contained in Bulik et al. (2009b). For an overall assessment of the predictive capacity of the approximate models, we computed mean NRMSD values by averaging across the individual NRMSD values for metabolites and fluxes (Table 6.3). In some cases, the approximate models failed to yield a stationary solution within a part of the full variation range of the perturbed load parameter. This is also depicted in the last four columns of Table 6.3.

6.2.4 Energetic load characteristics

Inspection of the NRMSD values in Table 6.3 (first and second columns) demonstrates that none of the approximate models provided a satisfactory reproduction of the true energetic load characteristics. The stoichiometric version of the LL yielded poor solutions. For the other approximate models, the average error in the prediction of stationary load characteristics ranged from 13.7% to 34.8% for small variations of the energetic load parameter, and from 22.3% to 50.9 for large variations. Considering that fixing all predicted fluxes and metabolite concentrations to zero gives an NRMSD value of 100%, we have to conclude that NRMSD value larger than 10% are unacceptably high. This conclusion is underpinned by the load characteristics for ATP shown in Figure 6.2. According to the exact model, the maximum of the ATP consumption rate appears at a 3.3-fold increased value of k_{ATPase} as compared to the value $k^0_{ATPase} = 1.6h^{-1}$. At values of k_{ATPase} exceeding seven-fold of its normal value, no stationary states can be found; that is, $k^{max} = 7k^0_{ATPase} = 11.2h^{-1}$ represents an upper threshold for the energetic load that still can be maintained by the glycolysis of the red cell. The nonmonotone shape of the load characteristics for ATP is accounted for by the kinetic properties of PFK, which is strongly controlled by the allosteric effectors AMP, ADP and ATP. The occurrence of a bifurcation at the critical value k^{max}_{ATPase} is an important feature of the energy metabolism of erythrocytes (Ataullakhanov et al., 1981). It is a consequence of the autocatalytic nature of glycolysis, which needs a certain amount of ATP for the 'sparking' reactions of HK and PFK in the upper part (Sel'kov, 1975). As shown in Figure 6.2, all approximate models completely failed to predict this important feature of the energetic load characteristics.

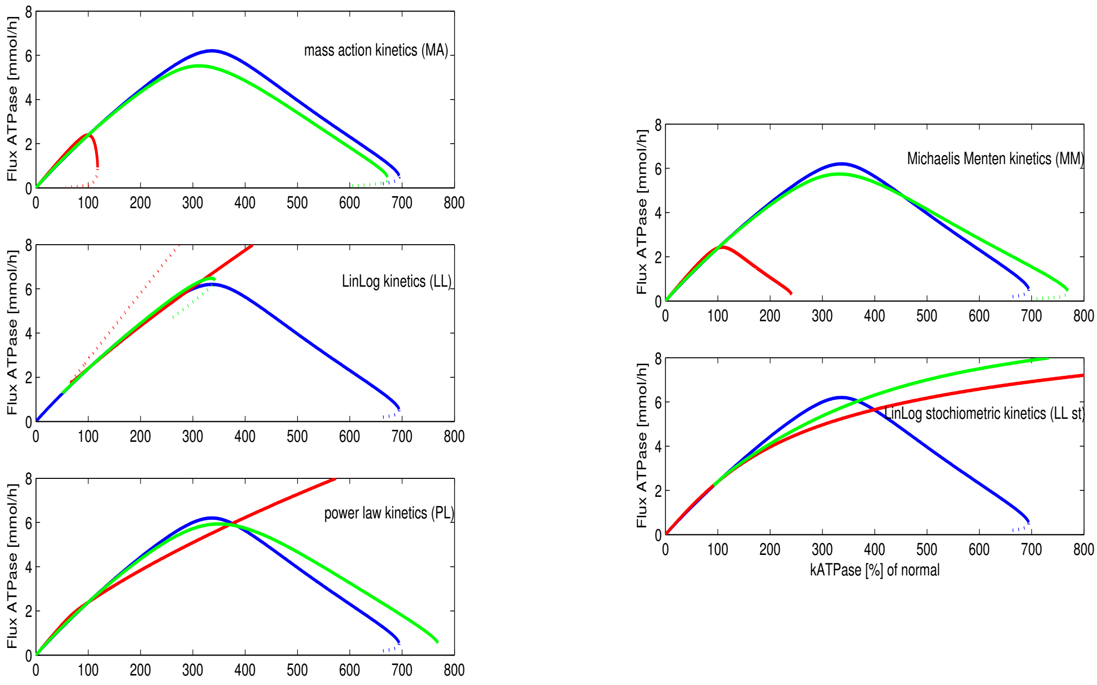


Figure 6.2: Erythrocyte energetic load characteristics. The diagrams show the total rate of ATP consumption versus the energetic load given as percentage of the energetic load $k_{ATPase} = 1.6$ mM of the reference state. Each diagram shows the load characteristics calculated by means of the mechanistic model (blue line), the approximate model fully based on simplified rate laws (red line), and the hybrid model (green line). Unstable steady states are indicated by dotted lines.

		Mean NRMSD					Range of load parameter values with stable solution (%)			
Simplified rate law	Variant of kinetic model	Energetic load 20-500% of normal	Energetic load 20-500% of normal	Oxidative load 2-5000% of normal	External glucose 3-15mM	External lactate 1-8mM	Energetic load 20-500% of normal	Oxidative load 2-5000% of normal	External glucose of normal	External lactate of normal
PL	**Hybrid**	**7.6**	**3.3**	**0.3**	**0.0**	**2.6**	**100**	**100**	**100**	**100**
	Fully simplified	38.0	23.9	5.0	0.5	5.1	100	100	100	100
MM	**Hybrid**	**8.9**	**3.4**	**1.4**	**0.1**	**2.6**	**100**	**100**	**100**	**100**
	Fully simplified	50.9	39.1	17.2	19.2	5.3	46	100	100	100
LL	**Hybrid**	**9.6**	**3.3**	**40.4**	**0.1**	**1.4**	**61**	**100**	**100**	**100**
	Fully simplified	22.3	13.7	41.0	0.4	5.9	84	100	100	100
MA	**Hybrid**	**14.2**	**3.7**	**16.2**	**0.1**	**3.4**	**100**	**91**	**100**	**100**
	Fully simplified	42.8	34.8	12.9	293.7	5.6	20	22	89	100
LLst	**Hybrid**	**95.9**	**40.1**	**98.9**	**1.9**	**10.6**	**100**	**100**	**100**	**100**
	Fully simplified	383.8	69.7	142.4	14.6	14.0	100	100	100	100

Table 6.3: Load characteristics. Mean NRMSD between the load characteristics calculated by means of the mechanistic kinetic model and the kinetic model either fully based on simplified rate laws (approximate model) or based on a mixture of simplified and detailed rate laws (hybrid model, values in bold). The heading designates the type of load parameter varied and the range of variation relative to the normal value of the reference state. The last four columns show the percentage of the total variation range of the load parameter where the simplified models yielded stable steady states. More detailed information is given in Bulik et al. (2009b). The mean NRMSD was obtained by averaging across the NRMSD values of all 29 metabolites and 30 fluxes of the model. NRMSD values were computed over the part of the variation range of the load parameter where the simplified model yielded a stable steady state.

Oxidative load characteristics

The true load characteristics are less complex than in the case of varying energetic load (Bulik et al., 2009b). Increasing rates of GSH consumption are paralleled by increasing rates of NADPH consumption. A decrease in the NADPH /NADP ratio activates G6PD and results in a monotone, quasilinear increase of the rate through the oxidative pentose pathway, whereas the much higher flux through glycolysis remains almost unaltered. Hence, those simplified rate equations capable of approximating reasonably well the kinetics of G6PD, the central regulatory enzyme in oxidative stress conditions, should also work reasonably well in the approximate kinetic model. Indeed, the NRMSD values in Table 6.3 (third column) clearly reflect the quality with which the simplified rate laws approximate the kinetics of G6PD (see Table 6.2): the approximate models based on PL-, MM- and MA-type rate equations provided a reasonably good reproduction of the exact load characteristics, whereas the approximate model based on LL-type rate equations performed poorly (mean NRMSD 41%).

Glucose characteristics

The approximate models performed generally better when external glucose levels were varied than for alterations of the energetic and oxidative load. The only exception is the model variant based on MA-type rate laws (mean RMSD = 293.7%). This is plausible because the linear MA-type rate law cannot describe substrate saturation. However, in the erythrocyte, the HK catalysing the first reaction step of glycolysis is completely saturated with glucose (Km value for glucose is about 0.1 mm); that is, even large variations in the blood level of glucose are hardly sensed by the cell. Indeed, the mechanistic rate law of the HK actually does not depend on the external glucose concentration, and thus the detailed network model yields identical flux patterns for the whole interval of external glucose concentrations studied. The nonlinear rate equations of the LL, MM and PL type are at least partially able to describe substrate saturation, and thus provide a reasonably good description of the HK kinetics.

Lactate characteristics

Increasing lactate concentrations in the blood and thus within the erythrocyte cause a 'back-pressure' to the lactate dehydrogenase (LDH) reaction, thus lowering the NAD /NADH ratio. This implies a decrease of the glycolytic flux, as NAD is a substrate of GAPD. The flux changes remain moderate even at high lactate concentrations, as GAPD has little control over glycolysis for a wide range of NAD concentrations. The induced changes in the flux pattern elicited by increasing lactate concentrations are small and monotone, and therefore can be predicted with sufficient quality by the approximate models, except for the variant based on stoichiometric LL-type rate laws.

