Institut für Biochemie und Biologie

Universität Potsdam

Mass-balanced randomization

A significance measure for metabolic networks

Dissertation

zur Erlangung des akademischen Grades
”doctor rerum naturalium” (Dr. rer. nat.)
in der Wissenschaftsdisziplin ”Bioinformatik”

eingereicht in kumulativer Form an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von

Georg Basler

Potsdam, den 26. Januar 2012

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2012/6203/
URN urn:nbn:de:kobv:517-opus-62037
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-62037

Abstract

Complex networks have been successfully employed to represent different levels of biological systems, ranging from gene regulation to protein-protein interactions and metabolism. Network-based research has mainly focused on identifying unifying structural properties, including small average path length, large clustering coefficient, heavy-tail degree distribution, and hierarchical organization, viewed as requirements for efficient and robust system architectures. Existing studies estimate the significance of network properties using a generic randomization scheme—a Markov-chain switching algorithm—which generates unrealistic reactions in metabolic networks, as it does not account for the physical principles underlying metabolism. Therefore, it is unclear whether the properties identified with this generic approach are related to the functions of metabolic networks.

Within this doctoral thesis, I have developed an algorithm for mass-balanced randomization of metabolic networks, which runs in polynomial time and samples networks almost uniformly at random. The properties of biological systems result from two fundamental origins: ubiquitous physical principles and a complex history of evolutionary pressure. The latter determines the cellular functions and abilities required for an organism's survival. Consequently, the functionally important properties of biological systems result from evolutionary pressure.

By employing randomization under physical constraints, the salient structural properties, i.e., the small-world property, degree distributions, and biosynthetic capabilities of six metabolic networks from all kingdoms of life are shown to be independent of physical constraints, and thus likely to be related to evolution and functional organization of metabolism. This stands in stark contrast to the results obtained from the commonly applied switching algorithm. In addition, a novel network property is devised to quantify the importance of reactions by simulating the impact of their knockout. The relevance of the identified reactions is verified by the findings of existing experimental studies demonstrating the severity of the respective knockouts. The results suggest that the novel property may be used to determine the reactions important for viability of organisms.

Next, the algorithm is employed to analyze the dependence between mass balance and thermodynamic properties of *Escherichia coli* metabolism. The thermodynamic landscape in the vicinity of the metabolic network reveals two regimes of randomized networks: those with thermodynamically favorable reactions, similar to the original network, and those with less favorable reactions. The results suggest that there is an intrinsic dependency between thermodynamic favorability and evolutionary optimization.

The method is further extended to optimizing metabolic pathways by introducing novel chemically feasibly reactions. The results suggest that, in three organisms of biotechnological importance, introduction of the identified reactions may allow for optimizing their growth. The approach is general and allows identifying chemical reactions which modulate the performance with respect to any given objective function, such as the production of valuable compounds or the targeted suppression of pathway activity. These theoretical developments can find applications in metabolic engineering or disease treatment.

The developed randomization method proposes a novel approach to measuring the significance of biological network properties, and establishes a connection between large-scale approaches and biological function. The results may provide important insights into the functional principles of metabolic networks, and open up new possibilities for their engineering.

Acknowledgement

I would like to deeply thank Joachim Selbig for extensive advice and help throughout the development of this thesis. I further greatly thank Zoran Nikoloski for extraordinary scientific and moral support at all times. I thank Oliver Ebenhöh for valuable support and discussions, particularly during the early stages.

I especially thank Sergio Grimbs for his explanations and fruitful discussions on many scientific and non-scientific topics. And I would like to thank the systems biology and bioinformatics working groups at the Max Planck Institute for Molecular Plant Physiology for a wonderful working atmosphere, particularly: Jan Lisec, Jost Neigenfind, Xiaoliang Sun, Nils Christian, Alexander Skupin, Sebastian Klie, Abdelhalim Larhlimi, Xiaoqing Li, Jan Hummel, Patrick May, Dirk Walther, Kathrin Jürchott, Marco Ende, Peter Krüger, and all the others.

Finally, I would like to deeply thank Azucena González Blanco for her care and support in difficult situations, my parents for their never-ending encouragement and generosity, and Amanda for her giggling and patience.

Contents

Chapter 1

Introduction

1.1 Background

The emerging field of systems biology aims at systematically analyzing and modeling biological systems at various levels of organization, from molecules and cells to tissues, organs, and entire organisms. The field is characterized by multidisciplinary approaches and highly depends on modern technologies facilitating high-throughput experiments. Computational methods play a central role, as they are fundamental in virtually all aspects of systems biology research, and enable the automated analysis and generation of hypotheses from an otherwise unmanageable plethora of experimental data.

The knowledge about complex biological systems is typically represented as networks of interacting components. In recent years, the accumulation of high-throughput data has led to increasingly accurate reconstructions of cellular networks. A central challenge in computational systems biology is the development of algorithmic methods facilitating predictions with biotechnological or clinical applications based on formal representations of biological systems.

1.1.1 Computational methods in biology

Classical applications of computational methods in biology were concerned with modeling metabolic processes (Chance et al., 1960) and automated assembly of DNA sequences (Staden, 1979). Their importance for addressing biological questions has ever since continuously increased. The development of high-throughput technologies for DNA sequencing, gene expression measurement, and metabolic profiling, has led to the accumulation of large amounts of data describing all levels of biological systems (Ideker et al., 2001). Nowadays, computational approaches are indispensable tools for establishing a connection between the plethora of experimental data and the functional principles of biological systems, with the aim of automatically generating biological knowledge.

Not only the amount of generated data, but also the inherent data complexity has drastically increased. Starting with the genome, which may be regarded as one-dimensional code, composed of four different nucleotides, data describing various levels of biological processes has been collected. Some examples include: the regulation of gene expression by transcription factors and micro-RNA, signal transduction and the formation of protein complexes, or the catalysis of metabolic reactions by enzymes. In contrast to the genetic code, all of these processes involve a

multitude of different molecular entities which perform their biological function through a complex network of mutual interactions. Consequently, our ability to harness the wealth of information contained in high-throughput experimental data depends ever more on the development of structured modeling approaches and efficient computational tools.

In the past decades, the need to view biological processes in the context of their complex interactions, instead of regarding molecular components independently, was increasingly recognized. Even biological processes which are often treated as independent, such as gene expression, signal transduction and metabolism, are closely interrelated (Gianchandani et al., 2006). Consequently, biological research has moved from reductionist to holistic approaches, which led to the reconstruction and analysis of large networks underlying diverse cellular processes (reviewed in Papin et al., 2005; Krogan et al., 2006; Karlebach and Shamir, 2008; Palsson, 2009). At the same time, computational power has rapidly increased, and more sophisticated algorithms for *in silico* investigations of biological processes have been developed. Thus, the promises of the emerging field of computational systems biology lie in obtaining predictive computational models of complex biological processes, which will ultimately lead to a deeper mechanistic understanding of biological systems and applications in biotechnology and medicine.

1.1.2 From data to knowledge

In its raw form, data from a single high-throughput experiment consists of thousands of data points. In addition, the publicly available knowledge of biological processes, such as gene regulation and metabolism, easily spans several thousands of experiments. Clearly, harnessing the information from these immense data sets requires that they are first brought into a structured and computationally accessible form.

Graphs provide both an intuitive and formal representation of a complex process of interacting components. The components, including genes, proteins, and metabolites, are represented as vertices, and their interactions are represented as edges (see Section 1.2.2). (In the following, the terms graph and network are used synonymously.) The representation of biological processes as graphs has several advantages. First, heterogeneous data types from several different experiments together with their relationships can be transparently represented in a unified fashion, as vertices and edges. Second, different levels of detail may be integrated coherently, which can be used to combine precise information, where available, with higher levels of abstraction, where details are lacking. For instance, in the same graph, one vertex may be annotated by molecular information of atomic detail, while another one may represent an abstraction of an entire biological process. Likewise, edges may be annotated by the type and weighted by the strength of the represented interaction. Finally, a large variety of established and efficient computational methods from other fields of research on networks exists, including graph theory, sociometry, statistical mechanics, or systems engineering, which may be useful in analyzing the functional principles of biological systems (Alon, 2003; Barabási and Oltvai, 2004).

Naturally, the abstraction of knowledge by means of graphs also comes with considerable drawbacks, complicating the potential applications in biology. For example, it is tempting to analyze large biological systems while neglecting levels of detail. However, the omission of necessary details may lead to erroneous results or limit the applicability of their findings, which must be considered when drawing conclusions. Further, it is not easy to represent time-variant, quantitative data in a graph. Therefore, one often cannot directly infer the dynamic behavior of physiological processes from a purely structural graph representation of a biological system. Finally, in order to be useful, any network-based analysis must find a trade-off between detailed molecular

representations and computational complexity. Thus, the challenge remains of developing efficient methods for network analysis to draw a connection between the structure and function of biological systems.

1.1.3 Complex networks

The first attempts to model complex processes by means of networks were made in sociometric studies, where a network represented social friendship relationships among a group of people (Moreno and Jennings, 1937). However, these studies were limited to very small networks, as the data were obtained from personal inquiries, while the large-scale structure of complex social or other networks remained unknown. Around the turn of the millennium, large amounts of data became available to describe complex systems ranging from technology to sociology and natural sciences (Newman, 2001; Liljeros et al., 2001; Albert et al., 1999; Pandey and Mann, 2000; Ito et al., 2001), which allowed the first computational analyses of the large-scale properties of complex systems.

The initial findings pointed out that a network of film actors, a power grid, and the neural network of *C. elegans* were dissimilar to both regular and random graphs. Despite their large sizes, the average length of the shortest paths connecting any two vertices was found to be relatively small, while most vertices were grouped into local clusters, which is known as the small-world phenomenon. In the following, the small-world property was found in different social, technological, metabolic, protein-protein interaction and brain networks (Wagner and Fell, 2001; Giot et al., 2003; Sporns and Zwi, 2004), indicating that different types of complex networks share some fundamental structural properties. Further, in many social, technological, and biological networks, the degree distributions are scale-free (follow a power law) (Barabási and Albert, 1999; Jeong et al., 2000, 2001), and vertices are organized hierarchically (Ravasz et al., 2002; Girvan and Newman, 2002). These studies aimed at discovering the universal principles of complex networks, and suggested that general frameworks may be developed to draw connections between the structure and function of networks across different research disciplines (Barabási and Oltvai, 2004).

However, subsequent studies revealed a clearer picture of the commonalities of and differences between the various types of networks. For example, the average path lengths and the scaling coefficients of degree distributions differ among social, technological and biological networks (Newman, 2003b), and some degree distributions are better described by a truncated power-law or an exponential distribution (Amaral et al., 2000). For metabolic networks, the small-world property crucially depends on the type of network representation and the corresponding definition of a path (Arita, 2004; Pitkänen et al., 2005, see Section 1.4.4). In most social networks, the degrees between neighboring vertices are positively correlated (assortativity), while negative degree-degree correlations prevail in technological and biological networks (dissortativity) (Maslov and Sneppen, 2002; Newman, 2003a). Further, different frequencies in the occurrence of motifs, small subnetworks with a particular connection pattern, allow to distinguish different types of biological and technological networks (Milo et al., 2002).

Particularly in biology, the impact of identifying the global properties of complex networks on experimental research was long and is still debated, mostly due to the difficulties in using structural properties for making predictions about the behavior of individual network components, such as genes or proteins, under *in vivo* conditions (Wolf et al., 2002; Kitano, 2002; Bray, 2003; Arita, 2004; Papp et al., 2009; Yamada and Bork, 2009; Lima-Mendez and van Helden, 2009). Clearly, it seems unlikely that a unified theory will eventually allow elucidating the underlying mechanisms of all types of biological processes. Instead, these analyses, aimed at developing a unifying

theory, can identify the emergent properties of a system, such as robustness or adaptability, which may give hints at their evolutionary history and facilitate classifications (e.g., of healthy and diseased states of cells or organisms). Finally, the central findings in complex network analyses strengthened the holistic view on biological systems, and inspired the development of more sophisticated methods for their analysis. Today, several successful applications of network analyses exist, ranging from the prediction of physiological metabolic states and metabolic engineering of microorganisms (Edwards et al., 2001; Famili et al., 2003; Almaas et al., 2004; Smid et al., 2005; Lee et al., 2005, 2007; Sohn et al., 2010) to the successful classification of diseases using protein-interaction networks (Chuang et al., 2007) and the prediction of drug targets using human metabolic (Folger et al., 2011) and functional brain networks (Sanz-Arigita et al., 2010).

1.1.4 Significance testing

A first step in establishing a connection between the large-scale structure and the function of biological systems is to identify the network properties which carry meaningful biological information. To this end, it is necessary to determine the importance, or significance, of a proposed network property with respect to biological function. Randomization is a classical statistical approach for determining the functionally important features in complex data sets (Fisher, 1925; Rubin, 1978; Lipman et al., 1984; Pearson and Lipman, 1988). The general idea of randomization is to determine the significance of an observation by estimating how likely the same observation could have been made by chance, i.e., by assuming that no functionally relevant principles are reflected in the data. For example, in order to identify co-regulated genes from complex gene expression data, a threshold has to be specified for detecting pairs of genes which are significantly co-regulated. By randomly reshuffling the underlying data points, one can obtain the probability that a set of genes has a similar expression pattern by chance. The threshold is then chosen such that the probability of identifying sets of genes which have a similar expression pattern by chance, i.e., without any functional importance, is reasonably small (lower than the *a priori* chosen significance level of usually 1 or 5%).

Randomization enables the extraction of the meaningful patterns from complex data sets by testing the null hypothesis that a pattern is observed by chance. If an observed pattern is significantly different in randomized data, then the null hypothesis can be rejected, and the pattern is assumed to be of functional importance. Unfortunately, it is not straightforward to devise an appropriate random background for structured data represented as networks. As complex networks are reconstructed from a large diversity of experiments, it is not obvious how to obtain randomized networks which lack any functionally important patterns. Consequently, complex networks have been compared to different types of random graphs, with differing results (Albert and Barabási, 2002). Nevertheless, virtually all network analyses rely on the generic Markov-chain switching algorithm for generating randomized networks (Maslov and Sneppen, 2002; Milo et al., 2002; Itzkovitz et al., 2003; Maslov et al., 2004; Milo et al., 2004; Nunes Amaral and Guimerà, 2006; Guimerà et al., 2007a,b; Marr et al., 2007; Sales-Pardo et al., 2007; Zhu et al., 2007; de la Fuente et al., 2008). Based on this algorithm, a random network is obtained by randomly reshuffling the edges of the original network while preserving the degrees of the vertices. The property of interest is then determined in the original network, and its significance is calculated by comparison to the distribution of values for the property in a large set of networks obtained from randomization.

The idea of preserving the degrees originates from the observation that the degree distributions are a ubiquitous feature, which constrains all classes of networks independently of their function (a different motivation was raised for ecological networks, see Cobb and Chen, 2003). By preserving this universal feature, the identified network properties are independent of the degrees,

and, therefore, assumed to be a result of other, functionally important constraints imposed on the network. However, when applied to metabolic networks, the algorithm generates physically impossible chemical reactions, since physical principles, such as mass balance and thermodynamics, are disregarded (see Section 2.1 and Figure 5.1 on page 46). In addition, it is unclear whether preservation of the degrees is sufficient for obtaining randomized networks void of biological function, as biological systems are shaped by complex physical and evolutionary constraints. Thus, it is arguable whether switch randomization can be used in identifying the functionally important features of biological networks.

In addition, some statistical tests assume that the values of an analyzed property follow a normal distribution in randomized samples, which is rarely verified in practice. The z-score based p-value relies on this assumption, as it derives the distance from the mean in standard deviations of a normal distribution as test statistic. This test is applied in Chapter 5 for estimating the significance of the small-world property. The corresponding randomized distributions were tested for normality (see Figures C.5-C.8).

A more general limitation of significance testing is that no conclusions can be drawn if the null hypothesis is not rejected, i.e., if a property is not significant. While it is frequently suggested in the literature that a "non-significant" property is a result of chance, or unimportant (e.g. Gionis et al., 2006), this conclusion is not statistically sound. The reason is that, in statistical hypothesis testing, the p-value gives an estimation of the probability, that an observation, D (such as a property in the original network), is made under the assumption that the null hypothesis, H_0, is true: $p \approx P(D|H_0)$. Here, H_0 represents the hypothesis that a network property is not functionally important, but a result of random events. If p is small, then the observation is likely to contradict H_0, and the null hypothesis may be rejected. However, if p is large, H_0 is not necessarily likely to be true, as this only reflects that D does not contradict H_0. Instead, the possible reasons for a large p-value are manifold: the significance threshold, which is usually chosen arbitrarily, may be too small, especially if the p-value only slightly missed the threshold. In addition, there may be errors in the experimental design (here, the reconstructed networks), or the number of random samples may be insufficient (see Nickerson, 2000 for a detailed discussion). More intuitively, a property may be related to an important function, but some randomized samples may exhibit the same property value "by chance", resulting in a large p-value. Therefore, one may not draw any conclusions from a non-significant property.

1.2 Metabolic networks

A metabolic network is a collection of chemical reactions involving molecules (metabolites) within an organism. Most metabolic reactions are catalyzed by enzymes, which are encoded in the genome. Thus, the metabolic capabilities of an organism are determined by its genotype. The biological functions of metabolism are diverse: uptake of nutrients from the environment, excretion of unnecessary or toxic compounds, conversion and harness of energy, and biosynthesis of cellular components. Therefore, metabolism is sometimes regarded as the *molecular phenotype* of a cell (Fiehn et al., 2000).

Research in metabolism has traditionally focused on bottom-up approaches, such as kinetic modeling. Therein, the time-dependent changes in metabolite concentrations are represented by a system of differential equations, which can be used to calculate the steady-state as well as the temporal trajectories of the concentrations within the pathway. Unfortunately, kinetic modeling approaches are limited to relatively small pathways due to their dependence on parameters whose

values are often unknown (Famili et al., 2005; Jamshidi and Palsson, 2008). In addition, statistical approaches, such as network analysis, are of limited application to small subnetworks, as their reliability depends on large amounts of available data. Therefore, I will focus in the following on the structure of genome-scale metabolic networks, which are the object of study of this doctoral thesis.

In Section 1.2.1, I will first describe how metabolic networks are reconstructed. Next, I will summarize different ways of formally representing a metabolic network in Section 1.2.2.

1.2.1 Reconstruction

There are two principally different yet complementary techniques for reconstructing a metabolic network. The more accurate (and more tedious) methods rely on identification of enzymes and metabolites in a biological sample using analytical techniques, such as mass spectrometry (Breitling et al., 2006), followed by the characterization of enzymes using biochemical methods (e.g., Martínez-Blanco et al., 1990), correlation analysis (Arkin et al., 1997), or, more recently, computational function prediction from the protein structure (Hermann et al., 2007). Such approaches can detect with high confidence the enzymes present in a particular sample and the relationships between metabolites. The faster and more widespread approach is to transfer knowledge on metabolic reactions from a well-annotated organism by identifying homologous enzyme-coding genes (Ma and Zeng, 2003a; Francke et al., 2005). Without further refinement, e.g., by using existing knowledge from the literature, this approach results in an incomplete network draft which may not be able to account for known metabolic functions. Ideally, knowledge transfer and computational modeling approaches should be combined with experimental data in order to obtain high confidence models (Ideker et al., 2001; Feist et al., 2009). A workflow for reconstructing a genome-scale metabolic network is shown in Figure 1.1.

A typical genome-scale metabolic network consists of several thousands of reactions and metabolites, and is regarded as a complex network. For example, a recent reconstruction of human metabolism includes 3731 unique reactions and 1469 metabolites (Duarte et al., 2007), while another one contains 2819 reactions and 2691 metabolites (Ma et al., 2007). Both networks include information on the reversibility of reactions and the genes encoding their catalyzing enzymes, while only the first includes information on the subcellular localization of metabolites and reactions (which may account for the larger number of reactions, Ma et al., 2007). A recent reconstruction of *A. thaliana* primary metabolism consists of 1567 unique reactions and 1748 metabolites, and contains information on reversible reactions, gene annotations, and subcellular location of reactions (de Oliveira Dal'Molin et al., 2010). A large number of metabolic network reconstructions for organisms throughout all kingdoms of life is publicly available in database collections such as KEGG (Ogata et al., 1999), Reactome (Joshi-Tope et al., 2005), BioCyc (Caspi et al., 2010), or their integrations (Schellenberger et al., 2010; Kumar et al., 2012). These networks were reconstructed primarily using homology transfer and information from the literature, while experimental validations are presently scarce.

There are some general limitations of the applications of genome-scale metabolic networks. First, as most reactions are included automatically based on knowledge transfer from other organisms, organism-specific metabolites and reactions are more likely to be missing than those which are shared with well-studied organisms (Breitling et al., 2008). Second, experimental discovery and knowledge transfer is not equally possible for all parts of metabolism, so that certain, less studied pathways, e.g., from secondary metabolism, may be incomplete or entirely missing. Third, even reactions included with high confidence do not necessarily occur under physiological conditions

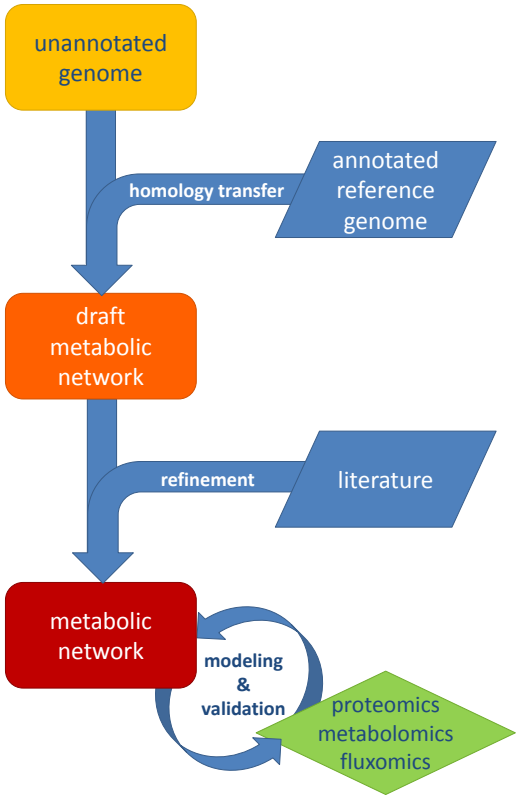


Figure 1.1: Workflow for reconstructing a genome-scale metabolic network. Starting from the unannotated genome for an organism of interest, homologous enzyme-coding genes are identified in a well-annotated reference genome. The identified reactions are transferred in order to obtain a first draft of a metabolic network. In the next step, information from the literature describing the metabolic activities of the target organism is used to add, modify or remove reactions, resulting in a refined network. Finally, the metabolic capacities of the reconstruction are modeled and validated iteratively using experimental techniques, resulting in a high-confidence model.

due to gene regulation, subcellular separation of substrates and enzymes, or differences between the *in vitro* measurements and *in vivo* conditions (Feist et al., 2009; Daily et al., 2007; Teusink et al., 2000). Finally, for multicellular organisms, most genome-scale network reconstructions do not specify in which tissue types individual reactions occur (see Section 7.4). Unless such information is integrated, the metabolic networks of higher organisms cannot be reliably used for physiological predictions on the organism-level (Shlomi et al., 2008).

Some of these problems can be alleviated by the aforementioned manual refinement and experimental validation of metabolic network reconstructions. Therefore, within this thesis, I relied primarily on reconstructed networks from dedicated publications, where particular effort was made in refining the model (summarized in Table B.1).

1.2.2 Graph representations

In the following, I will compare the most common forms of representing metabolic networks, and will present the bipartite graph representation which is used throughout the thesis. The most illustrative way of representing a network of metabolic reactions is to denote the metabolites by their names and connect them by arrows representing reactions (Figure 1.2a). The more detailed reaction diagrams allow to visualize the rearrangements of chemical groups (Figure 1.2b). In both representations, frequently occurring metabolites are repeatedly drawn, which allows for a clear layout. However, their purpose is limited to visualization, as the arbitrary repetition of metabolites does not coherently reflect the number of involved metabolic species, as required for modeling approaches.

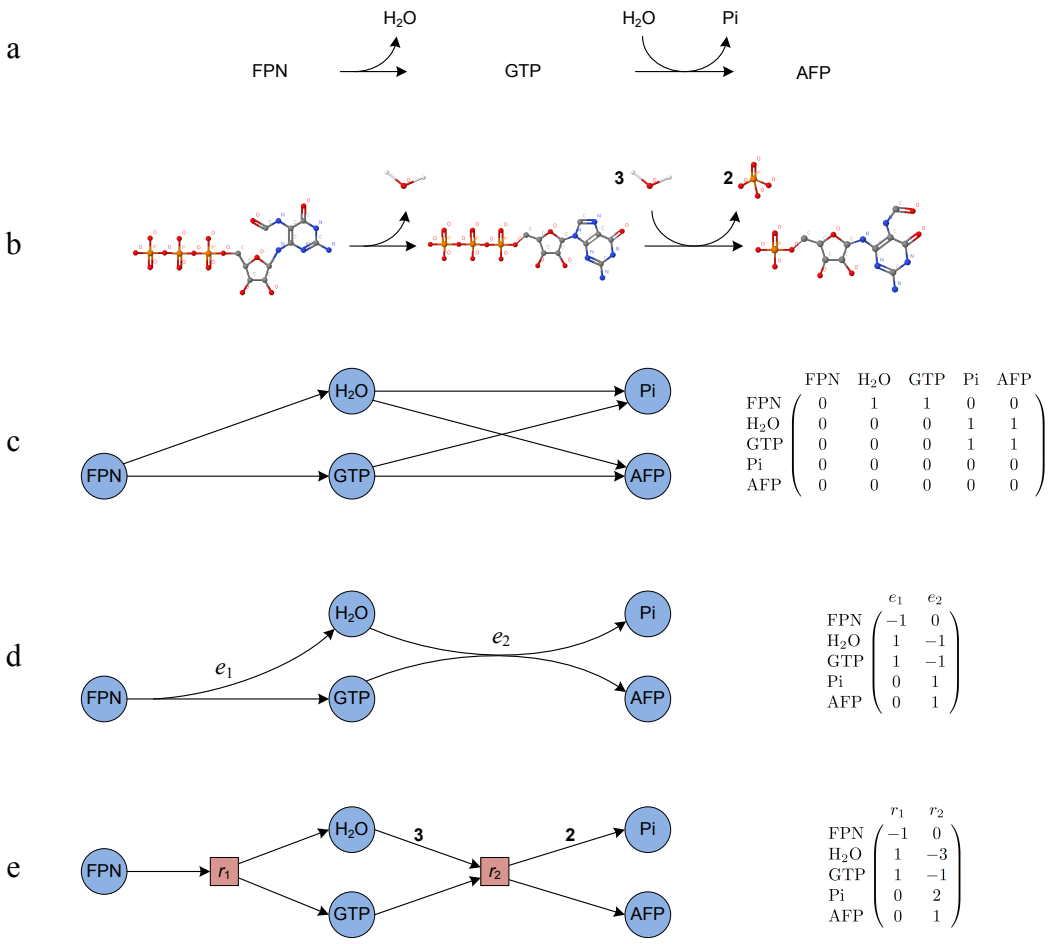


Figure 1.2: Different ways of representing metabolic networks, illustrated by two successive GTP 8,9-hydrolase reactions. (a) Pathway map commonly used in biology. Frequently occurring metabolites, here H_2O, are drawn repeatedly for each reaction. (b) Reaction diagram depicting the chemical structure and stoichiometric coefficients, commonly used in chemistry. The picture was created using Jmol (http://www.jmol.org). (c) Directed metabolite-metabolite network commonly used in complex network analysis (left) and the corresponding adjacency matrix (right). Metabolites are represented by vertices, two vertices are connected by a directed edge if the corresponding metabolites occur as substrate and product in the same reaction. (d) Directed metabolite hypergraph (left) and the corresponding adjacency matrix (right). Metabolites are represented by vertices, a directed hyperedge connects the substrates with the products of a reaction. (e) Directed, weighted bipartite network (left) and the corresponding stoichiometric matrix (right). Metabolites and reactions are represented by two types of vertices. Metabolite vertices are connected to reaction vertices by directed edges, which represent the substrate-reaction or product-reaction relationship, and are weighted by the stoichiometric coefficients. Note that, if a metabolite is allowed to occur as substrate and product of the same reaction, this cannot be represented in the stoichiometric matrix, which marks the difference between stoichiometric and adjacency matrix. Metabolite abbreviations: FPN: Formamidopyrimidine nucleoside triphosphate, GTP: Guanosine 5'-triphosphate, Pi: Phosphate, AFP: 2-Amino-5-formylamino-6-(5-phospho-D-ribosylamino)pyrimidin-4(3H)-one.

In complex network research, metabolic networks are frequently represented as directed metabolite-metabolite graphs. Therein, metabolites are represented by vertices, and reactions are represented by edges connecting the vertices of substrates with those of the products (Figure 1.2c). In this representation, each metabolic species corresponds exactly to one vertex. However, it is not possible to distinguish individual reactions, as a single edge may represent multiple reactions. Therefore, stoichiometric coefficients cannot be considered, and it is not possible to determine biosynthetic routes, where multiple substrates are converted into multiple products (Arita, 2004; Pitkänen et al., 2005).

A directed hypergraph is a generalization of a directed graph, where an edge may connect two sets of vertices (Gallo et al., 1993). Metabolites are represented as vertices, and a hyperedge connects the substrates of a reaction with its products (Figure 1.2d). This representation provides an accurate representation of the biosynthetic routes as hyperpaths. Stoichiometric coefficients of the substrates and products can be represented by assigning functions to hyperedges (Klamt et al., 2009). However, some algorithmic problems, such as average hyperpath length, are computationally prohibitive in large networks, and it is not straightforward to calculate the properties of the corresponding metabolite-metabolite graphs, as commonly analyzed in complex network research.

The directed, weighted bipartite graph used throughout the thesis combines the advantages of metabolite-metabolite graphs and hypergraphs. Metabolites and reactions are represented as two different types of vertices. Vertices representing metabolites are connected by directed edges to the corresponding reaction vertices. The direction of an edge indicates a substrate or product relationship, and its weight represents the stoichiometric coefficient (Figure 1.2e). Thus, the stoichiometric relationships between substrates and products are reflected properly and in a straightforward manner. In addition, this representation facilitates the annotation of both metabolites and reactions with additional information, such as the chemical structure of metabolites, or the enzyme catalyzing a reaction and its coding genes. The properties of the corresponding metabolite-metabolite graph may be calculated simply by neglecting the additional reaction vertices and edge weights. At the same time, this representation may be used to efficiently calculate biosynthetic routes of a network using the network expansion algorithm (Handorf et al., 2005) or metabolite fluxes through reactions using constraint-based approaches (Varma and Palsson, 1994).

1.3 Related work

In the following, I will explain the differences between random graph models and randomization approaches, and give some examples of studies which are related to the subject of this thesis. Random graph models clearly differ from randomization approaches in their formulation and objectives. In contrast to randomization approaches, random graphs are not obtained by modifying a network of interest, but consist of a set of vertices connected by a random set of edges. A classical problem in random graph theory is to study the probability of network properties, such as the frequency of subgraphs and the size of the largest connected component, emerging as a function of the number of vertices (Erdös and Rényi, 1959; Itzkovitz et al., 2003; Dorogovtsev et al., 2008). In complex network research, the structural properties of random graphs were compared to those in social, technological, and biological networks in order to study their evolutionary origins (Watts and Strogatz, 1998; Barabási and Albert, 1999; Newman et al., 2001; Oikonomou and Cluzel, 2006).

In contrast, the goal of network randomization approaches is not to model all aspects of a complex network as closely as possible. Instead, a network is randomized under certain constraints in

order to determine the properties which differ between the original network and its randomized variants. These properties are independent of the imposed constraints, and assumed to be of functional importance. Thus, a randomization algorithm can be considered as a method for estimating the significance of network properties. As discussed in Section 1.1.4, the widespread approach for estimating the significance of complex network properties is based on a Markov-chain switching algorithm (in the following referred to as *switch randomization*), which randomizes a given network while preserving the vertex degrees. Switch randomization generates a randomized network from a given network by repeatedly choosing a pair of edges, $a \rightarrow b, c \rightarrow d$ at random and replacing them by new edges $a \rightarrow d$ and $c \rightarrow b$, if these do not already exist (see Figure C.1). In Chapters 3 and 5, the number of iterations was chosen as the number of unique edge pairs in the network, which is conservative compared to other estimations for obtaining a properly randomized network (Milo et al., 2003). As the degrees of vertices are not changed, all switch randomized networks have the same degree sequence as the original network. By comparison of a property in the original network to the distribution of its values in switch randomized networks, the network properties which are independent of vertex degrees can be determined as significantly different. Since the degree distribution is assumed to be a ubiquitous feature of all networks independently of their function, a significant property is assumed to result from some non-arbitrary, functionally important constraint imposed on the network.

Recently, several enhancements of switch randomization have been proposed. In dense weighted networks, switch randomization may enable substitutions for only few edge pairs. To overcome this limitation, Zlatic et al. (2009) proposed a randomization method applicable to dense weighted networks. The goal was to measure the significance of the rich-club effect, i.e., the tendency of high-degree vertices to form cliques in social networks (Colizza et al., 2006), which are commonly dense. However, the method is of limited applicability to metabolic networks, which are inherently sparse.

Ying and Wu (2008) modified the switch randomization algorithm to minimizing the change in the graph spectra, i.e., the eigenvalues of the adjacency matrix and the corresponding Laplacian matrix. The graph spectra are directly linked to important topological properties of a graph, such as average path length, degree distribution, and hierarchical organization. By deriving the change in eigenvalues resulting from the switching of two edges, those switches which lead to an increase of the eigenvalues are alternated by switches which lead to a decrease. As a result, the graph spectra, and thus some of the fundamental topological properties, are less affected by the randomization procedure. While the original motivation of this randomization algorithm was to anonymize privacy information in social networks on the internet, the authors later extended the method to preserving arbitrary feature distributions, and suggested to use the method for testing the significance of topological properties (Ying and Wu, 2009). However, the algorithms have never been applied to real networks.

A similar approach was suggested by Hanhijärvi et al. (2009a), who developed a general framework for randomizing undirected unweighted graphs while preserving any user-defined set of graph statistics, as exemplified by the average path length and clustering coefficient. Additional operations for modifying graphs are proposed, which preserve the degree distribution and the connected components. While preservation of such properties may also be desirable for randomizing metabolic networks, the computational complexity increases, as the preserved properties are repeatedly calculated in each step of the algorithm. Therefore, when preserving the average path length, the algorithm is inefficient for large networks (Hanhijärvi et al., 2009a). In addition, its extension to directed, weighted biparite graphs is not straightforward.

The most similar approach to the present work is a recently proposed randomization method for

genome-scale metabolic networks (Samal and Martin, 2011). The aim of the approach is to analyze the dependence of topological properties on different functional constraints. The metabolic network of *E. coli* is randomized by exchanging randomly chosen reactions with reactions from the KEGG database. The authors study how the successive addition of constraints to the randomization algorithm affects the average path length, clustering coefficient, reachability of metabolites, and the size of the largest strongly connected component. The successively imposed constraints are: (1) Preservation of the number of reactions; (2) Limitation of the number of metabolites to be less or equal as in the original network; (3) Skipping of reactions which may not carry a flux under steady-state conditions; (4) Viability of the metabolic network under different nutritional environments, as predicted by flux balance analysis (Varma and Palsson, 1994). The authors find that, although a diverse set of randomized networks is obtained even when imposing all constraints ($<$50% of identical reactions), the topological properties become visually similar to the original network, particularly after imposing the constraints (2) and (4).

The approach differs from the present work in various aspects. Samal and Martin (2011) investigate to what extent the topological properties of *E. coli* depend on biological constraints. The authors rely on a metabolite-metabolite graph representation (Figure 1.2c), and remove metabolites with a large degree from the network before conducting their analyses. By randomly exchanging the reactions of the investigated network with reactions from a database, the authors obtain networks consisting only of reactions which are known to exist. While this approach may generate feasible networks, it is also restricted to those reactions which are annotated in the corresponding database, and thus have a known biological role. The topological similarities between the network of *E. coli* and the networks obtained from randomization under constraints are judged visually, and may thus reveal different results when subjected to a statistical test.