In summary, the LLst provided unsatisfactory results for all test cases. The four other variants of the approximate models clearly failed to reproduce with acceptable quality the true load characteristics for variations of the energetic and oxidative load. However, they performed significantly better for changes of the external metabolites glucose and lactate. Overall, using the NRMSD values and the relative range of stable model solutions as quality criteria, the approximate models based on PL-type rate laws performed best, followed by the LL variant. Except for the PL variant, all other variants of approximate models failed in some test cases to provide stationary solutions for all parameter variations.

6.2.5 Calculation of stationary system states calculated with kinetic hybrid models

In order to improve the quality of the approximate models, we tested a model variant (in the following referred to as hybrid model) in which we used detailed mechanistic rate equations for a small set of the most relevant regulatory enzymes but simplified rate equations for the remaining enzymes. The regulatory importance of the enzymes involved in the network was assessed by applying the method of structural kinetic modelling (see Experimental procedures). This method is based on a statistical resampling of the Jacobian matrix of the reaction network. It requires as input only the stoichiometric matrix of the network and measured metabolite concentrations, as well as fluxes in a specific working state of the system. The central entities of SKM are so-called saturation parameters. They quantify the impact of metabolites on enzyme activities. SKM provides a ranking of enzymes and related saturation parameters according to their relative influences on the stability of the network in the chosen reference state. Table 6.4 shows the 10 saturation parameters with the highest average rank in three different statistical tests. To keep the number of enzymes for which detailed rate equations have to be established as low as possible, we decided to designate only three enzymes as being of central regulatory importance: PFK, HK and PK. For these three enzymes, we used detailed rate equations, whereas for all other enzymes we used various types of simplified rate equations as listed in Table 6.1.

The NRMSD values in Table 6.3 demonstrate that the hybrid models yielded, in most cases, considerably better predictions of the true load characteristics than the full approximate models. The span of load parameter values for which a stationary solution was found also increased. To illustrate the improvements achieved, Figure 6.2 compares the load characteristics for ATP consumption obtained with the exact model, with the full approximate models, and with the hybrid models. Only the hybrid model based on LL rate laws failed to reproduce the shape of the true load characteristics.

Taking arbitrarily an NRMSD value of 10% as the upper threshold for a good prediction, the number of good predictions increased from only seven to 19. Intriguingly, the hybrid models based on PL- and MM-type rate laws now produced acceptable load characteristics for all five perturbation experiments tested. Only the stoichiometric variant of the LL-type rate laws still gave

Metabolite	Enzyme	Average rank
Fru1,6P$_2$	PFK	1.3
Glc6P	HK	3.3
PEP	PK	4.0
ADP	HK	4.0
Fru6P	PFK	6.3
1,3PG	DPGM	7.0
ADP	PFK	7.3
ATP	ATPase	9.0
2,3PG	DPGM	10.0
ADP	PK	10.7

Table 6.4: Ranking of saturation parameters for erythrocyte energy metabolism. Average ranking of saturation parameters according to their impact on the dynamic stability of the network assessed by analysis of the eigenvalues of the resampled Jacobian matrix using three different statistical measures: correlation coefficient (Pearson), mutual information, and P-value of the Kolmogorov-Smirnov test. Fru6P, fructose 6-phosphate; Fru1,6P$_2$, fructose 1,6-bisphosphate; PEP, phosphoenolpyruvate; 1,3PG, 1,3-bisphosphoglycerate; 2,3PG, 2,3-bisphosphoglycerate.

unacceptably poor predictions in four of the five perturbation experiments. In particular, much better reproduction of the energetic and oxidative load characteristics could be achieved.

6.2.6 Test case 2 - a metabolic network of the purine salvage in hepatocytes

As a second test case to check the feasibility of our hybrid modelling approach, we have chosen the purine nucleotide salvage metabolism of hepatocytes. This study has been confined to the use of the most simple types of simplified rate laws, the MA and the stoichiometric LL type. This choice was motivated by the fact that these two types of rate laws require a minimum of parameters and thus currently will certainly be the most frequently used ones in the kinetic modelling of complex metabolic networks.

Salvage metabolism plays an important role in the regulation of the purine nucleotide pool of the cell. The central metabolites here are AMP and GMP, which serve as sensors of the energetic status of the cell (Hardie, 2003). Under conditions of enhanced utilization or attenuated synthesis of ATP or GTP, the concentrations of the related monophosphates increase, due to the fast equilibrium maintained among the mononucleotides, dinucleotides and trinucleotides by adenylate kinase and guanylate kinase, respectively. This increase in AMP or GMP is accompanied by enhanced degradation of these metabolites by either deamination or dephosphorylation, giving rise to a reduction in the total pool of purine nucleotides. The physiological significance of this degradation is not fully understood. It can be argued that diminishing the concentration of AMP under conditions of energetic stress shifts the equilibrium of the adenylate kinase reaction towards AMP and ATP, and thus promotes the utilization of the energy-rich phosphate bond of ADP (Murray, 1971). Remarkably, some of the degradation products (adenosine, IMP, hypoxanthine, and guanine) can be salvaged, i.e. reconverted into AMP or GMP. Hence, under resting conditions, the depleted pool of purine nucleotides can be refilled without a notable rate increase of *de novo* synthesis.

The reaction scheme of this pathway (Figure 6.3) and the related kinetic model have been adopted from an earlier publication of our group (Bartel and Holzhütter, 1990).

We used the full mechanistic model to calculate the stationary reference state of the network at an ATP consumption rate of $20.8\mu\mathrm{M} \cdot s^{-1}$ and a GTP consumption rate of $0.19\mu\mathrm{M} \cdot s^{-1}$. On the basis of the stoichiometric matrix of the network and the flux rates and metabolite concentrations of the reference state, we applied the SKM method to identify those enzymes and reactants exerting the most significant influence on the stability of the system (Table 6.5). This analysis revealed the enzymes AMP deaminase and adenylosuccinate synthase to have the largest impact on the stability of the system. On the basis of this information, we constructed kinetic hybrid models, using,

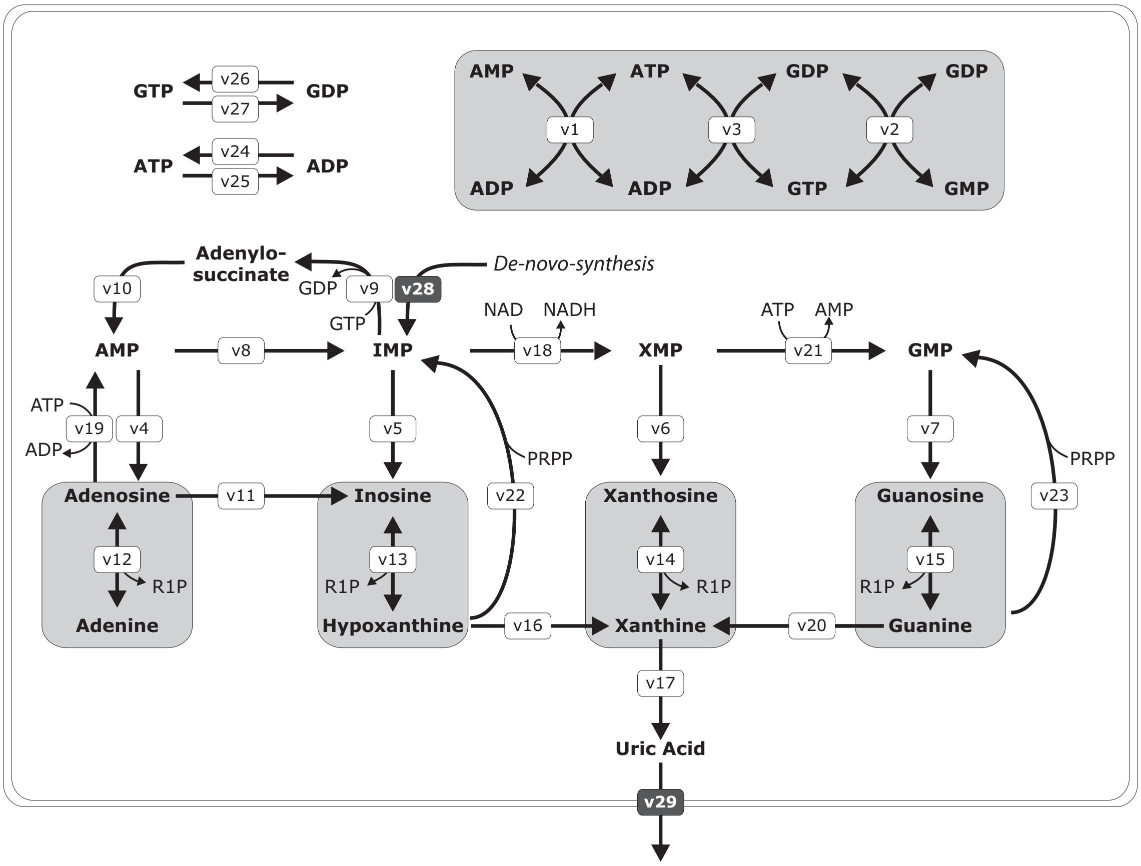


Figure 6.3: Hepatocyte purine metabolism. Reaction scheme of hepatocyte purine metabolism. The consumption and synthesis of ATP and GTP as well as the *de novo* synthesis of purines are overall reactions. Metabolites in grey boxes are in fast equilibrium. IMP, inosine monophosphate; XMP, xanthosine monophosphate; PRPP, phosphoribosyl pyrophosphate; R1P ribosyl 1-phosphate; v1, adenylate kinase; v2, guanylate kinase; v3, nucleotide diphosphate kinase; v4-v7, 5'-nucleotidase; v8, AMP deaminase; v9, adenylosuccinate synthetase; v10, adenylosuccinase; v11, adenosine deaminase; v12-v15, nucleoside phosphorylase; v16-v17, xanthine oxidase; v18, IMP dehydrogenase; v19 adenosine kinase; v20, guanine deaminase; v21, GMP synthetase; v22-v23, hypoxanthineguanine phosphoribosyltransferase; v24, ATP synthesis; v25, ATP consumption; v26, GTP synthesis; v27, GTP consumption; v28, purine *de novo* synthesis; v29, uric acid export.

for these two enzymes, the original mechanistic rate equations but modelling all other enzymes by simplified rate equations of either the MA type or the LL (stoichiometric) type, respectively. For comparison, we also constructed the fully reduced model by replacing all rate equations by their simplified counterparts. To check the performance of the simplified models, we simulated a physiologically relevant case where the cell is exposed to transient hypoxia 30 min in duration (e.g. owing to the complete occlusion of the hepatic artery) followed by a recovery period with a full oxygen supply. As shown in Figure 6.4, the fully approximated MA variant provides a reasonable description of adenine nucleotide behaviour during the anoxic period but completely fails to adequately describe the time-courses during the subsequent reoxygenation period. The LL (stoichiometric) approach describes the entire time-course quite well, even though the AMP concentration does not decline during the hypoxia period, and the depletion of the total pool of adenine nucleotides is clearly underestimated. Evidently, both types of simplified rate equations

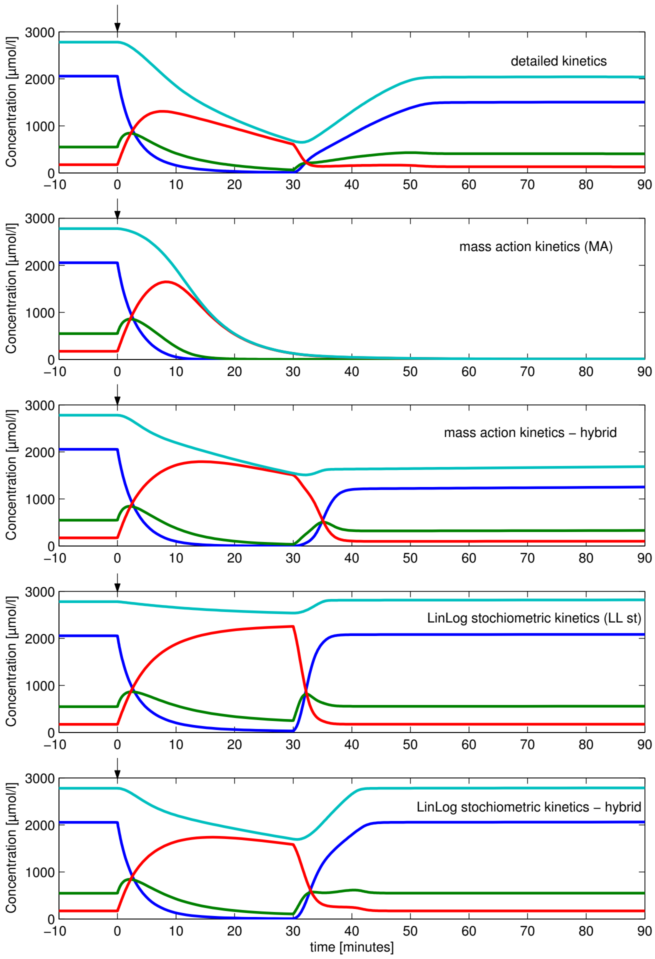


Figure 6.4: Hepatocyte anoxic simulation. The diagrams show the adenine nucleotides (ATP, blue; ADP, green; AMP, red) and the total adenine pool (turquoise) for hepatocyte purine metabolism. After a short initial period, the ATP synthesis is set to zero (indicated by arrow) for 30 min. After this anoxic interval, ATP synthesis is reset to its normal *in vivo* rate. The recovery of the adenine nucleotides for the next 60 min is also shown. Each panel displays a different model. The hybrid (third and fifth panel) models are closer to the full model (first panel) than the fully simplified models.

perform significantly better when incorporated into the hybrid model.

6.3 Experimental procedures

6.3.1 Distance measure

The distance between the paired values $\tilde{x}_i$ and x_i, $(i = 1, 2, ...n)$ of any variable X computed by the exact and the approximate model, respectively, was measured by the NRMSD:

$$NRMSD(X) = \left[\frac{\sum_{i=1}^{n}(x_i - \tilde{x}_i)^2}{\sum_{i=1}^{n}\tilde{x}_i^2}\right]^{1/2}$$

6.3.2 Parameterization of simplified rate equations

The concentrations of substrates [S], products [P] and allosteric effectors [E] of the corresponding enzyme were randomly varied within concentration intervals bounded by half and two-fold the reference concentrations. The conservation rules of the original model were kept. This parameterization procedure simulates an ideal situation where the flux rates through the individual reactions and the concentration values of the respective reactants are being measured within the intact network operating in its cellular environment (i.e. either in the intact cell or at least in a cell lysate) and adopting a sufficiently large spectrum of different states elicited by external perturbations. In

Flux	Enzyme	Metabolite	Average rank
v9	Adenylosuccinate synthetase	GDP	1.7
v8	AMP-deaminase	IMP	2.0
v9	Adenylosuccinate synthethase	GTP	3.0
v9	Adenylosuccinate synthethase	IMP	5.3
v18	IMP dehydrogenase	IMP	5.7
v18	IMP dehydrogenase	XMP	6.0
v21	GMP synthethase	XMP	7.0
v10	Adenylosuccinase	Adenylosuccinate	10.7
v8	AMP-deaminase	AMP	11.0
v21	GMP synthetase	ATP	11.3

Table 6.5: Ranking of saturation parameters for hepatocyte purine metabolism. Average ranking of saturation parameters according to their impact on the dynamic stability of the network assessed by analysis of the eigenvalues of the resampled Jacobian matrix using three different statistical measures: correlation coefficient (Pearson), mutual information, and P-value of the Kolmogorov-Smirnov test.

this case, the measured flux rates - here represented by the values of the mechanistic rate law - are influenced by allosteric effectors and other kinetic effects (e.g. reversible chemical modifications, and binding of enzymes to other proteins or membranes), although these regulatory influences are not explicitly considered in the simplified rate equations. Numerical values of the unknown parameters of a simplified rate equation were determined by minimizing the NRMSD given by the above equation of the predicted flux. Minimization was performed using the nonlinear optimization program solver 6.5 for excel. In these calculations, the random variation of the concentrations of reactants preserved the conservation rules of the system, e.g. constancy of the total concentration of adenine and pyridine nucleotides. Each reaction was trained separately and then corrected for the reference state of the erythrocyte network. For the LL rate law, we additionally tested a recently proposed variant (Smallbone et al., 2007) in which the coefficients are identical with the stoichiometric coefficient of the respective reactant; that is, for the monomolecular reaction S → P, the rate law simply reads $v = v^0 \cdot \left(1 + \log\left(S/S^0\right) - \log\left(P/P^0\right)\right)$.

6.3.3 Construction of load characteristics

A system is stationary when it satisfies the equation $dS/dt = 0$, with $dS/dt = Nv(S)$, N being the stoichiometric matrix, S the vector of metabolite concentrations, and $v(S)$ the vector of fluxes of the system. The load characteristics were calculated by varying a load parameter within a preset range of physiologically reasonable values. For each value of the load parameter, the steady state was computed by determining the metabolites S so that the above stationary condition is fulfilled. The numerical calculations were carried out with Matlab (MathWorks, Natick, MA, USA) Version 7.5.0.338. The stability of each solution was determined by evaluation of the eigenvalues of the Jacobian matrix $[J = dv(S)/dS]$.

6.3.4 Identification of regulatory enzymes by the SKM method

Quantification of the regulatory importance of the enzymes involved in the network was performed by applying the SKM method (Grimbs et al., 2007a; Steuer et al., 2006). This method is based on linearization of the kinetic equations with respect to a stationary working state of the system for which experimental data on fluxes and metabolite concentrations are available. The corresponding Jacobian matrix is decomposed into a product of two matrices, one depending on the flux rates and metabolite concentrations, and the other being constituted of so-called saturation parameters quantifying the influence that a small change in the concentration of an arbitrary metabolite has on the flux through a given reaction. If the change in the reaction rate is zero (meaning that the metabolite is neither a substrate nor an allosteric effector of the catalysing enzyme or, alternatively, that the enzyme is saturated with the metabolite), the corresponding saturation parameter is zero. If, at the other extreme, the change in the reaction rate is proportional to the change in the concentration of the metabolite, the saturation parameter equals unity. The saturation parameter thus has a strong similarity to the so-called elasticity coefficient used in metabolic control theory (Heinrich and Schuster, 1996; Fell, 1996).