Further, by including functional constraints, such as the viability under nutritional environments, networks with increasingly complex biological functions are generated. Therefore, the approach does not allow to evaluate the importance of network properties with respect to biological function. By contrast, the present thesis aims at developing a significance measure for metabolic networks, which preserves the physical constraints in order to identify properties related to biological function. Hence, it is not desired to impose constraints regarding complex biological functions, as these may impede functionally important properties from being identified as significant.

Finally, all randomization approaches differ from the present work in the addressed problem and analyzed networks. The existing methods aim at preserving topological constraints, such as the degree distribution, clustering coefficient, or the number of connected components. The approach by Samal and Martin (2011) is not suitable for measuring the significance of network properties, and has so-far only been applied to one metabolic network. The central idea of the present work is to preserve only the fundamental physical constraints which are irrevocable and govern the feasibility of metabolic processes, as detailed in the following section. The approach was applied to the directed, weighted bipartite graph representations of seven genome-scale metabolic networks.

1.4 Definitions

In the following, I will give the definitions of mass balance and the most important analyzed network properties, which are used throughout the thesis.

1.4.1 Mass balance

The term *mass balance* is used frequently in the literature with a different meaning as used here. All constraint-based (Burgard et al., 2003; Varma and Palsson, 1994), elementary mode (Schuster and Hilgetag, 1994), and extreme current analyses (Prigogine and Rice, 1980) rely on the assumption that the analyzed system is at steady state, which is defined as

$$M \cdot v = 0, \tag{1.1}$$

where M is the stoichiometric matrix (see Figure 1.2e), and v is the vector of reaction fluxes. As a consequence, the concentrations of all metabolites are balanced, i.e., they are not allowed to change over time, so that the accumulation or depletion of any metabolic species is avoided. Therefore, the steady state assumption is also referred to as the requirement of "mass balance" or "flux balance".

Here, a different but related definition of mass balance is used. In any realistic chemical reaction r, all atomic elements must be balanced, i.e., the number of involved atoms must be equal on both sides of the reaction. This definition will be referred to as *mass balance* in the following, and is given by

$$\sum_{s \in S} a_{s,r} \cdot m_s = \sum_{p \in P} a_{p,r} \cdot m_p, \tag{1.2}$$

where S is the set of substrates, P is the set of products, m_s, m_p are the vectors of sum formulas of s and p, respectively, and $a_{s,r}$, $a_{p,r}$ their stoichiometric coefficients. For example, the reaction equation $2{\cdot}CH_2 + 2{\cdot}C \rightarrow C_4H_2 + H_2$ is mass balanced, as the number of atoms sums to four carbon and four hydrogen atoms on each side. In contrast to flux balance, this definition of mass balance is a fundamental physical principle which must always be satisfied for a reaction in order to be possible. Although obvious, it is rarely taken into account that, if any reaction in the network violates equation 1.2, then equation 1.1 becomes meaningless, as it is not sufficient to ensure that the system is at steady state. Therefore, the definition used here may be regarded as a lower level physical constraint, which is a requirement for any analysis involving the steady-state assumption.

Mass balance is the fundamental physical constraint which is explicitly preserved by the proposed randomization algorithm (see Chapter 2). Note that the algorithm preserves balances, but does not establish balance of previously unbalanced reactions, so that all reactions in randomized networks will only be balanced if this holds for the original network. Notably, a large proportion of the reactions contained in metabolic databases as well as manually refined network reconstructions is not mass balanced (see Poolman et al., 2006 and Table B.1). Mostly, the imbalances are due to hydrogen atoms missing or in excess, which is a consequence of varying sum formulas depending on the pH state of the surrounding medium. Only few reconstructions consider sum formulas under different pH states, which would allow to balance most reactions, e.g., Feist et al. (2007). Here, this problem was largely resolved by replacing phosphate, hydrogen phosphate, dihydrogen phosphate, or phosphoric acid in unbalanced reactions by a form which establishes balance, by increasing the stoichiometric coefficient of an involved hydrogen atom, or by adding hydrogen to the corresponding side of an unbalanced reaction. This simple yet biochemically reasonable procedure allows to obtain networks where nearly all of the reactions are mass balanced (Table B.1).

By preserving mass balance, the generated randomized networks consist of physically possible metabolic reactions. This has at least two advantages: (1) the null model can be used as a biologically meaningful measure of significance for estimating the independence of network properties from principal physical constraints, and thus their evolutionary importance (Chapter 5), and (2) previously unknown, yet physically possible metabolic reactions are generated, which may provide a novel impetus for metabolic engineering and drug design (Chapter 6 and Section 7.6).

1.4.2 Thermodynamic favorability

Aside from mass balances, another fundamental physical requirement is the thermodynamic favorability of reactions, which can be estimated by the Gibbs free energy change under standard conditions, denoted by $\Delta_r G^0$ (Mavrovouniotis, 1991). The thermodynamic favorability of metabolic reactions is directly affected by the constraint of mass balance. The values of $\Delta_r G^0$ obtained for randomized reactions under the constraint of mass balance are strikingly more similar to the values of reactions in the original network, and thus more realistic, as compared to the reactions obtained from switch randomization (Figure 5.2). Nevertheless, the distribution of $\Delta_r G^0$ significantly differs between reactions in the metabolic network of *E. coli* and its mass-balanced randomized variants. The randomized networks, though structurally distant from the original network, may be classified in two types according to their $\Delta_r G^0$: those with thermodynamically favorable reactions, similar to the original network, and those with less favorable thermodynamics. This finding suggests an evolutionary optimization with respect to the thermodynamic favorability of reactions (Chapter 4). Detailed definitions and applications of thermodynamic favorability are given in Section 4.1.

1.4.3 Distance measure for metabolic networks

The thermodynamic landscape analyzed in Chapter 4 relies on a distance measure for pairwise comparison of randomized metabolic networks. Most existing distance measures aim at reconstructing phylogenetic relationships by comparison of metabolic networks from different organisms, and rely on the functional similarity of enzymes, as reflected by their EC number (Heymans and Singh, 2003), or on the number of shared enzymes (Ma and Zeng, 2004). More recent approaches rely on topological graph kernels, neglecting substrate-product relationships and stoichiometry (Oh et al., 2006; Kuchaiev et al., 2010).

The aforementioned approaches are not suitable for comparing randomized networks, as they either rely on biological knowledge of existing reactions, or they do not account for the substrates and products shared by reactions. Therefore, the distance μ between two metabolic networks is defined here as the number of shared substrate-reaction and product-reaction relationships[1]. For a network M^t obtained by randomizing the original network M^0, the relative complement between the edge sets E^0 and E^t is determined as follows:

$$\mu(M^t) = |E^0 \backslash E^t| = |E^0| - |E^0 \cap E^t|. \tag{1.3}$$

Note that $|E^0| = |E^t|$, as the randomization algorithm does not modify the number of edges (see Section 2.3.1). Thus, μ measures the number of substrates and products, by which the reactions in M^t and M^0 differ. This accounts for the substrate-product relationship of the underlying bipartite representation, and allows for comparison of randomized networks lacking any previous biological knowledge. The measure could be used to construct a multidimensional landscape by calculating the distances between any pair of randomized networks. However, due to the large number of 10^{10} networks analyzed in Section 4.3, a two-dimensional landscape representation is employed to illustrate the dependency between network distance and thermodynamic favorability.

[1] After developing our distance measure, Chang et al. (2011) proposed an approach to decompose reactions into pairs of substrates and products with similar atomic substructures. Therein, metabolic networks are compared by the signatures of their substrate-product pairs, which is applicable to mass-balanced randomized networks, as they contain existing metabolites with known structures.

1.4.4 Average path length

Within this doctoral thesis, I have applied the method of mass-balanced randomization to estimating the evolutionary significance of several topological properties, which have been extensively studied in complex network research. The average path length, referred to as diameter in the physics community, is defined as the average length of all directed shortest paths connecting any pair of reachable metabolites in a network. In Chapter 5, the length of a path is defined as the number of reactions involved in the path. For example, the shortest path length from FPN to AFP in Figure 1.2e is 2. This corresponds to the classical definition of path length used in complex network analysis (Watts and Strogatz, 1998; Albert et al., 2000; Jeong et al., 2000) (note that Wagner and Fell, 2001 represent metabolic networks as undirected graphs, resulting in shorter average path lengths).

Path length should not be confounded with the length of a biosynthetic pathway, or the number of reactions involved in producing one metabolite from another, as a simple path does not take into account the dependencies between substrates and products of a reaction. For example, the path length between H_2O and AFP in Figure 1.2e is 1, although the production of AFP requires both GTP and H_2O. Instead, the average path length characterizes the global structure of a network, and is related to robustness and the ability of a network to generate complex dynamic patterns (Lago-Fernández et al., 2000; Albert et al., 2000). Methods for determining biosynthetic pathways in metabolic networks were proposed elsewhere (Arita, 2004; Pitkänen et al., 2005; Handorf et al., 2005, see Section 1.4.6). The evolutionary importance of the average path lengths of six genome-scale metabolic networks is analyzed in Section 5.2.3.

1.4.5 Clustering coefficient

The clustering coefficient, also referred to as cluster index, is a measure of the local clustering of vertices in a network. In an undirected metabolite-metabolite graph, the clustering coefficient is defined as the average ratio of mutually connected neighbors of metabolite vertices: two metabolite vertices are connected, if they take part on opposing sides of a reaction (see Figure 1.2c). Thus, the neighbors $N(m)$ of a metabolite m are all metabolites, which form a substrate-product relationship with m. Let $G = (V, E)$ be an undirected metabolite-metabolite graph, where V is the set of (metabolite) vertices, and E the set of edges. Further, let $c(m) = |\{(a, b) \in E : a, b \in N(m)\}|$, the number of mutually connected neighbors of m. Then,

$$C(G) = \frac{1}{|V|} \cdot \sum_{\substack{m \in V \\ |N(m)|>1}} \frac{2 \cdot c(m)}{|N(m)|^2 - |N(m)|} \tag{1.4}$$

is the clustering coefficient of G. Thus, for each metabolite vertex m with $|N(m)| > 1$, the number of mutually connected neighbors $c(m)$ is divided by the number of possible connections between the neighbors. This number is averaged over all metabolite vertices.

Equation 1.4 corresponds to the classical definition of the clustering coefficient used in complex network analysis (Watts and Strogatz, 1998; Wagner and Fell, 2001; Albert and Barabási, 2002). In order to calculate the clustering coefficient in directed bipartite graphs, used throughout the thesis, reaction vertices and the directionality of edges are neglected (in other words, the metabolite-metabolite graph corresponding to the bipartite graph, also referred to as the corresponding unipartite graph, is used). A large clustering coefficient is associated with an overlap in functionally segregated networks (Sporns and Zwi, 2004), and, in conjunction with small average path

length, promotes complex dynamic patterns, efficient routing, and robustness (Lago-Fernández et al., 2000; Latora and Marchiori, 2001; Amaral et al., 2004; Albert et al., 2000). The clustering coefficients of six genome-scale networks are analyzed in Section 5.2.3.

1.4.6 Biosynthetic capabilities

The network expansion algorithm is a computationally efficient method for calculating the biosynthetic capabilities of a metabolic network (Handorf et al., 2005). By taking the substrate-product relationships of bipartite networks into account (Figure 1.2e), this method can be employed to determine the set of metabolites which can be synthesized from a specified subset of nutrients. The set of synthesizable metabolites, also referred to as scope, is calculated as follows: (1) From the given set of nutrients, the reactions for which all substrates are contained in the nutrient set are determined; (2) The products of these reactions are added to the nutrient set; (3) The procedure is repeated, until no more products can be added (see Algorithm A.3 on page 74).

The distribution of scope sizes, obtained by determining the scope sizes for a large number of random nutrient sets, characterizes the biosynthetic capability of a network and was shown to be correlated with the evolutionary history of organisms (Borenstein et al., 2008; Ebenhöh and Handorf, 2009). Therefore, the scope size distributions were used to validate the ability of the randomization method to identify the evolutionary importance of network properties (Section 5.2.2).

1.5 Thesis statement

The goal of network-based research in biology is to draw conclusions about the function of large-scale biological systems from their network structure. Due to the inherent complexity of biological processes, it is prohibitive to develop mathematical representations which precisely model all physical details at a molecular level, while accounting for all interactions on a system level (e.g., cell or organism). Two fundamentally different approaches allow for reducing the complexity in the analysis of biological systems. Bottom-up approaches, such as kinetic modeling, attempt to explicitly model dynamic processes at the molecular level, while restricting to well-characterized subsystems of manageable size. On the other hand, top-down approaches, such as complex network analyses, include all known interactions of the system, while neglecting dynamic information on the molecular level. Eventually, both paradigms will have to be unified by addressing only the dynamical details and large-scale interactions which are necessary in order to precisely model the dynamics of a process, cell, or an entire organism. Consequently, the challenge for bottom-up approaches is to increase their scope in order to allow modeling of larger systems (Bulik et al., 2009), while top-down approaches should aim at including more detailed information on the molecular level in order to allow for biologically more meaningful predictions.

The aim of this doctoral thesis was to develop an efficient computational method for assessing the biological importance of network properties in genome-scale metabolic networks. By accounting for a fundamental level of physical detail—mass balance—the method aims at determining the relation between topological properties and biological function. The method is based on the widely accepted hypothesis that biological systems and their properties evolve under physical constraints and evolutionary pressure (Lotka, 1922). In metabolic networks, the physical constraints (such as fundamental mass-balance and thermodynamic laws) are well-understood, while the biologically relevant functional properties arise from a long history of evolutionary pressure, which in turn depends on hardly understood complex interactions on various molecular, cellular, organismal, and

population levels. Consequently, a biologically meaningful significance measure should account for the fundamental physical principles underlying metabolic networks, and aim at identifying properties which are a consequence of evolutionary pressure.

It has been suggested that, in order to discover novel important properties, a randomization method should preserve the already known, lower-level properties, such as the degree distribution or modularity (Maslov, 2007; Hanhijärvi et al., 2009b). The method developed here further extends this idea with the aim of identifying the functionally important properties in biological networks. With this respect, in order to discover the properties which are a result of evolutionary pressure, and thus of functional importance, one should preserve the physical constraints imposed on the network. Consequently, properties identified as statistically significant are independent of basic physical principles, and thus likely to be a result of evolutionary pressure. Randomization constrained by physical laws may help to identify the network properties which are of functional importance, and thus of particular interest for modeling biological systems and generating biologically meaningful hypotheses.

In this spirit, an efficient method for randomizing metabolic networks was developed, which preserves mass balance, a fundamental physical principle constraining metabolic networks. A question which may arise is how the method behaves if a network property is a result of both physical principles as well as evolutionary pressure. Under the abovementioned hypothesis, any such property would be identified as significant by the method, if its value is affected by the dependence on evolutionary pressure. Thus, the method should also be able to detect properties which have evolved from both evolutionary pressure and physical constraints, which is certainly desirable, as they may be related to an important biological function.

The method was analyzed and applied to detecting the evolutionary importance of the salient network properties in six genome-scale metabolic networks. The results of this doctoral thesis are: (1) Development of a computationally feasible method for randomizing genome-scale metabolic networks and analysis of its complexity and uniformity properties (Chapter 2), published as Basler et al. (2011a). (2) Implementation of a user-friendly tool for mass-balanced randomization and calculation of several topological properties in metabolic networks and their randomized variants (Chapter 3), published as Basler and Nikoloski (2011). (3) Analysis of the dependency between mass balance constraints and thermodynamic favorability of reactions (Chapter 4), published as Basler et al. (2010). (4) Validation of the method and its application to determining the evolutionary importance of the salient properties in six genome-scale metabolic networks (Chapter 5), published as Basler et al. (2011b). (5) Extension of the method to generating feasible reactions which are predicted to facilitate improvements in biomass production (Chapter 6, unpublished manuscript). Finally, the results are summarized and future developments are proposed in Chapter 7.

Chapter 2

Mass-balanced randomization of metabolic networks

Authors: Georg Basler, Oliver Ebenhöh, Joachim Selbig, Zoran Nikoloski
Published as: Mass-balanced randomization of metabolic networks. *Bioinformatics*, 27(10):1397–1403. (Basler et al., 2011a)

Abstract

Motivation: Network-centered studies in systems biology attempt to integrate the topological properties of biological networks with experimental data in order to make predictions and posit hypotheses. For any topology-based prediction, it is necessary to first assess the significance of the analyzed property in a biologically meaningful context. Therefore, devising network null models, carefully tailored to the topological and biochemical constraints imposed on the network, remains an important computational problem.

Results: We first review the shortcomings of the existing generic sampling scheme—switch randomization—and explain its unsuitability for application to metabolic networks. We then devise a novel polynomial-time algorithm for randomizing metabolic networks under the (bio)-chemical constraint of mass balance. The tractability of our method follows from the concept of mass equivalence classes, defined on the representation of compounds in the vector space over chemical elements. We finally demonstrate the uniformity of the proposed method on seven genome-scale metabolic networks, and empirically validate the theoretical findings. The proposed method allows a biologically meaningful estimation of significance for metabolic network properties.

2.1 Introduction

The advances in omics technologies and algorithmic techniques for analysis of high-throughput data have placed network-based integrative studies in the focus of systems biology (Yamada and Bork, 2009; Albert, 2005). The promise of network analyses lies in the possibility to devise genome-scale representations of biological systems for predictive analyses. However, the statistical significance of any prediction must be validated in a biologically meaningful context using an appropriate null model.

The seminal work of (Barabási and Albert, 1999) directed complex networks research toward revealing the unifying properties of biological networks, starting from metabolic (Jeong et al., 2000) to gene-regulatory (Shen-Orr et al., 2002) to protein-protein networks (Maslov and Sneppen, 2002) and their integrated variants (Yamada and Bork, 2009). Despite the identification of simple mechanisms by which these networks may arise and evolve, such as the preferential attachment of newly added nodes (representing genes, proteins, reactions, or metabolites) to already highly connected ones, the advantage of such approaches to answering biological questions remains debatable.

Nevertheless, this direction in network research has resulted in the discovery of salient properties of biological networks, *i.e.*, properties which show similar trends for a wide variety of networks from different cells, tissues, and species. Some of these properties include: scale-free (*i.e.*, power-law) degree distribution, large clustering coefficient, small average path length, degree-degree correlation, different behavior of various centrality measures, and the distribution and over-representation of subnetworks, known as motifs (Milo et al., 2002; Barabási and Oltvai, 2004).

The studies following the work of Barabási and Albert have attempted to relate the salient properties of biological networks to their functionality (Jeong et al., 2001; Ma and Zeng, 2003b; Stuart et al., 2003; Albert and Albert, 2004; Papin et al., 2005; Marr et al., 2007)). However, it is often the case that the detection of novel salient properties of complex biological networks and determination of their statistical significance is based on a generic null model, which may result in misleading conclusions and, consequently, in inappropriate biological reasoning (Artzy-Randrup et al., 2004; Bernhardsson and Minnhagen, 2010).

Network null models are essential for establishing the significance of any prediction obtained from a network representation of a biological system. A randomization procedure allows for sampling from the (usually large) space of networks from a null model, and for estimating the statistical significance empirically. A p-value of a given property is usually calculated based on the following procedure: (1) determine the chosen property from an investigated biological network, (2) sample a large number of random networks which have a *similar* structure to that of the analyzed network, and (3) estimate the mean and variance of the property from the simulated networks to calculate a z-score and p-value under the assumption of normal distribution. Without this assumption, in principle, step (3) requires determining the distribution of values for the property under the considered network null model.

Clearly, the p-value of a property strongly depends on the sampling procedure and structure of the network null model. Therefore, any network-based analysis is prone to detecting statistically significant properties due to an ill-posed null model (Artzy-Randrup et al., 2004).

Finally, a null model strongly and ultimately depends on the type of analyzed network. For instance, gene-regulatory networks include directionality, while protein-protein interaction networks are undirected; signal transduction and metabolic networks are directed hypergraphs (representable as bipartite graphs) (Klamt et al., 2009), whereas metabolic networks include stoichiometry and biologically meaningful node-labels (representing chemical structure). Thus, a common

randomization procedure, which samples from a generic network null model, is unlikely to resolve the problem of relating the properties of different classes of networks to their biological function.

Despite these observations, many network-based studies (*e.g.*, Maslov and Sneppen, 2002; Milo et al., 2002; Guimerà et al., 2007a; Sales-Pardo et al., 2007) do rely on a common reference frame for all biological networks, called *switch randomization*. According to switch randomization, a randomized network is obtained from a given network by shuffling its edges while ensuring that the number of (incoming and outgoing) edges of every node remains unchanged. This can be achieved by the *switch* operation, whereby a randomly chosen pair of edges, (u, v) and (x, y), is replaced by two other edges, (u, y) and (x, v), provided that they do not already exist in the network. Switch randomization ensures that the probability of two nodes being connected is effectively independent of their distance in the original network. However, there are contradicting results with regard to whether the generated networks are sampled uniformly from the ensemble of networks with preserved degree distribution (Milo et al., 2003; Artzy-Randrup and Stone, 2005; Picard et al., 2008).

The underlying assumption of switch randomization is that the distribution of incoming and outgoing edges sufficiently characterizes the constraints under which networks of the analyzed type evolve. While this assumption may be valid on, *e.g.*, gene-regulatory networks, where the number of regulatory targets of a gene is a principle constraint, completely different constraints permeate the evolution of metabolic networks. For illustration, consider the following two metabolic reactions: glucose isomerase (Glucose $\rightarrow$ Fructose) and maleate isomerase (Maleate $\rightarrow$ Fumarate). After applying switch randomization we may obtain: Glucose $\rightarrow$ Fumarate and Maleate $\rightarrow$ Fructose, which is chemically infeasible due to the violation of the preservation of mass, since the corresponding chemical equations are $C_6H_{12}O_6 \rightarrow C_4H_2O_4$ and $C_4H_2O_4 \rightarrow C_6H_{12}O_6$. In the metabolic networks we analyzed, 99.8% of the reactions are unbalanced after applying switch randomization. By disregarding this fundamental principle, the generated networks are able to consume and produce matter out of nothing, yielding them incomparable to metabolic networks.

Establishing the statistical significance of a network property, mediated through a common, yet inappropriate reference frame, may result in the erroneous detection of significant properties, leading to questionable biological hypotheses. Therefore, the techniques for establishing suitable null models and randomization procedures need to be developed further, before making any statements about their biological importance. Recent work of (Picard et al., 2008) on estimating the overrepresentation of motifs is a first step toward a network null model tailored to a particular set of real-world biological networks (therein, protein-protein interaction networks).

Motivated by the shortcomings of the switch randomization and the lack of a network null model for metabolic networks which includes directionality, topological salient properties, and biochemical constraints (*e.g.*, reaction degrees and preservation of mass in biochemical reactions), here we present a method for randomizing metabolic networks. Our randomization procedure is based on the notion of mass equivalence classes for compounds and can be used to estimate the significance of a given topological property with respect to its importance in chemically constrained biological systems. Moreover, we show that our procedure samples a randomized network uniformly at random, which is another important requirement for any network sampling scheme. For the empirical validation of our results, we use the metabolic networks of seven organisms from all kingdoms of life: (1) *Bacillus subtilis* (Oh et al., 2007), (2) *Saccharomyces cerevisiae* (Herrgård et al., 2008), (3) *Escherichia coli* from iAF1260 (Feist et al., 2007) and (4) EcoCyc (Keseler et al., 2009), (5) *Chlamydomonas reinhardtii* (May et al., 2008), (6) *Arabidopsis thaliana* (Swarbreck et al., 2008), and (7) *Homo sapiens* (Ma et al., 2007) (network properties are shown in Table B.1).

2.2 Approach

A metabolic network is represented as a directed bipartite graph $G = (V_c \cup V_r, E)$, where V_c is the set of compound nodes, V_r the set of reaction nodes, and $E \subseteq (V_c \times V_r) \cup (V_r \times V_c)$ is the set of *directed* edges denoting substrate-reaction and product-reaction relationships. For a compound $c \in V_c$, we denote by $m_c \in \mathbb{N}^n$ its *mass vector*, *i.e.*, the vector representation of c over n chemical elements. For instance, one may consider only the six most abundant elements in biological systems (Dobson, 2004): carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorus (P), and sulfur (S). The mass vector of water is then $m_{H_2O} = (0, 2, 0, 1, 0, 0) \cdot (C, H, N, O, P, S)^T$. For a given reaction r, $r_{in} = \{c \in V_c \mid (c, r) \in E\}$ denotes the set of substrates, and $r_{out} = \{c \in V_c \mid (r, c) \in E\}$, the set of products. We abbreviate the expression $c \in r_{in} \cup r_{out}$ by $c \in r$, and write $d(r) = |r_{in}| + |r_{out}|$ for the degree of r (we omit the definition of compound degree, as it is not required for our purpose). Reversible reactions are represented by one reaction node for each direction: r^+ and r^-, where $r^+_{in} = r^-_{out}$ and $r^+_{out} = r^-_{in}$. Furthermore, let $s_{c,r} \in \mathbb{N}^+$ be the stoichiometric coefficient of a substrate (product) c of reaction r. A reaction is *mass balanced*, *i.e.*, chemically feasible with respect to the conservation of mass, if and only if the sum of its substrate atoms equals the sum of its product atoms:

$$\sum_{c \in r_{in}} s_{c,r} \cdot m_c = \sum_{k \in r_{out}} s_{k,r} \cdot m_k. \tag{2.1}$$

In order to uniformly randomize a network while preserving mass balance, each possible mass balanced network has to be generated with equal probability. This requires enumeration of all possible sets of substrates and products, for which Equation (2.1) is satisfied. A special case of this problem is to find all possible partitions of a set of integers, which sum up to 0 (which, in turn, is a special case of the Knapsack problem, see (Horowitz and Sahni, 1974)). As a consequence, the number of possible mass balanced networks is at least exponential in the number of compounds.

We approach the complexity of the general problem by restricting the set of possible solutions to Equation (2.1) twofold: (1) the in- and out-degrees of reactions are preserved, and (2) the substitution of compounds is limited to certain subsets, as detailed below, which allows to easily find a solution for Equation (2.1). The first restriction is in line with the observation that reaction degrees are biochemically constrained by the number of interacting compounds. The second allows to divide the randomization procedure into a precalculation step and an actual randomization. As a result, the generation of a large set of mass balanced randomized networks becomes computationally feasible.

We now move to the description of our randomization procedure including the abovementioned restrictions. Our procedure depends on determining the classes of linearly dependent mass vectors. Two compounds $c, k \in V_c$ will be called *mass equivalent* if and only if their respective mass vectors m_c and m_k are linearly dependent. Moreover, two pairs of compounds, denoted by (c, k) and (c', k'), will be called mass equivalent if and only if the corresponding sums of mass vectors $m_c + m_k$ and $m_{c'} + m_{k'}$ are linearly dependent. Note that mass equivalence is an equivalence relation, which follows from the reflexivity, symmetry, and transitivity of linear dependence for vectors in $\mathbb{N}^n$. As a result, the mass equivalence relation partitions the set of compounds and pairs of compounds (see Tables 2.1 and 2.2 for examples, and Figures C.2, C.3 for the class size distributions).

The inclusion of linear dependent triplets of mass vectors is straightforward and may further increase the sample space. However, due to the computational restrictions imposed by the size of

Compound	C	H	N	O	P	S
Allose	6	12	0	6	0	0
Alpha-d-galactose	6	12	0	6	0	0
Alpha-glucose	6	12	0	6	0	0
Arabinose	5	10	0	5	0	0
Cpc-10774	5	10	0	5	0	0
Cpd0-1108	5	10	0	5	0	0
Cpd0-1110	5	10	0	5	0	0
D-arabinose	5	10	0	5	0	0
D-ribulose	5	10	0	5	0	0
D-xylulose	5	10	0	5	0	0
Dihydroxyacetone	3	6	0	3	0	0
Formaldehyde	1	2	0	1	0	0
Galactose	6	12	0	6	0	0
Glc	6	12	0	6	0	0
Glycolaldehyde	2	4	0	2	0	0
L-lyxose	5	10	0	5	0	0
L-ribulose	5	10	0	5	0	0
L-xylulose	5	10	0	5	0	0
Mannose	6	12	0	6	0	0
Myo-inositol	6	12	0	6	0	0
Xylose	5	10	0	5	0	0

Table 2.1: Example of a mass equivalence class for individual compounds and their mass vectors. Each mass vector is a multiple of a scalar and the basis vector $(1, 2, 0, 1, 0, 0)$.

Compound pair	C	H	N	O	P	S
2-Ketoglutarate	5	4	0	5	0	0
D-beta-D-heptose-17-diphosphate	7	12	0	13	2	0
2-pg	3	4	0	7	1	0
Methyl-glyoxal	3	4	0	2	0	0
3-p-hydroxypyruvate	3	2	0	7	1	0
Acetol	3	6	0	2	0	0
3-p-hydroxypyruvate	3	2	0	7	1	0
Hydroxypropanal	3	6	0	2	0	0
3-p-hydroxypyruvate	3	2	0	7	1	0
Lactald	3	6	0	2	0	0
3OH-4P-OH-alpha-ketobutyrate	4	4	0	8	1	0
Acetald	2	4	0	1	0	0
Ascorbate	6	6	0	6	0	0
Fructose-16-diphosphate	6	10	0	12	2	0
Ascorbate	6	6	0	6	0	0
Tagatose-1-6-diphosphate	6	10	0	12	2	0
Cpd0-1063	9	14	0	12	1	0
Phospho-enol-pyruvate	3	2	0	6	1	0
Formate	1	1	0	2	0	0
Cpd-10551	5	7	0	7	1	0
Dihydroxy-butanone-p	4	7	0	6	1	0
Glyox	2	1	0	3	0	0
Dihydroxyacetone	3	6	0	3	0	0
Phospho-enol-pyruvate	3	2	0	6	1	0
Dihydroxy-acetone-phosphate	3	5	0	6	1	0
Pyruvate	3	3	0	3	0	0
Gap	3	5	0	6	1	0
Pyruvate	3	3	0	3	0	0
G3P	3	4	0	7	1	0
Methyl-glyoxal	3	4	0	2	0	0
Hydrogen-molecule	0	2	0	0	0	0
L-ascorbate-6-phosphate	6	6	0	9	1	0
L-glyceraldehyde-3-phosphate	3	5	0	6	1	0
Pyruvate	3	3	0	3	0	0
OH-pyr	3	3	0	4	0	0
Propionyl-P	3	5	0	5	1	0
Propionyl-P	3	5	0	5	1	0
Tartronate-S-ald	3	3	0	4	0	0

Table 2.2: Example of a mass equivalence class for pairs of compounds and their mass vectors. The sum of mass vectors for each pair is a multiple of a scalar and the basis vector $(6, 8, 0, 9, 1, 0)$.

genome-scale metabolic networks, we rely only on substitutions of individual and pairs of compounds. Finally, our approach is in line with the observations that some fundamental properties should be fixed while carrying out the randomization—here, these are the degrees of the reaction nodes and mass balance.

2.3 Methods

In this section, we present the details of the proposed algorithm for randomizing metabolic networks together with its computational complexity, and show the main result about the uniformity of the method for network randomization.

2.3.1 Randomization algorithm

The algorithm consists of two steps: In the first step, for a given metabolic network G, the mass equivalence classes are generated from the set of compounds $V_c(G)$. This step is to be executed only once for all subsequent randomizations of the same network. In the second step, the reactions of G are randomized while preserving mass balance. To randomize a reaction, chosen uniformly at random from $V_r(G)$, substrates and products are replaced by randomly chosen substitutes from their corresponding mass equivalence classes. In addition, this substitution entails recalculation of the stoichiometric coefficients to guarantee the preservation of mass balance. The output from this step is a network in which stoichiometric coefficients are changed, edges are replaced, and, consequently, the degrees of the compounds are altered, while the reaction degrees and mass balance of all reactions are preserved (see Figure C.1 for an overview and comparison to switch randomization).

Let $\sigma(c)$ denote the mass equivalence class of a compound c and $\sigma(c,k)$, the mass equivalence class of a pair of compounds (c,k). Given a reaction r, a substrate (product) c of r will be called *substitutable in* r by a compound $c' \in V_c$, denoted by $c \sim_r c'$, if and only if the following two conditions are satisfied:

- (S1) the compounds are mass equivalent, *i.e.*, $c' \in \sigma(c)$,
- (S2) the substitute c' is not already a substrate (product) of r.

Similarly, we define a pair of substrates (products) $(c,k) \in (r_{in} \times r_{in}) \cup (r_{out} \times r_{out})$, $c \neq k$, to be substitutable in r by a pair of compounds (c',k'), $c' \neq k'$, denoted by $(c,k) \sim_r (c',k')$, if and only if the following three conditions hold:

- (P1) (c,k) is mass equivalent to (c',k'), *i.e.*, $(c',k') \in \sigma(c,k)$,
- (P2) neither c' nor k' is already a substrate (product) of r,
- (P3) there are stoichiometric coefficients $s_{l,r'} \in \mathbb{N}^+, l \in r'$ for the new reaction r', such that Equation (2.1) is satisfied.

Note that substitutability, in contrast to mass equivalence, is defined over substrates and products of a reaction, such that a substitution only affects either the substrates or the products of one reaction. In addition, conditions (S2) and (P2) imply $c' \neq c$, such that each substitution results in a reaction $r' \neq r$ (*i.e.*, substitutability is irreflexive).

In order to choose a particular substitution for a given reaction r uniformly at random, the set of all possible substitutions for r has to be determined. Let the set of substitutions of individual

compounds be denoted by $\Psi_s(r)$, and the set of substitutions of pairs of compounds be denoted by $\Psi_p(r)$. According to the above definitions, these sets are then given by

$$\begin{aligned}\Psi_s(r) &= \{(c, c') \mid c \sim_r c',\ c \in r\}, \\ \Psi_p(r) &= \{(c, k, c', k') \mid (c, k) \sim_r (c', k'), \\ &\qquad (c, k) \in (r_{in} \times r_{in}) \cup (r_{out} \times r_{out})\},\end{aligned} \tag{2.2}$$

where $c, k, c', k' \in V_c$. The combined set of all possible substitutions for r is then given by $\Psi(r) = \Psi_s(r) \cup \Psi_p(r)$. Note that substitutability is symmetric, *i.e.*, any substitution can be reversed, as we can always replace the substitutes and their stoichiometric coefficients by those of the original reaction.

Proposition 2.3.1. For a given reaction r, each substitution results in a unique reaction.

Proof. Suppose the substitutions of individual compounds (c, c') and (k, k') in r both result in the same reaction r'. Then, $c' \in r'$ and $k' \in r'$ imply that $c' \in r$ and $k' \in r$, which contradicts condition (S2). By condition (P2), this holds analogously for the substitution of pairs of compounds. Suppose the substitution of individual compounds (c, c') results in the same reaction r' as the substitution of a pair of compounds (k, l, k', l'). Then, either $k' \in r$ or $l' \in r$, both contradicting condition (P2). □

In the following, we analyze the algorithm for randomizing metabolic networks: For a reaction r, chosen uniformly at random, the set of possible substitutions for all substrates, products, and pairs of substrates or products in r is generated, in order to then choose one substitution uniformly at random (see Algorithm 2.1). The stoichiometric coefficients in r are recalculated (line 6) by finding positive integers $s_{l,r} \in \mathbb{N}^+, l \in r$ satisfying Equation (2.1). For the substitution of an individual compound (c, c'), such coefficients can always be found, due to the linear dependence of the mass vectors: $s_{c',r}$ is obtained as $\frac{1}{m_{c'}} \cdot s_{c,r} m_c$. If $s_{c',r}$ is a non-integer a/b, then all coefficients of r are multiplied by b. Recalculation of the stoichiometric coefficients for the substitution of pairs of compounds requires solving a system of n linear equations with two unknowns. In case there is no solution, the substitution is not carried out. Table 2.3 shows examples of possible substitutions (see Algorithms A.1 and A.2 for more details).

Note that the number of reactions in G as well as the in- and out-degrees of perturbed reactions are not changed by the algorithm. Since both directions of a reversible reaction are considered independently, reversibilities can optionally easily be preserved by choosing only forward reactions in line 1, and updating the reversed reaction accordingly after line 6.

Due to the consideration of all pairs of compounds, the time complexity for precalculating the mass equivalence classes is in $O(|V_c|^2)$. However, this step is executed only once for any (usually large) number of subsequent randomizations of the same network.

For the randomization procedure, choosing a reaction and a substitution uniformly at random (lines 1 and 3), and replacing edges (lines 4 and 5) can be performed in constant time. Determining all possible substitutions for a reaction r (line 2) requires retrieving the precalculated mass equivalence class of each substrate, product and each pair of substrates or products, which is in $O(d(r)^2)$. Then, for each mass equivalent compound or pair of compounds, one has to determine whether they are already substrates or products in r, and whether there exist stoichiometric coefficients satisfying Equation (2.1), in order to obtain $\Psi(r)$. The latter requires solving a system of n linear equations with two unknowns, which is in $O(n)$, such that the solution can be used in line 6. Hence, line 2 is in $O(d(r)^2 \cdot \sigma^{max} \cdot n)$, where σ^{max} is the size of the largest mass equivalence

Input:
Mass balanced metabolic network, $G = (V_c \cup V_r, E)$,
Mass equivalence classes, $\sigma = \sigma(c) \cup \sigma(c,k),\ (c,k) \in V_c \times V_c,\ c \neq k$,
Number of iterations, $t \in \mathbb{N}^+$
Output:
Randomized mass balanced network
Repeat t times: ;

1 Choose a reaction $r \in V_r$ uniformly at random
2 Determine the set of possible substitutions $\Psi(r)$ from σ
3 Choose a substitution $d \in \Psi(r)$ with probability $1/|\Psi(r)|$
4 **if** d *is an individual substitution* (c, c') **then**
  if c *is a substrate of* r **then**
    replace the edge (c, r) by (c', r)
  else
    replace the edge (r, c) by (r, c')
5 **else if** d *is a pair substitution* (c, k, c', k') **then**
  if c,k *are substrates of* r **then**
    replace the edges (c, r) and (k, r) by (c', r) and (k', r)
  else
    replace the edges (r, c) and (r, k) by (r, c') and (r, k')
6 Recalculate the stoichiometric coefficient(s) in r

Algorithm 2.1: Mass-balanced randomization of metabolic networks

	Dihydroxyacetone C3 H6 O3	+	Phospho-enol-pyruvate C3 H2 O6 P1	→	Dihydroxy-acetone-phosphate C3 H5 O6 P1	+	Pyruvate C3 H3 O3
3	**Formaldehyde** **C1 H2 O1**	+	Phospho-enol-pyruvate C3 H2 O6 P1	→	Dihydroxy-acetone-phosphate C3 H5 O6 P1	+	Pyruvate C3 H3 O3
3	**Glycolaldehyde** **C2 H4 O2**	+ **2**	Phospho-enol-pyruvate C3 H2 O6 P1	→ **2**	Dihydroxy-acetone-phosphate C3 H5 O6 P1	+ **2**	Pyruvate C3 H3 O3
	G3P **C3 H4 O7 P1**	+	**Methyl-glyoxal** **C3 H4 O2**	→	Dihydroxy-acetone-phosphate C3 H5 O6 P1	+	Pyruvate C3 H3 O3
	Ascorbate **C6 H6 O6**	+	**Fructose-16-diphosphate** **C6 H10 O12 P2**	→ **2**	Dihydroxy-acetone-phosphate C3 H5 O6 P1	+ **2**	Pyruvate C3 H3 O3

Table 2.3: Phosphoenolpyruvate-glycerone phosphotransferase reaction in *E. coli* (EcoCyc) (row 1) and examples of possible substitutions for individual substrates (rows 2 and 3) and pairs of substrates (rows 4 and 5). The mass vectors are given below the compound names, modified stoichiometric coefficients and compounds are shown in bold.

class, and line 6 can be executed in constant time. Therefore, the algorithm has time complexity in $O(t \cdot (\Delta^2 \cdot \sigma^{max} \cdot n))$, where Δ is the maximum reaction degree of G. Note that Δ and n are bounded by small constants: $\Delta \leq 17$, $n \leq 23$, and $\sigma^{max} \leq 780$ in the investigated networks.

2.3.2 Uniformity of sampling

Any algorithm for randomizing a combinatorial structure should guarantee that every random instance is generated with equal probability. In other words, the probability distribution over the space of possible combinatorial structures must converge to the uniform probability distribution. Otherwise, the properties of the sample space would be biased towards those of more frequently generated networks, and, consequently, the significance assigned to any property would be questionable. Here, we show that our proposed algorithm for randomizing metabolic networks indeed has this property on the class of metabolic networks randomized via substitutions of single compounds and pairs of compounds (with mild assumption for the latter).

To establish this result, we rely on a transition graph Σ_G, in which a node represents a network that can be generated by our algorithm, and two nodes are connected by an edge (u, v), if there exists a substitution in u generating v. The given metabolic network to be randomized is denoted by $G^0 \in V(\Sigma_G)$. The set of networks obtained after applying t substitutions to G^0 is denoted by $\Gamma^t = \{G_i^t \mid i = 1, \ldots m, m \in \mathbb{N}^+\}$. Note that, due to the symmetry of the substitutability relation, Σ_G is undirected (*i.e.*, each edge corresponding to a substitution can be traversed in both directions). Moreover, since each node in the transition graph Σ_G corresponds to a network obtained after applying t substitutions starting from G^0, the transition graph Σ_G is connected.

Applying the randomization algorithm is equivalent to a random walk on Σ_G, starting at G^0. Therefore, we use the existing results from the theory of random walks on graphs. The classical theorem for uniformity of random walks on graphs (see Lovász, 1993) states that, for any non-bipartite regular transition graph Σ_G, a random walk using transition probabilities, $1/d(u)$, $u \in V(\Sigma_G)$, is stationary, *i.e.*, the probabilities for stopping the random walk at a node after any number t of transitions do not change with $t \to \infty$. Therefore, to prove the uniformity, we show that Σ_G is (almost) regular, *i.e.*, the degree distribution of Σ_G is (almost) uniform.

We first show the uniformity of our method if only individual compounds are allowed to be substituted. Given a metabolic network G^0, for any reaction $r \in V_r$ the number of possible substitutions of individual compounds in r is $|\Psi_s(r)|$ (see Equation 2.2). From Proposition 2.3.1, it follows that each substitution corresponds to a unique edge in Σ_G. Therefore, the degree of G_0 in the transition graph is

$$d_s(G^0) = \sum_{r \in V_r(G^0)} |\Psi_s(r)|. \tag{2.3}$$

Theorem 1. If only individual compounds are allowed to be substituted, then Σ_G is regular.

Proof. To establish the claim, we need to show that $d(G^0) = d(G), G \in \Gamma^t$, for any number of substitutions $t \in \mathbb{N}$. Note that the number of reactions $|V_r|$ and their degrees remain unchanged. Therefore, it suffices to show that the number of possible substitutions for a reaction r does not change after substituting a compound.

Let x be a substrate (product) of a reaction r and let $x \sim_r y$, *i.e.*, $y \in \sigma(x)$ and y is not already a substrate (product) of r. The symmetry of mass equivalence implies $x \in \sigma(y)$. The possible substitutions for x are then the same as the possible substitutions for y after replacing x in r by y,

except that $x \sim_r y$ is replaced by $y \sim_{r'} x$ in the new reaction r'. For any substrate (product) $z \neq x$, if $z \in \sigma(x)$, then the transitivity of mass equivalence implies $z \in \sigma(y)$. Thus, the substitutions for z do not change, except that $z \sim_r y$ is replaced by $z \sim_{r'} x$ (as y is a substrate (product) of the new reaction r'). On the other hand, if $z \notin \sigma(x)$, then $z \notin \sigma(y)$ implies that the substitutions for z do not change after substituting x in r by y. Thus, we have $d(G^0) = d(G)$, and the sampling is uniform. □

The more general case, on which our algorithm is based, considers substitutions of both individual compounds and pairs of compounds. In this case, due to changes after applying a substitution, Σ_G may not be regular. To illustrate this point, for a reaction r, if a substrate c is substituted by a compound x, we may subsequently substitute the pair of substrates (x, k), where k is any other substrate of r. The possible substitutions for (c, k) in r, $\{(c, k, c', k') \mid (c, k) \sim_r (c', k')\}$, may be different from the possible substitutions for (x, k) in the new reaction r', $\{(x, k, x', k'') \mid (x, k) \sim_{r'} (x', k'')\}$. Similarly, the possible substitutions for individual compounds may change after substituting a pair of compounds. Consequently, the sizes of substitutability classes $\Psi_s(r)$ and $\Psi_p(r)$ may differ from the sizes of $\Psi_s(r')$ and $\Psi_p(r')$, so that two nodes in Σ_G may have different degrees.

In the following, we analyze the probability that the algorithm samples nodes from Σ_G almost uniformly at random. Let us consider a random walk $\{G^0, G^1, \ldots, G^t\}$ on Σ_G, starting at node G^0. Let Y_i be the non-negative random variable whose value is the absolute value of difference of degrees between two neighbors G^i and G^{i+1} on the walk, *i.e.*, $Y_i = |d(G^i) - d(G^{i+1})|$, $0 \leq i < t$. We assume that all Y_i are independent and identically distributed variables, with probability density function $P(Y_i = k) = P(Y = k) = Ck^{-\gamma}$ for a positive constant C. Since all networks and the number of possible substitutions are finite, this distribution exhibits a finite mean.

A sequence of random variables $X_0, X_1, \ldots, X_t$, where the expected value of X_t is determined by X_{t-1}, is called a martingale (Williams, 1991). Then, the sequence $X_j = \sum_{k=0}^{j-1} Y_k + \sum_{k=j}^{t-1} E[Y_k]$, $0 \leq j \leq t$, forms a martingale, and, in particular, $X_0 = E\left[\sum_{k=0}^{t-1} Y_k\right]$ and $X_t = Y_0 + Y_1 + \ldots Y_{t-1}$ (Chung and Lu, 2006). Furthermore, let B_j denote the event that $|X_j - X_{j+1}| > c_j$, $c_j > 0$, $0 \leq j < t$; then, $P(B_j) = P(|E[Y_j] - Y_j| > c_j)$ is the probability that the absolute difference between expected and actual degree changes in step j of the random walk on Σ_G exceeds some $c_j > 0$. By a result of Chung and Lu, 2003 (Theorem 8.3), the following generalized Azuma inequality holds for the probability that degree changes differ at least by λ from the expected degree changes after t steps:

$$P\left(|X_t - X_0| \geq \lambda\right) \leq \exp\left(\frac{-\lambda^2}{2\sum_{j=1}^{t} c_j^2}\right) + P(B), \tag{2.4}$$

where $B = B_t$.

Let δ denote the expected degree difference of adjacent nodes, *i.e.*, $\delta = E[Y] = E[|d(G^i) - d(G^{i+1})|]$, $0 \leq i < t$. Given that $P(Y = k) = Ck^{-\gamma}$, the cumulative probability distribution is given by $P(Y > k) = C'k^{1-\gamma}$ (Li et al., 2005). Therefore, the probability that the degree difference between neighbors is larger than the expected difference can be expressed as $P(B) = P(Y > \delta) \sim \delta^{1-\gamma}$. We then have the following claim:

Theorem 2. If the distribution of differences in degrees between neighboring nodes follows a power-law $P(Y = k) \sim k^{-\gamma}$ and $P\left(|X_j - X_{j+1}| > \delta\right) \sim \delta^{1-\gamma}$, $\delta = E[Y]$, then the probability that the accumulated degree difference between any two nodes, sampled by a random walk,

exceeds the number of steps t is bounded by:

$$P\left(|X_t - X_0| \geq t\right) \leq \exp\left(\frac{-t}{2\delta^2}\right) + \delta^{1-\gamma}.$$

Proof. By invoking Equation (2.4) with $c_j = \delta$, $2\sum_{j=1}^{t} c_j^2 = 2t \cdot \delta^2$, we get the probability that, after t steps, the accumulated difference between expected and actual degree differences is at least t:

$$P\left(|X_t - X_0| \geq t\right) \leq \exp\left(\frac{-t^2}{2t \cdot \delta^2}\right) + \delta^{1-\gamma}.$$

As X_t and X_0 are the sums of absolute differences in degrees, the above expression represents the maximum difference in degrees between any two nodes reachable within t steps, *i.e.*, $|X_t - X_0| \geq |d(G^t) - d(G^0)|$. □

The proof relies on the assumption that the distribution of differences in degrees of neighboring nodes in Σ_G follows a power-law distribution. This is confirmed in Figure 2.1A for *E. coli* (see Figure C.4 for the remaining organisms).

Let $\overline{d}(\Sigma_G)$ denote the average degree of Σ_G. We call Σ_G *almost regular* if, for any two nodes $G, H \in V(\Sigma_G)$, the following holds:

$$\frac{|d(G) - d(H)|}{\overline{d}(\Sigma_G)} \leq 1.$$

We then have the following corollary:

Corollary. The probability that the algorithm samples nodes from Σ_G almost uniformly at random is bounded by:

$$P\left(\frac{|X_t - X_0|}{\overline{d}(\Sigma_G)} < 1\right) \geq 1 - \exp\left(\frac{-\overline{d}(\Sigma_G)^2}{2t \cdot \delta^2}\right) - \delta^{1-\gamma}.$$

Proof. Since $|X_j - X_{j+1}| = |E[Y_j] - Y_j| \leq |d(G^j) - d(G^{j+1})| + E\left[|d(G^j) - d(G^{j+1})|\right]$, from Equation (2.4) we can establish the probability that $|X_t - X_0| \geq \lambda$ with $\lambda = \overline{d}(\Sigma_G)$, as in the proof of Theorem 2. We then have $P\left(|X_t - X_0| \geq \overline{d}(\Sigma_G)\right) \leq e^{\frac{-\overline{d}(\Sigma_G)^2}{2t \cdot \delta^2}} + \delta^{1-\gamma}$, which is equivalent to

$$1 - P\left(\frac{|X_t - X_0|}{\overline{d}(\Sigma_G)} < 1\right) \leq \exp\left(\frac{-\overline{d}(\Sigma_G)^2}{2t \cdot \delta^2}\right) + \delta^{1-\gamma}.$$

□

As an example of the corollary, for the case of *E. coli*, we obtain $P(\Delta = k) \sim k^{-1.87}$, $\delta \approx 7.14$, and $\overline{d}(\Sigma_G) \approx 19490$ from sampling 10^4 random walks. Then the probability, that the algorithm samples nodes from Σ_G uniformly at random within $t = 10^6$ steps is bounded by:

$$P\left(\frac{|d(G^t) - d(G^0)|}{\overline{d}(\Sigma_G)} < 1\right) \geq 1 - e^{\frac{-19490^2}{2 \cdot 10^6 \cdot 7.14^2}} - 7.14^{1-1.87} \approx 0.80$$

(Table B.3 shows the results for the remaining organisms). Note that these probabilities represent a rare worst-case, since all X_j are the sums of absolute differences in degrees. In practice, the

cumulative degree changes of sampled nodes are likely to be smaller due to positive and negative changes in degree.

Finally, we briefly analyze some practical implications of these findings. First, we determine the size of the sample space, *i.e.*, the number of distinct randomized networks which can be generated from a given metabolic network G, if only individual compounds are substituted. Let $\Phi_s(r)$ denote the set of all mass equivalence classes, which contain a substrate or product of r. From each such equivalence class $e_s \in \Phi_s(r)$, we may choose any subset with the size of the number of substrates (products) of r contained in $\Phi_s(r)$; let ϕ denote this number. Then, there are $\binom{|e|}{\phi}$ possible reactions for each mass equivalence class e_s, where the original reaction may be obtained by reversing any previous substitutions. Therefore, the number of distinct networks which can be generated from G by substituting only individual compounds is

$$\Omega_{G,s} = \prod_{r \in V_r(G)} \prod_{e \in \Phi_s(r)} \binom{|e|}{\phi}. \tag{2.5}$$

For the model organism *E. coli*, the size of the sample space is $\Omega_s \approx 2.97 \cdot 10^{957}$ (see Table B.3 for the remaining organisms). The large sample spaces, again, illustrate the importance of uniform sampling.

As shown before, the number of distinct networks which can be generated by substituting pairs of compounds does not merely depend on the reactions in the original network, as the number of possible substitutions may change after applying substitutions. Therefore, we are unable to give a precise expression for the sample size in this case. Nevertheless, it is clear that, for the case of substituting individual compounds and pairs of compounds, the sample space is at least as large as $\Omega_{G,s}$.

In order to confirm the result of uniformity empirically, we analyze a random walk on the transition graph of the TCA cycle, a central respiratory metabolic pathway consisting of only 8 reactions and 20 compounds. For this network, the sample spaces are $\Omega_{TCA,s} = 256$, $\Omega_{TCA,p} = 1024$, with a combined total of 1024 possible randomized networks (*i.e.*, all networks generated by a sequence of individual compound substitutions can also be generated by pair substitutions). We observe that the sojourn frequencies, *i.e.*, the number of times each network is visited by the random walk, indeed converge towards the uniform distribution (see Figure 2.1B), confirming our theoretical claims.

2.4 Conclusion

The advances in high-throughput omics technologies require developing algorithmic techniques for the analysis of large-scale biological networks. However, the significance of any network-based prediction must be validated using a realistic null model. While the method based on switch randomization has been extensively used to study the significance of topological properties in many different types of networks, we argued that it is unsuitable for the analysis of metabolic networks.

We presented a new method for randomizing metabolic networks under the constraint of mass balance. We observed that a null model should satisfy two important requirements: preservation of ubiquitous constraints characterizing the class of analyzed networks, and uniformity of the sampling procedure. We demonstrated the uniformity of the proposed method theoretically and empirically on seven metabolic networks from all kingdoms of life.

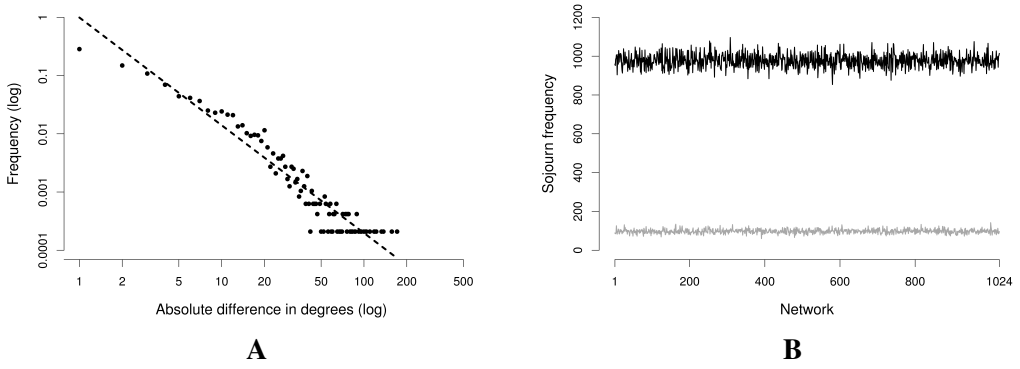


Figure 2.1: **(A)** Distribution of absolute differences in degrees between neighbors, sampled by a random walk on the transition graph of *E. coli* (EcoCyc). The dashed line shows the power-law fit with a scaling coefficient of $\gamma \approx 1.87$. The mean difference is $\delta \approx 7.14$ (see Figure C.4 for the remaining organisms). **(B)** Sojourn frequencies of a random walk on the transition graph of the TCA cycle (equivalent to a randomization of the TCA cycle). For 10^5 steps, the standard deviation of sojourn frequencies is $\sigma \approx 10.8$, yielding a coefficient of variation of 0.113 (grey line); after 10^6 steps, we have $\sigma \approx 34.6$ and a coefficient of variation of 0.038 (black line), confirming that the probability distribution over the 1024 networks converges towards the uniform distribution.

By integrating the (bio)chemical constraint of mass balance into a network null model, our method allows for a more realistic measure of significance. In addition, the proposed approach can be used for identifying network properties which are independent of mass balance constraints, and thus are likely to relate to the evolutionary history of metabolic networks. For instance, in a recent study, we applied the method to assess the evolutionary significance of thermodynamic favorability of metabolic reactions (Basler et al., 2010). We believe the integration of mass balance constraints is a necessary first step toward extracting biologically meaningful properties of genome-scale metabolic networks.

Chapter 3

JMassBalance: mass-balanced randomization and analysis of metabolic networks

Authors: Georg Basler, Zoran Nikoloski
Published as: *JMassBalance*: mass-balanced randomization and analysis of metabolic networks. *Bioinformatics*, 27(19):2761–2762. (Basler and Nikoloski, 2011)

Abstract

Summary: Analysis of biological networks requires assessing the statistical significance of network-based predictions by using a realistic null model. However, the existing network null model, *switch randomization*, is unsuitable for metabolic networks, as it does not include physical constraints and generates unrealistic reactions. We present *JMassBalance*, a tool for mass-balanced randomization and analysis of metabolic networks. The tool allows efficient generation of large sets of randomized networks under the physical constraint of mass balance. In addition, various structural properties of the original and randomized networks can be calculated, facilitating the identification of the salient properties of metabolic networks with a biologically meaningful null model.

Availability and Implementation: *JMassBalance* is implemented in Java and freely available on the web at http://mathbiol.mpimp-golm.mpg.de/massbalance/.

3.1 Introduction

Network-based studies of biological systems attempt to relate topological properties to biological function. The first step in drawing this connection involves determining the network properties which do not arise by chance. To this end, a network null model can be used to assess the statistical significance of network properties.

The common approach for determining the statistical significance of a given property is to determine a p-value based on the following procedure: (1) determine the chosen property from an investigated biological network, (2) sample a large number of random networks under biologically meaningful constraints, and (3) estimate the mean and variance of the property from the simulated networks to calculate a z-score (with the corresponding p-value) under the assumption of normal distribution.

Clearly, the significance of a network property strongly depends on the null model. The commonly used method, *switch randomization* (Milo et al., 2002; Guimerà et al., 2007a; Sales-Pardo et al., 2007), does not account for physical constraints, and thus generates unrealistic biochemical reactions (see Figure C.1 for an example). Thus, it is questionable whether the significance determined by this generic randomization scheme helps to elucidate the relation between network properties and biological functions.

Motivated by the lack of a biologically meaningful null model for metabolic networks, we developed a method for randomizing metabolic networks under the constraint of mass balance, and analyzed its computational complexity and uniformity of sampling (Basler et al., 2011a). Here, we present a tool which can be run via a graphical user interface (GUI) or from the command line, and implements mass-balanced randomization of metabolic networks provided in one of three standard data formats: (1) BioCyc (http://www.biocyc.org), (2) Systems Biology Markup Language (SBML, http://sbml.org), or (3) a customizable text file format.

3.2 Method

A metabolic network is represented as a weighted directed bipartite graph $G = (V_c \cup V_r, E)$, where V_c is the set of compound nodes, V_r the set of reaction nodes, and $E \subseteq (V_c \times V_r) \cup (V_r \times V_c)$ is the set of weighted, directed edges denoting stoichiometric substrate-reaction and product-reaction relationships. For example, an edge (c, r) specifies that compound c is a substrate of reaction r, while the stoichiometric coefficient $s_{c,r}$ of c in r is represented as the weight of (c, r).

A compound node is uniquely represented by a name, a compartment, and a mass vector, $m_c \in \mathbb{N}^n$, *i.e.*, the vector representation of the compound c over n chemical elements. For instance, when considering the six most abundant elements in biological systems: carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorus (P), and sulfur (S), then the mass vector of water is $m_{H_2O} = (0, 2, 0, 1, 0, 0) \cdot (C, H, N, O, P, S)^T$. The set of considered chemical elements can be specified in a configuration file (see online Reference Manual).

For a reaction r, r_{in} denotes the set of substrates, and r_{out} the set of products. A reaction node is uniquely represented by a name and its direction: reversible reactions are represented by one reaction node for each direction, r^+ and r^-, where $r^+_{in} = r^-_{out}$ and $r^+_{out} = r^-_{in}$. A reaction is *mass balanced*, *i.e.*, chemically feasible with respect to the conservation of mass, if the sum of its

substrate atoms equals the sum of its product atoms:

$$\sum_{c \in r_{in}} s_{c,r} \cdot m_c = \sum_{k \in r_{out}} s_{k,r} \cdot m_k. \tag{3.1}$$

The randomization procedure consists of a pre-calculation step, which classifies the compounds from the network according to their chemical sum formula (see Basler et al., 2011a), followed by the actual randomization. The precalculation is executed only once for all subsequent randomizations of the same network, and renders the method applicable to large networks. A network is randomized by replacing the substrates and products of randomly chosen reactions by compounds from within the same network, and choosing their stoichiometric coefficients, such that Equation 3.1 is satisfied (Figure 3.1). The polynomial-time algorithm generates randomized networks uniformly at random and clearly outperforms switch randomization (see Table B.1).

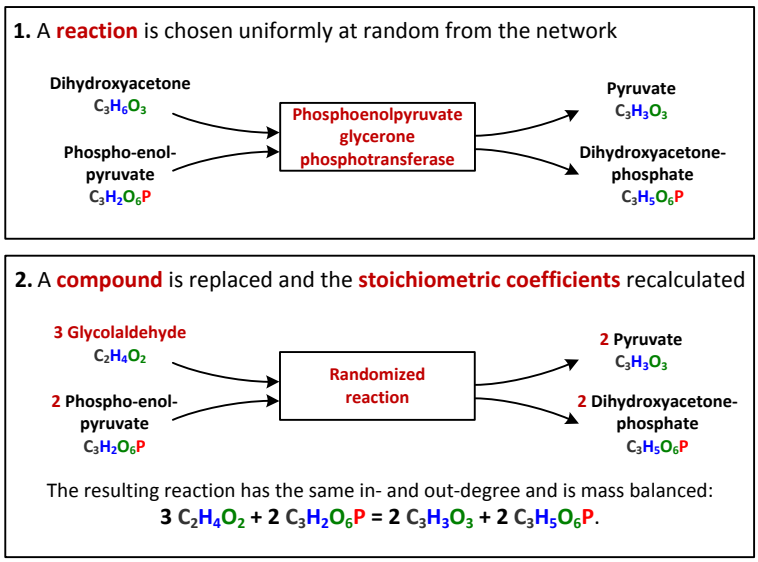


Figure 3.1: Mass-balanced substitution of a substrate. A large number of substitutions is applied in order to obtain fully randomized networks.

3.3 Application

JMassBalance is written in Java and comes with all required libraries. Hence, an installation is not required, and it can be used on any operating system with installed Java (http://www.oracle.com).

The randomization procedure accepts network files in BioCyc, SBML, or a customizable text format. Additional optional parameters allow specifying whether unbalanced reactions in the original network should be fixed, whether compartments should be considered, the randomization depth and probability, and the number of randomized networks to generate. All calculations can easily be parallelized by executing the program multiple times with different network indices (see online Reference Manual). Switch randomization is also implemented, and can be applied to compare the results of the two null models.

In addition to randomization, the following structural properties can be calculated for the original and randomized networks, respectively, which allows to determine their statistical significance in a biologically meaningful context:

- Average path length: the average number of reactions on the shortest path between two compounds.
- Clustering coefficient: average fraction of mutually connected neighbors of a node in the corresponding (unipartite) metabolite-metabolite network.
- Assortativity: correlation coefficient of the in-/out-degree of a node and the average in-/out-degree of its predecessors/successors in the corresponding (unipartite) metabolite-metabolite network.
- n-cycles: the number of directed cycles of length n in the corresponding (unipartite) metabolite-metabolite network.
- Path: test whether the given compounds constitute a path.
- Connectedness: test whether the given compounds are connected via paths.
- Transition degree: the number of possible mass-balanced substitutions.
- Local essentiality: the ratio of successor reactions affected by the knockout of a reaction.
- Reaction centrality: the ratio of reactions globally affected by the knockout of a reaction.
- Knockout set: the set of reactions globally affected by the knockout of a given reaction.
- Degree distribution: the compound degree distribution.
- Weight distribution: the distribution of edge weights.
- Scope size distribution (Handorf et al., 2005): the distribution of the number of compounds producible from a random set of seed compounds of the given size.
- Distribution of $\Delta_r^0 G$ (Mavrovouniotis, 1991): the distribution of the standard Gibbs free energy change of reactions.

The randomized networks may be printed as stoichiometric matrices or as text files, thus enabling subsequent investigations, such as constraint-based analysis (Feist et al., 2010).

3.4 Conclusion

JMassBalance is a flexible and efficient tool for assessing the significance of metabolic network properties through a biologically meaningful null model. It can be used to determine the salient structural properties of metabolic networks and to identify new properties, which are statistically significant and independent of basic physical constraints. Thus, we believe the tool is useful for the initial analysis of reconstructed metabolic networks, as well as subsequent network-based research.

Chapter 4

Thermodynamic landscapes of randomized large-scale metabolic networks

Authors: Georg Basler, Sergio Grimbs, Joachim Selbig, Zoran Nikoloski
Published as: Thermodynamic landscapes of randomized large-scale metabolic networks. In *Proceedings of the 7th International Workshop on Computational Systems Biology, WCSB 2010*, Tampere, Finland. Tampere International Center for Signal Processing. (Basler et al., 2010)

Abstract

Genome-scale metabolic network models are valuable tools for deciphering the evolutionary principles of metabolism. However, the effect of evolutionary pressure on basic thermodynamic properties of such models is not well understood. We analyze the thermodynamic favorability of reactions in the metabolic network of *E. coli* and its variants obtained by randomization under physical constraints, but free of evolutionary pressure. We find that the reactions of *E. coli* exhibit a characteristic pattern of Gibbs free energies, and are energetically more favorable compared to the reactions of randomized networks. This indicates that the prevalence of thermodynamically favorable reactions in metabolism is a result of evolutionary pressure, and not of physical constraints imposed on the network. As a consequence, thermodynamic patterns of species might provide interesting insights into evolutionary optimization principles.

4.1 Introduction

The staggering importance of Systems Biology studies, directed at improving the understanding of cellular processes, strongly depends on the integration of omics data with biological knowledge about the physical and biochemical principles of genome-scale gene, protein, and metabolite interactions. Thermodynamics, adding an important piece in the puzzle of biological systems, has already proven valuable in providing additional constraints to further confine the solution space in constraint-based analyses (Beard et al., 2002; Henry et al., 2006, 2007; Hoppe et al., 2007; Nagrath et al., 2007). Thermodynamic data, represented by the feasible ranges for the Gibbs free energy change of biochemical reactions, has been consequently applied to the reconstruction, curation, and kinetic modeling of metabolic networks (Feist et al., 2007; Ederer and Gilles, 2007), and assessing the degree of reaction reversibility (Kümmel et al., 2006).

With regard to the thermodynamic properties of metabolic networks, it is not yet known to what extent the energetic favorability of certain reactions is related to the evolution of metabolism. In particular, it is unclear whether the observed patterns of Gibbs free energy of metabolic reactions are a result of evolutionary pressure, or basic physical principles, such as the conservation of mass. One way of addressing this problem is to compare the thermodynamic properties of high quality metabolic networks to those obtained by randomization under physical constraints. In particular, we compare the distributions of Gibbs free energy changes in reactions, and analyze the thermodynamic landscape of randomized networks, defined by a simple topological distance measure.

The degree of thermodynamic favorability of a biochemical reaction r can be quantified by $\Delta_r G$, its standard Gibbs free energy change. One of the most prominent methods for estimating $\Delta_r G$ is the group contribution method of Mavrovouniotis (1991), which is based on rapid calculation of accurate estimations for $\Delta_c G$, the standard Gibbs free energy, for a wide variety of biological compounds. In group contribution methods, the molecular structure of a single compound c is decomposed into a set of smaller molecular substructures, based on the hypothesis that $\Delta_c G$ and $\Delta_r G$ can be estimated using a linear model. Each model parameter is associated with one of the constituent molecular substructures (or groups) that combine to form the compound. To estimate $\Delta_c G$ of the entire compound, the contributions of each of the groups are summed as follows:

$$\Delta_c G_{est} = \sum_{j \in gr} n_j \Delta_{gr} G_j,$$

where gr is the set of groups in c for which $\Delta_{gr} G_j$ is known and n_j is the number of occurrences of the group in the molecular structure. Similarly, $\Delta_r G$ can be estimated as:

$$\Delta_r G_{est} = \sum_{c \in r} s_{c,r} \Delta_c G_{est}, \tag{4.1}$$

where $s_{c,r}$ is the stoichiometric coefficient of compound c in reaction r. The method has already been employed to estimate $\Delta_c G$ and $\Delta_r G$ for the majority of the compounds and reactions contained in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Tanaka et al., 2003) and in a genome-scale model of *E. coli* (Henry et al., 2006).

It is worth mentioning that the standard Gibbs free energy only quantifies the thermodynamic favorability of reactions under standard conditions, *i.e.*, a pH of 7, temperature of 298.15 K, and concentrations of 1 mole per liter. However, for physiologically relevant conditions, the Gibbs free energy change of a reaction also depends on the pressure, magnesium ion concentration (pMg), and other factors (Vojinovi and von Stockar, 2009). Nonetheless, as a genome-scale metabolic

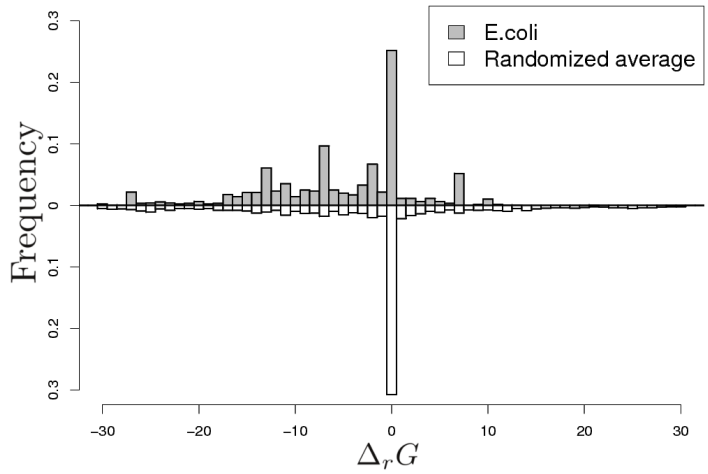


Figure 4.1: Distribution of $\Delta_r G$ for metabolic reactions in *E. coli* (grey), and distribution of $\Delta_r G$ averaged over 10^4 networks obtained after applying 10^6 randomization steps (white). In *E. coli*, characteristic $\Delta_r G$ are close to -13, -7, -2, 0, and 7, while most reactions in randomized networks have $\Delta_r G \sim 0$, with no further characteristic peaks.

network reflects the evolutionary history of metabolism, including various possible physiological conditions, here we only consider the standard values of $\Delta_r G$. In the rest of the paper, we focus on a genome-scale metabolic model of *E. coli*, where the standard Gibbs free energy of formation is known for 872 of 1039 compounds (84%), which also allows the estimation of the Gibbs free energy over networks obtained by randomization, as detailed in the next section.

4.2 Methods

4.2.1 Mass-balanced randomization

We apply a recently developed method for randomization of metabolic networks (Basler et al., 2011a)[1] in order to generate biochemically feasible networks not constrained by evolutionary pressure. A metabolic network is represented as a directed bipartite graph $M = (V_c \cup V_r, E)$, where V_c is the set of compound nodes, V_r the set of reaction nodes, and $E \subseteq (V_c \times V_r) \cup (V_r \times V_c)$ is the set of *directed* edges representing substrate-reaction and product-reaction relationships.