At given values of the saturation parameters, one may compute the eigenvalues of the Jacobian matrix that determine the kinetic modes of the system elicited by small perturbations of the chosen working state. In particular, the largest eigenvalue indicates whether or not the working state is (locally) stable. The basic idea of SKM is to generate in a random fashion a large set of putative saturation parameter values for each enzyme. This results in an equally large set of Jacobian matrices containing the information on the stability of the system. As the interaction of nonreactant metabolites with enzymes in the system is generally unknown, the respective entries in the matrix are fixed to zero to reduce complexity and computational costs. The nonzero entries of the saturation matrix were sampled in the range $[0, x_{st}]$, with x_{st} being the stoichiometric coefficient of the metabolite in the catalysed reaction. Various statistical methods, such as correlation analysis, mutual information analysis, or the Kolmogorov-Smirnov test, can be used to assign a statistical measure to each possible saturation parameter, evaluating its linkage with changes in the largest eigenvalue of the Jacobian matrix. Fixing a reasonable threshold value for the statistical measure used, one arrives at a restricted list of potential regulatory enzymes and relevant metabolites controlling their rate (for further details, see Grimbs et al. (2007a)).

6.4 Discussion

Complex cellular functions such as growth, aging, spatial movement and excretion of chemical compounds are brought about by a giant network of molecular interactions. Kinetic models of cellular reaction networks still represent the only elaborated mathematical framework that allows temporal changes and spatial distribution of the constituting molecules to be related to the underlying chemical conversions and transport processes in a causal manner. With the establishment of systems biology as a new field of study, a tremendous effort has been made to develop high-throughput screening methods enabling the simultaneous monitoring of huge sets of different molecules (mRNAs, proteins, and organic metabolites). These methods have revealed unexpectedly vivid dynamics of gene products and related metabolites. However, in most cases, these dynamics appear to be enigmatic and hardly explicable in a causal manner, because up to now not enough effort has been made to elucidate and kinetically characterize the biochemical processes behind the observed changes in levels of molecule. In contrast, enzyme kinetics – a field that has shaped the face of biochemistry over decades – is currently considered to be old-fashioned. As a result, kinetic modelling of cellular reaction pathways is today seriously hampered by the unavailability of reliable rate laws for the processes involved in a network under consideration. For lack of anything better, it is common practice in the contemporary literature to base kinetic models on simplified rate laws. Such an approach may work reasonably well for small perturbations of a well-characterized working state. This conclusion is almost trivial, as sufficiently close to a steady state, the complex nonlinear kinetic rate laws can be reasonably well approximated even by simple linear rate laws of the MA type. Indeed, most of the studied approximate models of the erythrocyte network performed sufficiently well for changes of the external concentrations of glucose and lactate. The reason is that the metabolism of this cell is controlled by the demand for energy and redox equivalents, and not by the offer of substrates. Even larger variations in the concentrations of glucose and lactate give rise to only small changes in the activity of the sensing enzymes HK and LDH, and thus represent small metabolic challenges.

The point is, however, that in most biological, medical and biotechnological applications, small perturbations are not of great interest. Instead, one wants to make predictions about how the system behaves in cases of large perturbations, e.g. a sudden increase in the ATP demand when starting muscular work, the pharmacological inhibition of an enzyme, a sudden change in pH, the depletion of an essential substrate, or the presence of a toxic compound. As revealed by our analysis, under such conditions, kinetic models composed of simplified rate laws may lead to completely wrong predictions of the system's response, because the kinetic properties of those enzymes with decisive regulatory impact are not adequately captured. One may argue that this disappointingly poor performance is due to the fact that the simplified rate equations used in our analysis do not capture regulatory effects as exerted, for example, by allosteric effectors. First, such knowledge is currently available for only a small percentage of enzymes. Second, it is not acceptable to fill the gaps in our knowledge of regulatory properties by making the assumption that the same regulatory effectors are operative at isoenzymes in different species or different compartments of the same cell type. For example, the glycolytic enzyme PK can be isolated from mammalian tissues as four isoenzymes (L, R, M1 and M2). Each isoenzyme exhibits different kinetic properties that reflect the particular metabolic requirements of the expressing tissue (Bond et al., 2000; Jurica et al., 1998). Finally, if regulatory effectors have been elucidated by careful kinetic characterization of an enzyme, there are sufficient data available to set up a mechanistic rate law instead of a simplified one. Therefore, our decision to incorporate into the simplified rate laws only the chemistry of the reaction appears to be justified.

As a feasible compromise between the use of kinetic models fully based on either simplified or mechanistic rate laws, we propose here the use of hybrid models composed of simplified rate equations for the majority of reactions but detailed rate equations for a limited set of regulatory enzymes. Our approach relies on biochemically substantiated evidence that kinetic control of cellular metabolism is not democratically distributed across all participating enzymes and transporters. Rather, in all pathways hitherto studied in more detail, there exists a narrow set of key regulatory enzymes that are targeted by allosteric effectors and often also regulated by reversible phosphorylation. Accordingly, kinetic models should incorporate the kinetic properties of these central regulatory enzymes with sufficient accuracy, whereas the majority of the other 'workhorse' enzymes can be modelled with simplified rate equations.

To demonstrate the feasibility of the proposed hybrid approach, we have applied it to two well-studied metabolic systems, the redox and energy metabolism of erythrocytes, and the purine salvage metabolism of hepatocytes. In both cases, existing comprehensive kinetic models have been used as reference standards irrespective of the problem of to what extent these reference standards actually recapitulate all available biochemical knowledge of the considered networks. In fact, both reference models do not include all reactions that have been reported in the KEGG database, and some of the parameter values, in particular those of the thermodynamic equilibrium constants, need revision in the light of new measurements. Nevertheless, both reference models have been shown to correctly reflect basic dynamic features of the underlying pathways. Several elaborated kinetic models of erythrocyte metabolism have been recently compared (du Preez et al., 2008) and shown to adequately describe stationary load characteristics despite the use of different sets of parameters for the involved enzymes.

In the investigation of the salvage metabolism, we anticipated a typical situation when only a minimal amount of data is available. The SKM method requires data on metabolite concentrations and fluxes for one working state of the system. Both of the simplified approaches used (MA and LLst) can be parameterized with such data, whereas the more advanced models (LL, PL and MM) require more data to train the rate law parameters. Importantly, even the two most simple hybrid approaches yielded satisfactory results, and the more sophisticated models should perform even better.

The crucial problem in our approach is to identify the key regulatory enzymes and their main effectors. This problem is closely related to the determination of flux control coefficients and elasticity coefficients defined in metabolic control analysis (Heinrich and Schuster, 1996). Experimentally, this task can be tackled by measuring stationary load characteristics recorded upon inhibitor titration of individual enzymes (Fell, 1996; Small, 1993). Alternatively, one may apply a dynamic approach to estimate control coefficients from transient metabolite trajectories elicited by perturbations of the network (Delgado et al., 1993; Kresnowati et al., 2005). Whereas these methods are very expensive from the experimental point of view, the concept of structural kinetic modelling (Grimbs et al., 2007a) requires as input only the stoichiometry of the network and data on metabolite concentrations and fluxes in a typical working state of the network. Using this method, we identified the three glycolytic enzymes HK, PFK and PK as the putative most relevant regulatory enzymes of the network. This insight is not new, but here it was derived just from the topology of the network and metabolic data of a single reference state, whereas it took decades of biochemical research combined with mathematical modelling to unravel the central regulatory role of these enzymes. It has to be noted, however, that selecting a limited set of relevant regulatory enzymes from a ranked list of statistical scores is, to some extent, arbitrary. One way to remove this arbitrariness might be to include a successively enlarged set of putative regulatory enzymes in the construction of the kinetic hybrid model and to stop the procedure if there is no significant change of the computed trajectories and load characteristics relevant to the questions addressed by the model.

For the kinetic characterization of selected regulatory enzymes, *in vitro* experiments still seem to be the method of choice, because they allow a systematic search for allosteric effectors and a variation of the enzyme ligands in a sufficiently broad concentration range. In some cases, the derivation of a detailed rate law can be facilitated by searching enzyme databases (Barthelmes et al., 2007; Wittig et al., 2006) for rate laws already established for the same enzyme from other cell types. If the three-dimensional protein structures are known, it is even possible to estimate numerical values of kinetic constants for structurally and mechanistically similar enzymes (Gabdoulline et al., 2007).

Taken together, the development of hybrid models could be a realistic strategy to speed up the kinetic analysis of cellular reaction networks.

Chapter 7

General conclusion

The presented work used mathematical and computational approaches to cover various aspects of metabolic network modelling, especially regarding the limited availability of detailed kinetic knowledge on reaction rates. It was shown that precise mathematical formulations of problems are needed i) to find appropriate and, if possible, efficient algorithms to solve them, and ii) to determine the quality of the found approximate solutions. Furthermore, some means were introduced to gain insights on dynamic properties of metabolic networks either directly from the network structure or by additionally incorporating steady-state information. Finally, an approach to identify key reactions in a metabolic networks was introduced, which helps to develop simple yet useful kinetic models.

The rise of novel techniques renders genome sequencing increasingly fast and cheap (Rothberg and Leamon, 2008). In the near future, this will allow to analyze biological networks not only for species but also for individuals (Hood et al., 2004). Hence, automatic reconstruction of metabolic networks provides itself as a means for evaluating this huge amount of experimental data. This was discussed in Chapter 2. A mathematical formulation as an optimization problem was presented, taking into account existing knowledge and experimental data as well as the probabilistic predictions of various bioinformatical methods. The reconstructed networks are optimized for having large connected components of high accuracy, hence avoiding fragmentation into small isolated subnetworks. The usefulness of this formalism was exemplified on the reconstruction of the sucrose biosynthesis pathway in *Chlamydomonas reinhardtii*. However, the problem was shown to be computationally demanding and therefore necessitates efficient approximation algorithms. The development of algorithms with provable approximation quality remains as an open problem.