For a compound $c \in V_c$, we denote by $m_c \in \mathbb{N}^n$ its *mass vector*, *i.e.*, the vector representation of c over n chemical elements. Here, we restrict our approach to the six most common elements in organic compounds: carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorus (P), and sulfur (S). For instance, $m_{H_2O} = (0, 2, 0, 1, 0, 0) \cdot (C, H, N, O, P, S)^T$. For a given reaction $r \in V_r$, $r_{in} = \{c \in V_c \mid (c, r) \in E\}$ denotes the set of substrates, and $r_{out} = \{c \in V_c \mid (r, c) \in E\}$ the set of products.

We say that a reaction is *mass balanced*, *i.e.*, biochemically feasible with respect to the conservation of mass, if and only if the total number of substrate atoms equals the number of product atoms:

[1]The original publication references an unpublished manuscript.

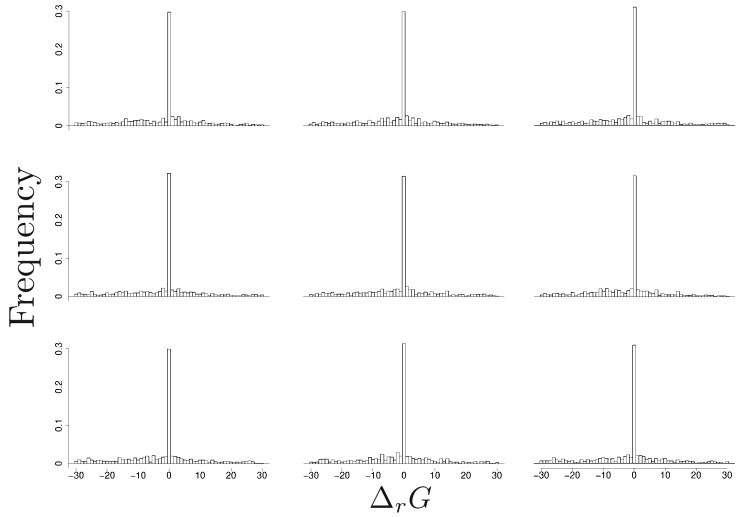


Figure 4.2: Distributions of $\Delta_r G$ for 9 arbitrarily chosen randomized networks obtained after applying 10^6 randomization steps. The networks have a characteristic $\Delta_r G \sim 0$, higher absolute values and higher variance, which clearly distinguishes them from the original network of *E. coli*.

$$\sum_{c \in r_{in}} s_{c,r} \cdot m_c = \sum_{k \in r_{out}} s_{k,r} \cdot m_k. \tag{4.2}$$

A metabolic network M is then randomized by repeating the following steps: (1) choose a reaction r at random, and (2) substitute edges of r uniformly at random, such that the in- and out-degrees of reactions are preserved, and Equation (4.2) is satisfied. This substitution entails recalculation of the stoichiometric coefficients to guarantee the preservation of mass balance. After applying a reasonably large number of randomization steps we obtain a uniformly randomized network, in which stoichiometric coefficients are changed, the degrees of the compounds are altered, while the reaction degrees and mass balance are preserved in all reactions.

4.2.2 Thermodynamic favorability

The amount of free energy released during the occurrence of a metabolic reaction r is given by $\Delta_r G$, where low values indicate a high energetic contribution to metabolism. Therefore, the thermodynamic favorability of a metabolic network can be captured by the following summary statistic:

$$\overline{\Delta_r G} = \frac{1}{|R|} \cdot \sum_{r \in R} \Delta_r G, \tag{4.3}$$

where $R \subseteq V_r$ is the subset of reactions with known $\Delta_r G$. Note that only irreversible reactions contribute to Equation (4.3), as the contribution of reversible reactions, according to Equation (4.1), is 0. While in *E. coli*, the $\Delta_r G$ are known for 84.5% of reactions, this set reduces to an average of 76.5% in the randomized networks, due to random substitutions involving compounds with unknown $\Delta_c G$.

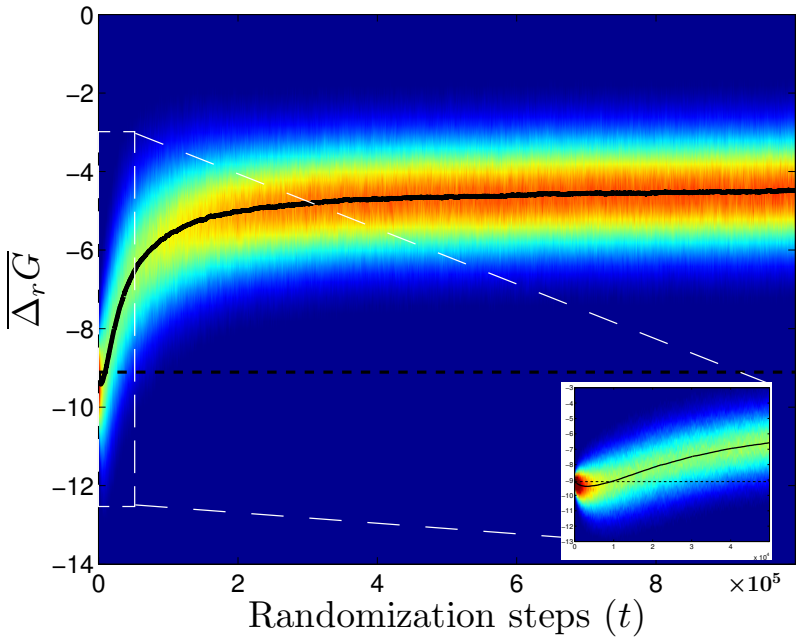


Figure 4.3: At each randomization step t a histogram of $\overline{\Delta_r G}$ values is calculated and plotted on a gray scale. The dashed line marks the $\overline{\Delta_r G}$ for the unperturbed network, while the solid line gives the average $\overline{\Delta_r G}$ for the randomized networks. The inlay magnifies the first $5 \cdot 10^4$ randomization steps. The average $\overline{\Delta_r G}$ drops even below the initial value before it sharply increases, and finally reaches a plateau.

4.2.3 Randomization of the metabolic network of *E. coli*

Starting from the initial metabolic network of *E. coli* from (Feist et al., 2007), denoted as $M^0 = (V_c \cup V_r, E^0)$, we apply $t = 10^6$ randomization steps, leading to a trajectory of randomized networks $M^t = (V_c \cup V_r, E^t)$. At each step, first $\Delta_r G$ is calculated for every reaction in R, and then the average $\overline{\Delta_r G}$ is determined. We repeat this procedure 10^4 times starting from the initial network M^0, giving a total of 10^{10} randomized networks.

In order to quantify the distance between a randomized network M^t and the initial network M^0 we define a function μ, which accounts for the differences between the substrates and products in reactions of M^t, and the corresponding reactions in M^0, from which they were derived by randomization. For this purpose, we determine the relative complement between the edge sets E^0 and E^t as follows:

$$\mu(M^t) = |E^0 \backslash E^t| = |E^0| - |E^0 \cap E^t|.$$

4.3 Results

To get a general idea of how randomization affects the Gibbs free energy of metabolic reactions, we compare the distribution of $\Delta_r G$ in *E. coli* with the distributions of 10^4 randomized networks, which were obtained after completing 10^6 randomization steps. Figure 4.1 shows that the distribution averaged over the $\overline{\Delta_r G}$ of randomized networks shows fewer modes, a tendency to higher absolute values and higher variance, and is more symmetrically distributed around zero compared

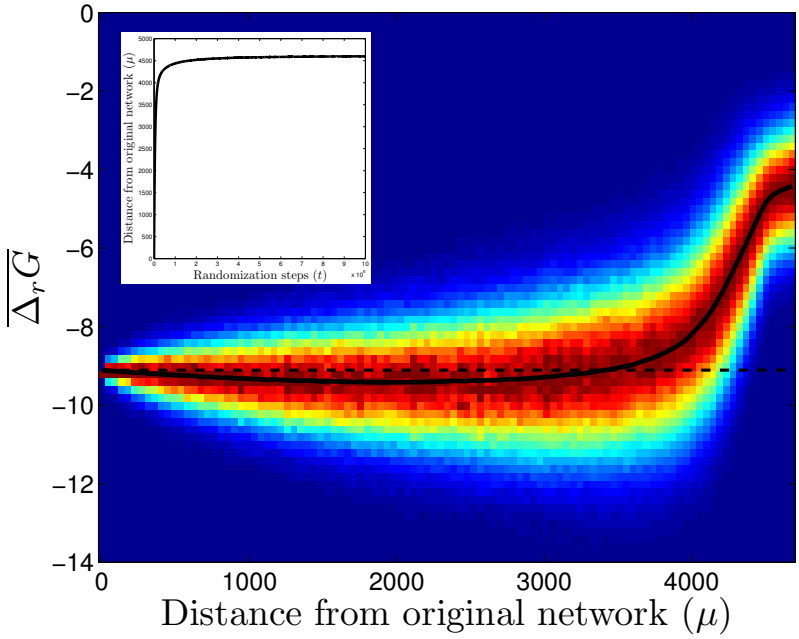


Figure 4.4: Randomized networks are binned according to their distance $\mu(M^t)$, and for each bin a histogram of $\overline{\Delta_r G}$ is shown on a gray scale. Again, the dashed line marks the $\overline{\Delta_r G}$ for the unperturbed network, while the solid line gives the average $\overline{\Delta_r G}$ for the randomized networks. For a wide range of μ the average $\overline{\Delta_r G}$ is well below the initial value. Only at large μ the average $\overline{\Delta_r G}$ increases dramatically and stays at a high level. The inlay shows μ as a function of randomization steps t. After only $2.6 \cdot 10^4$ steps μ is at 90% of its largest observed value, which explains the limiting upper bound of μ in the main figure.

to the initial network. It is further confirmed by Figure 4.2, which shows the distributions of $\Delta_r G$ for 9 arbitrarily chosen randomized networks. This initial observation points out that, after 10^6 randomization steps, the randomized networks can be clear-ly separated from the initial network with respect to the distribution of $\Delta_r G$.

In the following, we analyze how the randomization process influences $\Delta_r G$ and, hence, $\overline{\Delta_r G}$. At each randomization step we calculate a histogram of $\overline{\Delta_r G}$ over the 10^4 randomized networks (Figure 4.3). In very close proximity to the initial network the average $\overline{\Delta_r G}$ even decreases below the initial value of $-9.1 \frac{kcal}{mol}$. However, the average $\overline{\Delta_r G}$ increases sharply after approximately 3000 randomization steps and remains on a high level of about $-4.6 \frac{kcal}{mol}$.

Next, we generate a landscape of randomized networks using our previously defined function μ to describe the topological distances to M^0 (Figure 4.4). We group all randomized networks M^t according to the value of $\mu(M^t)$ into bins of size 50 and calculate a histogram of $\overline{\Delta_r G}$ for each such bin. Since the value for μ grows much faster than the number of randomization steps, not every bin contains the same number of randomized networks. For instance, about 80% of all randomized networks have a distance of $\mu > 4500$. For $\mu < 3000$, the average $\overline{\Delta_r G}$ is even lower than in the initial network. In this region, also the variance of $\overline{\Delta_r G}$ increases with increasing μ. In contrast, for large values of μ, $\overline{\Delta_r G}$ increases and remains at a high level, while the variance decreases. Interestingly, after $3 \cdot 10^5$ randomization steps, there are four networks with a distance $\mu \sim 4500$ and $\Delta_r G \sim -9.4 \frac{kcal}{mol}$ (data not shown). Even after $9.9 \cdot 10^5$ steps one network maintains a low $\Delta_r G = -9.23 \frac{kcal}{mol}$, while attaining a high distance of $\mu = 4551$. Such networks represent artificial reaction systems with highly favorable thermodynamics, possibly of great interest for

synthetic biology studies.

Regardless of whether we consider the changes of $\overline{\Delta_r G}$ over the randomization steps or the landscape defined by the topological measure, we observe two clearly separated regions of attraction. The first region contains those networks which are sufficiently similar to the metabolic network of *E. coli*, yielding equally low or even lower $\overline{\Delta_r G}$. The second region contains randomized networks topologically dissimilar to the initial network and with significantly increased $\overline{\Delta_r G}$.

4.4 Conclusion

We applied a recently developed algorithm for physically constrained randomization of metabolic networks in order to analyze the relation between thermodynamic properties and the evolutionary history of *E. coli*. Our results demonstrate that the Gibbs free energy change of reactions is significantly more favorable as compared to the randomized networks. This first finding may point out directions for future research aimed at discovering optimization principles of thermodynamic favorability in metabolic networks.

In addition, we define and analyze a thermodynamic landscape for the randomized network ensemble, which is based on the average Gibbs free energy of reactions and the distance to the curated network of *E. coli*. This landscape reveals a clear separation into two classes of networks: those with favorable thermodynamics, similar to and including the original network of *E. coli*, and those with less favorable thermodynamics and topologically dissimilar to *E. coli*.

While networks of the second class are obtained after relatively few randomization steps, they are characterized by a large topological distance to the original network. This indicates that the applied randomization method generates networks with distinct coherent thermodynamic characteristics in a defined and limited number of randomization steps. Moreover, the method itself allows for a clear distinction between the thermodynamic favorability of the original network of *E. coli* from those of the randomized networks obtained by exceeding the indicated number of steps.

To summarize, our results demonstrate that thermodynamic properties, in particular the Gibbs free energy of reactions, are a product of evolutionary constraints imposed on the genome-scale metabolic network of *E. coli*. As a consequence, thermodynamic favorability of metabolic reactions represents a biologically meaningful pattern, and may therefore provide novel biological insights when applied to extending the existing structural approaches. Furthermore, as the evolutionary history is reflected in such patterns, one may, as a next step, perform biologically motivated functional analyses based on thermodynamic properties.

The thermodynamic landscapes of this single-species study could be extended to include metabolic networks from multiple organisms. In such a multi-dimensional thermodynamic landscape the local minima would represent the present species, while elevated regions would reflect evolutionary unfavorable chemical reaction systems. Regions within the vicinity of known species could give hints at how biochemical reaction systems evolve in an evolutionary and physically constrained environment.

We believe that our study provides a first necessary step towards assessing the relationship between thermodynamic favorability and evolutionary constraints in genome-scale metabolic networks, and may serve as a starting point for analyzing thermodynamic properties on a large scale.

Chapter 5

Evolutionary significance of metabolic network properties

Authors: Georg Basler, Sergio Grimbs, Oliver Ebenhöh, Joachim Selbig, Zoran Nikoloski

Published as: Evolutionary significance of metabolic network properties. *Journal of The Royal Society Interface*. (Basler et al., 2011b)

Abstract

Complex networks have been successfully employed to represent different levels of biological systems, ranging from gene regulation to protein-protein interactions and metabolism. Network-based research has mainly focused on identifying unifying structural properties, such as small average path length, large clustering coefficient, heavy-tail degree distribution, and hierarchical organization, viewed as requirements for efficient and robust system architectures. However, for biological networks, it is unclear to what extent these properties reflect the evolutionary history of the represented systems. Here we show that the salient structural properties of six metabolic networks from all kingdoms of life may be inherently related to the evolution and functional organization of metabolism by employing network randomization under mass balance constraints. Contrary to the results from the common Markov-chain switching algorithm, our findings suggest the evolutionary importance of the small-world hypothesis as a fundamental design principle of complex networks. The approach may help to determine the biologically meaningful properties which result from evolutionary pressure imposed on metabolism, such as the global impact of local reaction knockouts. Moreover, the approach can be applied to test to what extent novel structural properties can be used to draw biologically meaningful hypothesis or predictions from structure alone.

5.1 Introduction

The central findings in network-based research suggest that there exist simple mechanisms directing the evolution of both engineered and natural networks (Ciliberti et al., 2007; Kuchaiev et al., 2010; Hyduke and Palsson, 2010; Duarte et al., 2007; Newman, 2003b; Ravasz et al., 2002; Guimerà and Amaral, 2005; Jeong et al., 2000; Dorogovtsev and Mendes, 2003; Barabási and Albert, 1999). However, the relation between the functions of a biological system and its network properties is hardly understood. Therefore, the advantage of using network representations for positing meaningful hypotheses about biological systems remains largely debatable (Yamada and Bork, 2009).

Properties of biological systems arise from two fundamental origins: *physical principles*, universally constraining the feasibility of biochemical processes, and *evolutionary pressure*, bearing the specific functional abilities required for an organism's vitality (Lotka, 1922). The former comprise well-understood physical laws, such as mass balance and thermodynamics, which constitute the basic requirements imposed on all living systems. In contrast, evolution depends on the interplay of complex phenomena, such as adaptation to environmental changes, symbiosis, and biodiversity of populations (Caetano-Anolls et al., 2009; Fani and Fondi, 2009), leading to diverse cellular functions. Consequently, the unique properties related to the functions of a biological system are a result of its evolutionary history.

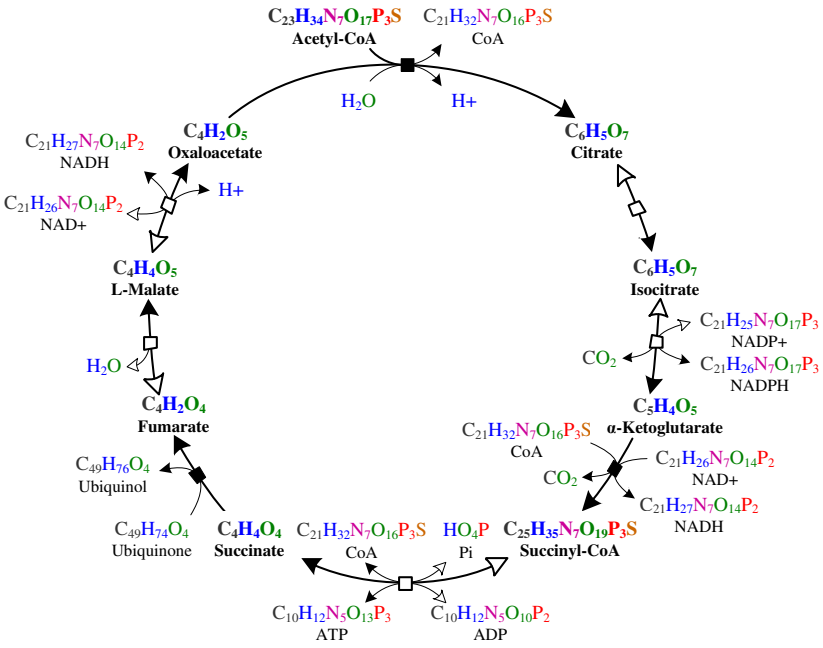
Explaining cellular behavior through network representations and their properties is a key challenge of modern biology. While many structural properties of metabolic networks are similar to those of other complex networks (Wagner and Fell, 2001)[1], it is unclear whether they are a consequence of the evolutionary history or merely arise as a result of general physical principles. Here, we apply a randomization method to determine which properties of metabolic networks, represented as bipartite metabolite-reaction graphs, may result from evolutionary pressure. This is an essential step in understanding the relation between the functional characteristics of biological systems and their network representations.

The common approach for estimating the relevance of a network property is to determine the statistical significance (p-value) by comparing the value of the property in the investigated network to those in the null-model distribution obtained from randomized networks (Casella and Berger, 1990). Clearly, the significance of a property strongly depends on the chosen null model, which should be constrained to preserve universal network properties (Maslov, 2007; Serrano et al., 2006). Since the p-value is the probability that the value of a property originates from the null-model distribution, a statistically significant property is likely to have emerged from some non-arbitrary process influencing network evolution independently of the imposed constraints.

In virtually all network-based studies (Guimerà et al., 2007a; Milo et al., 2002; Maslov and Sneppen, 2002; Sales-Pardo et al., 2007; de la Fuente et al., 2008; Milo et al., 2004; Marr et al., 2007), a Markov-chain switching algorithm, *switch randomization*, has been employed to determine the significance of network properties by generating randomized networks with preserved degree sequence. Its motivation stems from the finding that heavy-tail degree distributions are a universal feature of complex networks. This generic null model can be applied to any type of network, and guarantees the independence of an identified property from vertex degrees. We demonstrate how switch randomization affects the citric acid (TCA) cycle, a central respiratory metabolic pathway of outstanding importance for aerobic organisms (Figure 5.1a): two reactions $substrate_1 \rightarrow product_1$ and $substrate_2 \rightarrow product_2$ are substituted with new reactions

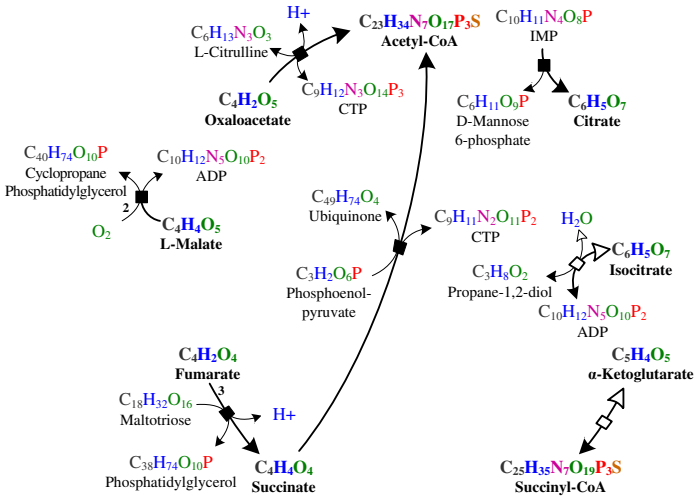
[1] In the original publication, the reference points erroneously to Barabási and Albert (1999).

(a) **TCA cycle**



(b) **Switch randomization**

$C_5H_4O_5 \leftrightarrow C_{25}H_{35}N_7O_{19}P_3S$
α-Ketoglutarate ↔ Succinyl-CoA



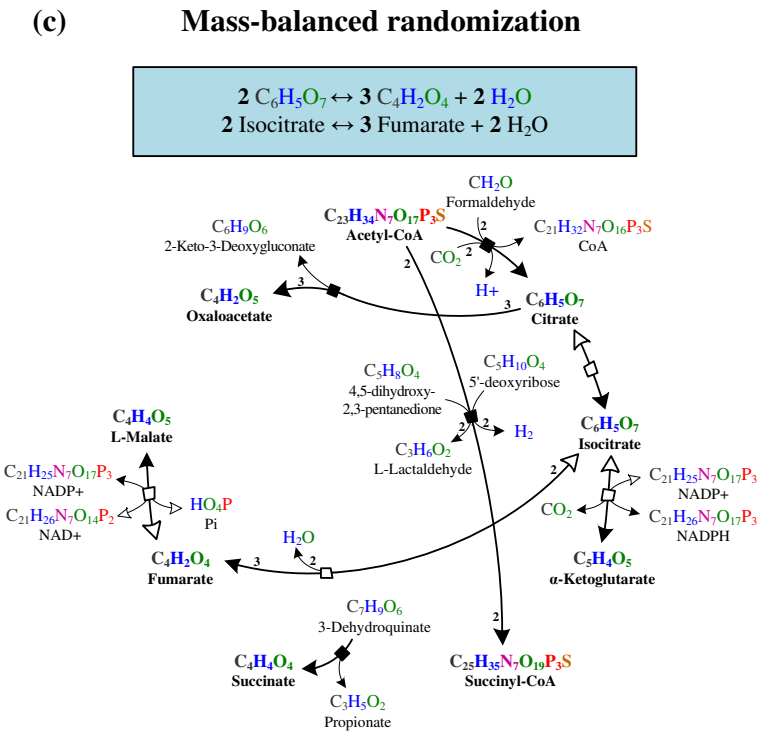


Figure 5.1: Illustration of how switch and mass-balanced randomization of the genome-scale metabolic network of *Escherichia coli* affect the TCA cycle. (a) The TCA cycle in *Escherichia coli*, consisting of 8 reactions and 22 compounds. Compound names are shown with corresponding sum formulas, irreversible reactions are represented by solid squares, and reversible reactions by blank squares. Internally, a reversible reaction is represented by one vertex for each direction, in order to adequately model the substrate-product relationships (see Methods). (b) Reactions involving metabolites from the TCA cycle (bold arrows and names) after applying switch randomization. The degrees of compounds and reactions are preserved, but the generated reactions violate fundamental physical constraints (see inlay). Note that the shown reactions are obtained from randomization of the entire network of *Escherichia coli*; the degrees therefore do not correspond to those shown in a. (c) All reactions obtained by mass-balanced randomization are chemically feasible due to balanced atom masses and realistic thermodynamic energy ranges, as indicated by the sum formulas and stoichiometric coefficients (thermodynamic data not shown).

$substrate_1 \rightarrow product_2$ and $substrate_2 \rightarrow product_1$, ensuring that the vertex degrees remain unchanged (Figure 5.1b). Since chemical feasibility is disregarded, a reaction that converts α-*Ketoglutarate* into *Succinyl-CoA* may be generated, where several atoms are created out of nothing. Hence, it remains hypothetical to what extent the properties, identified as significant with this method, relate to the function of the network, as they could well result from universal physical constraints imposed during network evolution.

5.2 Results

5.2.1 Measuring evolutionary significance

To identify the properties which originate from evolutionary pressure, a network should be compared to random networks which evolved free of evolutionary pressure, but persistently satisfy all relevant physical constraints. As this is practically impossible to simulate, we apply our recent method for randomizing metabolic networks while preserving mass balance of the biochemical reactions (Basler et al., 2011a). A reaction r with substrate set S and product set P is mass balanced if the number of substrate atoms equals the number of product atoms:

$$\sum_{s \in S} a_{s,r} \cdot m_s = \sum_{p \in P} a_{p,r} \cdot m_p. \tag{5.1}$$

where m_s, m_p are the vectors of sum formulas of s and p, respectively, and $a_{s,r}$, $a_{p,r}$ their stoichiometric coefficients (see Methods). The mass-balanced randomization of the TCA cycle does not violate this basic physical constraint, as shown in Figure 5.1c.

Thermodynamic properties, reflecting the energy change of reactions, constitute another important physical requirement for metabolic networks. As shown in Figure 5.2, the reactions generated by mass-balanced randomization of the *Escherichia coli* network are characterized by plausible Gibbs free energy changes under standard conditions (pH=7, T=298.15K, see Section 4.1) (Feist et al., 2007). In contrast, switch randomization results in unrealistic energy ranges. By preserving mass balance and thermodynamic properties during randomization, our null model imposes realistic physical constraints on the generated randomized networks. This ensures that the significant properties are independent of the fundamental physical requirements, and instead are likely to result from evolutionary pressure. Therefore, we refer to the statistically significant properties under the proposed null model as *evolutionary significant*.

For illustration, consider a landscape formed by the values of any given property over all randomized networks (Figure 5.3). The constrained networks, obtained by mass-balanced randomization, carve out a region in the vicinity of the original network which is embedded in the region of unconstrained networks resulting from switch randomization. As these regions exhibit different distributions of values, illustrated by different magnitudes of the peaks, an evolutionary significant property may only be identified when comparing the property of the original network to the constrained region.

5.2.2 Biosynthetic capabilities

To verify our approach, first we determine the evolutionary significance of the scope size distribution in the genome-scale metabolic networks of six model organisms: *Bacillus subtilis*, *Escherichia coli* (bacteria), *Saccharomyces cerevisiae* (fungi), *Chlamydomonas reinhardtii* (protista), *Arabidopsis thaliana* (plantae), and *Homo sapiens* (animalia) (see Methods). The scope (Handorf et al., 2005) represents the set of compounds which can be produced in a metabolic network from a given set of initial nutrients. We determine the scope size distribution of each network by repeatedly calculating the scope for 5000 randomly chosen sets of nutrient compounds, one set at a time, according to the following procedure: (1) from the initial set of nutrients, determine the reactions for which all substrates are contained in the nutrient set; (2) add the products of these reactions; (3) repeat the procedure, until no more products can be added (see Algorithm A.3 on page 74).

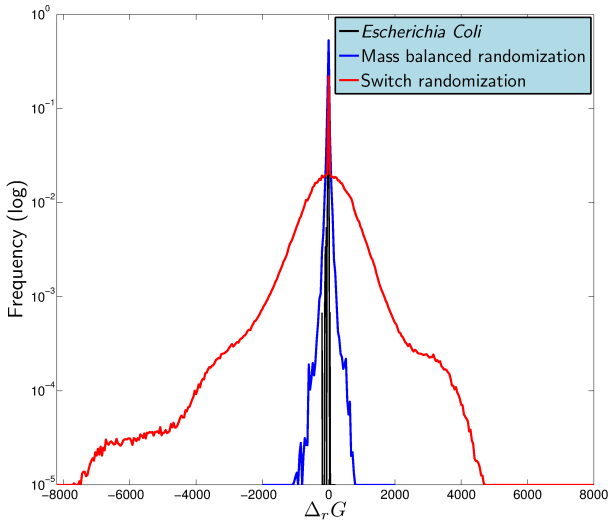


Figure 5.2: Distributions of Gibbs free energy changes under standard conditions ($\Delta_r G^0$) in *Escherichia coli* (black), and averaged over 10^4 mass-balanced (blue) and switch randomized (red) networks. Energy changes in *Escherichia coli* have a mean of 7.5 and standard deviation 15.1, mass-balanced randomized networks have a similar mean of 6.5 and standard deviation 53.5. In contrast, switch randomization generates implausible energy ranges with a mean of 32.5 and standard deviation 847.3.

The scope size distribution characterizes the biosynthetic capability of a network and has been shown to exhibit a strong correlation with the evolutionary history of organisms (Ebenhöh and Handorf, 2009; Borenstein et al., 2008). After applying mass-balanced randomization to the six networks, we compare the scope size distributions of each organism and its randomized network ensemble, and determine p-values using the Kolmogorov-Smirnov test (see Methods). We find the scope size distributions to be evolutionary significant for all studied organisms (p-values $< 10^{-49}$, Table B.2 and Figures C.9, C.10), which demonstrates that our method correctly identifies the interdependence of the network property and its evolutionary background.

5.2.3 Small-world property

In the following, we focus on determining the evolutionary significance of salient network properties which have been extensively studied in complex network research and are prominently applied in biological studies. In particular, we analyze the small-world property (Wagner and Fell, 2001), defined by a large clustering coefficient in conjunction with small average path length (see Sections 1.4.4 and 1.4.5), and the metabolite degree distribution (Jeong et al., 2000). We find that the clustering coefficient is significant in all species (p-values $< 10^{-5}$), regardless of the applied null model. On the other hand, the average path length is evolutionary significant with p-values < 0.025 in all species. With switch randomization, this property is significant (p-values $< 10^{-5}$) in all but *Saccharomyces cerevisiae* (p-value $= 0.77$, Table B.2).

More importantly, we may now assess the importance of the small-world phenomenon by determining whether this property is more pronounced in the analyzed networks as compared to their randomized variants. Interestingly, in each species we find that the average path length is smaller and the clustering coefficient is greater than the values of the respective properties obtained from

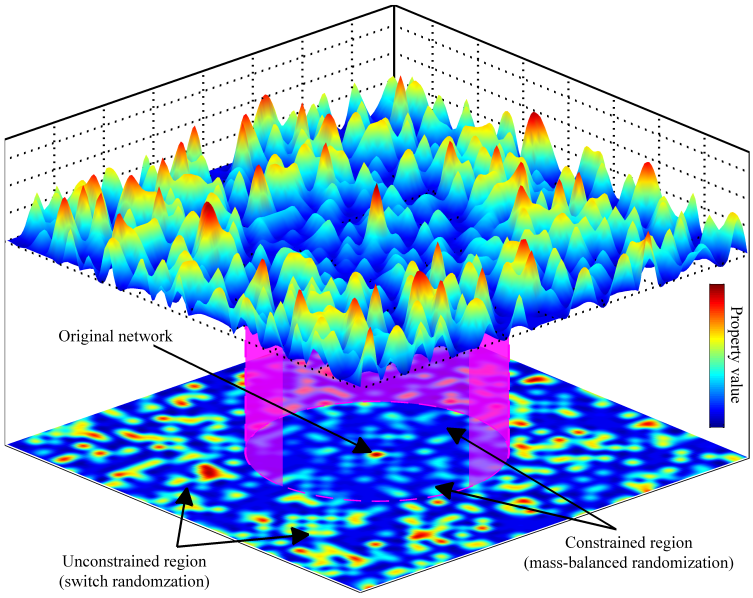


Figure 5.3: Illustration of a landscape of property values over all randomized networks. The property of the original network stands out from the inner region of constrained networks, but becomes inconspicuous in the outer region of unconstrained networks. Therefore, only by comparison with the constrained networks one may detect the evolutionary significance of the property.

mass-balanced randomization (Figure 5.4). This finding indicates that the small-world property is independent of physical constraints, and thus likely to be of evolutionary importance for metabolic networks. By contrast, when comparing the networks with their switch randomized ensembles, we arrive at a contrary conclusion–larger average path lengths and smaller clustering coefficients are prominent in real-world metabolic networks. Therefore, the results from switch randomization suggest that metabolic networks are the opposite of small worlds, disproving the small-world hypothesis. Moreover, this finding hints at two major hazards of network null models: (1) the obtained results crucially depend on the chosen model, and (2) the application of a generic null model which provides an unrealistically constrained environment may lead to counterintuitive results.

5.2.4 Degree distributions

Next, we analyze the metabolite degree distributions, where the degree of a metabolite is the number of reactions it is involved in either as substrate or product. The degree can be interpreted as metabolite specificity, with highly specific metabolites occurring in only few reactions. To our knowledge, the significance of degree distributions was never studied, since switch randomization is unsuited for this task. The degree distributions of all six organisms are evolutionary significant (p-values $< 10^{-17}$, Table B.2 and Figure C.13), suggesting that the patterns of metabolite specificities across different organisms emerge as a consequence of their evolutionary history, and not from the imposed physical constraints. This finding complements the well-known evolutionary requirement of a network architecture which is robust to random errors, as exhibited by the heavy-tail degree distributions (Jeong et al., 2000).

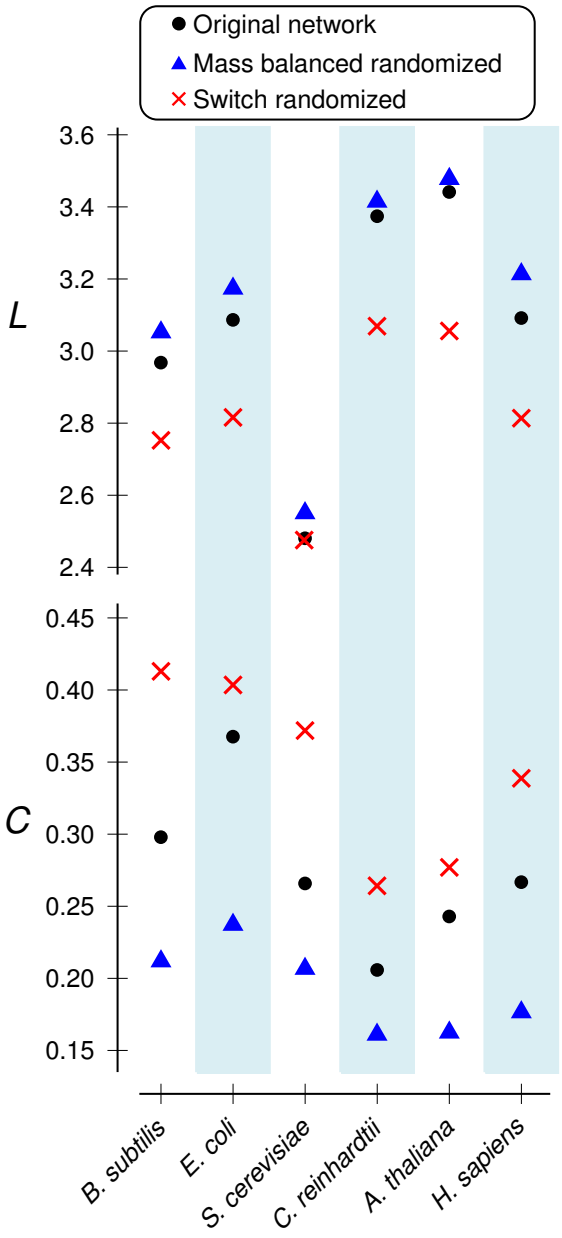


Figure 5.4: Characteristic path lengths (L) and clustering coefficients (C) of the six investigated metabolic networks (black dots) and averaged values of their mass-balanced (blue triangles) and switch randomized (red crosses) ensembles. Compared to the mass-balanced null model, characteristic path lengths are small and clustering coefficients large in all six organisms, confirming the small world hypothesis. Contrarily, in comparison to the switch based null model, characteristic path lengths are large and clustering coefficients small. The standard deviation is below 0.02 for each randomized distribution.

5.2.5 Reaction centrality

Finally, we propose a measure for determining the global importance of individual reactions, which is based on a centrality index previously used in sociological studies (Hubbell, 1965) (also referred to as Hubbell Index). For two reactions r_i and r_j, we define the dependence of r_j on r_i as the largest ratio by which r_i contributes to the overall production of an intermediary c (*i.e.*, a compound which is produced by r_i and consumed by r_j): $\omega(r_i, r_j) = \max_c d_{in}(c)^{-1}$, where $d_{in}(c)$ is the in-degree of c, which is the total number of reactions producing c. Note that this definition corresponds to the strength of impact of a knockout of r_i on r_j, where $\omega(r_i, r_j) = 1$, if r_j becomes inoperable upon knockout of r_i (*e.g.* if r_i and r_j are neighbours in a linear chain of reactions), and $\omega(r_i, r_j) \sim 0$, if the intermediaries required by r_j can be produced by many other reactions in the network.

The global impact of the knockout of a reaction on the entire network, which we call *reaction*

centrality, is

$$\nu(r_i) = \sum_{r_j \in R} \nu(r_j) \cdot \omega(r_i, r_j), \tag{5.2}$$

where R is the set of all reactions in the network, and $\omega(r_i, r_j) = 0$, if r_i and r_j do not share any intermediary compound (*i.e.*, r_i and r_j are not directly connected). This measure accounts for the direct dependencies between reactions through their intermediary compounds, as well as the global importance of the affected reactions: a knockout may affect only few other reactions directly, but can still have a large impact on the network, if an important reaction is affected indirectly (*e.g.* the knockout of a reaction at the beginning of a linear chain which leads to a reaction producing many important compounds).

Equation 5.2 can be written in matrix form as $A\nu = \nu$, where $A_{i,j} = \omega(r_i, r_j)$. In order to solve this eigenvalue problem, we need to ensure the inverse of A exists, which can be achieved by the PageRank transformation (Langville and Meyer, 2003). In particular, the transformed matrix A' is obtained by normalizing the columns of A and applying a damping factor d:

$$A'_{i,j} = d \cdot A_{i,j} / \sum_i A_{i,j} + (1 - d)/|R|,$$

which yields the Markov chain represented by A' ergodic, as the corresponding network is strongly connected, and ensures the largest eigenvalue is 1. In order to minimize the diluting effect of the damping factor on the topology of A, we choose $d = 0.99$. The eigenvector ν corresponding to the eigenvalue 1 of A' then contains the global centrality values of the reactions in the network, where $\nu(i)$ corresponds to the reaction centrality of the i-th reaction. The calculation for large networks is tractable using a Fortran implementation of the Implicitly Restarted Arnoldi Method (Lehoucq et al., 1998).

We determine a p-value for each reaction by comparing its centrality value in the original network with those obtained from mass-balanced randomized networks while preserving the reaction itself. In order to estimate the effect of evolutionary pressure toward high centrality values, we apply a one-sided test with the null hypothesis, that the values obtained from randomization are at least as large as the values of the original reactions (see Methods).

Table 5.1 shows the reactions which have a significant centrality (p-value ≤ 0.025) in at least three of the analyzed species (see Table S7 in the online Supplementary Material for a complete list). The references provide evidence that each reaction is of outstanding importance for metabolism, as demonstrated by their evolutionary ubiquity, severity of knockout or inhibition effects, and clinical applications. For instance, catalase (EC 1.11.1.6) inactivation was shown to have severe effects on the life span of *Saccharomyces cerevisiae* cells (Mesquita et al., 2010). Superoxide dismutase (EC 1.15.1.1) is essential for defense against oxygen toxicity and aerobic growth in eukaryotes (van Loon et al., 1986; Gralla and Valentine, 1991), and is involved in a multitude of diseases (Noor et al., 2002). Carbonic anhydrase (EC 4.2.1.1) fulfills diverse metabolic functions in organelles, tissues, and membranes of virtually all species, is used as a drug target for various diseases, and is one of the evolutionary oldest enzymes (Tashian, 1989; Henry, 1996; Smith et al., 1999; Smith and Ferry, 2000; Ferreira et al., 2008; Duanmu et al., 2009; Gilmour, 2010). The numerous experimental corroborations suggest that the proposed centrality index, in conjunction with the evolutionary significance determined by using our null model, could be used to predict enzymes responsible for maintaining organismal viability solely from the network structure.

For comparison, when repeating the analysis using switch randomization, the picture is less clear. In *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Homo sapiens*, 89%, 27%, and 14% of the reactions have a p-value of 0.0099, rendering the analysis useless at least for the first two species.

Enzyme	EC no.	BS	EC	SC	CR	AT	HS	Ref.
Catalase	1.11.1.6	✓	✓	✓	-	✓	n/a	[1-2]
Superoxide dismutase	1.15.1.1	✓	✓	n/a	-	✓	✓	[3-5]
Carbonic anhydrase	4.2.1.1	✓	-	✓	✓	✓	n/a	[6-12]
L-Arabinose isomerase	5.3.1.4	✓	✓	n/a	n/a	n/a	✓	[13-14]
Phosphoglycerate mutase	5.4.2.1	✓	-	✓	-	-	✓	[15-18]

Table 5.1: All reactions with centrality p-values ≤ 0.025 in at least three of the following species: *Bacillus subtilis* (**BS**), *Escherichia coli* (**EC**), *Saccharomyces cerevisiae* (**SC**), *Chlamydomonas reinhardtii* (**CR**), *Arabidopsis thaliana* (**AT**), and *Homo sapiens* (**HS**). A checkmark indicates that the reaction catalyzed by the enzyme has a significant centrality in the corresponding species; a hyphen indicates not significant; n/a indicates the corresponding enzyme is not annotated for the species. References: **[1-2]**: Mesquita et al. (2010); Zamocky et al. (2008); **[3-5]**: van Loon et al. (1986); Gralla and Valentine (1991); Noor et al. (2002); **[6-12]**: Tashian (1989); Henry (1996); Smith et al. (1999); Smith and Ferry (2000); Ferreira et al. (2008); Duanmu et al. (2009); Gilmour (2010); **[13-14]**: Novotny and Englesberg (1966); Schleif (2010); **[15-18]**: Irani and Maitra (1974); Oh and Freese (1976); Lam and Marmur (1977); Papini et al. (2010).

Five reactions have a significant centrality in at least two of the remaining three analyzed species (see Table B.4 on page 78 and Table S8 in the online Supplementary Material). We omit a detailed statistical analysis of these initial results, which will be necessary to draw further conclusions.

5.3 Discussion

To conclude, we proposed a novel method to reveal the relation between network properties and their evolutionary background by preserving the universal physical principles which constrain the design of metabolic networks. Any property which originates from evolutionary pressure, and thus relates to an important biological function, should not be observed in artificial metabolic networks, which evolved free of evolutionary pressure, but satisfy all relevant physical constraints. This should even hold for properties evolved from complex time-dependent phenomena, if they are reflected in the ultimately observed network.

We recognize that the proposed method only preserves mass balance and thermodynamic constraints, while other physical principles, such as electric charges, may also be relevant for metabolic network properties. Nevertheless, the considered physical constraints are the most fundamental and ubiquitous ones. Therefore, we believe that the method is a reasonable first approach to extract the biological importance of metabolic network properties. Accounting for additional physical constraints is complicated by the lack of reliable data for genome-scale metabolic networks; however, we expect such extensions to become possible in the future, which should further improve the biological relevance of the significance measure and the accuracy of the resulting predictions.

In contrast to the commonly applied switch randomization, our approach provides a realistic network background, and attributes an important evolutionary role to the small-world property and heavy-tail degree distributions. Our findings shed new light on the conclusions of previous studies, and suggest that the salient network properties are indeed a product of evolutionary pressure. Therefore, these properties carry important biological information, and can be justifiably used to generate meaningful hypotheses for experimental research.

We demonstrate that the proposed centrality index is one such network property which determines reactions important for viability of organisms. The method could therefore be used to identify candidate reactions for metabolic engineering and drug development. The results provide an impetus for addressing the long-standing doubts concerning the biological relevance of network properties. In addition, the proposed null model could be employed to verify the evolutionary assumptions in constraint-based approaches (Feist et al., 2010) and to provide an interface to synthetic biology studies.

Finally, we envision that, similar to the proposed approach for metabolic networks, specifically designed null models will be developed for other physically constrained systems, represented by gene-regulatory, protein-protein interaction and signaling networks. For instance, transcription factors depend on cis-elements and DNA binding domains, which constrain the sequence of genes by which they are encoded. Likewise, protein interactions and signaling interactions depend on functional domains and binding sites. Development of null models which integrate the governing physical constraints of such systems will likely stimulate novel insights into the structure-function relationship in complex biological networks.

5.4 Methods

5.4.1 Genome-scale metabolic networks

We conduct our analyses on the most widely used genome-scale metabolic networks of six model organisms from all kingdoms of life: *Bacillus subtilis* (Oh et al., 2007), *Escherichia coli* (Feist et al., 2007), *Saccharomyces cerevisiae* (Herrgård et al., 2008), *Chlamydomonas reinhardtii* (May et al., 2008), *Arabidopsis thaliana* (Rhee et al., 2003), and *Homo sapiens* (Ma et al., 2007). The sizes of the networks vary according to the complexity of the represented organisms, ranging from 855 reactions and 766 compounds (*Bacillus subtilis*) to 2819 reactions and 2691 compounds (*Homo sapiens*). Resulting from the bipartite graph reconstruction, detailed in the next section, the number of vertices and edges varies accordingly, from 1877 vertices and 5368 edges (*Bacillus subtilis*) to 7059 vertices and 19651 edges (*Homo sapiens*). The networks further differ in their quality regarding mass balance of reactions, availability of information on reversible reactions, and the number of (strongly) connected components (see Table B.1): only the network of *Escherichia coli* is fully balanced and consists of one connected component.

5.4.2 Mass-balanced randomization

To estimate the evolutionary significance of network properties, we generated 10^4 mass-balanced randomized networks for each of the six analyzed genome-scale metabolic networks. A metabolic network is represented as a directed bipartite graph $G = (V_c \cup V_r, E)$, where V_c is the set of compound vertices, V_r the set of reaction vertices, and $E \subseteq (V_c \times V_r) \cup (V_r \times V_c)$ is the set of *directed* edges denoting substrate-reaction and product-reaction relationships. For a compound $c \in V_c$, we denote by $m_c \in \mathbb{N}^n$ its *mass vector*, *i.e.*, the vector representation of c over n chemical elements. Here, we consider only the six most abundant elements in biological systems (Dobson, 2004): carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorus (P), and sulfur (S). The mass vector of water is then $(0, 2, 0, 1, 0, 0) \cdot (C, H, N, O, P, S)^T$. Reversible reactions are represented by one reaction vertex for each direction: r^+ and r^-, such that $r^+_{in} = r^-_{out}$ and $r^+_{out} = r^-_{in}$.

In order to uniformly randomize a network while preserving mass balance, each possible mass-balanced network has to be generated with equal probability. This requires enumeration of all possible sets of substrates and products, for which Equation (5.1) is satisfied. As this problem is a special case of the Knapsack problem (Horowitz and Sahni, 1974), the number of possible mass-balanced networks is at least exponential in the number of compounds.

We approach the complexity of the general problem by applying a new method for mass-balanced randomization, introduced in (Basler et al., 2011a). The set of possible solutions to Equation (5.1) is restricted twofold: (1) the in- and out-degrees of reactions are preserved, and (2) the substitution of compounds is limited to certain subsets, as detailed below, which allows to easily find a solution for Equation (5.1). The first restriction is in line with the observation that reaction degrees are biochemically constrained by the number of interacting compounds. The second allows to divide the randomization procedure into a precalculation step and an actual randomization. As a result, the generation of a large set of mass-balanced randomized networks becomes computationally feasible.

The randomization procedure consists of two steps: In the first step, for a given metabolic network G, we determine the classes of mass equivalent compounds and pairs of compounds from $V_c(G)$. Two compounds are called mass equivalent, if the mass vector of one compound is a multiple of the other (e.g. CO_2 and C_2O_4). Two pairs of compounds are called mass equivalent, if the sum of mass vectors of one pair is a multiple of the sum of mass vectors of the other pair (e.g. (CH_2O, CO_2) and ($C_4H_2O_4$, H_2O_2)). This definition ensures that the mass vectors of two compounds (and the sums of the mass vectors of two pairs of compounds) from the same mass equivalence class differ only by rational factors (e.g. $2\cdot$ CH_2O + $2\cdot$ CO_2 = $C_4H_2O_4$ + H_2O_2). The precalculation of mass equivalent compounds is to be executed only once for all subsequent randomizations of the same network and renders the generation of large sets of mass-balanced randomized networks computationally feasible (see Table B.1 for a performance comparison to switch randomization).

In the second step, the reactions of G are randomized while preserving mass balance. To randomize a reaction chosen uniformly at random from $V_r(G)$, its substrates and products are replaced by randomly chosen substitutes from their corresponding mass equivalence classes. When substituting an individual substrate or product, the stoichiometric coefficients of the new reaction are obtained by multiplying the corresponding previous coefficients with the abovementioned factor, such that Equation (5.1) is satisfied. For the substitution of a pair of substrates or products, the stoichiometric coefficients satisfying Equation (5.1) are determined by solving a system of n linear equations with two unknowns (see Table 2.3 on page 24 for examples). In case there is no solution, the substitution is not carried out. The output from this step is an (almost) uniformly randomized network in which stoichiometric coefficients are changed, edges are replaced, and, consequently, the degrees of the compounds are altered (Basler et al., 2011a). The approach is in line with the observation that some fundamental properties should be fixed while carrying out the randomization—here, these are the degrees of the reaction vertices and mass balance.

5.4.3 Calculation of *p*-values

The analyzed properties are calculated in the original metabolic network and in each of the 10^4 randomized networks. For the average path length and clustering coefficient, we derive a z-score, $z = \frac{x-\overline{y}}{\sigma}$, from the original value x, the average randomized value $\overline{y}$, and the standard deviation of randomized values σ. The two-sided p-value is determined as $p = 2 \cdot \int_{|z|}^{\infty} \mathcal{N}(0,1)$.

For comparing the metabolite degree and scope size distributions of the metabolic networks with their randomized versions we apply the two-sample Kolmogorov-Smirnov test. From the cumula-

tive distribution F_n of the property in the original network, and the joint cumulative distribution $F_{n'}$ of the randomized networks, a test statistic is derived as $d_{n,n'} = \sup_x |F_n(x) - F_{n'}(x)|$, where n and n' are the number of values in the original, respectively the joint randomized distributions. The p-value is $p = \sqrt{\frac{nn'}{n+n'}} d_{n,n'}$.

For each reaction vertex r, we determine its centrality, $\nu(r)$, in the original network and in 100 randomized networks, which are obtained by preserving r and randomizing the remaining reactions. The p-value of r is $p_r = \frac{q'_r+1}{n'+1}$, where q'_r is the number of randomized networks, in which the centrality of r is at least as large as in the original network, and $n' = 100$.

Chapter 6

Optimizing metabolic pathways by screening for feasible synthetic reactions

Authors: Georg Basler, Sergio Grimbs, Zoran Nikoloski
Under review at: *BioSystems* (Elsevier Science)

Abstract

Background: Reconstruction of genome-scale metabolic networks has resulted in models capable of reproducing experimentally observed biomass yield/growth rates and predicting the effect of alterations in metabolism for biotechnological applications. The existing studies rely on modifying the metabolic network of an investigated organism by removing or adding reactions taken either from evolutionary similar organisms or from databases of chemical reactions (e.g., KEGG). A potential disadvantage of these knowledge-driven approaches is that the result is biased towards known reactions, as they do not account for the possibility of including novel enzymes, together with the reactions they catalyze.

Results: Here, we explore the alternative of increasing biomass yield in three model organisms: *Bacillus subtilis*, *Escherichia coli*, and *Hordeum vulgare*, by applying small, chemically feasible network modifications. We use the predicted and experimentally confirmed growth rates of the wild-type networks as reference values and determine the effect of replacing existing reactions by mass-balanced, thermodynamically feasible reactions on the predicted growth rate by using flux balance analysis.

Conclusions: While many replacements of existing reactions naturally lead to a decrease or complete cease of biomass production, in each of the three organisms we find feasible modifications which facilitate a significant increase in this biological function. We focus on modifications with feasible chemical properties and a significant increase in biomass yield. The results demonstrate that small modifications are sufficient to substantially alter biomass yield in the three organisms. The method can be used to predict the effect of targeted modifications on the yield of any set of metabolites, thus providing a computational framework for metabolic engineering.

6.1 Background

Current estimates on the completeness of the knowledge of proteomes suggest that the function of approximately 50-70% of the proteins expressed in organisms is known (Hanson et al., 2010). Consequently, large fractions of enzymatic pathways remain undiscovered, and genome-scale metabolic network reconstructions are inherently incomplete (Breitling et al., 2008; Yamada and Bork, 2009). Nevertheless, such models include sufficient level of detail to computationally predict the growth of microorganisms under different environmental conditions (Edwards et al., 2001), detect missing reactions (Quek and Nielsen, 2008), or discover genetic modifications resulting in a desired phenotype with biotechnological applications (Oliveira et al., 2005; Sohn et al., 2010).

Metabolic engineering aims at designing or re-programming the genetic information of an organism with clinical and biotechnological applications. The existing approaches for modifying an organism's metabolic system rely either on knocking out enzyme-coding genes (Burgard et al., 2003; Alper et al., 2005; Lee et al., 2007) or introducing genes, together with the corresponding reactions, from other organisms (Sohn et al., 2010; Bar-Even et al., 2010; Wang et al., 2011). The advantage of relying on the entire set of known biochemical reactions lies in the potential to use them in various experimental applications, since the introduced reactions are known to be chemically feasible. However, given the large fraction of unknown chemical reactions in biological systems, and the immense space of possible macromolecules potentially catalyzing chemical reactions (Dobson, 2004), the most promising targets for metabolic engineering may not be found among the already known and characterized enzymes, but by systematic screening for chemically feasible, but so-far unknown reactions.

We present an approach for systematic generation of novel reactions and evaluate its ability to modulate, and particularly increase, biomass yield. The introduced reactions are chemically feasible, as they satisfy mass conservation and thermodynamic constraints under standard conditions, and make use only of the compounds already present in the analyzed network. We investigate the metabolic networks of three model organisms, for which growth was predicted *in silico* and experimentally validated: *Bacillus subtilis* (Oh et al., 2007), *Escherichia coli* (Feist et al., 2007), and seeds of *Hordeum vulgare* (Grafahrend-Belau et al., 2009). Each of these organisms has several important agricultural or biotechnological applications: *B. subtilis* is used for food and enzyme production and has been genetically engineered for producing riboflavin and polyhydroxyalkanoates (Schallmey et al., 2004; Perkins et al., 1999; Wang et al., 2006); *E. coli* has a long history of biotechnological applications, such as: production of insulin, lycopene, and succinic acid (Goeddel et al., 1979; Alper et al., 2005; Lee et al., 2005), and is currently explored for its use in producing polymers and biofuels (Atsumi et al., 2008; Bond-Watts et al., 2011; Yim et al., 2011); *Hordeum vulgare* has been genetically engineered for enhanced breeding properties, protein synthesis, food and cellulose production (Horvath et al., 2000, 2001; von Wettstein et al., 2000; Patel et al., 2000).

We point out that our approach is not restricted to optimizing biomass yield, thus allowing the detection of reactions which, when introduced into the respective network, improve any metabolic objective of interest.

6.2 Methods

6.2.1 Generating chemically feasible reactions

We modified the reactions in the wild-type networks by replacing one reaction at a time by a new reaction. In order to obtain realistic chemical reactions, we extended the recent method for mass-balanced randomization of metabolic networks (Basler et al., 2011a). The method generates a new reaction from an existing reaction by replacing its substrates and products by compounds from within the network, while preserving the mass-balance equation, i.e., the number of substrate atoms equals the number of product atoms:

$$\sum_{e \in E_r} s_{e,r} \cdot m_e = \sum_{p \in P_r} s_{p,r} \cdot m_p, \tag{6.1}$$

where E_r is the set of substrates and P_r the set of products of r, m_e, m_p are the vectors of sum formulas of e and p, respectively, and $s_{e,r}$, $s_{p,r}$ their stoichiometric coefficients. As an example for replacing an individual substrate, consider the Aldose 1-epimerase reaction in *E. coli*: β-D-Galactose $\rightarrow$ D-Galactose, with $m_{\beta-D-Galactose} = m_{D-Galactose} = \mathrm{C_6H_{12}O_6}$. Then, as Glyceraldehyde with $m_{Glyceraldehyde} = \mathrm{C_3H_6O_3}$ participates in the network, the method may generate the mass-balanced reaction 2 Glyceraldehyde $\rightarrow$ D-Galactose, which satisfies Equation 6.1, as $2\ \mathrm{C_3H_6O_3} = \mathrm{C_6H_{12}O_6}$. In addition to substituting individual substrates or products, the method also allows more complex substitutions involving pairs of substrates or products, yielding a large number of possible substitutions.

While the presence of the involved compounds and the requirement of mass-balance ensure the reaction can in principal take place, it may still be thermodynamically infeasible. Unfortunately, the current genome-scale metabolic networks do not contain information about the physiological conditions under which individual reactions may occur. However, under the assumption of standard conditions (pH=7, T=298.15K), the thermodynamic feasibility of a reaction can still be estimated from the chemical structure of the involved molecules using the group contribution method (Mavrovouniotis, 1991; Tanaka et al., 2003; Henry et al., 2006). The estimated standard Gibbs free energy change of a reaction, $\Delta_r G^0_{est}$, can be calculated from the corresponding estimates of the Gibbs free energy of formation of its substrates, $\Delta_f G^0_{est}(e)$, and products, $\Delta_f G^0_{est}(p)$, as follows:

$$\Delta_r G^0_{est} = \sum_{p \in P_r} s_{p,r} \cdot \Delta_f G^0_{est}(p) - \sum_{e \in E_r} s_{e,r} \cdot \Delta_f G^0_{est}(e). \tag{6.2}$$

Thus, the thermodynamic feasibility of a reaction can be estimated only from its stoichiometry and the chemical structure of its substrates and products. We obtained $\Delta_f G^0_{est}$ for all compounds in KEGG (Vassily Hatzimanikatis, personal communication) and mapped them to the compounds of the three analyzed metabolic networks. This further facilitated the prediction of the thermodynamic feasibility for the newly generated reactions.

Given the lack of information on physiological conditions for the reactions of a network, we consider a generated reaction infeasible, if its Gibbs free energy change, $\Delta_r G^0_{est}$, is larger than the energy change of any other reaction in the network. In this case, it is unlikely that the organism is able to provide sufficient energy for its activation, and we discard it in the further analysis.

6.2.2 Calculating biomass yield

Most existing studies rely on flux balance analysis (FBA) (Varma and Palsson, 1994) in order to predict the effect of pathway modifications. FBA allows to calculate an optimal flux of metabolic species through the network for a given objective function under the assumption that the system is at steady state.

In order to apply FBA, one needs to specify the stoichiometric matrix, S, containing the stoichiometric coefficients $s_{i,j}$ of compound i in reaction j and including the reaction stoichiometries of import and export reactions. Under the assumption that the network operates at steady state, one can then calculate the optimal flux distribution by solving the linear program:

$$\begin{aligned} \max \quad & v_b \\ \text{s.t.} \quad & S \cdot v = 0 \end{aligned}$$

where v is the vector of reaction fluxes, and v_b the entry in v corresponding to the metabolic objective function.

In order to predict the growth rate of an organism, v_b is chosen such that the b-th column in S describes the consumption of biomass precursors. By applying FBA to the metabolic networks with experimentally validated nutrient uptake and growth rates, we can calculate the growth rates of the wild-type networks and their modified variants.

6.2.3 Screening for feasible synthetic reactions

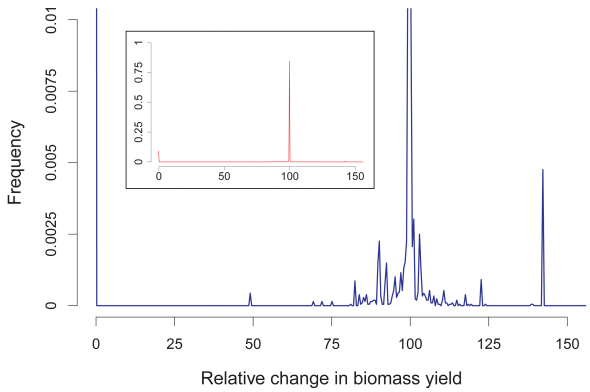
First, we calculate the optimal biomass yield for the metabolic networks of each analyzed organism, in order to reproduce the experimental results and obtain a reference value of the wild-type networks. We then generate all possible networks obtained by substituting individual substrates or products, and pairs thereof, in each reaction (see Basler et al., 2011a). Thermodynamically infeasible reactions are neglected, as described in Section 6.2.1. The total number of remaining modified networks is 20579 for *B. subtilis*, 23835 for *E. coli*, and 722 for the smaller network of *H. vulgare* seeds. For these networks, we calculate the biomass yield by using the import/export fluxes and biomass precursors of the corresponding original network.

6.3 Results and Discussion

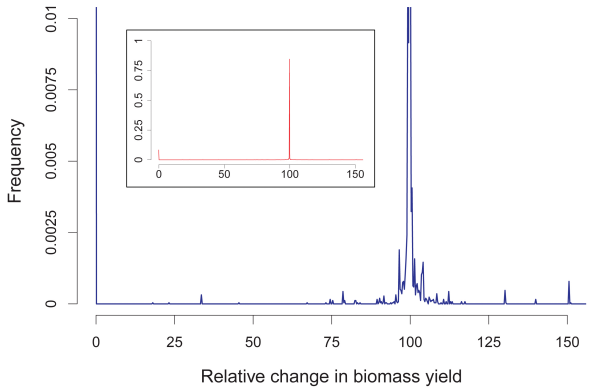
We determine the distribution of biomass yield in the modified networks relative to the wild-type network (Figure 6.1). We find that most modifications do not affect the optimal biomass yield of the three analysed organisms, indicating that only few reactions contribute to optimal growth. This is most pronounced in the genome-scale metabolic networks, where only 12.8% (*B. subtilis*), respectively 11% (*E. coli*) of the modifications affect biomass yield, suggesting that most reactions may be needed for other objectives, such as response to environmental changes or stress. In contrast, the smaller and more specialized, organ-specific network of *H. vulgare* seeds requires a larger fraction of the reactions for producing biomass, as it is affected by 31.3% of the modifications.

Similarly, in *B. subtilis* and *E. coli*, 8-9% of modifications result in a complete loss of the capacity to produce biomass, which corresponds to a non-viable phenotype. In *H. vulgare*, this ratio is 24.4%, indicating this metabolic network is highly sensitive to small modifications.

(a)



(b)



(c)

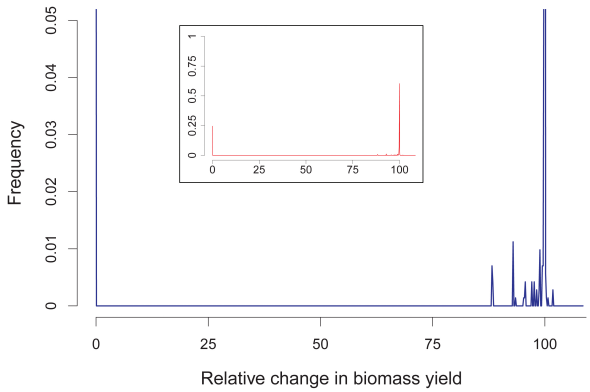


Figure 6.1: Distributions of the relative change in biomass yield after modifying the wild-type networks of (a) *B.subtilis*, (b) *E. coli*, and (c) *H. vulgare* seeds. In each organism, most modifications do not affect biomass yield, while several lead to a complete loss, and some to a significant decrease or increase. The main panels are scaled for clarity; the inlays show the full frequency ranges on the y-scale.

Modifications with largest increase in biomass yield		
Original equation	Modified equations	Increase
(a) L-Malate + NADP → Pyruvate + CO_2 + NADPH	2 L-Malate + 2 NADP → D-Malate + Oxaloacetate + 2 NADPH	42.5%
(b) Fumarate + H_2O ↔ L-Malate	Methylisocitrate + Bicarbonate ↔ 2 L-Malate	51.3%
(c) Glutamate + NAD ↔ 2-Oxoglutarate + NH_3 + NADH	Glutamate + NAD ↔ Glycine + Pyruvate + NADH	1.9%

Table 6.1: Equations of the original and modified reactions with the highest increase in biomass yield. (a) Malic enzyme reaction in *B. subtilis* (EC 1.1.1.40), (b) Fumarate hydratase reaction in *E. coli* (EC 4.2.1.2), (c) Glutamate dehydrogenase reaction in *H. vulgare* (EC 1.4.1.2). In *B. subtilis*, 98 other modifications give the same biomass yield, and in *H. vulgare* one more.

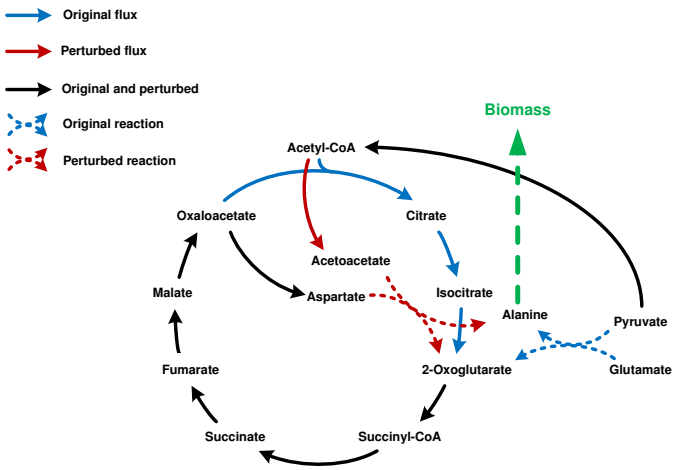


Figure 6.2: Modification of the Alanine transaminase reaction in the TCA cycle leading to a 5.9% increase in biomass yield in *B. subtilis*. From the original reaction, Pyruvate + Glutamate ↔ 2-Oxoglutarate + L-Alanine (dashed blue), we have generated the reaction Acetoacetate + Aspartate ↔ 2-Oxoglutarate + L-Alanine (dashed red). Calculated fluxes in the original network are shown as blue arrows, fluxes in the modified network as red arrows. Black arrows indicate a flux in both networks.

While several modifications lead to a decrease or total loss of biomass yield (*B. subtilis*: 11% *E. coli*: 9.5% *H. vulgare*: 29.6%), only few modifications allow a significant increase (*B. subtilis*: 1.8% *E. coli*: 1.4% *H. vulgare*: 0.3%) . This indicates that the wild-type networks are strongly optimized, though not optimal, with respect to biomass yield. Interestingly, the highest level of optimization is observed in the organ-specific network of *H. vulgare*, where only two modifications allow for an increase in biomass yield by more than 1%. Table 6.1 shows the modifications with the highest increase in biomass yield for each organism.

To illustrate the underlying mechanism leading to increased biomass yield, we show how a modification of the Alanine transaminase reaction (EC 2.6.1.2) affects the metabolite fluxes of the TCA cycle and leads to a 5.9% increase in biomass yield in *B. subtilis* (Figure 6.2). Instead of using Pyruvate and Glutamate as substrates for producing L-Alanine, a direct biomass precursor, the modified network uses Acetoacetate and Aspartate in order to produce L-Alanine more efficiently.

For comparison, we repeated the analysis by adding instead of replacing reactions, i.e., the original reaction and its modified variant are both part of the modified network. Naturally, when adding a reaction, biomass yield may only increase, but never decrease. Interestingly, we find that, in *B.*

Modifications with largest biomass yield when added to *E. coli*	
Original equation	Modified equations
D-Lactate + NAD ↔ Pyruvate + NADH + H	3 D-Lactate + 3 NAD ↔ Pyruvate + 3 NADH + 2-dehydro-3-deoxy-D-galactonate
	3 D-Lactate + 3 NAD ↔ Pyruvate + 3 NADH + 2-Dehydro-3-deoxy-D-gluconate
	4 D-Lactate + 4 NAD ↔ 5-Dehydro-4-deoxy-D-glucarate + 4 NADH + Glycogen

Table 6.2: D-Lactate dehydrogenase (EC 1.1.1.28) and three modified reactions, each resulting in a 214% increase in biomass yield when added to the network.

subtilis and *H. vulgare*, none of the modifications by adding a reaction allow for a higher biomass yield compared to the previously analyzed replacements of reactions. Only in *E. coli*, we find three modifications with a further increased biomass yield (+214%) when adding the generated reactions to the network (Table 6.2).

Most of the identified reactions are not present in KEGG nor Brenda, but might well occur in a less characterized or unknown organism, as they are biochemically feasible. In addition, the corresponding enzymes could be artificially synthesized and introduced into the organism in order to experimentally validate the predicted growth effect. We point out that the identified reactions could not have been found using a standard knowledge-based approach.

6.4 Conclusions

We have presented a new approach for systematic detection of novel feasible reactions which alter biomass yield. We found that the three analyzed metabolic networks are strongly optimized, particularly the more specialized network of *H. vulgare* seeds. Nevertheless, we identified several reactions which, when introduced into the organism, are predicted to further increase biomass yield. By using a different objective function, the same approach may be directly applied to generate reactions facilitating the improved production of valuable compounds, or a parallel suppression of the production of toxic compounds.

As all reactions satisfy basic mass conservation and thermodynamic constraints, they may be potentially catalyzed by suitable enzymes. Strategies for their design and synthesis may be developed in three ways: (1) targeted search for a gene encoding the enzyme in a less characterized organism; (2) discovery of the enzyme in the environment; (3) targeted synthesis of the enzyme using methods of chemical engineering (Qi et al., 2001).

We further point out that, in order to obtain the predicted effect of introducing a reaction in the network, one does not necessarily need to catalyze the exact stoichiometry of the reaction. Alternatively, concatenated reactions having the same net consumption and production as the identified reaction will have the same effect, broadening the possibilities for obtaining suitable enzymes.

Another possible application is the automated curation of metabolic networks. If an analyzed network is not capable of reproducing the experimental observations, it is likely to be incomplete. The generation of novel reactions which allow a model to better fit experimental data might then give hints to which important reactions are missing. We thus believe that the approach provides a valuable tool for reconstruction of metabolic networks and their use in metabolic engineering or drug development.

Chapter 7

Conclusions & Outlook

7.1 Summary

Motivated by the lack of a biologically meaningful null model, I have developed a computational method for mass-balanced randomization of metabolic networks. In comparison, the existing approaches either disregard basic physical principles or do not aim at estimating the significance of network properties (see Section 1.3). The presented method is based on the hypothesis that metabolic network properties arise as a consequence of evolutionary pressure and physical constraints. By preserving the basic physical constraints, the method can be used in demonstrating the independence of a network property from physical principles, and thus in estimating its significance with regard to evolutionary pressure. The method may therefore be a reasonable proxy for quantifying functional importance in metabolic networks.