The problem of minimal nutrient requirements for genome-scale metabolic networks was analyzed in Chapter 3. Given a metabolic network and a set of target metabolites, the *inverse scope problem* has as it objective determining a minimal set of metabolites that have to be provided in order to produce the target metabolites. These target metabolites might stem from experimental measurements and therefore are known to be produced by the metabolic network under study, or are given as the desired end-products of a biotechological application. The inverse scope problem was shown to be computationally hard to solve. However, we assume that the complexity strongly depends on the number of directed cycles within the metabolic network. This might guide the development of efficient approximation algorithms. Furthermore, in the mathematical framework of the inverse scope problem, a reaction is assumed to take place if all substrates are available in principle, without taking into account the stoichiometry of the reaction. Hence, mass-balance is not considered. A way to overcome this problem might be achieved by using Petri nets (Heiner and Koch, 2004) as models of metabolic networks.

Assuming mass-action kinetics, chemical reaction network theory (CRNT), introduced in

Chapter 4, allows for eliciting conclusions about multistability directly from the structure of metabolic networks. Although CRNT is based on mass-action kinetics originally, it was also shown how to incorporate further reaction schemes by emulating molecular enzyme mechanisms. CRNT was used to compare several models of the Calvin cycle, which differ in size and level of abstraction. Definite results were obtained for small models, but the available set of theorems and algorithms provided by CRNT could not be applied to larger models due to the computational limitations of the currently available implementations of the provided algorithms. Therefore, implementing improved versions of these algorithms would be a compulsive step to apply CRNT to metabolic networks of medium to large size.

Structural kinetic modelling was presented in Chapter 5. Given the stoichiometry of a metabolic network together with steady-state fluxes and concentrations, this sampling approach allows to analyze the dynamic behavior of the metabolic network, even if the explicit rate equations are not known. In particular, this approach was used to study the stabilizing effects of allosteric regulation in a model of human erythrocytes. Furthermore, the reactions of that model could be ranked according to their impact on stability of the steady state. The most important reactions in that respect were identified as hexokinase, phosphofructokinase and pyruvate kinase, which are known to be highly regulated and almost irreversible. In general, the sampling technique could possibly be improved by restricting the parameter space to account for possible thermodynamic dependencies between kinetic parameters within a metabolic pathway (Liebermeister and Klipp, 2006).

In Chapter 6 kinetic modelling approaches using standard rate equations have been compared and evaluated against reference models for erythrocytes and hepatocytes. The results from this simplified kinetic models could simulate acceptably the temporal behavior for small changes around a given steady state, but failed to capture important characteristics for larger changes. The aforementioned approach to rank reactions according to their influence on stability was used to identify a small number of key reactions. These reactions were modelled in detail, including knowledge about allosteric regulation, while all other reactions were still described by simplified reaction rates. These so-called *hybrid* models could capture the characteristics of the reference models significantly better than the simplified models alone. The resulting hybrid models might serve as a good starting point for kinetic modelling of genome-scale metabolic networks, as they provide reasonable results in the absence of experimental data, regarding, for instance, allosteric regulations, for a vast majority of enzymatic reactions. As the identification of important reactions is crucial for constructing hybrid models, further approaches to identify such key reactions could also be included to cover additional (structural) network properties apart from stability.

Implementation of the introduced algorithms was carried out in the free statistical programming language R and in MATLAB. To allow for further applications besides the examples discussed in Chapter 5 and 6 and to provide a comprehensive software package, an interface to the *Systems Biology Markup Language* (Hucka et al., 2003) should be developed as a next step.

The presented approaches and methods contribute to alleviating the complex and challenging task of modelling metabolic networks and improve the quality of simulations and predictions. Naturally, the proposed methods can be further improved, as already mentioned above and in the discussion and conclusion of the individual chapters. However, incorporating these improvements is beyond the scope of this work and will be addressed in future works.

Bibliography

Acuña, V., Chierichetti, F., Lacroix, V., Marchetti-Spaccamela, A., Sagot, M.-F., and Stougie, L. (2008). Modes and cuts in metabolic networks: Complexity and algorithms. *Biosystems*.

Albert, R. (2005). Scale-free networks in cell biology. *J Cell Sci*, 118(21):4947–4957.

Alm, E. and Arkin, A. P. (2003). Biological networks. *Curr Opin Struct Biol*, 13(2):193–202.

Alon, U. (2003). Biological networks: the tinkerer as an engineer. *Science*, 301(5641):1866–1867.

Aloy, P. and Russell, R. B. (2006). Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol*, 7(3):188–197.

Andre, M., Ijaz, K., Tillinghast, J. D., Krebs, V. E., Diem, L. A., Metchock, B., Crisp, T., and McElroy, P. D. (2007). Transmission network analysis to complement routine tuberculosis contact investigations. *Am J Public Health*, 97(3):470–477.

Angeli, D., Ferrell, J. E., and Sontag, E. D. (2004). Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. *Proc Natl Acad Sci U S A*, 101(7):1822–1827.

Ataullakhanov, F. I., Vitvitsky, V. M., Zhabotinsky, A. M., Pichugin, A. V., Platonova, O. V., Kholodenko, B. N., and Ehrlich, L. I. (1981). The regulation of glycolysis in human erythrocytes. the dependence of the glycolytic flux on the ATP concentration. *Eur J Biochem*, 115(2):359–365.

Bagga, J., Gewali, L., and Glasser, D. (1996). The complexity of illuminating polygons by alpha-flood-lights. In *Proceedings of the 8th Canadian Conference on Computational Geometry*, pages 337–342. Carleton University Press.

Bailey, J. E. (1991). Toward a science of metabolic engineering. *Science*, 252(5013):1668–1675.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A., and Gifford, D. K. (2003). Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, 21(11):1337–1342.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113.

Bartel, T. and Holzhütter, H. G. (1990). Mathematical modelling of the purine metabolism of the rat liver. *Biochim Biophys Acta*, 1035(3):331–339.

Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I., and Schomburg, D. (2007). BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res*, 35(Database issue):D511–D514.

Becker, J., Klopprogge, C., Zelder, O., Heinzle, E., and Wittmann, C. (2005). Amplified expression of fructose 1,6-bisphosphatase in *Corynebacterium glutamicum* increases in vivo flux through the pentose phosphate pathway and lysine production on different carbon sources. *Appl Environ Microbiol*, 71(12):8587–8596.

Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc*, 2(11):2692–2703.

Berg, J., Tymoczko, J., and Stryer, L. (2002). *Biochemistry*. W.H. Freeman and Company, New York, 5 edition.

Bhan, A., Galas, D. J., and Dewey, T. G. (2002). A duplication growth model of gene expression networks. *Bioinformatics*, 18(11):1486–1493.

Bisswanger, H. (2002). *Enzyme Kinetics : Principles and Methods*. Wiley-VCH.

Bölling, C. and Fiehn, O. (2005). Metabolite profiling of *Chlamydomonas reinhardtii* under nutrient deprivation. *Plant Physiol*, 139(4):1995–2005.

Bonarius, H. P., Schmid, G., and Tramper, J. (1997). Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends Biotechnology*, 15:308–314.

Bond, C. J., Jurica, M. S., Mesecar, A., and Stoddard, B. L. (2000). Determinants of allosteric activation of yeast pyruvate kinase and identification of novel effectors using computational screening. *Biochemistry*, 39(50):15333–15343.

Borenstein, E., Kupiec, M., Feldman, M. W., and Ruppin, E. (2008). Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proc Natl Acad Sci U S A*, 105(38):14482–14487.

Buck, M. J. and Lieb, J. D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360.

Bulik, S., Grimbs, S., Huthmacher, C., Selbig, J., and Holzhütter, H. G. (2009a). Kinetic hybrid models composed of mechanistic and simplified enzymatic rate laws - a promising method for speeding up the kinetic modelling of complex metabolic networks. *FEBS Journal*, 276:410–424.

Bulik, S., Grimbs, S., Huthmacher, C., Selbig, J., and Holzhütter, H. G. (2009b). *Supporting information*. `http://www3.interscience.wiley.com/journal/121588609/suppinfo`.

Burchhardt, G. and Ingram, L. O. (1992). Conversion of xylan to ethanol by ethanologenic strains of *Escherichia coli* and *Klebsiella oxytoca*. *Appl Environ Microbiol*, 58(4):1128–1133.

Burton, S. G., Cowan, D. A., and Woodley, J. M. (2002). The search for the ideal biocatalyst. *Nat Biotechnol*, 20(1):37–45.

Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., Miller-Graziano, C., Moldawer, L. L., Mindrinos, M. N., Davis, R. W., Tompkins, R. G., Lowry, S. F., and Inflamm and Host Response to Injury Large Scale Collab. Res. Program (2005). A network-based analysis of systemic inflammation in humans. *Nature*, 437(7061):1032–1037.

Caspi, R., Foerster, H., Fulcher, C. A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S. Y., Tissier, C., Zhang, P., and Karp, P. D. (2006). MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*, 34(Database issue):D511–D516.

Chang, A., Scheer, M., Grote, A., Schomburg, I., and Schomburg, D. (2009). BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res*, 37:588–592.

Chien, C. T., Bartel, P. L., Sternglanz, R., and Fields, S. (1991). The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci U S A*, 88(21):9578–9582.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *N Engl J Med*, 357(4):370–379.

Conradi, C., Flockerzi, D., Raisch, J., and Stelling, J. (2007). Subnetwork analysis reveals dynamic features of complex (bio)chemical networks. *Proc Natl Acad Sci U S A*, 104(49):19175–19180.

Craciun, G. and Feinberg, M. (2005). Multiple equilibria in complex chemical reaction networks I: The injectivity property. *SIAM J. Appl. Math.*, 65(5):1526–1546.

Craciun, G. and Feinberg, M. (2006a). Multiple equilibria in complex chemical reaction networks: extensions to entrapped species models. *Syst Biol (Stevenage)*, 153(4):179–186.

Craciun, G. and Feinberg, M. (2006b). Multiple equilibria in complex chemical reaction networks II: The species-reaction graph. *SIAM J. Appl. Math.*, 66(4):1321–1338.

Craciun, G., Tang, Y., and Feinberg, M. (2006). Understanding bistability in complex enzyme-driven reaction networks. *Proc Natl Acad Sci U S A*, 103(23):8697–8702.

Dariyerli, N., Toplan, S., Akyolcu, M. C., Hatemi, H., and Yigit, G. (2004). Erythrocyte osmotic fragility and oxidative stress in experimental hypothyroidism. *Endocrine*, 25(1):1–5.

de Atauri, P., Ramírez, M. J., Kuchel, P. W., Carreras, J., and Cascante, M. (2006). Metabolic homeostasis in the human erythrocyte: *in silico* analysis. *Biosystems*, 83(2-3):118–124.

de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574.

Delgado, J. and Liao, J. C. (1992). Determination of flux control coefficients from transient metabolite concentrations. *Biochem J*, 282 (Pt 3):919–927.

Delgado, J., Meruane, J., and Liao, J. C. (1993). Experimental determination of flux control distribution in biochemical systems: In vitro model to analyze transient metabolite concentrations. *Biotechnol Bioeng*, 41(11):1121–1128.

Deutscher, D., Meilijson, I., Kupiec, M., and Ruppin, E. (2006). Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet*, 38(9):993–998.

Deville, Y., Gilbert, D., van Helden, J., and Wodak, S. J. (2003). An overview of data models for the analysis of biochemical pathways. *Brief Bioinform*, 4(3):246–259.

du Preez, F. B., Conradie, R., Penkler, G. P., Holm, K., van Dooren, F. L. J., and Snoep, J. L. (2008). A comparative analysis of kinetic models of erythrocyte glycolysis. *J Theor Biol*, 252(3):488–496.

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(6):1777–1782.

Duarte, N. C., Herrgård, M. J., and Palsson, B. Ø. (2004). Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res*, 14(7):1298–1309.

Duffy, D. A. (1991). *Principles of automated theorem proving*. John Wiley & Sons.

Ebenhöh, O., Handorf, T., and Heinrich, R. (2004). Structural analysis of expanding metabolic networks. *Genome Inform*, 15(1):35–45.

Ebenhöh, O., Handorf, T., and Heinrich, R. (2005). A cross species comparison of metabolic network functions. *Genome Inform*, 16(1):203–213.

Ebenhöh, O., Handorf, T., and Kahn, D. (2006). Evolutionary changes of metabolic networks and their biosynthetic capacities. *Syst Biol (Stevenage)*, 153(5):354–358.

Ebenhöh, O. and Liebermeister, W. (2006). Structural analysis of expressed metabolic subnetworks. *Genome Inform*, 17(1):163–172.

Edmonds, J. and Johnson, E. L. (1970). Matching: A well-solved class of integer linear programs. In Haim Hanani, Norbert Sauer, J. S., editor, *Combinatorial Structures and Their Applications. Proceedings of the Calgary International Conference on Combinatorial Structures and Their Applications*, pages 89–92. Gordon and Breach.

Edwards, J. S., Ibarra, R. U., and Palsson, B. Ø. (2001). *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol*, 19(2):125–130.

Edwards, J. S. and Palsson, B. Ø. (2000a). The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A*, 97(10):5528–5533.

Edwards, J. S. and Palsson, B. Ø. (2000b). Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnol Prog*, 16(6):927–939.

Ellison, P. R. (1998). *The Advanced Deficiency Algorithm and Its Applications to Mechanism Discrimination*. PhD thesis, University of Rochester, Rochester, New York.

Feinberg, M. (1987). Chemical reaction network structure and the stability of complex isothermal reactors - I. The deficiency zero and deficiency one theorems. *Chem Eng Sci*, 42:2229–2268.

Feinberg, M. (1988). Chemical reaction network structure and the stability of complex isothermal reactors - II. Multiple steady states for networks of deficiency one. *Chem Eng Sci*, 43:1–25.

Feinberg, M. (1995a). The existence and uniqueness of steady states for a class of chemical reaction networks. *Arch Rational Mech Anal*, 132:311–370.

Feinberg, M. (1995b). Multiple steady states for chemical reaction networks of deficiency one. *Arch Rational Mech Anal*, 132:371–406.

Feinberg, M. and Ellison, P. (2000). The chemical reaction network toolbox (CRNT). *http://www.che.eng.ohio-state.edu/ feinberg/crnt/*, version 1.1.

Fell, D. (1996). *Understanding the Control of Metabolism*. Portland Press, London.

Fernau, H. and Manlove, D. F. (2006). Vertex and edge covers with clustering properties: Complexity and algorithms. In *Proceedings of ACiD 2006: The 2nd Algorithms and Complexity in Durham Workshop, Texts in Algorithmics*, volume 7, pages 69–84. College Publications.

Fernie, A. R., Trethewey, R. N., Krotzky, A. J., and Willmitzer, L. (2004). Metabolite profiling: from diagnostics to systems biology. *Nat Rev Mol Cell Biol*, 5(9):763–769.

Ferrante, A., Pandurangan, G., and Park, K. (2008). On the hardness of optimization in power-law graphs. *Theor Comp Sci*, 393:220–230.

Flügge, U. I., Freisl, M., and Heldt, H. W. (1980). Balance between metabolite accumulation and transport in relation to photosynthesis by isolated spinach chloroplasts. *Plant Physiol*, 65(4):574–577.

Forsythe, R. G., Karp, P. D., and Mavrovouniotis, M. L. (1997). Estimation of equilibrium constants using automated group contribution methods. *Comput Appl Biosci*, 13(5):537–543.

Gabdoulline, R. R., Stein, M., and Wade, R. C. (2007). qPIPSA: relating enzymatic kinetic parameters and interaction fields. *BMC Bioinformatics*, 8:373.

Gallai, T. (1959). Über extreme Punkt- und Kantenmengen. *Annales Univ. Sci. Budapest*, 2:133–138.

Gatermann, K. and Wolfrum, M. (2005). Bernstein's second theorem and Viro's method for sparse polynomial systems in chemistry. *Adv Appl Math*, 34:252–294.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003). A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–1736.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–7826.

Goldberg, R. N. (1999). Thermodynamics of enzyme-catalyzed reactions: part 6 - 1999 update. *J Phys Chem Ref Data*, 28:931–965.

Gombert, A. K. and Nielsen, J. (2000). Mathematical modelling of metabolism. *Curr Opin Biotechnol*, 11(2):180–186.

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., and Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol*, 22(5):245–252.

Goto, S., Okuno, Y., Hattori, M., Nishioka, T., and Kanehisa, M. (2002). LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res*, 30(1):402–404.

Green, M. L. and Karp, P. D. (2004). A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5:76.

Grimbs, S., Selbig, J., Bulik, S., Holzhütter, H.-G., and Steuer, R. (2007a). The stability and robustness of metabolic states: identifying stabilizing sites in metabolic networks. *Mol Syst Biol*, 3:146.

Grimbs, S., Selbig, J., Bulik, S., Holzhütter, H.-G., and Steuer, R. (2007b). *Supplementary information*. `http://www.nature.com/msb/journal/v3/n1/suppinfo/msb4100186_S1.html`.

Guberman, J. M. (2003). Mass action reaction networks and the deficiency zero theorem. Master's thesis, Harvard University, Cambridge, MA.

Guimerà, R. and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900.

Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2007a). Classes of complex networks defined by role-to-role connectivity profiles. *Nat Phys*, 3(1):63–69.

Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2007b). A network-based method for target selection in metabolic networks. *Bioinformatics*.

Gunawardena, J. (2003). Chemical reaction network theory for *in-silico* biologists. *http://www.jeremy-gunawardena.com/papers/crnt.pdf*.

Hakes, L., Pinney, J. W., Robertson, D. L., and Lovell, S. C. (2008). Protein-protein interaction networks and biology–what's the connection? *Nat Biotechnol*, 26(1):69–72.

Handorf, T., Christian, N., Ebenhöh, O., and Kahn, D. (2008). An environmental perspective on metabolism. *J Theor Biol*, 252(3):530–537.

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

Handorf, T., Ebenhöh, O., Kahn, D., and Heinrich, R. (2006). Hierarchy of metabolic compounds based on their synthesising capacity. *Syst Biol (Stevenage)*, 153(5):359–363.

Hardie, D. G. (2003). Minireview: the AMP-activated protein kinase cascade: the key sensor of cellular energy status. *Endocrinology*, 144(12):5179–5183.

Heiner, M. and Koch, I. (2004). Petri net based model validation in systems biology. *Lecture Notes in Computer Science*, 3099:216–237.

Heinrich, R. and Rapoport, T. A. (1974). A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur J Biochem*, 42(1):89–95.

Heinrich, R. and Schuster, S. (1996). *The regulation of cellular system*. Chapman & Hall, New York.

Heldt, H. W., Chon, C. J., and Maronde, D. (1977). Role of orthophosphate and other factors in the regulation of starch formation in leaves and isolated chloroplasts. *Plant Physiol*, 59(6):1146–1155.