The randomization algorithm was presented and its complexity and uniformity properties analyzed (Chapter 2). The method was further implemented as a user-friendly tool, which allows efficient randomization and calculation of several important network properties (Chapter 3). It was then applied to assessing the dependence of the thermodynamic properties on mass balance (Chapter 4). A principal goal of the thesis was to address the question whether the salient network properties, analyzed frequently in complex network research, can be justifiably used as determinants of metabolic network function, and thus for drawing biologically meaningful conclusions. The presented results confirm this claim in six genome-scale metabolic networks. Based on the proposed randomization algorithm, a novel network property was developed, which promises to reveal the importance of metabolic reactions for viability of organisms, as measured by the global impact of their knockout (Chapter 5). Finally, the method was extended to the identification of previously unknown metabolic reactions, which improve the predicted growth rates of biotechnologically important organisms, when introduced into the network (Chapter 6).

7.2 Uniform mass-balanced randomization

In Section 2.3.2, we demonstrated the uniformity of the algorithm for mass-balanced randomization, if only individual compounds are substituted. The general algorithm allows for substituting individual compounds and pairs of compounds, which greatly increases the sample space, but does not result in a strictly uniform randomization. Therefore, we derived worst-case bounds for the probabilities that networks are sampled almost uniformly at random after a given number of steps,

and empirically validated the results. Uniformity of sampling is an important requirement for any randomization approach, as it provides the basis for an unbiased measure of significance.

Switch randomization in its general form is not uniform, which can be seen by a simple example. Consider the directed graph $G^0 = (V, E^0)$ with $V = \{a, b, c, d\}$ and $E^0 = \{(a, b), (b, c), (c, d), (c, a)\}$. Then, from G^0, only the graph $G^1 = (V, E^1)$ with $E^1 = \{(a, d), (b, c), (c, b), (c, a)\}$ can be generated by switching (a, b) with (c, d) (we do not allow self-edges, such as (a, a)). However, from G^1, two graphs can be generated: either by switching (a, d) with (c, b) to generate G^0, or by switching (a, d) with (b, c) to obtain a new graph, G^2. Thus, the transition degree of G^0 is 1, while the transition degree of G^1 is 2. Consequently, the transition graph is not regular and the corresponding Markov chain does not converge to a uniform stationary distribution, as G^1 is sampled with a higher probability than G^0. Hence, switch randomization is not uniform for directed graphs.

It is easy to circumvent this problem by allowing for "void" switches, which do not change the graph but are counted as an iteration (corresponding to self-loops in the transition graph): in the example, a void switch is created for G^0, so that the transition degrees of both G^0 and G^1 are 2, and the randomization becomes uniform (Cobb and Chen, 2003; Ying and Wu, 2009). Unfortunately, this approach requires prior knowledge of all degrees in the transition graph, which is unrealistic for the mass-balanced randomization algorithm, as the sample spaces are prohibitively large (see equation 2.5 on page 28 and Table B.3).

A more general approach for obtaining a uniform sampling is given by the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). Before applying a randomization step $t \rightarrow t+1$, the transition degree $d(G^t)$ of the graph G^t is compared to the transition degree of the new graph, $d(G^{t+1})$. The probability for applying the modification is chosen as $p_{t \rightarrow t+1} = \min\{1, \frac{d(G^t)}{d(G^{t+1})}\}$. Thereby, the probabilities for obtaining graphs with a larger transition degree are adjusted to yield a uniform stationary distribution of the Markov-chain, and thus a uniform sampling. The mass-balanced randomization algorithm with single and pair substitutions (Algorithm 2.1 on page 24) can be easily adjusted in accordance with this argument to yield a strictly uniform randomization. To this end, the total number of possible substitutions of the original network G^0, $d(G^0) = \sum_{r \in V_r} \Psi(r)$, where $\Psi(r)$ is the set of all possible substitutions for a reaction r, must be determined once at the beginning of the algorithm. The time complexity of this calculation is in $O(|V_r| \cdot \Delta^2 \cdot \sigma^{max} \cdot n)$ (Δ is the maximum reaction degree of G, σ^{max} is the size of the largest mass equivalence class, and n the number of considered chemical elements, see Section 2.3.1). In every subsequent randomization step, the number of possible substitutions, $\Psi(r')$, must be calculated for the new reaction r', before applying the substitution, in order to adjust the probability as mentioned above. This does not require any additional calculation time, as the original algorithm also calculates the possible substitutions for a chosen reaction in each step. Thus, the computational complexity of the algorithm would increase to $O((t + |V_r|) \cdot \Delta^2 \cdot \sigma^{max} \cdot n)$. However, note that we have chosen $t \gg |V_r|$ (Algorithm A.2 on page 73), so that the applicability of the algorithm to genome-scale networks should not be hampered (as can be estimated from Table B.1).

7.3 Preserving thermodynamic constraints

The presented randomization algorithm preserves mass balance of reactions, which is a fundamental physical requirement constraining metabolic networks. Thermodynamic laws constitute further important requirements for the feasibility of biochemical processes. The Gibbs free energy change of reactions under standard conditions, $\Delta_r G$, was shown to remain in a realistic range

in mass-balanced randomized networks (Section 5.2.1). Nevertheless, the distribution of $\Delta_r G$ does change (Figure 5.2 on page 48). An extension of the randomization algorithm to preserving thermodynamic constraints could aim at preserving the distribution of $\Delta_r G$.

Preserving the distribution of $\Delta_r G$ exactly is not straightforward, as the randomized reactions in general have different values of $\Delta_r G$ from the reactions in the original network. However, the distribution can be easily approximated by choosing the direction of randomized reactions accordingly, as the direction determines the sign of its $\Delta_r G$. To this end, first the $\Delta_r G$ of all reactions in the original network are partitioned into intervals $X_i = [x_i, x_{i+1}), x_i \in \mathbb{R}, i = 1, ..., n$. Then, for a reaction r' obtained by randomizing an existing reaction, its $\Delta_r G$, denoted as $\delta(r')$, is estimated from the molecular structures of the substrates and products and their stoichiometric coefficients (Mavrovouniotis, 1991, Section 4.1). Let $\delta(r') \in X_i$ and $-\delta(r') \in X_j$, and let $|X_i|$, $|X_j|$ denote the numbers of reactions in the corresponding partitions. The direction of r' is then reversed (i.e., substrates and products are swapped) with probability

$$p_{rev} = \frac{|X_j|}{|X_j| + |X_i|}, \tag{7.1}$$

Thus, the direction of r' is chosen according to the probabilities of a reaction r in the original network to have a $\Delta_r G$ similar to $\delta(r')$, respectively $-\delta(r')$. Consequently, when choosing proper interval bounds for X_i, the distribution of $\Delta_r G$ is approximated as closely as possible by choosing the direction of randomized reactions.

Note that, for brevity, the abovementioned approach only applies to irreversible reactions. The distribution of $\Delta_r G$ for reversible reactions may be approximated likewise by deriving the probabilities for choosing the reversibility of randomized reactions. In addition, randomized reactions with $\delta(r')$ far from the $\Delta_r G$ of any reaction in the original network may be skipped in order to avoid generating reactions with unrealistic thermodynamic properties. Note that these considerations may affect the uniformity of the randomization algorithm, as reactions which are thermodynamically similar to the reactions in the original network are more likely to be generated. Nevertheless, generating randomized networks while preserving thermodynamic properties may yield an interesting new tool for future analyses.

7.4 Compartments

Within this work, mass-balanced randomization was applied to seven genome-scale metabolic networks, most of which were obtained from dedicated publications, where particular effort was made in refining the model (see Table B.1). Nevertheless, the networks were continuously improved and extended by additional details. For example, in 2009, only two of the seven analyzed networks contained information on the subcellular localization of metabolites and reactions in compartments (Herrgård et al., 2008; Feist et al., 2007), while the remaining five did not (Oh et al., 2007; Keseler et al., 2009; Swarbreck et al., 2008; May et al., 2008; Ma et al., 2007). Meanwhile, several metabolic network reconstructions include this information (Duarte et al., 2007; de Oliveira Dal'Molin et al., 2010, and more recent versions of Keseler et al., 2009; Swarbreck et al., 2008; May et al., 2008). Ongoing empirical analyses revealed that, in a recent genome-scale metabolic network of *E. coli* (Orth et al., 2011), the average path length increases, while the clustering coefficient decreases when taking compartments into account. Thus, it would be intriguing to re-assess the significance of the small-world and other properties using more recent network reconstructions. The implementation of mass-balanced randomization already allows for randomizing compartmentalized networks (Chapter 3), rendering this a straightforward task.

7.5 Network properties

In Chapter 5, mass-balanced randomization was applied to estimating the evolutionary importance of the average path length, clustering coefficient, biosynthetic capabilities, and to determining reactions important for viability of organisms. The method can be directly applied to testing the significance of any property defined on the structure of metabolic networks. For example, small subnetwork patterns, also called motifs, represent the elementary building blocks which may facilitate specific molecular functions in biological networks (Milo et al., 2002; Alon, 2003). Analyzing the evolutionary importance of different motifs by mass-balanced randomization would be an interesting future study. However, care must be taken when analyzing patterns which are sensitive to local changes in the networks, as switch randomization was found to generate different network motifs in a highly correlated manner (Ginoza and Mugler, 2010). Therefore, identification of correlations in the local structure of mass-balanced randomized networks must precede the analysis of local patterns, such as motifs or cycles of a given length, in order to remove artificial redundancies.

Flux balance analysis is a widespread computational approach for determining metabolic fluxes in microorganisms (Varma and Palsson, 1994). Under the steady-state assumption (Equation 1.1 on page 12), the method determines a flux distribution which optimizes a given objective function. Usually, the optimization of biomass production is used as objective function, as this reflects the optimization of an organism for growth (Edwards et al., 2001). Clearly, growth is a complex biological function, which results from a long history of evolutionary optimization. Thus, one would expect the optimal flux distribution determined using flux balance analysis to be of outstanding evolutionary importance.

A possible application of mass-balanced randomization is therefore to test the evolutionary importance of the growth rate by applying flux balance analysis to randomized networks. An initial observation is that, in fully randomized networks of *H. vulgare* seeds (Grafahrend-Belau et al., 2009), there is no possible steady-state flux, and, thus, growth is impossible. At an average, when applying only 10% of the usual number of substitutions, the ratio of randomized networks allowing for growth sharply drops below 10% (Figure 7.1). While this result may seem disappointing, it is not unexpected when recalling the hypothesis underlying mass-balanced randomization (Section 1.5): the randomized networks satisfy basic physical principles, but lack any biological function. Consequently, complex biological functions such as growth are not found in randomized networks, which may also be interpreted as a confirmation of the relation between evolutionary optimization and growth. It is thus of more interest to analyze network properties for which the relation to an important biological function is less obvious, such as alternative definitions of feasible metabolic pathways (Pitkänen et al., 2005). Nevertheless, these results inspired the extension of the randomization algorithm to identifying novel reactions, which improve the growth rate in organisms of biotechnological importance, presented in Chapter 6.

7.6 Application to metabolic engineering and disease treatment

In Chapter 6, the randomization algorithm was extended to identifying feasible reactions, which improve the growth rates of *E. coli*, *B. subtilis*, and *H. vulgare*, as predicted by flux balance analysis. For a biotechnological application, however, an enzyme catalyzing the identified reaction must be introduced into the organism of interest. Therefore, the next step is to identify suitable enzymes catalyzing the identified reactions. This may be achieved by three consecutive strategies: (1) Search for an identified reaction and its catalyzing enzyme in public databases. If such an

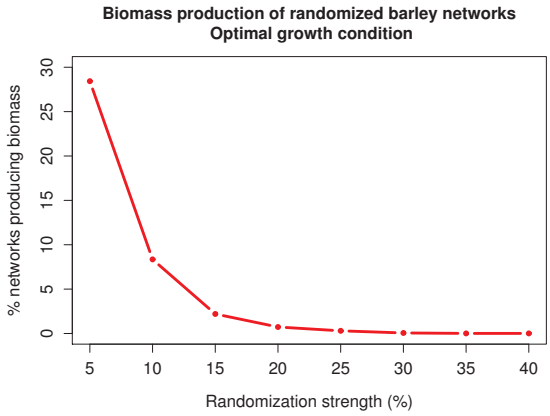


Figure 7.1: Application of flux balance analysis to randomized networks of *H. vulgare* seeds under optimal growth conditions. Randomization is applied gradually, demonstrating that the average number of randomized networks capable of carrying a steady-state flux drops sharply with progressing randomization. Randomization strength refers to the percentage of the default number of applied substitutions, here $t = \lceil |V_r| \cdot \overline{d}(V_r)^2 \rceil = 3395$ (see Algorithm A.2).

enzyme is found, the coding gene may be a promising candidate for transformation into the analyzed organism and experimentally testing the predicted increase in growth; (2) If no such enzyme is found, screen for suitable enzymes by predicting the function of protein structures contained in protein databases (e.g., Berman et al., 2000). A computational method for screening protein structures could be adopted based on Hermann et al. (2006); (3) Finally, if no matching protein structure is found for a promising candidate reaction, protein databases could be searched for proteins which are suitable for designing the enzyme by chemical engineering approaches (Qi et al., 2001).

Notably, the approach presented in Chapter 6 can easily be applied to any metabolic objective function of interest. Thus, not only reactions improving growth could be identified, but also those that improve the efficient production compounds valuable for agricultural or medical purposes. In addition, the approach could be applied to identifying novel enzymatic drugs. The applicability of a genome-scale, tissue-specific cancer metabolic network to the prediction of cancer drug targets was recently demonstrated (Folger et al., 2011). Enzymes, in contrast to small molecules, are promising drug candidates due to their high affinity and specificity, and they have been successfully applied to the treatment of different cancers (Vellard, 2003). Thus, metabolic reactions and their catalyzing enzymes, predicted to suppress growth in cancer cells, may be promising candidates for novel anticancer drugs.

7.7 Extension to other biological networks

The presented method was developed specifically for the analysis of metabolic networks, as the mass balance principle applies to chemical reactions. However, the general idea of physically constrained randomization is applicable to all kinds of biological networks, such as gene-regulatory, protein-protein-interaction, and signaling networks, as they are all shaped by basic physical principles and evolutionary pressure. Thus, similar null models could be specifically designed for other biological networks by preserving their governing physical principles, with the aim of assessing the functional importance of network properties.

For example, in gene-regulatory networks, genes are represented by vertices, and edges represent regulatory relationships between genes, such as transcriptional activation or inhibition (Karlebach and Shamir, 2008). Clearly, edges should be directed and weighted, depending on the direction and type of regulation. The key entities are the transcription factors, i.e., proteins which contain one or more DNA-binding domains. The number and type of binding domains of a transcription factor determines the cis-regulatory elements on the DNA it can bind to and its specificity. Thus, the number of binding domains and their target DNA sequences may be regarded as physical principles. On the other hand, the particular genes located downstream of a regulatory site may be regarded a result of evolutionary pressure, as they are determined by evolutionary events, such as gene duplication. Consequently, a randomization approach for gene-regulatory networks could aim at preserving the number of binding domains of transcription factors and their target DNA strands, while randomizing the location of target genes.

In signaling networks, vertices represent proteins and protein complexes in different phosphorylation states, while edges represent the phosphorylation of proteins and the formation of protein complexes. Thus, signaling networks can be regarded as chemical reaction networks for which the principles of mass balance apply, similar to metabolic networks (Papin and Palsson, 2004). Consequently, a randomization approach for signaling networks could impose similar constraints of mass balance, where phosphate groups are transferred among proteins in a balanced fashion, and protein complexes are formed by the participating individual proteins.

To conclude, the careful design of null models which account for the physical principles of a particular class of biological networks may give important insights into their structure-function relationships and open up valuable new strategies for their modification.

Appendix A

Algorithms

Algorithm: Mass equivalence class calculation

Input:
Set of compounds, V_c

Output:
Mass equivalence classes, $\sigma = \{\sigma(c), \sigma(c,k)\},\ (c,k) \in V_c \times V_c,\ c \neq k$

σ := $\{\}$

$\forall c, k \in V_c$: $\sigma(c)$:= $\{\}$, $\sigma(c,k)$:= $\{\}$

foreach $c \in V_c$ **do**

1 **if** $\sigma(c) \notin \sigma$ **then**

2 add $\sigma(c)$ to σ

3 add c to $\sigma(c)$

4 **foreach** $\sigma(x) \in \sigma$ **do**

5 **foreach** $k \in \sigma(x), k \neq c$ **do**

6 **if** $\sigma(c,k) \notin \sigma$ **then**

7 add $\sigma(c,k)$ to σ

8 add (c,k) to $\sigma(c,k)$

Algorithm A.1: Algorithm for calculating the mass equivalence classes for all individual compounds and pairs of compounds. Lines 1 and 6 involve testing whether a mass equivalent compound, respectively pair of compounds, is already in σ; likewise, lines 3 and 8 require retrieving the corresponding mass equivalence class from σ. Both can be done in constant time when using a hash map for σ, with the basis of the mass vector(s) as hash key, as the basis uniquely identifies a mass equivalence class (see Tables 1 and 2 in the main manuscript). Thus, the time complexity of the algorithm is in $O(|V_c|^2)$, as we iterate over each pair of compounds exactly once in line 5.

Algorithm: Mass balanced randomization of metabolic networks

Input:
Mass balanced metabolic network, $G = (V_c \cup V_r, E)$,
Mass equivalence classes, $\sigma = \sigma(c) \cup \sigma(c,k),\ (c,k) \in V_c \times V_c,\ c \neq k$,
Set of preserved compounds, $D \subset V_c$,
Number of iterations, $t \in \mathbb{N}^+$
Output:
Randomized mass balanced network
Repeat t times:

1 Choose a reaction $r \in V_r$ uniformly at random
2 **foreach** $c \in r \setminus D$ **do**
3 **foreach** $c' \in \sigma(c), c' \notin r \cup D$ **do**
4 add (c, c') to $\Psi_s(r)$
5 **foreach** $(c,k) \in (r_{in} \times r_{in}) \cup (r_{out} \times r_{out}), c, k \notin D$ **do**
6 **foreach** $(c', k') \in \sigma(c,k), c', k' \notin r \cup D$ **do**
7 Let $A = (m_{c'}, m_{k'})$ be the $(n \times 2)$ matrix of mass vectors of length n
8 Solve $As = b$ with $b = s_{c,r} \cdot m_c + s_{k,r} \cdot m_k$
9 **if** *there is a solution* $s_1, s_2 \in \mathbb{N}^+$ **then**
10 add $(c, k, c', k', s_1, s_2, 1)$ to $\Psi_p(r)$
11 **else if** *there is a solution* $s_1, s_2 \in \mathbb{Q}^+$ **then**
12 Let $f > 0$ be the smallest integer, such that $fs_1, fs_2 \in \mathbb{N}^+$
13 add $(c, k, c', k', s_1, s_2, f)$ to $\Psi_p(r)$
14 $\Psi(r) := \Psi_s(r) \cup \Psi_p(r)$
15 Choose a number $u \in \mathbb{N}^+$ uniformly at random from $[1, |\Psi(r)|]$
16 Let d_u be the u-th substitution in $\Psi(r)$
17 **if** d_u *is an individual substitution* (c, c') **then**
18 **if** c *is a substrate of* r **then**
19 replace the edge (c, r) by (c', r)
20 **else**
21 replace the edge (r, c) by (r, c')
22 Let $f > 0$ be the smallest integer, such that $\frac{f}{m_{c'}} \cdot s_{c,r} m_c \in \mathbb{N}^+$
23 $s_{c',r} := \frac{1}{m_{c'}} \cdot s_{c,r} m_c$
24 Multiply the stoichiometric coefficients of r by f
25 **else if** d_u *is a pair substitution* $(c, k, c', k', s_1, s_2, f)$ **then**
26 **if** c,k *are substrates of* r **then**
27 replace the edges (c, r) and (k, r) by (c', r) and (k', r)
28 **else**
29 replace the edges (r, c) and (r, k) by (r, c') and (r, k')
30 $s_{c',r} := s_1$
31 $s_{k',r} := s_2$
32 Multiply the stoichiometric coefficients of r by f

Algorithm A.2: Detailed algorithm for mass-balanced randomization of a metabolic network. For a randomly chosen reaction, the set of individual compound substitutions (lines 2-4) and the set of pair substitutions (lines 5-13) are determined from the mass equivalence classes. Optionally, a set of preserved compounds D may be specified, *e.g.* cofactors, which remain unmodified. For pair substitutions, it is necessary to determine whether there are stoichiometric coefficients satisfying mass balance (lines 7-8). If there is no rational solution, the pair substitution is neglected. In lines 14-16, a substitution is chosen uniformly at random from the set of all possible substitutions. In lines 18-21 and 26-29, the edges corresponding to the chosen substituted compounds are replaced by new edges connecting the substitutes. For an individual compound substitution, this involves determining the new stoichiometric coefficients of the reaction, which can always be found due to the linear dependence of mass vectors (lines 22-24). For a pair a substitution, the previously determined solution is used (lines 30-32). Note that the stoichiometric coefficients of compounds other than the substitutes are modified only if $f > 1$, which is the case if the substituted (sum of) mass vector(s) is no integer multiple of the new (sum of) mass vector(s) (see Table 3 in the main manuscript). For a full randomization, the number of iterations, t, should be chosen as the number of compounds and pairs of compounds available for substitutions. We use $t = \lceil |V_r| \cdot \overline{d}(V_r)^2 \rceil$ as an upper approximation, where $\overline{d}(V_r)$ is the average (undirected) reaction degree.

Algorithm: Network expansion

Input:
Metabolic network, $G = (V_c \cup V_r, E)$,
Set of initial nutrients, $N \subset V_c$.
Output:
Set of compounds which can be produced from N in G.

```
repeat
1   N* := N
2   foreach r ∈ V_r do
3       if r_in ⊆ N* then
4           N := N ∪ r_out
until N* = N
```

Algorithm A.3: Network expansion algorithm for determining the set of producible compounds (scope) from a set of nutrients in a metabolic network. In order to determine the biosynthetic capabilities of a network, we generated 5000 sets of randomly chosen initial nutrients, and determined their scope. The scope size is the number of producible compounds, $|N^*|$. The scope size distribution (Figures C.9 to C.12) gives the probability $S(m, n)$, that n compounds can be produced from a random set of m nutrients, and is determined empirically by calculating the scope sizes for the 5000 sets of nutrients. We used $m = 8, 16$, and 32, and applied the procedure to the original networks, and each of the randomized networks.

Appendix B

Tables

Network	$\|\mathbf{V_r}\|$	$\|\mathbf{V_c}\|$	$\|\mathbf{V}\|$	$\|\mathbf{E}\|$	CHNOPS	U	U*	rev.	CC	SCC	$\mathbf{T_c}$	$\mathbf{T_{rand}}$	$\mathbf{T_{switch}}$	$\|\sigma_\mathbf{s}\|$	$\|\sigma_\mathbf{p}\|$
B. subtilis (Oh et al., 2007)	855	777	1877	5368	736	6	4	✓	1 (766)	172 (596)	15.4s	75.6m	490.7m	567	114116
S. cerevisiae (Herrgård et al., 2008)	1203	995	3398	11908	596	388	6	–	32 (721)	32 (721)	9.5s	176.1m	2228.6m	474	76076
E. coli (iAF1260) (Feist et al., 2007)	1481	1039	2731	7963	887	0	0	✓	1 (1034)	183 (858)	28.4s	124.9m	1809.5m	702	190230
E. coli (EcoCyc) (Keseler et al., 2009)	1622	1088	4166	10159	785	34	1	✓	42 (1694)	450 (1312)	23.5s	124.1m	1887.2m	611	133933
C. reinhardtii (May et al., 2008)	1541	1377	3772	9391	1054	404	9	✓	21 (1595)	321 (1356)	43.5s	120.1m	1735.8m	780	192586
A. thaliana (Swarbreck et al., 2008)	2508	2190	5647	14363	1790	716	87	✓	20 (2327)	445 (1974)	191.9s	308.2m	5927.0m	1280	482925
H. sapiens (Ma et al., 2007)	2819	2690	7059	19651	1953	574	153	✓	9 (2670)	578 (2122)	165.2s	537.9m	10981.4m	1214	386315

Table B.1: Summary statistics for the seven analyzed genome-scale metabolic networks. $|V_r|$: number of reactions; $|V_c|$: number of compounds; $|\mathbf{V}|$: number of vertices in the bipartite graph; $|\mathbf{E}|$: number of edges in the bipartite graph; CHNOPS: number of compounds consisting only of carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur; **U**: number of mass unbalanced reactions in the original network; **U***: number of mass unbalanced reactions after fixing imbalances due to phosphate or hydrogen; **rev**.: availability of information on the reversibility of reactions; **CC**: number of connected components (number of compounds contained in the largest component); **SCC**: number of strongly connected components (number of compounds contained in the largest component); T_c: runtime for calculating the mass equivalence classes; T_{rand}: runtime for generating 1000 mass balanced randomized networks; T_{switch}: runtime for generating 1000 switch randomized networks; $|\sigma_s|$, $|\sigma_p|$: number of mass equivalence classes for individual and pairs of compounds, respectively. All runtime calculations were performed on a single core of an Intel Xeon Processor E5345 with 2.33GHz and 16GB RAM running Fedora 13 Linux 32-bit and Sun Java SE 1.6.0 update 6 in JVM server mode.

Mass-balanced randomization	L	μ_L	σ_L	p_L	C	μ_C	σ_C	p_C	$p_{D(k)}$	$p_{S(8,n)}$	$p_{S(16,n)}$	$p_{S(32,n)}$
B. subtilis	2.97	3.05	0.02	$2 \cdot 10^{-5}$	0.3	0.21	0.01	$3.5 \cdot 10^{-31}$	$1.9 \cdot 10^{-30}$	$7.5 \cdot 10^{-174}$	0	0
E. coli	3.09	3.17	0.02	$3 \cdot 10^{-6}$	0.37	0.24	0.01	$1.3 \cdot 10^{-100}$	$2.1 \cdot 10^{-27}$	0	0	0
S. cerevisae	2.48	2.55	0.02	$1.5 \cdot 10^{-5}$	0.27	0.21	0.01	$1.1 \cdot 10^{-24}$	$1.3 \cdot 10^{-18}$	$1.5 \cdot 10^{-50}$	$1.3 \cdot 10^{-198}$	$1.1 \cdot 10^{-157}$
C. reinhardtii	3.37	3.42	0.02	0.02	0.21	0.16	0.01	$5.9 \cdot 10^{-19}$	$6.1 \cdot 10^{-26}$	$3.5 \cdot 10^{-52}$	$2.5 \cdot 10^{-254}$	0
A. thaliana	3.44	3.48	0.02	0.02	0.24	0.16	0.01	$1.3 \cdot 10^{-81}$	$1.4 \cdot 10^{-45}$	$3.6 \cdot 10^{-96}$	0	0
H. sapiens	3.09	3.21	0.01	$1.6 \cdot 10^{-22}$	0.27	0.18	0.004	$1.2 \cdot 10^{-141}$	$1.1 \cdot 10^{-86}$	$1.9 \cdot 10^{-84}$	0	0
Switch randomization												
B. subtilis	2.97	2.75	0.02	$2.2 \cdot 10^{-35}$	0.3	0.41	0.01	$1 \cdot 10^{-22}$	n/a	0	0	0
E. coli	3.09	2.82	0.02	$4.5 \cdot 10^{-69}$	0.37	0.4	0.01	$7 \cdot 10^{-5}$	n/a	0	0	0
S. cerevisae	2.48	2.49	0.02	0.77	0.27	0.37	0.01	$1.1 \cdot 10^{-23}$	n/a	0	0	$1.2 \cdot 10^{-261}$
C. reinhardtii	3.37	3.07	0.02	$3 \cdot 10^{-67}$	0.21	0.26	0.01	$6.9 \cdot 10^{-12}$	n/a	0	0	0
A. thaliana	3.44	3.06	0.01	$6.3 \cdot 10^{-169}$	0.24	0.28	0.01	$2.4 \cdot 10^{-6}$	n/a	0	0	0
H. sapiens	3.09	2.81	0.01	$3.5 \cdot 10^{-120}$	0.27	0.34	0.01	$3.3 \cdot 10^{-25}$	n/a	0	0	0

Table B.2: Significance of network properties from mass-balanced randomization (a) and switch randomization (b). The values of the average path length and the clustering coefficient for the genome-scale metabolic networks of six investigated species appear in the columns titled L and C, respectively. The p-values, p_L and p_C, are calculated from the means, μ_L, μ_C, and standard deviations, σ_L, σ_C, of these properties in randomized networks via z-scores. The p-values $p_{D(k)}$ and $p_{S(m,n)}$ for the degree distribution and the scope size distributions are obtained from the two-sample Kolmogorov-Smirnov test applied to the original and joint cumulative distributions of the randomized networks. In each case, the values are obtained from 10,000 randomizations of the original network.

Network	γ	δ	$\overline{d}(\Sigma_G)$	P_{uni}	Ω_s
B. subtilis	1.85	6.91	15641	0.73	$4.4 \cdot 10^{645}$
S. cerevisiae	2.04	6.15	16032	0.82	$3.8 \cdot 10^{610}$
E. coli (iAF1260)	1.89	6.52	21769	0.81	$4.3 \cdot 10^{1089}$
E. coli (EcoCyc)	1.87	7.14	19490	0.80	$3.0 \cdot 10^{957}$
C. reinhardtii	1.77	7.38	25047	0.78	$6.3 \cdot 10^{972}$
A. thaliana	1.63	12.50	54201	0.80	$1.2 \cdot 10^{2033}$
H. sapiens	1.24	22.55	158138	0.53	$2.8 \cdot 10^{2210}$

Table B.3: Scaling coefficient γ of the distribution of differences in degrees of adjacent nodes and their expected degree difference δ were obtained from a random walk, average degree of the transition graph $\overline{d}(\Sigma_G)$ was obtained from sampling 10^3 random walks on each network. P_{uni}: lower bound for the worst-case probability, that nodes from Σ_G are sampled almost uniformly at random after $t = 10^6$ steps; Ω_s: size of the sample space for individual substitutions.

Enzyme	EC no.	BS	EC	CR
L-arab	5.3.1.4	✓	✓	n/a
P-amino	5.3.1.6	✓	✓	-
S-ribose	5.3.1.23	✓	✓	n/a
Gluta	5.4.3.8	✓	✓	✓
Iso	5.4.99.6	✓	✓	n/a

Table B.4: All reactions with centrality p-values ≤ 0.025, obtained from switch randomization, in at least two of *Bacillus subtilis* (**BS**), *Escherichia coli* (**EC**), and *Chlamydomonas reinhardtii* (**CR**). A checkmark indicates that the reaction catalyzed by the enzyme has a significant centrality in the corresponding species; a hyphen indicates not significant; n/a indicates the corresponding enzyme is not annotated for the species. The remaining three species are excluded due to unreasonably large numbers of significant reactions: 2266 (94.2%) in *Saccharomyces cerevisiae*, 1349 (41.2%) in *Arabidopsis thaliana*, 1022 (23.4)% in *Homo sapiens* (both directions of reversible reactions are considered independently). Abbreviations: L-arab: L-Arabinose isomerase, P-amino: Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase, S-ribose: S-methyl-5-thioribose-1-phosphate isomerase, Gluta: Glutamate-1-semialdehyde 2,1-aminomutase, Iso: Isochorismate synthase

Appendix C

Figures

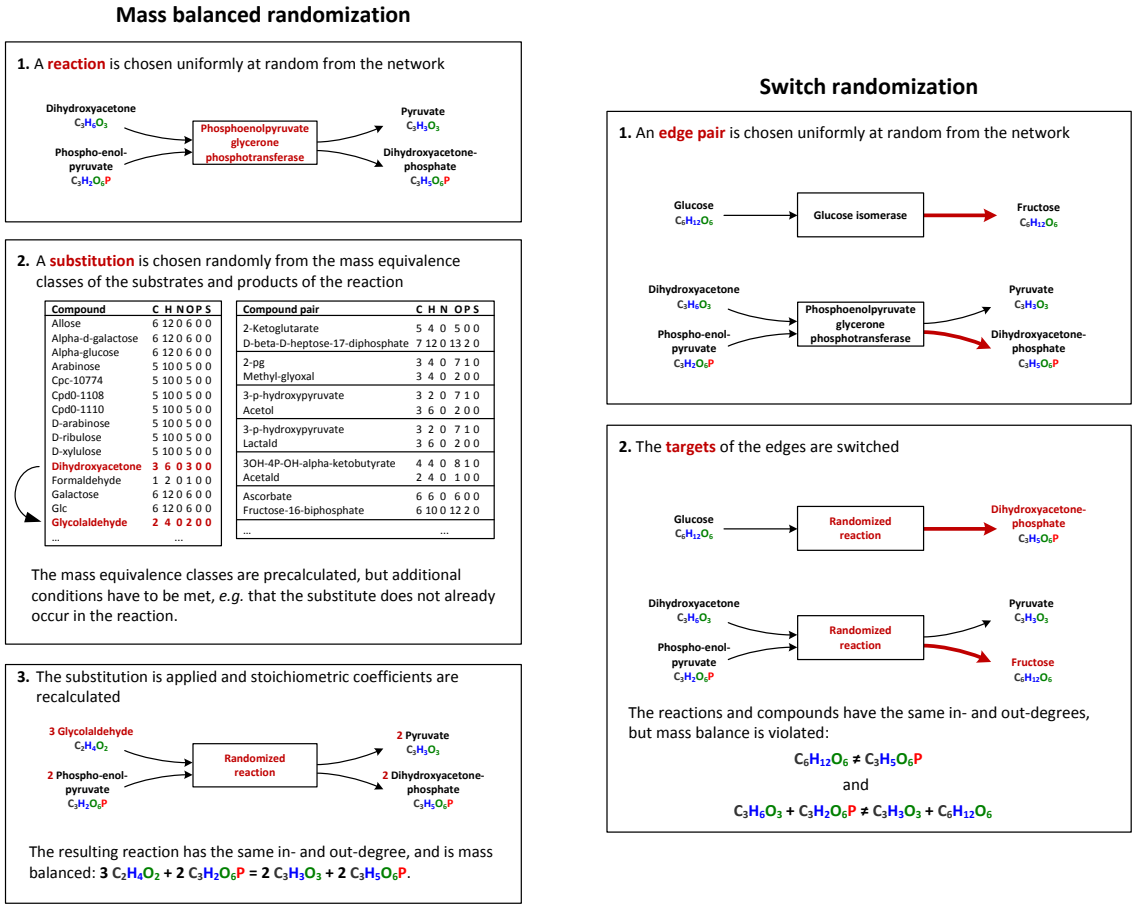


Figure C.1: Workflow schemes depicting the mass balanced (left) and switch (right) randomization methods. The procedures are repeated a large number of times in order to obtain fully randomized networks.