Henry, C. S., Jankowski, M. D., Broadbelt, L. J., and Hatzimanikatis, V. (2006). Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. *Biophys J*, 90(4):1453–1461.

Heymans, M. and Singh, A. K. (2003). Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19 Suppl 1:i138–i146.

Holzhütter, H.-G. (2004). The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur J Biochem*, 271(14):2905–2922.

Holzhütter, H.-G., Jacobasch, G., and Bisdorff, A. (1985). Mathematical modelling of metabolic pathways affected by an enzyme deficiency. A mathematical model of glycolysis in normal and pyruvate-kinase-deficient red blood cells. *Eur J Biochem*, 149(1):101–111.

Hood, L., Heath, J. R., Phelps, M. E., and Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. *Science*, 306(5696):640–643.

Hoppe, A., Hoffmann, S., and Holzhütter, H.-G. (2007). Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Syst Biol*, 1:23.

Horn, F. and Jackson, R. (1972). General mass action kinetics. *Arch. Rational Mech. Anal.*, B47:81–116.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Novre, N. L., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J., and Forum, S. B. M. L. (2003). The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, 23(2):254–267.

Hwang, D., Rust, A. G., Ramsey, S., Smith, J. J., Leslie, D. M., Weston, A. D., de Atauri, P., Aitchison, J. D., Hood, L., Siegel, A. F., and Bolouri, H. (2005). A data integration methodology for systems biology. *Proc Natl Acad Sci U S A*, 102(48):17296–17301.

Jacobasch, G. and Rapoport, S. M. (1996). Hemolytic anemias due to erythrocyte enzyme deficiencies. *Mol Aspects Med*, 17(2):143–170.

Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.

Jonsson, P. (1998). Near-optimal nonapproximability results for some NPO PB-complete problems. *Information Processing Letters*, 68(5):249–253.

Joshi, A. and Palsson, B. Ø. (1989). Metabolic dynamics in the human red cell. part I–A comprehensive kinetic model. *J Theor Biol*, 141(4):515–528.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, 33(Database issue):D428–D432.

Jurica, M. S., Mesecar, A., Heath, P. J., Shi, W., Nowak, T., and Stoddard, B. L. (1998). The allosteric regulation of pyruvate kinase by fructose-1,6-bisphosphate. *Structure*, 6(2):195–210.

Kacser, H. and Burns, J. (1973). The control of flux. *Symp Soc Exp Biol*, 27:65 – 104.

Kahlem, P. and Birney, E. (2006). Dry work in a wet world: computation in systems biology. *Mol Syst Biol*, 2:40.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue):D277–D280.

Karp, R. M. (1972). Reducibility among combinatorial problems. In Miller, R. E. and Thatcher, J. W., editors, *Complexity of computer computations*, Proceedings of a Symposium in the Complexity of Computer Computations, pages 85–103. Plenum Press, New York and London.

Kell, D. B. (2004). Metabolomics and systems biology: making sense of the soup. *Curr Opin Microbiol*, 7(3):296–307.

Khanin, R. and Wit, E. (2006). How scale-free are biological networks. *J Comput Biol*, 13(3):810–818.

Kirkpatrick, D. G. and Hell, P. (1978). On the completeness of a generalized matching problem. In *Proceedings of the tenth annual ACM symposium on Theory of computing*, Annual ACM Symposium on Theory of Computing, pages 240–245. ACM, New York.

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

Kitano, H. (2002b). Systems biology: a brief overview. *Science*, 295(5560):1662–1664.

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

Klamt, S. and Stelling, J. (2003). Two approaches for metabolic pathway analysis? *Trends Biotechnol*, 21(2):64–69.

Klipp, E., Herwig, R., Kowald, A., Wierling, C., and Lehrach, H. (2005). *Systems Biology in Practice: Concepts, Implementation and Application*. Wiley-VCH.

Kodícek, M. (1986). Enhanced glucose consumption in erythrocytes under mechanical stress. *Cell Biochem Funct*, 4(2):153–155.

Kompala, D. S., Ramkrishna, D., and Tsao, G. T. (1984). Cybernetic modeling of microbial growth on multiple substrates. *Biotechnology and Bioengineering*, 26(11):1272–1281.

Kotera, M., McDonald, A. G., Boyce, S., and Tipton, K. F. (2008). Functional group and substructure searching as a tool in metabolomics. *PLoS ONE*, 3(2):e1537.

Kresnowati, M. T. A. P., van Winden, W. A., and Heijnen, J. J. (2005). Determination of elasticities, concentration and flux control coefficients from transient metabolite data using linlog kinetics. *Metab Eng*, 7(2):142–153.

Kumar, P. and Shoukri, M. (2007). Copula based prediction models: an application to an aortic regurgitation study. *BMC Med Res Methodol*, 7(1):21.

Kümmel, A., Panke, S., and Heinemann, M. (2006). Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol Syst Biol*, 2:2006.0034.

Larhlimi, A. and Bockmayr, A. (2008). A new constraint-based description of the steady-state flux cone of metabolic networks. *Discrete Applied Mathematics*, in press.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 14(6):1085–1094.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804.

Li, H.-Y. (1997). The determination of multiple steady-states for a family of catalytic reactions in an isothermal CFSTR. *Chem. Eng. Technol.*, 20:212–219.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J.-F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Heuvel, S. V. D., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science*, 303(5657):540–543.

Liebermeister, W. and Klipp, E. (2006). Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor Biol Med Model*, 3:41.

Llaneras, F. and Picó, J. (2008). Stoichiometric modelling of cell metabolism. *J Biosci Bioeng*, 105(1):1–11.

Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., and Goryanin, I. (2007). The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*, 3:135.

Ma, H. and Zeng, A.-P. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277.

May, P., Wienkoop, S., Kempa, S., Usadel, B., Christian, N., Rupprecht, J., Weiss, J., Recuenco-Munoz, L., Ebenhöh, O., Weckwerth, W., and Walther, D. (2008). Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of *Chlamydomonas reinhardtii*. *Genetics*, 179(1):157–166.

McIntyre, L. M., Thorburn, D. R., Bubb, W. A., and Kuchel, P. W. (1989). Comparison of computer simulations of the F-type and L-type non-oxidative hexose monophosphate shunts with ^{31}P-NMR experimental data from human erythrocytes. *Eur J Biochem*, 180(2):399–420.

McMillan, D. C., Bolchoz, L. J., and Jollow, D. J. (2001). Favism: effect of divicine on rat erythrocyte sulfhydryl status, hexose monophosphate shunt activity, morphology, and membrane skeletal proteins. *Toxicol Sci*, 62(2):353–359.

Monnot, J. and Toulouse, S. (2007). The P_k partition problem and related problems in bipartite graphs. *Lecture Notes in Computer Science*, 4362:422–433.

Morohashi, M., Winn, A. E., Borisuk, M. T., Bolouri, H., Doyle, J., and Kitano, H. (2002). Robustness as a measure of plausibility in models of biochemical networks. *J Theor Biol*, 216(1):19–30.

Mulquiney, P. J. and Kuchel, P. W. (1999). Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: computer simulation and metabolic control analysis. *Biochem J*, 342 Pt 3:597–604.

Murray, A. W. (1971). The biological significance of purine salvage. *Annu Rev Biochem*, 40:811–826.

Murty, K. G. and Perin, C. (1982). A 1-matching blossom-type algorithm for edge covering problems. *Networks*, 12(4):379–391.

Nakayama, Y., Kinoshita, A., and Tomita, M. (2005). Dynamic simulation of red blood cell metabolism and its application to the analysis of a pathological condition. *Theor Biol Med Model*, 2(1):18.

Ni, T. C. and Savageau, M. A. (1996a). Application of biochemical systems theory to metabolism in human red blood cells. Signal propagation and accuracy of representation. *J Biol Chem*, 271(14):7927–7941.

Ni, T. C. and Savageau, M. A. (1996b). Model assessment and refinement using strategies from biochemical systems theory: application to metabolism in human red blood cells. *J Theor Biol*, 179(4):329–368.

Nicholson, J. K. and Wilson, I. D. (2003). Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nat Rev Drug Discov*, 2(8):668–676.

Nielsen, J. (1997). Metabolic control analysis of biochemical pathways based on a thermokinetic description of reaction rates. *Biochem J*, 321 (Pt 1):133–138.

Nikoloski, Z., Grimbs, S., May, P., and Selbig, J. (2008a). Metabolic networks are NP-hard to reconstruct. *J Theor Biol*, 254(4):807–816.

Nikoloski, Z., Grimbs, S., Selbig, J., and Ebenhöh, O. (2008b). Hardness and approximability of the inverse scope problem. *Lecture Notes in Bioinformatics*, 5251:99–112.

Nissen, T. L., Kielland-Brandt, M. C., Nielsen, J., and Villadsen, J. (2000). Optimization of ethanol production in *Saccharomyces cerevisiae* by metabolic engineering of the ammonium assimilation. *Metab Eng*, 2(1):69–77.

Novère, N. L., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J. L., and Hucka, M. (2006). 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.

Olçer, H., Lloyd, J. C., and Raines, C. A. (2001). Photosynthetic capacity is differentially affected by reductions in sedoheptulose-1,7-bisphosphatase activity during leaf development in transgenic tobacco plants. *Plant Physiol*, 125(2):982–989.

Olivier, B. G. and Snoep, J. L. (2004). Web-based kinetic modelling using JWS Online. *Bioinformatics*, 20(13):2143–2144.