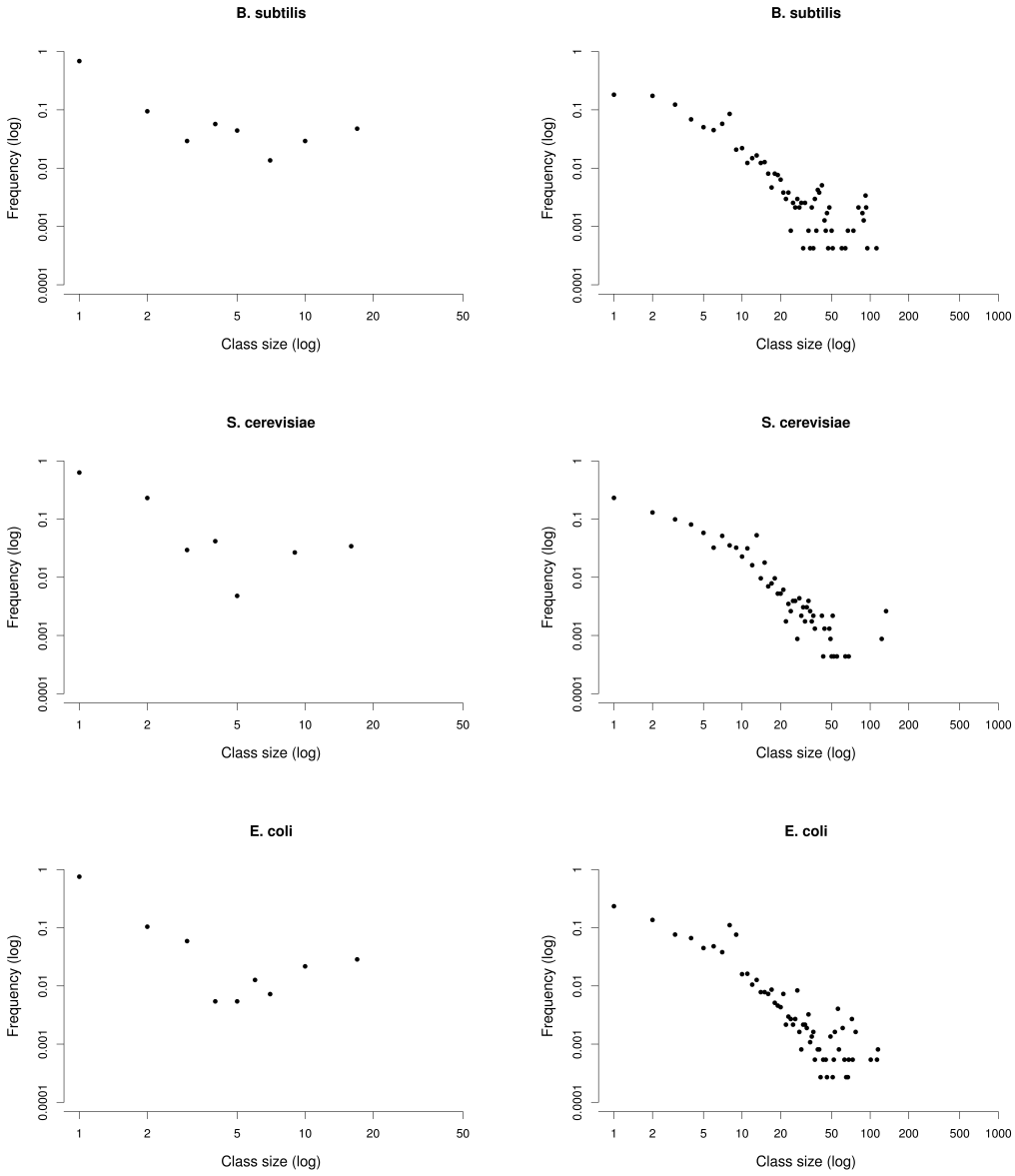


Figure C.2: Mass equivalence class size distributions for individual compounds (left column) and pairs of compounds (right column) in *B. subtilis*, *S. cerevisiae*, and *E. coli* (iAF1260).

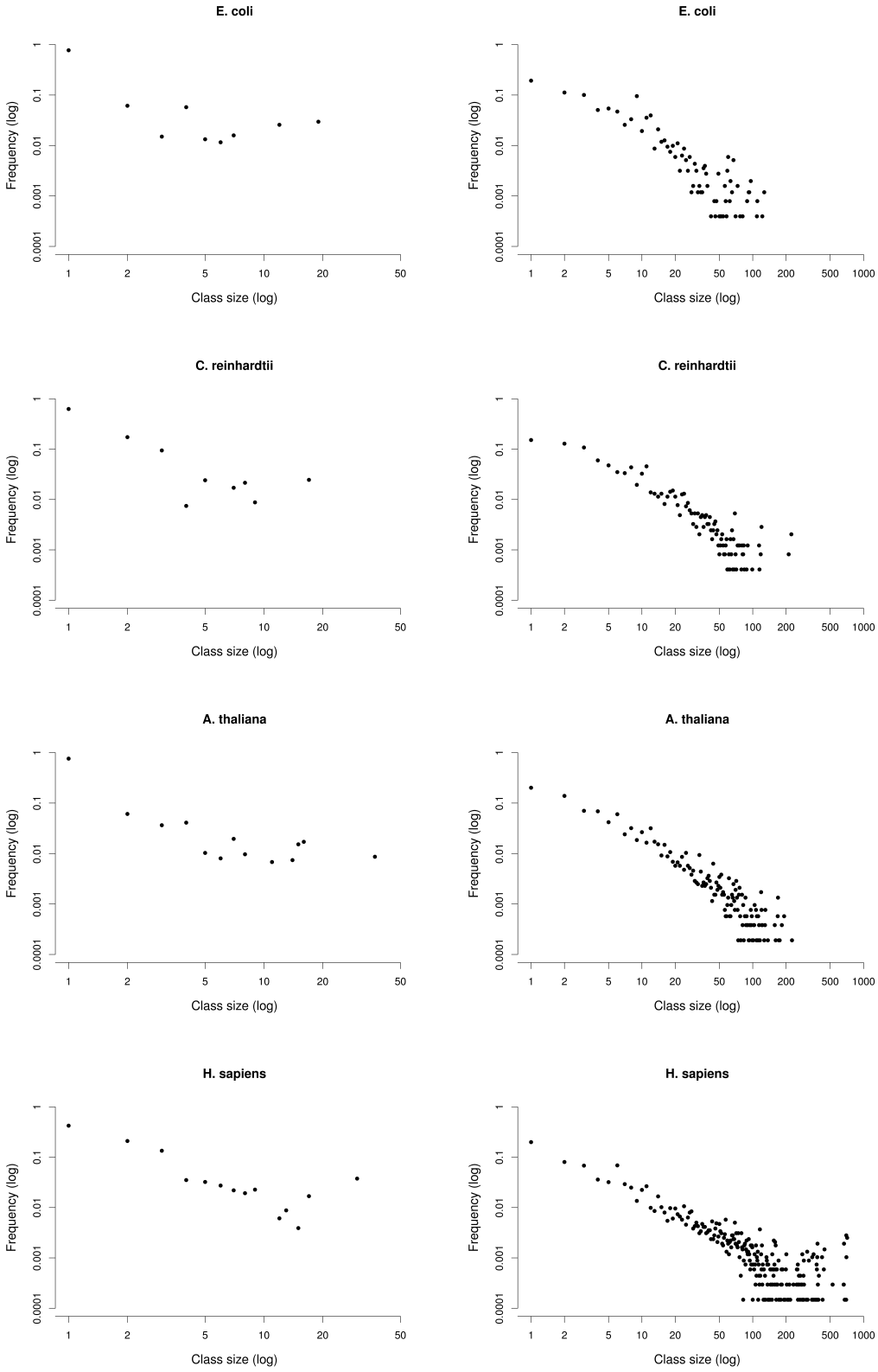


Figure C.3: Mass equivalence class size distributions for individual compounds (left column) and pairs of compounds (right column) in *E. coli* (EcoCyc), *C. reinhardtii*, *A. thaliana*, and *H. sapiens*.

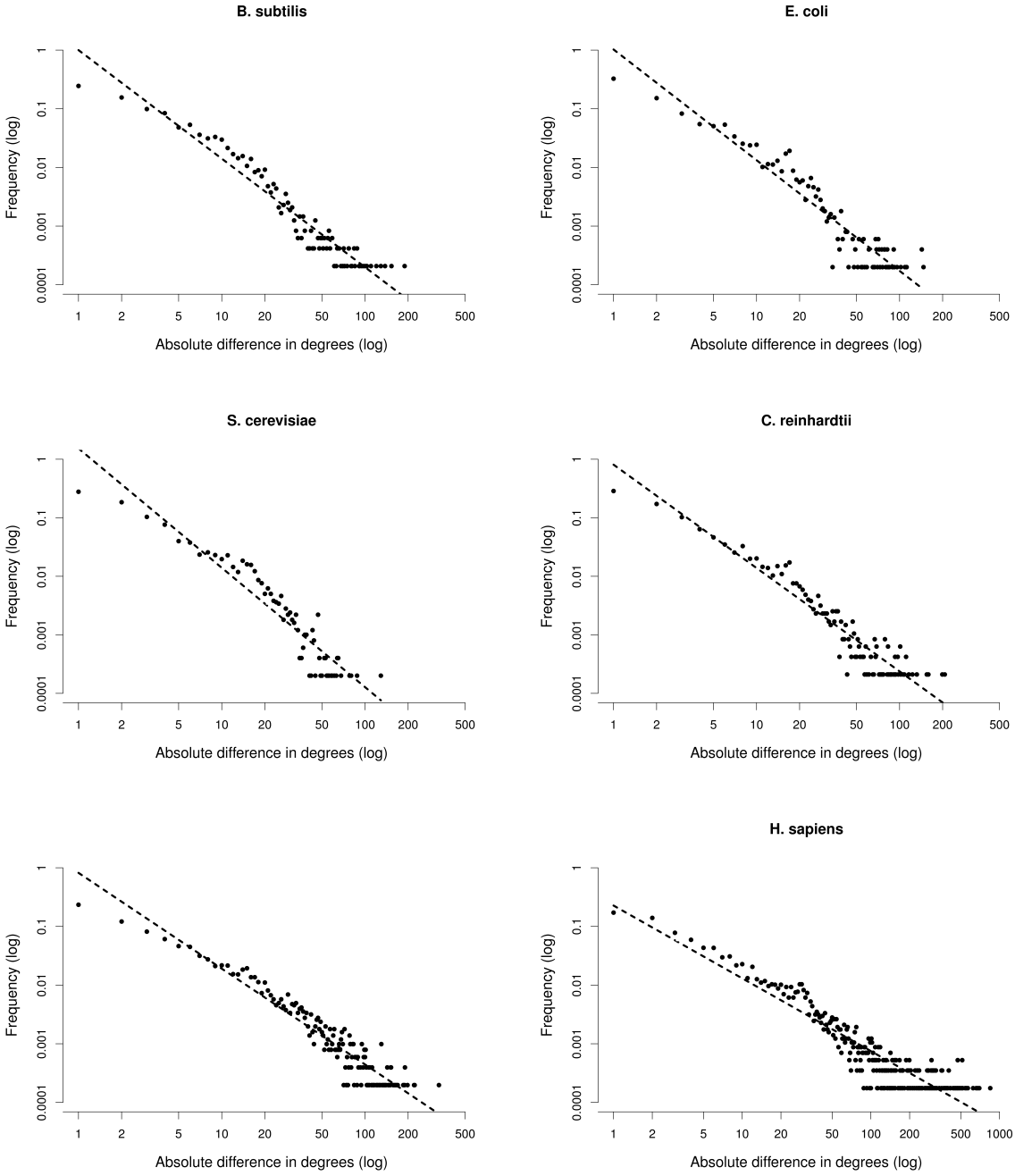


Figure C.4: Distributions of absolute differences in degrees between neighbors, sampled by random walks on the transition graphs of *B. subtilis*, *S. cerevisiae*, *E. coli* (iAF1260), *C. reinhardtii*, *A. thaliana*, and *H. sapiens*. The dashed lines show the power-law fit. Scaling coefficients and mean differences are given in Table B.3.

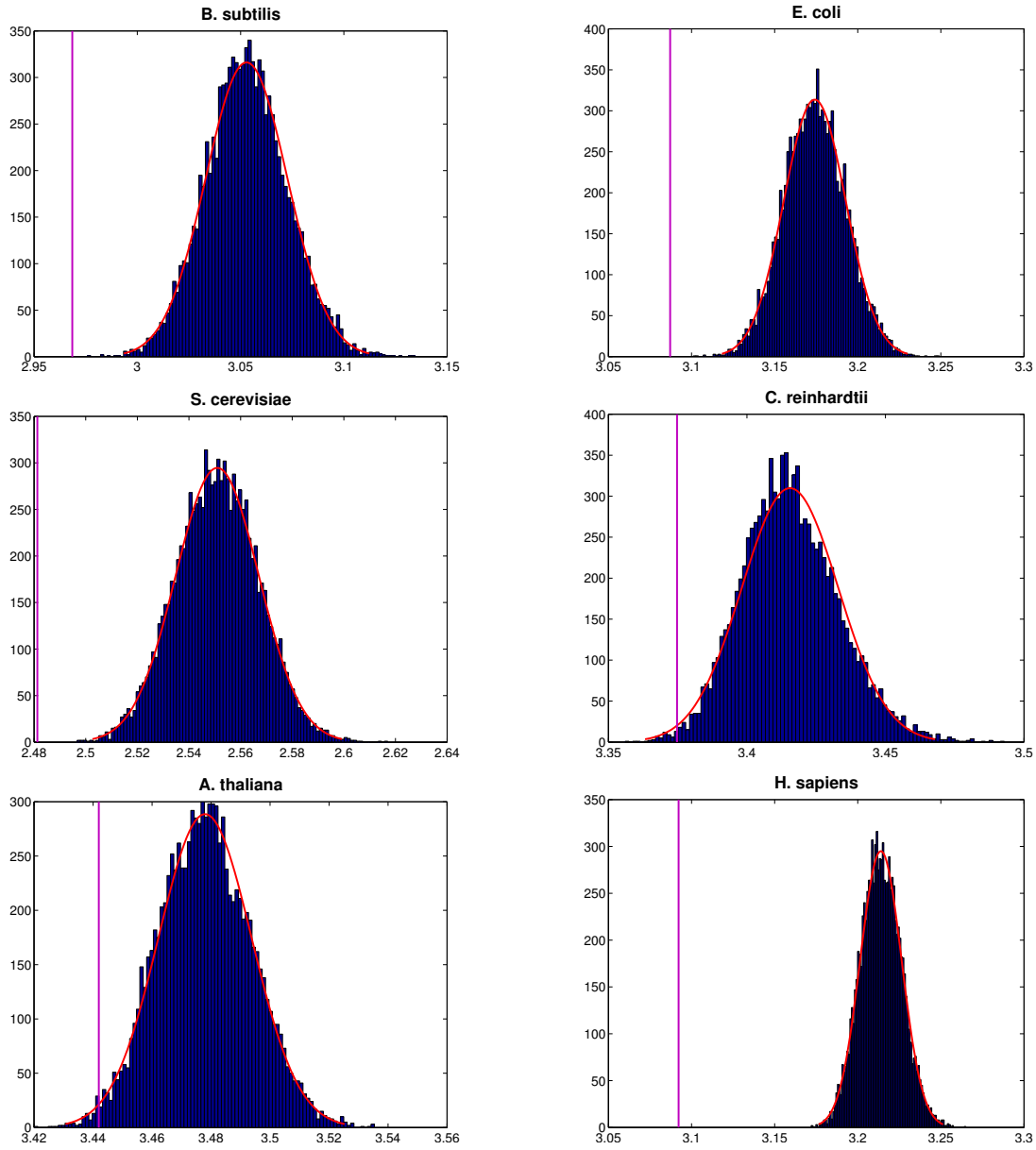


Figure C.5: Distributions of average path lengths in 10^4 mass-balanced randomized networks of six genome-scale metabolic networks. The red curve indicates the normal distribution with identical mean and standard deviation, the purple line shows the average path length in the original network. The distribution in randomized networks of *C. reinhardtii* (middle right) is rejected by the one-sample Kolmogorov-Smirnov test as normally distributed at a significance level of 0.05 (Massey, 1951). Strictly, one can therefore not rely on the significance attributed to the average path length of *C. reinhardtii* in Section 5.2.3. Nevertheless, the distribution is visually similar to a normal distribution, and the average path length is more than two standard deviations different from the mean of the distribution (Figure C.5), supporting the result that this property is evolutionary significant.

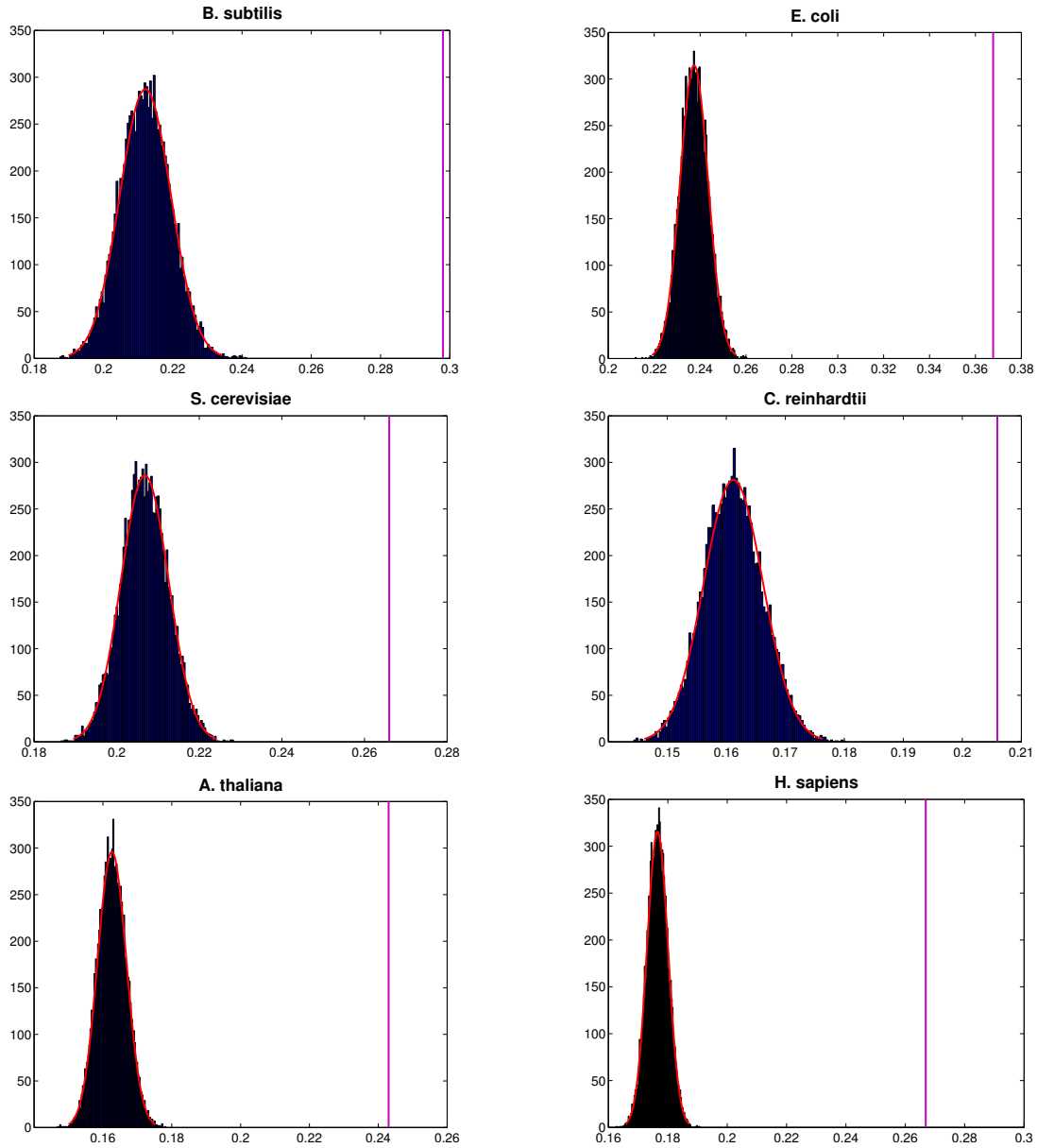


Figure C.6: Distributions of clustering coefficients in 10^4 mass-balanced randomized networks of six genome-scale metabolic networks. The red curve indicates the normal distribution with identical mean and standard deviation, the purple line shows the clustering coefficient in the original network. All distributions are accepted by the one-sample Kolmogorov-Smirnov test as normally distributed at a significance level of 0.05

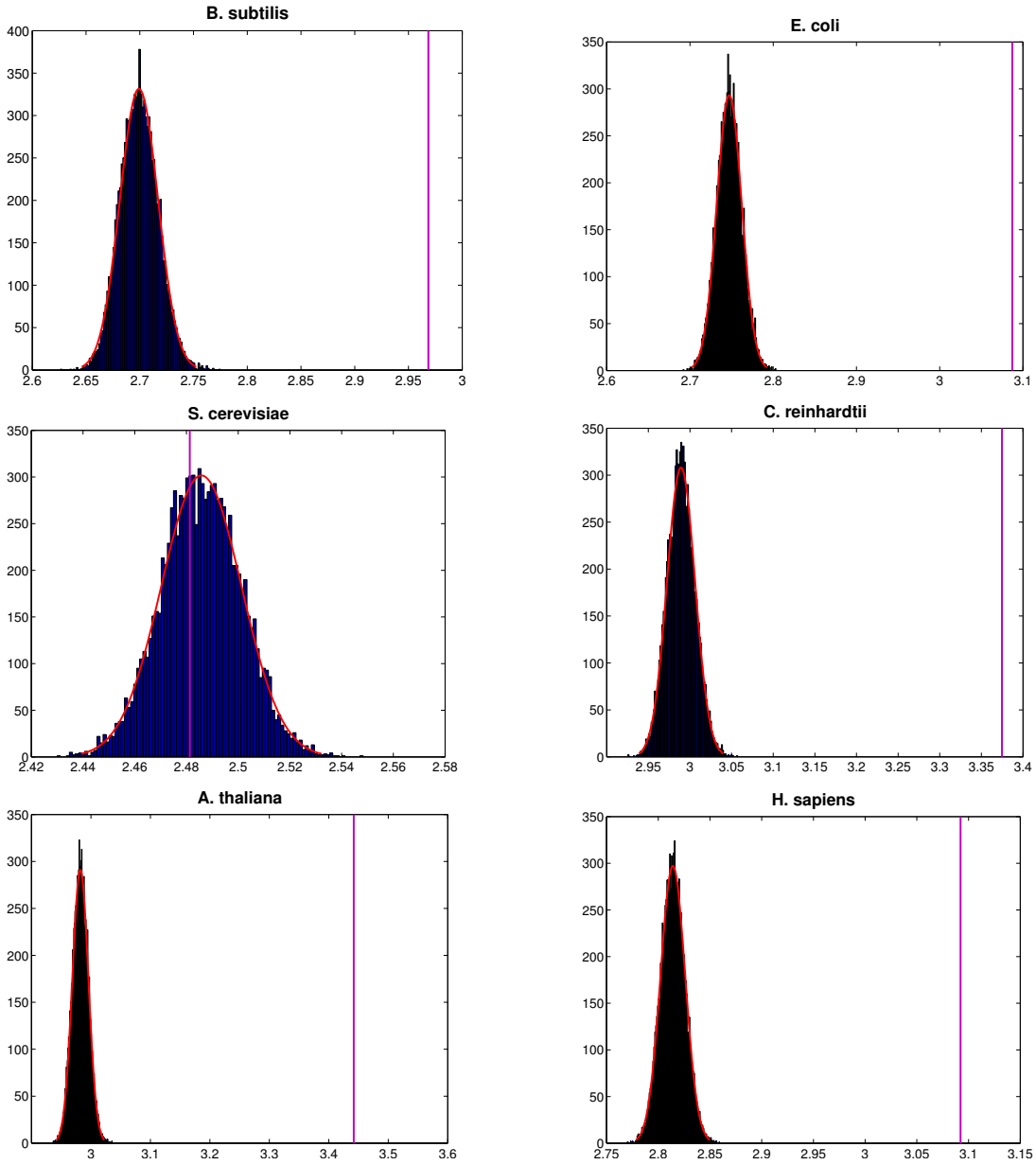


Figure C.7: Distributions of average path lengths in 10^4 switch randomized networks of six genome-scale metabolic networks. The red curve indicates the normal distribution with identical mean and standard deviation, the purple line shows the average path length in the original network. All distributions are accepted by the one-sample Kolmogorov-Smirnov test as normally distributed at a significance level of 0.05. The average path length of *S. cerevisiae* (middle left) is not significant with a p-value of 0.77 (see Section 5.2.3)

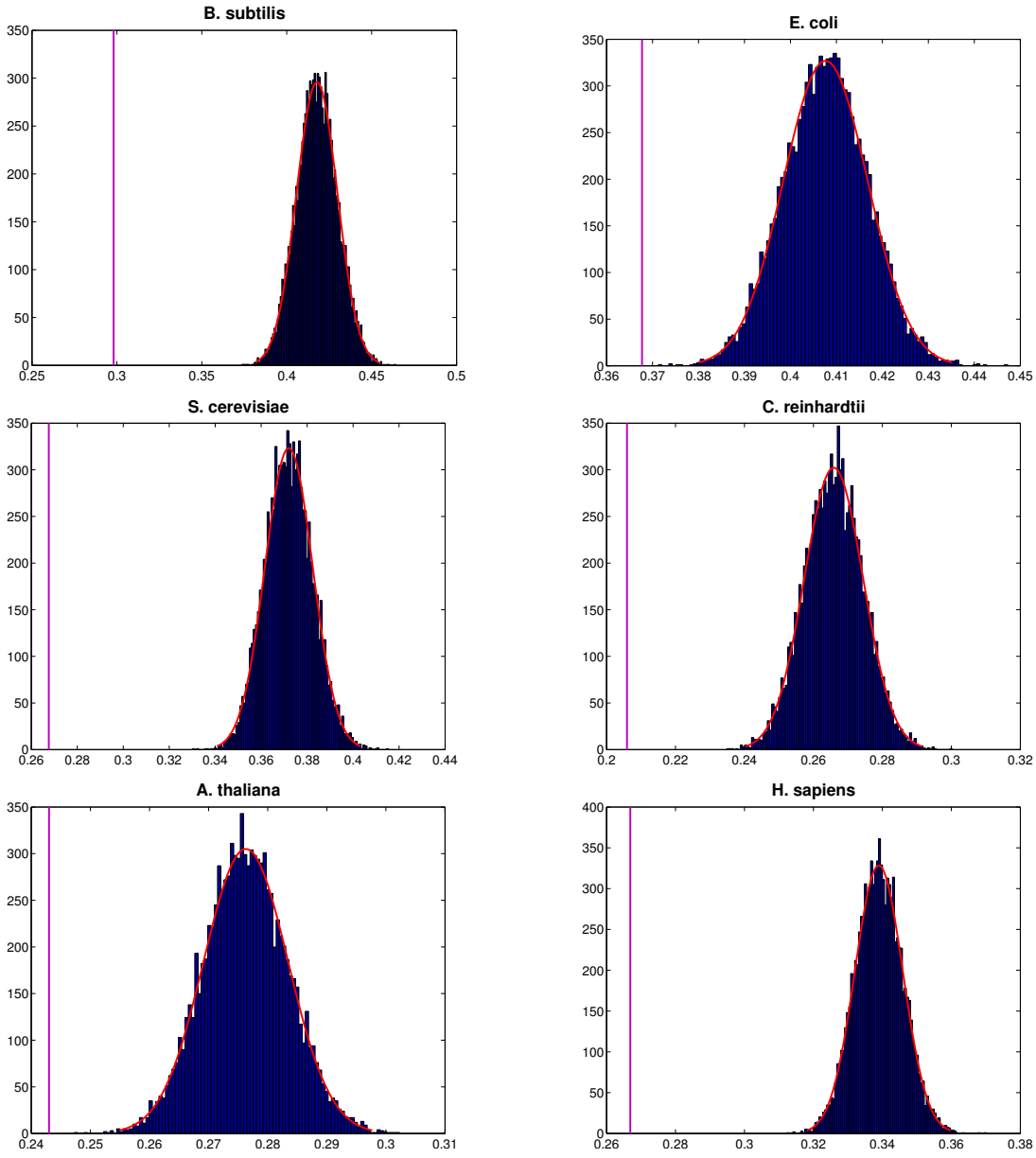


Figure C.8: Distributions of clustering coefficients in 10^4 switch randomized networks of six genome-scale metabolic networks. The red curve indicates the normal distribution with identical mean and standard deviation, the purple line shows the clustering coefficient in the original network. All distributions are accepted by the one-sample Kolmogorov-Smirnov test as normally distributed at a significance level of 0.05

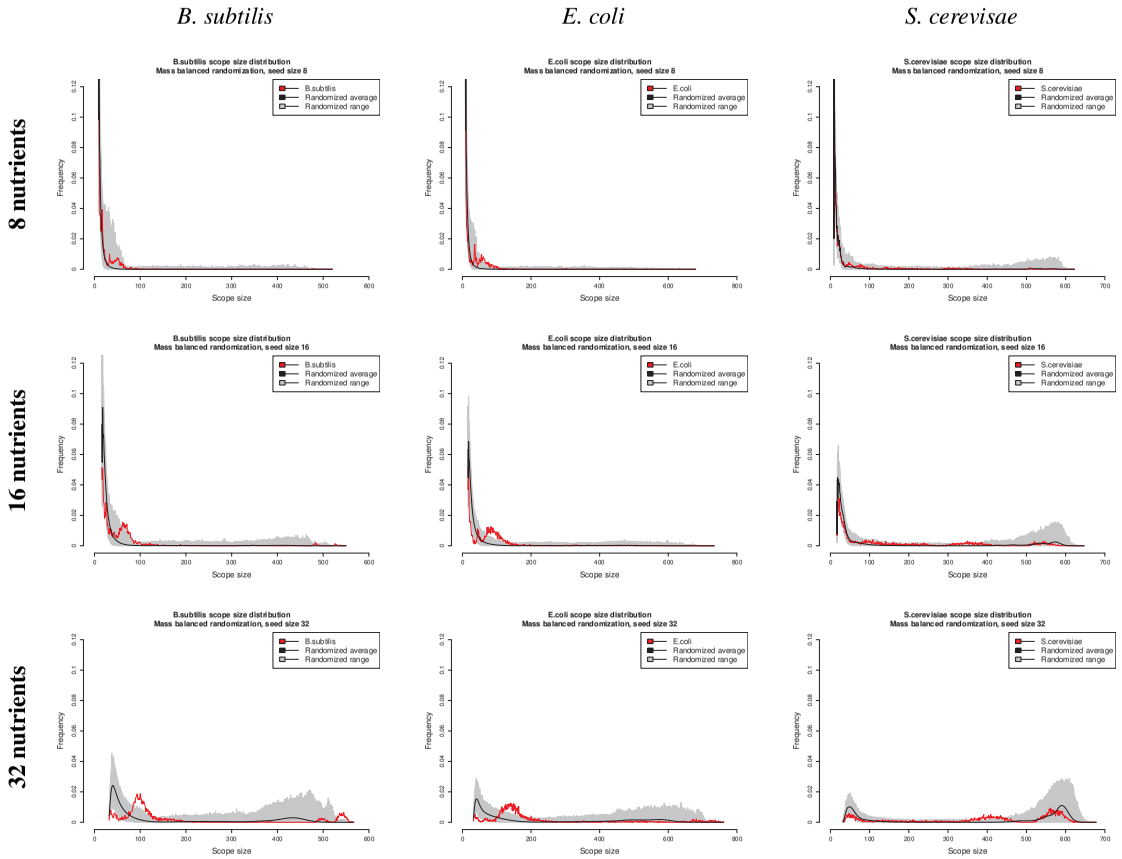


Figure C.9: Scope size distributions of *B. subtilis*, *E. coli*, and *S. cerevisae* (red lines), the frequency range over all scope size distributions in 10^4 mass-balanced randomized networks (gray areas) and their averaged distributions (black lines). The distributions were determined by calculating 5000 scopes in each randomized network from 8, 16, and 32 randomly chosen nutrients.

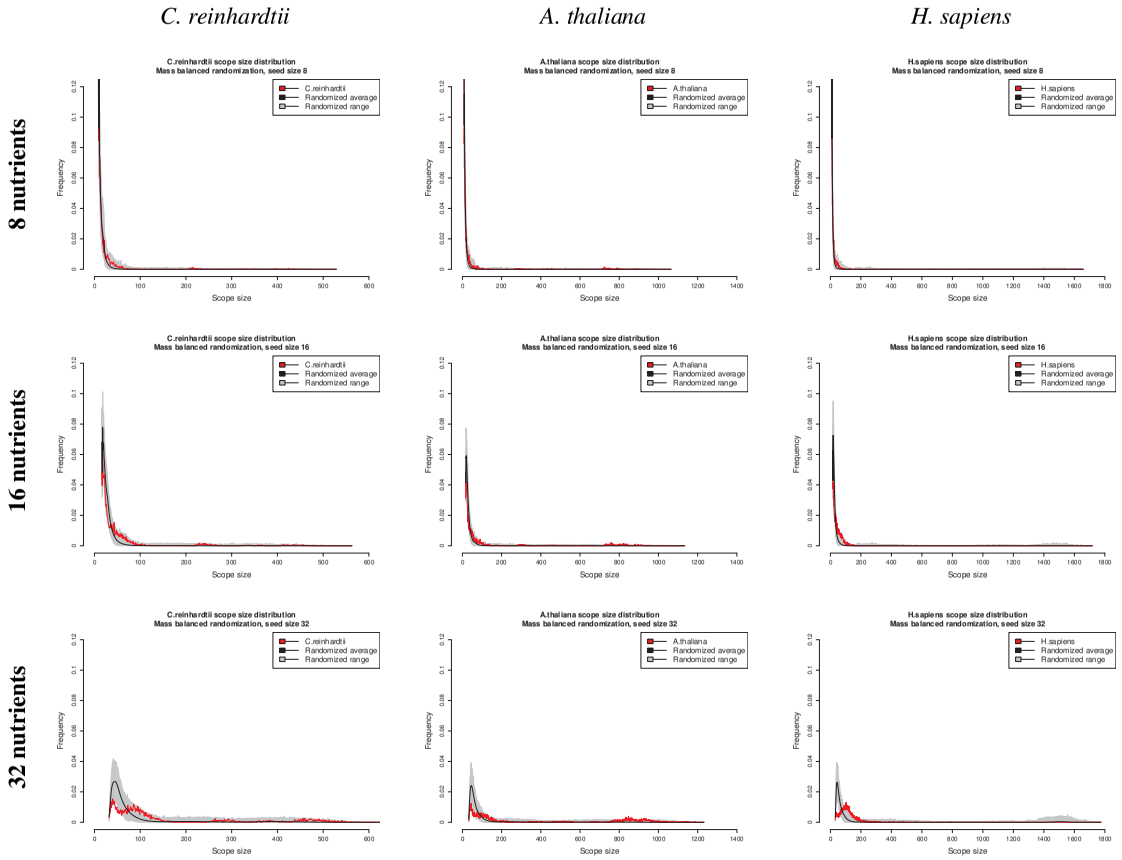


Figure C.10: Scope size distributions of *C. reinhardtii*, *A. thaliana*, and *H. sapiens* (red lines), the frequency range over all scope size distributions in 10^4 mass-balanced randomized networks (gray areas) and their averaged distributions (black lines). The distributions were determined by calculating 5000 scopes in each randomized network from 8, 16, and 32 randomly chosen nutrients.

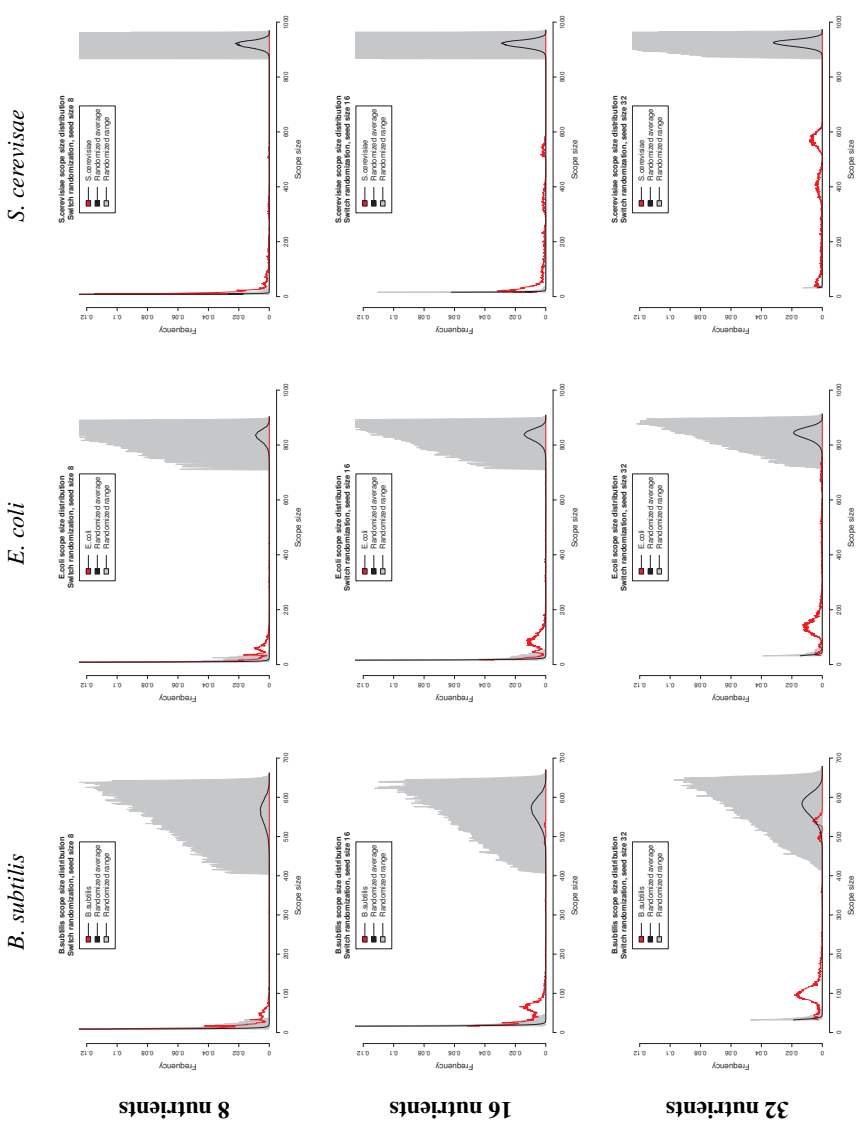


Figure C.11: Scope size distributions of *B. subtilis*, *E. coli*, and *S. cerevisae* (red lines), the frequency range over all scope size distributions in 10^4 switch randomized networks (gray areas) and their averaged distributions (black lines). The distributions were determined by calculating 5000 scopes in each randomized network from 8, 16, and 32 randomly chosen nutrients.