Papadimitriou, C. H. and Yannakakis, M. (1991). Optimization, approximation, and complexity classes. *Journal of Computer and System Sciences*, 43:425–440.

Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S., and Palsson, B. Ø. (2004). Comparison of network-based pathway analysis methods. *Trends Biotechnol*, 22(8):400–405.

Peregrin-Alvarez, J. M., Tsoka, S., and Ouzounis, C. A. (2003). The phylogenetic extent of metabolic enzymes and pathways. *Genome Res*, 13(3):422–427.

Petibois, C. and Deleris, G. (2004). Oxidative stress effects on erythrocytes determined by FT-IR spectrometry. *Analyst*, 129(10):912–916.

Pettersson, G. and Ryde-Pettersson, U. (1988). A mathematical model of the Calvin photosynthesis cycle. *Eur J Biochem*, 175(3):661–672.

Poolman, M. G., Bonde, B. K., Gevorgyan, A., Patel, H. H., and Fell, D. A. (2006). Challenges to be faced in the reconstruction of metabolic networks from public databases. *Syst Biol (Stevenage)*, 153(5):379–384.

Poolman, M. G., Fell, D. A., and Thomas, S. (2000). Modelling photosynthesis and its control. *J Exp Bot*, 51 Spec No:319–328.

Poolman, M. G., Olçer, H., Lloyd, J. C., Raines, C. A., and Fell, D. A. (2001). Computer modelling and experimental evidence for two steady states in the photosynthetic Calvin cycle. *Eur J Biochem*, 268(10):2810–2816.

Price, N. D., Reed, J. L., Papin, J. A., Wiback, S. J., and Palsson, B. Ø. (2003). Network-based analysis of metabolic regulation in the human red blood cell. *J Theor Biol*, 225(2):185–194.

Pulleyblank, W. R. (1996). *Handbook of combinatorics*, volume 1, chapter Matchings and extensions, pages 179–232. MIT Press.

Rapoport, T. A., Heinrich, R., and Rapoport, S. M. (1976). The regulatory principles of glycolysis in erythrocytes *in vivo* and *in vitro*. a minimal comprehensive model describing steady states, quasi-steady states and time-dependent processes. *Biochem J*, 154(2):449–469.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555.

Redestig, H., Weicht, D., Selbig, J., and Hannah, M. A. (2007). Transcription factor target prediction using multiple short expression time series from *Arabidopsis thaliana*. *BMC Bioinformatics*, 8:454.

Reed, J. L., Famili, I., Thiele, I., and Palsson, B. Ø. (2006). Towards multidimensional genome annotation. *Nat Rev Genet*, 7(2):130–141.

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

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., and Karp, P. D. (2005). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol*, 6(1):R2.

Rooney, N., McCann, K., Gellner, G., and Moore, J. C. (2006). Structural asymmetry and the stability of diverse food webs. *Nature*, 442(7100):265–269.

Rothberg, J. M. and Leamon, J. H. (2008). The development and impact of 454 sequencing. *Nat Biotechnol*, 26(10):1117–1124.

Rottenberg, H. (1973). The thermodynamic description of enzyme-catalyzed reactions. The linear relation between the reaction rate and the affinity. *Biophys J*, 13(6):503–511.

Sauer, U. (2004). High-throughput phenomics: experimental methods for mapping fluxomes. *Curr Opin Biotechnol*, 15(1):58–63.

Sauer, U. (2006). Metabolic networks in motion: ^{13}C-based flux analysis. *Mol Syst Biol*, 2:62.

Savageau, M. A. (1969). Biochemical systems analysis I. Some mathematical properties of the rate law for the component enzymatic reactions. *J Theor Biol*, 25(3):365–369.

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(3):229–248.

Schilling, C. H., Schuster, S., Palsson, B. Ø., and Heinrich, R. (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol Prog*, 15(3):296–303.

Schuster, R. and Holzhütter, H. G. (1995). Use of mathematical models for predicting the metabolic effect of large-scale enzyme activity alterations. Application to enzyme deficiencies of red blood cells. *Eur J Biochem*, 229(2):403–418.

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(2):53–60.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat Biotechnol*, 18(12):1257–1261.

Segrè, D., Vitkup, D., and Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A*, 99(23):15112–15117.

Sel'kov, E. E. (1975). Stabilization of energy charge, generation of oscillations and multiple steady states in energy metabolism as a result of purely stoichiometric regulation. *Eur J Biochem*, 59(1):151–157.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet*, 31(1):64–68.

Shlomi, T., Berkman, O., and Ruppin, E. (2005). Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A*, 102(21):7695–7700.

Small, J. R. (1993). Flux control coefficients determined by inhibitor titration: the design and analysis of experiments to minimize errors. *Biochem J*, 296 (Pt 2):423–433.

Smallbone, K., Simeonidis, E., Broomhead, D. S., and Kell, D. B. (2007). Something from nothing: bridging the gap between constraint-based and kinetic modelling. *FEBS J*, 274(21):5576–5585.

Speers, A. E. and Cravatt, B. F. (2004). Profiling enzyme activities in vivo using click chemistry methods. *Chem Biol*, 11(4):535–546.

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

Steuer, R., Gross, T., Selbig, J., and Blasius, B. (2006). Structural kinetic modeling of metabolic networks. *Proc Natl Acad Sci U S A*, 103(32):11868–11873.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18 Suppl 2:S231–S240.

Steuer, R., Kurths, J., Fiehn, O., and Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19(8):1019–1026.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255.

Teichmann, S. A. and Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nat Genet*, 36(5):492–496.

Tekir, S. D., Cakir, T., and Ülgen, K. Ö. (2006). Analysis of enzymopathies in the human red blood cells by constraint-based stoichiometric modeling approaches. *Comput Biol Chem*, 30(5):327–338.

Teusink, B., Passarge, J., Reijenga, C. A., Esgalhado, E., van der Weijden, C. C., Schepper, M., Walsh, M. C., Bakker, B. M., van Dam, K., Westerhoff, H. V., and Snoep, J. L. (2000). Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur J Biochem*, 267(17):5313–5329.

Tsantili, I. C., Karim, M. N., and Klapa, M. I. (2007). Quantifying the metabolic capabilities of engineered *Zymomonas mobilis* using linear programming analysis. *Microb Cell Fact*, 6:8.

Tyson, J. J., Chen, K., and Novak, B. (2001). Network dynamics and cell physiology. *Nat Rev Mol Cell Biol*, 2(12):908–916.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627.

Varma, A. and Palsson, B. Ø. (1994). Metabolic flux balancing : basic concepts, scientific and pratical use. *Nat Biotech*, 12:994 – 998.

Voit, E. O. and Radivoyevitch, T. (2000). Biochemical systems analysis of genome-wide expression data. *Bioinformatics*, 16(11):1023–1037.

von Dassow, G., Meir, E., Munro, E. M., and Odell, G. M. (2000). The segment polarity network is a robust developmental module. *Nature*, 406(6792):188–192.

Wagner, A. and Fell, D. A. (2001). The small world inside large metabolic networks. *Proc Biol Sci*, 268(1478):1803–1810.

Wang, L., Birol, I., and Hatzimanikatis, V. (2004). Metabolic control analysis under uncertainty: framework development and case studies. *Biophys J*, 87(6):3750–3763.

Wang, Z., Zhu, X.-G., Chen, Y., Li, Y., Hou, J., Li, Y., and Liu, L. (2006). Exploring photosynthesis evolution by comparative analysis of metabolic networks between chloroplasts and photosynthetic bacteria. *BMC Genomics*, 7:100.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Weed, R. I., LaCelle, P. L., and Merrill, E. W. (1969). Metabolic dependence of red cell deformability. *J Clin Invest*, 48(5):795–809.

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

Wiechert, W. (2002). Modeling and simulation: tools for metabolic engineering. *J Biotechnol*, 94(1):37–63.

Wiechert, W., Schweissgut, O., Takanaga, H., and Frommer, W. B. (2007). Fluxomics: mass spectrometry versus quantitative imaging. *Curr Opin Plant Biol*, 10(3):323–330.

Wildermuth, M. C. (2000). Metabolic control analysis: biological applications and insights. *Genome Biol*, 1(6):1031.1–1031.5.

Wilkinson, S. J., Benson, N., and Kell, D. B. (2008). Proximate parameter tuning for biochemical networks with uncertain kinetic parameters. *Mol Biosyst*, 4(1):74–97.

Williams, R. J., Berlow, E. L., Dunne, J. A., Barabási, A.-L., and Martinez, N. D. (2002). Two degrees of separation in complex food webs. *Proc Natl Acad Sci U S A*, 99(20):12913–12916.

Wittig, U., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann, A., Anstein, S., Saric, J., and Rojas, I. (2006). SABIO-RK: Integration and curation of reaction kinetics data. *Lecture Notes in Bioinformatics*, 4075:94–103.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N., and Suzek, B. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue):D187–D191.

Zhu, X.-G., Alba, R., and de Sturler, E. (2008). A simple model of the Calvin cycle has only one physiologically feasible steady state under the same external conditions. *Nonl Anal RWA*.

Zhu, X.-G., de Sturler, E., and Long, S. P. (2007). Optimizing the distribution of resources between enzymes of carbon metabolism can dramatically increase photosynthetic rate: a numerical simulation using an evolutionary algorithm. *Plant Physiol*, 145(2):513–526.

Zwolak, J. W., Tyson, J. J., and Watson, L. T. (2004). Finding all steady state solutions of chemical kinetic models. *Nonl Analysis R W Appl*, 5:801–814.