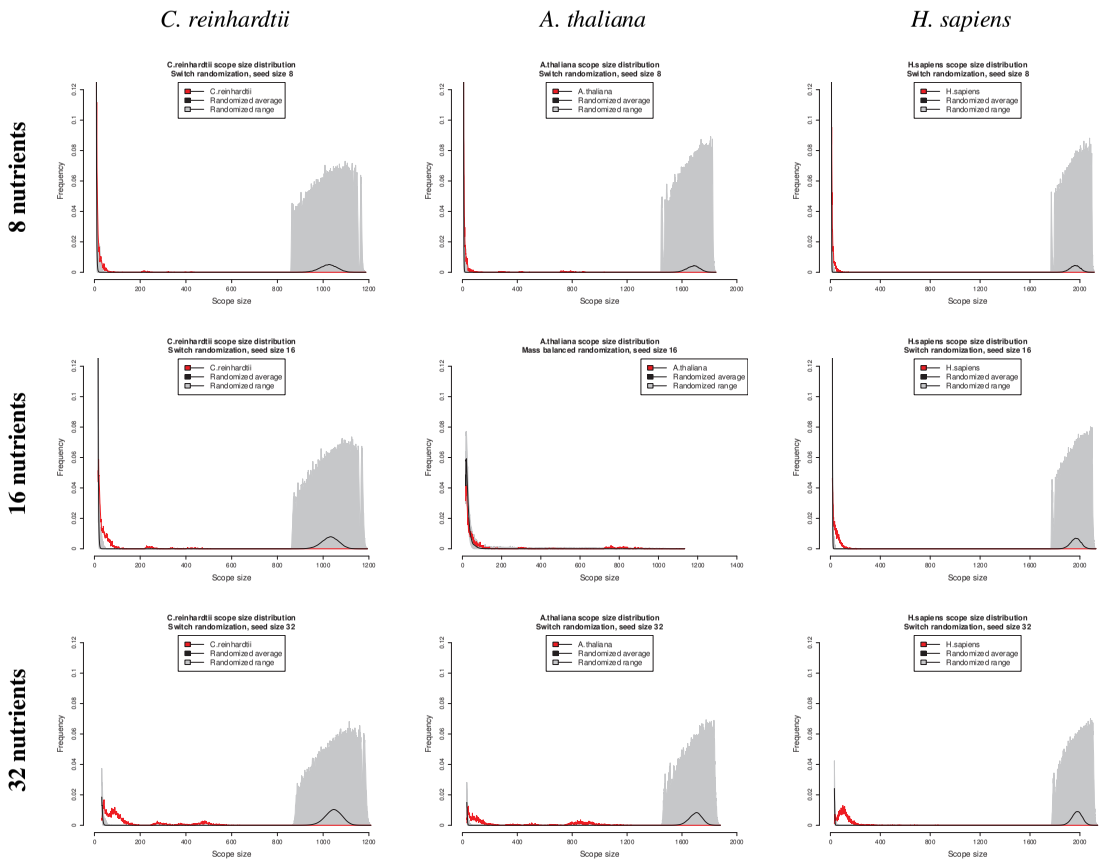


Figure C.12: Scope size distributions of *C. reinhardtii*, *A. thaliana*, and *H. sapiens* (red lines), the frequency range over all scope size distributions in 10^4 switch randomized networks (gray areas) and their averaged distributions (black lines). The distributions were determined by calculating 5000 scopes in each randomized network from 8, 16, and 32 randomly chosen nutrients.

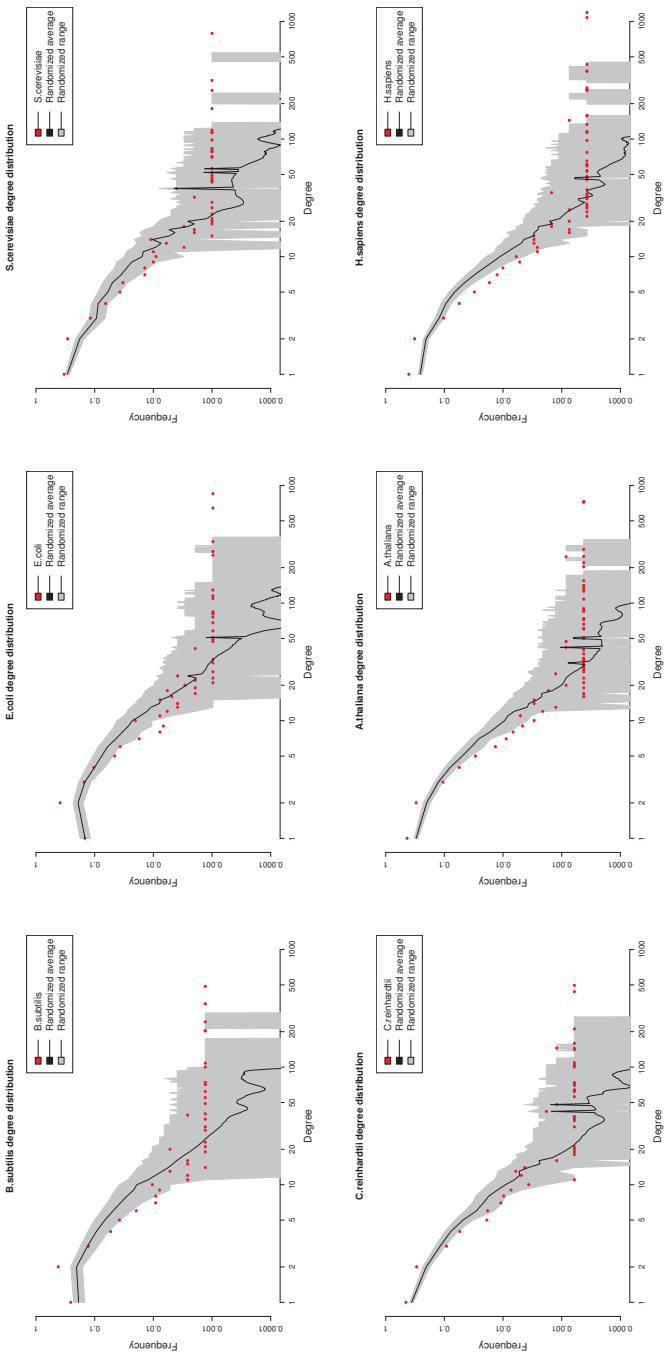


Figure C.13: Degree distributions of *B. subtilis*, *E. coli*, *S. cerevisae*, *C. reinhardtii*, *A. thaliana*, and *H. sapiens* (red lines), the frequency range over all degree distributions in 10^4 mass-balanced randomized networks (gray areas) and their averaged distributions (black lines) on a log-log scale.

Selbständigkeitserklärung

Hiermit erkläre ich, daß die vorliegende Arbeit an keiner anderen Hochschule eingereicht und von mir selbständig, nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt wurde.

Potsdam, den 26. Januar 2012

Georg Basler

Bibliography

Albert, I. and Albert, R. (2004). Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, 20(18):3346–3352.

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

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97.

Albert, R., Jeong, H., and Barabási, A.-L. (1999). Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131.

Albert, R., Jeong, H., and Barabasi, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382.

Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N., and Barabási, A.-L. (2004). Global organization of metabolic fluxes in the bacterium Escherichia coli. *Nature*, 427(6977):839–843.

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

Alper, H., Miyaoku, K., and Stephanopoulos, G. (2005). Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets. *Nat Biotechnol*, 23(5):612–616.

Amaral, L. A., Scala, A., Barthélémy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proc Natl Acad Sci U S A*, 97(21):11149–11152.

Amaral, L. A. N., Díaz-Guilera, A., Moreira, A. A., Goldberger, A. L., and Lipsitz, L. A. (2004). Emergence of complex dynamics in a simple model of signaling networks. *Proc Natl Acad Sci U S A*, 101(44):15551–15555.

Arita, M. (2004). The metabolic world of Escherichia coli is not small. *Proc Natl Acad Sci U S A*, 101(6):1543–1547.

Arkin, A., Shen, P., and Ross, J. (1997). A test case of correlation metric construction of a reaction pathway from measurements. *Science*, 277(5330):1275–1279.

Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N., and Stone, L. (2004). Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science*, 305(5687):1107.

Artzy-Randrup, Y. and Stone, L. (2005). Generating uniformly distributed random networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 72(5 Pt 2):056708.

Atsumi, S., Hanai, T., and Liao, J. C. (2008). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature*, 451(7174):86–89.

Bar-Even, A., Noor, E., Lewis, N. E., and Milo, R. (2010). Design and analysis of synthetic carbon fixation pathways. *Proc Natl Acad Sci U S A*, 107(19):8889–8894.

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.

Basler, G., Ebenhöh, O., Selbig, J., and Nikoloski, Z. (2011a). Mass-balanced randomization of metabolic networks. *Bioinformatics*, 27(10):1397–1403.

Basler, G., Grimbs, S., Ebenhöh, O., Selbig, J., and Nikoloski, Z. (2011b). Evolutionary significance of metabolic network properties. *Journal of The Royal Society Interface*.

Basler, G., Grimbs, S., and Nikoloski, Z. (2010). Thermodynamic landscapes of randomized large-scale metabolic networks. In *Proceedings of the 7th International Workshop on Computational Systems Biology, WCSB 2010*, Tampere, Finland. Tampere International Center for Signal Processing.

Basler, G. and Nikoloski, Z. (2011). JMassBalance: mass-balanced randomization and analysis of metabolic networks. *Bioinformatics*, 27(19):2761–2762.

Beard, D. A., Liang, S.-d., and Qian, H. (2002). Energy balance for analysis of complex metabolic networks. *Biophys J*, 83(1):79–86.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res*, 28(1):235–242.

Bernhardsson, S. and Minnhagen, P. (2010). Selective pressure on metabolic network structures as measured from the random blind-watchmaker network. *New Journal of Physics*, 12(10):103047.

Bond-Watts, B. B., Bellerose, R. J., and Chang, M. C. Y. (2011). Enzyme mechanism as a kinetic control element for designing synthetic biofuel pathways. *Nat Chem Biol*, 7(4):222–227.

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.

Bray, D. (2003). Molecular networks: the top-down view. *Science*, 301(5641):1864–1865.

Breitling, R., Pitt, A. R., and Barrett, M. P. (2006). Precision mapping of the metabolome. *Trends Biotechnol*, 24(12):543–548.

Breitling, R., Vitkup, D., and Barrett, M. P. (2008). New surveyor tools for charting microbial metabolic maps. *Nat Rev Microbiol*, 6(2):156–161.

Bulik, S., Grimbs, S., Huthmacher, C., Selbig, J., and Holzhtter, H. G. (2009). 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 J*, 276(2):410–424.

Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng*, 84(6):647–657.

Caetano-Anolls, G., Yafremava, L. S., Gee, H., Caetano-Anolls, D., Kim, H. S., and Mittenthal, J. E. (2009). The origin and evolution of modern metabolism. *Int J Biochem Cell Biol*, 41(2):285–297.

Casella, G. and Berger, R. (1990). *Statistical Inference*. Duxbury Press, Belmont, California.

Caspi, R., Altman, T., Dale, J. M., Dreher, K., Fulcher, C. A., Gilham, F., Kaipa, P., Karthikeyan, A. S., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Paley, S., Popescu, L., Pujar, A., Shearer, A. G., Zhang, P., and Karp, P. D. (2010). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, 38(Database issue):D473–D479.

Chance, B., Garfinkel, D., Higgins, J., and Hess, B. (1960). Metabolic control mechanisms. 5. a solution for the equations representing interaction between glycolysis and respiration in ascites tumor cells. *J Biol Chem*, 235:2426–2439.

Chang, C.-W., Lyu, P.-C., and Arita, M. (2011). Reconstructing phylogeny from metabolic substrate-product relationships. *BMC Bioinformatics*, 12 Suppl 1:S27.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140.

Chung, F. and Lu, L. (2006). Concentration inequalities and martingale inequalities: A survey. *Internet Mathematics*, 3(1):79–127.

Chung, F. R. K. and Lu, L. (2003). Coupling online and offline analyses for random power law graphs. *Internet Mathematics*, 1(4):409–461.

Ciliberti, S., Martin, O. C., and Wagner, A. (2007). Innovation and robustness in complex regulatory gene networks. *Proc Natl Acad Sci U S A*, 104(34):13591–13596.

Cobb, G. W. and Chen, Y.-P. (2003). An application of markov chain monte carlo to community ecology. *The American Mathematical Monthly*, 110(4):pp. 265–288.

Colizza, V., Flammini, A., Serrano, M. A., and Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nat Phys*, 2(2):110–115.

Daily, J. P., Scanfeld, D., Pochet, N., Le Roch, K., Plouffe, D., Kamal, M., Sarr, O., Mboup, S., Ndir, O., Wypij, D., Levasseur, K., Thomas, E., Tamayo, P., Dong, C., Zhou, Y., Lander, E. S., Ndiaye, D., Wirth, D., Winzeler, E. A., Mesirov, J. P., and Regev, A. (2007). Distinct physiological states of plasmodium falciparum in malaria-infected patients. *Nature*, 450(7172):1091–1095.

de la Fuente, A., Fotia, G., Maggio, F., Mancosu, G., and Pieroni, E. (2008). Insights into biological information processing: structural and dynamical analysis of a human protein signalling network. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224013.

de Oliveira Dal'Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., and Nielsen, L. K. (2010). AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol*, 152(2):579–589.

Dobson, C. M. (2004). Chemical space and biology. *Nature*, 432(7019):824–828.

Dorogovtsev, S. and Mendes, J. (2003). *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford, United Kingdom.

Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F. (2008). Critical phenomena in complex networks. *Rev. Mod. Phys.*, 80:1275–1335.

Duanmu, D., Wang, Y., and Spalding, M. H. (2009). Thylakoid lumen carbonic anhydrase (CAH3) mutation suppresses air-dier phenotype of LCIB mutant in Chlamydomonas reinhardtii. *Plant Physiol*, 149(2):929–937.

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.

Ebenhöh, O. and Handorf, T. (2009). Functional classification of genome-scale metabolic networks. *EURASIP J. Bioinformatics Syst. Biol.*, 2009:1–13.

Ederer, M. and Gilles, E. D. (2007). Thermodynamically feasible kinetic models of reaction networks. *Biophys J*, 92(6):1846–1857.

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.

Erdös, P. and Rényi, A. (1959). On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297.

Famili, I., Forster, J., Nielsen, J., and Palsson, B. Ø. (2003). Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc Natl Acad Sci U S A*, 100(23):13134–13139.

Famili, I., Mahadevan, R., and Palsson, B. Ø. (2005). k-cone analysis: determining all candidate values for kinetic parameters on a network scale. *Biophys J*, 88(3):1616–1625.

Fani, R. and Fondi, M. (2009). Origin and evolution of metabolic pathways. *Physics of Life Reviews*, 6(1):23–52.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. Ø. (2007). A genome-scale metabolic reconstruction for Escherichia coli k-12 mg1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol*, 3:121.

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., and Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol*, 7(2):129–143.

Feist, A. M., Zielinski, D. C., Orth, J. D., Schellenberger, J., Herrgård, M. J., and Palsson, B. Ø. (2010). Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli. *Metab Eng*, 12(3):173–186.

Ferreira, F. J., Guo, C., and Coleman, J. R. (2008). Reduction of plastid-localized carbonic anhydrase activity results in reduced Arabidopsis seedling survivorship. *Plant Physiol*, 147(2):585–594.

Fiehn, O., Kopka, J., Drmann, P., Altmann, T., Trethewey, R. N., and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat Biotechnol*, 18(11):1157–1161.

Fisher, R. (1925). *Statistical methods for research workers*. Oliver & Boyd, Edinburgh, United Kingdom.

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., and Shlomi, T. (2011). Predicting selective drug targets in cancer through metabolic networks. *Mol Syst Biol*, 7:501.

Francke, C., Siezen, R. J., and Teusink, B. (2005). Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol*, 13(11):550–558.

Gallo, G., Longo, G., Pallottino, S., and Nguyen, S. (1993). Directed hypergraphs and applications. *Discrete Applied Mathematics*, 42(2-3):177 – 201.

Gianchandani, E. P., Brautigan, D. L., and Papin, J. A. (2006). Systems analyses characterize integrated functions of biochemical networks. *Trends Biochem Sci*, 31(5):284–291.

Gilmour, K. M. (2010). Perspectives on carbonic anhydrase. *Comp Biochem Physiol A Mol Integr Physiol*, 157(3):193–197.

Ginoza, R. and Mugler, A. (2010). Network motifs come in sets: correlations in the randomization process. *Phys Rev E Stat Nonlin Soft Matter Phys*, 82(1 Pt 1):011921.

Gionis, A., Mannila, H., Mielikinen, T., and Tsaparas, P. (2006). Assessing data mining results via swap randomization. In *In KDD (2006*, pages 167–176.

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, Jr, R., 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.

Goeddel, D. V., Kleid, D. G., Bolivar, F., Heyneker, H. L., Yansura, D. G., Crea, R., Hirose, T., Kraszewski, A., Itakura, K., and Riggs, A. D. (1979). Expression in Escherichia coli of chemically synthesized genes for human insulin. *Proc Natl Acad Sci U S A*, 76(1):106–110.

Grafahrend-Belau, E., Schreiber, F., Koschtzki, D., and Junker, B. H. (2009). Flux balance analysis of barley seeds: a computational approach to study systemic properties of central metabolism. *Plant Physiol*, 149(1):585–598.

Gralla, E. B. and Valentine, J. S. (1991). Null mutants of Saccharomyces cerevisiae Cu,Zn superoxide dismutase: characterization and spontaneous mutation rates. *J Bacteriol*, 173(18):5918–5920.

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. *Nature Physics*, 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*, 23(13):1616–1622.

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

Hanhijärvi, S., Garriga, G. C., and Puolamäki, K. (2009a). Randomization techniques for graphs. In *SDM*, pages 780–791. SIAM.

Hanhijärvi, S., Ojala, M., Vuokko, N., Puolamäki, K., Tatti, N., and Mannila, H. (2009b). Tell me something i don't know: randomization strategies for iterative data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 379–388, New York, NY, USA. ACM.

Hanson, A. D., Pribat, A., Waller, J. C., and de Crcy-Lagard, V. (2010). 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list–and how to find it. *Biochem J*, 425(1):1–11.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. *Biophys J*, 92(5):1792–1805.

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.

Henry, R. P. (1996). Multiple roles of carbonic anhydrase in cellular transport and metabolism. *Annu Rev Physiol*, 58:523–538.

Hermann, J. C., Ghanem, E., Li, Y., Raushel, F. M., Irwin, J. J., and Shoichet, B. K. (2006). Predicting substrates by docking high-energy intermediates to enzyme structures. *J Am Chem Soc*, 128(49):15882–15891.

Hermann, J. C., Marti-Arbona, R., Fedorov, A. A., Fedorov, E., Almo, S. C., Shoichet, B. K., and Raushel, F. M. (2007). Structure-based activity prediction for an enzyme of unknown function. *Nature*, 448(7155):775–779.

Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., Blüthgen, N., Borger, S., Costenoble, R., Heinemann, M., Hucka, M., Novère, N. L., Li, P., Liebermeister, W., Mo, M. L., Oliveira, A. P., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasić, I., Weichart, D., Brent, R., Broomhead, D. S., Westerhoff, H. V., Kırdar, B., Penttilä, M., Klipp, E., Palsson, B. Ø., Sauer, U., Oliver, S. G., Mendes, P., Nielsen, J., and Kell, D. B. (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol*, 26(10):1155–1160.

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

Hoppe, A., Hoffmann, S., and Holzhtter, 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.

Horowitz, E. and Sahni, S. (1974). Computing partitions with applications to the knapsack problem. *J. ACM*, 21(2):277–292.

Horvath, H., Huang, J., Wong, O., Kohl, E., Okita, T., Kannangara, C. G., and von Wettstein, D. (2000). The production of recombinant proteins in transgenic barley grains. *Proc Natl Acad Sci U S A*, 97(4):1914–1919.

Horvath, H., Jensen, L. G., Wong, O. T., Kohl, E., Ullrich, S. E., Cochran, J., Kannangara, C. G., and von Wettstein, D. (2001). Stability of transgene expression, field performance and recombination breeding of transformed barley lines. *TAG Theoretical and Applied Genetics*, 102:1–11. 10.1007/s001220051612.

Hubbell, C. H. (1965). An Input-Output Approach to Clique Identification. *Sociometry*, 28(4):377–399.

Hyduke, D. R. and Palsson, B. Ø. (2010). Towards genome-scale signalling network reconstructions. *Nat Rev Genet*, 11(4):297–307.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2:343–372.

Irani, M. and Maitra, P. K. (1974). Isolation and characterization of Escherichia coli mutants defective in enzymes of glycolysis. *Biochem Biophys Res Commun*, 56(1):127–133.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–4574.

Itzkovitz, S., Milo, R., Kashtan, N., Ziv, G., and Alon, U. (2003). Subgraphs in random networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(2 Pt 2):026127.

Jamshidi, N. and Palsson, B. Ø. (2008). Formulating genome-scale kinetic models in the post-genome era. *Mol Syst Biol*, 4:171.

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., and Barabási, A.-L. (2000). The large–scale organization of metabolic networks. *Nature*, 407:651–654.

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.

Karlebach, G. and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol*, 9(10):770–780.

Keseler, I. M., Bonavides-Martínez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S., Paulsen, I. T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A. G., and Karp, P. D. (2009). EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res*, 37(Database issue):D464–D470.

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

Klamt, S., Haus, U.-U., and Theis, F. (2009). Hypergraphs and cellular networks. *PLoS Comput Biol*, 5(5):e1000385.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrn-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637–643.

Kuchaiev, O., Milenkovic, T., Memisevic, V., Hayes, W., and Przulj, N. (2010). Topological network alignment uncovers biological function and phylogeny. *J R Soc Interface*, 7(50):1341–1354.

Kumar, A., Suthers, P. F., and Maranas, C. D. (2012). MetRxn: A knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics*, 13(1):6.

Kümmel, A., Panke, S., and Heinemann, M. (2006). Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics*, 7:512.

Lago-Fernández, L. F., Huerta, R., Corbacho, F., and Sigenza, J. A. (2000). Fast response and temporal coherent oscillations in small-world networks. *Phys Rev Lett*, 84(12):2758–2761.

Lam, K. B. and Marmur, J. (1977). Isolation and characterization of saccharomyces cerevisiae glycolytic pathway mutants. *J Bacteriol*, 130(2):746–749.

Langville, A. N. and Meyer, C. D. (2003). Deeper Inside PageRank. *Internet Mathematics*, 1(3):335–380.

Latora, V. and Marchiori, M. (2001). Efficient behavior of small-world networks. *Phys Rev Lett*, 87(19):198701.

Lee, K. H., Park, J. H., Kim, T. Y., Kim, H. U., and Lee, S. Y. (2007). Systems metabolic engineering of Escherichia coli for L-threonine production. *Mol Syst Biol*, 3:149.

Lee, S. J., Lee, D.-Y., Kim, T. Y., Kim, B. H., Lee, J., and Lee, S. Y. (2005). Metabolic engineering of Escherichia coli for enhanced production of succinic acid, based on genome comparison and in silico gene knockout simulation. *Appl Environ Microbiol*, 71(12):7880–7887.

Lehoucq, R. B., Sorensen, D. C., and Yang, C. (1998). *ARPACK Users' Guide - Solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods.* Software, environments, tools. Society for Industrial & Applied Mathematics (SIAM), Philadelphia, USA.

Li, L., Alderson, D., Doyle, J., and Willinger, W. (2005). Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4):431–523.

Liljeros, F., Edling, C. R., Amaral, L. A., Stanley, H. E., and Aberg, Y. (2001). The web of human sexual contacts. *Nature*, 411(6840):907–908.

Lima-Mendez, G. and van Helden, J. (2009). The powerful law of the power law and other myths in network biology. *Mol Biosyst*, 5(12):1482–1493.

Lipman, D. J., Wilbur, W. J., Smith, T. F., and Waterman, M. S. (1984). On the statistical significance of nucleic acid similarities. *Nucleic Acids Res*, 12(1 Pt 1):215–226.

Lotka, A. J. (1922). Natural selection as a physical principle. *Proc Natl Acad Sci U S A*, 8(6):151–154.

Lovász, L. (1993). Random walks on graphs: a survey. *Combinatorics, Paul Erdős is Eighty*, 2:1–46.

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

Ma, H.-W. and Zeng, A.-P. (2003b). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11):1423–1430.

Ma, H.-W. and Zeng, A.-P. (2004). Phylogenetic comparison of metabolic capacities of organisms at genome level. *Mol Phylogenet Evol*, 31(1):204–213.

Marr, C., Müller-Linow, M., and Hütt, M.-T. (2007). Regularizing capacity of metabolic networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 75(4 Pt 1):041917.

Martínez-Blanco, H., Reglero, A., Rodriguez-Aparicio, L. B., and Luengo, J. M. (1990). Purification and biochemical characterization of phenylacetyl-CoA ligase from Pseudomonas putida. A specific enzyme for the catabolism of phenylacetic acid. *J Biol Chem*, 265(12):7084–7090.

Maslov, S. (2007). Complex networks: Role model for modules. *Nat Phys*, 3(1):18–19.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913.

Maslov, S., Sneppen, K., and Zaliznyak, A. (2004). Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A: Statistical and Theoretical Physics*, 333:529–540.

Massey, Frank J., J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):pp. 68–78.

Mavrovouniotis, M. L. (1991). Estimation of standard gibbs energy changes of biotransformations. *J Biol Chem*, 266(22):14440–14445.

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.

Mesquita, A., Weinberger, M., Silva, A., Sampaio-Marques, B., Almeida, B., Leo, C., Costa, V., Rodrigues, F., Burhans, W. C., and Ludovico, P. (2010). Caloric restriction or catalase inactivation extends yeast chronological lifespan by inducing h2o2 and superoxide dismutase activity. *Proc Natl Acad Sci U S A*, 107(34):15123–15128.

Metropolis, A. W., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542.

Milo, R., Kashtan, N., Itzkovitz, S., Newman, M. E. J., and Alon, U. (2003). On the uniform generation of random graphs with prescribed degree sequences. *eprint arXiv:cond-mat/0312028*. Available at http://arxiv.org/abs/cond-mat/0312028v2.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.

Moreno, J. and Jennings, H. (1937). Statistics of social configurations. *Sociometry*, 1:342–374.

Nagrath, D., Avila-Elchiver, M., Berthiaume, F., Tilles, A. W., Messac, A., and Yarmush, M. L. (2007). Integrated energy and flux balance based multiobjective framework for large-scale metabolic networks. *Ann Biomed Eng*, 35(6):863–885.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proc Natl Acad Sci U S A*, 98(2):404–409.

Newman, M. E., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Phys Rev E Stat Nonlin Soft Matter Phys*, 64(2 Pt 2):026118.

Newman, M. E. J. (2003a). Mixing patterns in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 67(2 Pt 2):026126.

Newman, M. E. J. (2003b). The structure and function of complex networks. *SIAM Review*, 45:167–256.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods*, 5(2):241–301.

Noor, R., Mittal, S., and Iqbal, J. (2002). Superoxide dismutase–applications and relevance to human diseases. *Med Sci Monit*, 8(9):RA210–RA215.

Novotny, C. P. and Englesberg, E. (1966). The l-arabinose permease system in Escherichia coli b/r. *Biochim Biophys Acta*, 117(1):217–230.

Nunes Amaral, L. A. and Guimerà, R. (2006). Complex networks: Lies, damned lies and statistics. *Nat Phys*, 2(2):75–76.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27(1):29–34.

Oh, S. J., Joung, J.-G., Chang, J.-H., and Zhang, B.-T. (2006). Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks. *BMC Bioinformatics*, 7:284.

Oh, Y. K. and Freese, E. (1976). Manganese requirement of phosphoglycerate phosphomutase and its consequences for growth and sporulation of bacillus subtilis. *J Bacteriol*, 127(2):739–746.

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

Oikonomou, P. and Cluzel, P. (2006). Effects of topology on network evolution. *Nat Phys*, 2(8):532–536.

Oliveira, A. P., Nielsen, J., and Frster, J. (2005). Modeling lactococcus lactis using a genome-scale flux model. *BMC Microbiol*, 5:39.

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., and Palsson, B. Ø. (2011). A comprehensive genome-scale reconstruction of escherichia coli metabolism–2011. *Mol Syst Biol*, 7:535.

Palsson, B. (2009). Metabolic systems biology. *FEBS Lett*, 583(24):3900–3904.

Pandey, A. and Mann, M. (2000). Proteomics to study genes and genomes. *Nature*, 405(6788):837–846.

Papin, J. A., Hunter, T., Palsson, B. Ø., and Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol*, 6(2):99–111.

Papin, J. A. and Palsson, B. Ø. (2004). Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J Theor Biol*, 227(2):283–297.

Papini, M., Nookaew, I., Scalcinati, G., Siewers, V., and Nielsen, J. (2010). Phosphoglycerate mutase knock-out mutant saccharomyces cerevisiae: physiological investigation and transcriptome analysis. *Biotechnol J*, 5(10):1016–1027.

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

Patel, M., Johnson, J. S., Brettell, R. I., Jacobsen, J., and Xue, G.-P. (2000). Transgenic barley expressing a fungal xylanase gene in the endosperm of the developing grains. *Molecular Breeding*, 6:113–124. 10.1023/A:1009640427515.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–2448.

Perkins, J. B., Sloma, A., Hermann, T., Theriault, K., Zachgo, E., Erdenberger, T., Hannett, N., Chatterjee, N. P., Williams II, V., Jr, G. R., Hatch, R., and Pero, J. (1999). Genetic engineering of Bacillus subtilis for the commercial production of riboflavin. *Journal of Industrial Microbiology & Biotechnology*, 22:8–18. 10.1038/sj.jim.2900587.

Picard, F., Daudin, J.-J., Koskas, M., Schbath, S., and Robin, S. (2008). Assessing the exceptionality of network motifs. *J Comput Biol*, 15(1):1–20.

Pitkänen, E., Rantanen, A., Rousu, J., and Ukkonen, E. (2005). Finding feasible pathways in metabolic networks. In Bozanis, P. and Houstis, E., editors, *Advances in Informatics*, volume 3746 of *Lecture Notes in Computer Science*, pages 123–133. Springer Berlin / Heidelberg.

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.

Prigogine, I. and Rice, S. (1980). *Advances in chemical physics*. John Wiley & Sons, New York.

Qi, D., Tann, C. M., Haring, D., and Distefano, M. D. (2001). Generation of new enzymes via covalent modification of existing proteins. *Chem Rev*, 101(10):3081–3111.

Quek, L.-E. and Nielsen, L. K. (2008). On the reconstruction of the Mus musculus genome-scale metabolic network model. *Genome Inform*, 21:89–100.

Ravasz, E., Somera, A., Mongru, D., Oltvai, Z., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–5.

Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. (2003). The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*, 31(1):224–228.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):pp. 34–58.

Sales-Pardo, M., Guimer, R., Moreira, A. A., and Amaral, L. A. N. (2007). Extracting the hierarchical organization of complex systems. *Proc Natl Acad Sci U S A*, 104(39):15224–15229.

Samal, A. and Martin, O. C. (2011). Randomizing genome-scale metabolic networks. *PLoS One*, 6(7):e22295.

Sanz-Arigita, E. J., Schoonheim, M. M., Damoiseaux, J. S., Rombouts, S. A. R. B., Maris, E., Barkhof, F., Scheltens, P., and Stam, C. J. (2010). Loss of 'small-world' networks in Alzheimer's disease: graph analysis of FMRI resting-state functional connectivity. *PLoS One*, 5(11):e13788.

Schallmey, M., Singh, A., and Ward, O. P. (2004). Developments in the use of bacillus species for industrial production. *Can J Microbiol*, 50(1):1–17.

Schellenberger, J., Park, J. O., Conrad, T. M., and Palsson, B. Ø. (2010). Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11:213.

Schleif, R. (2010). AraC protein, regulation of the l-arabinose operon in Escherichia coli, and the light switch mechanism of AraC action. *FEMS Microbiol Rev*, 34(5):779–796.

Schuster, S. and Hilgetag, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2(2):165–182.

Serrano, M. A., Bogu, M., and Pastor-Satorras, R. (2006). Correlations in weighted networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 74(5 Pt 2):055101.

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

Smid, E. J., Molenaar, D., Hugenholtz, J., de Vos, W. M., and Teusink, B. (2005). Functional ingredient production: application of global metabolic models. *Curr Opin Biotechnol*, 16(2):190–197.

Smith, K. S. and Ferry, J. G. (2000). Prokaryotic carbonic anhydrases. *FEMS Microbiol Rev*, 24(4):335–366.

Smith, K. S., Jakubzick, C., Whittam, T. S., and Ferry, J. G. (1999). Carbonic anhydrase is an ancient enzyme widespread in prokaryotes. *Proc Natl Acad Sci U S A*, 96(26):15184–15189.

Sohn, S. B., Kim, T. Y., Park, J. M., and Lee, S. Y. (2010). In silico genome-scale metabolic analysis of Pseudomonas putida KT2440 for polyhydroxyalkanoate synthesis, degradation of aromatics and anaerobic survival. *Biotechnol J*, 5(7):739–750.

Sporns, O. and Zwi, J. D. (2004). The small world of the cerebral cortex. *Neuroinformatics*, 2(2):145–162.

Staden, R. (1979). A strategy of dna sequencing employing computer programs. *Nucleic Acids Res*, 6(7):2601–2610.

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.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*, 36(Database issue):D1009–D1014.

Tanaka, M., Okuno, Y., Yamada, T., Goto, S., Uemura, S., and Kanehisa, M. (2003). Extraction of a Thermodynamic Property for Biochemical Reactions in the Metabolic Pathway. *Genome Informatics*, 14:370–371.

Tashian, R. E. (1989). The carbonic anhydrases: widening perspectives on their evolution, expression and function. *Bioessays*, 10(6):186–192.

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.

van Loon, A. P., Pesold-Hurt, B., and Schatz, G. (1986). A yeast mutant lacking mitochondrial manganese-superoxide dismutase is hypersensitive to oxygen. *Proc Natl Acad Sci U S A*, 83(11):3820–3824.

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

Vellard, M. (2003). The enzyme as drug: application of enzymes as pharmaceuticals. *Curr Opin Biotechnol*, 14(4):444–450.

Vojinovi, V. and von Stockar, U. (2009). Influence of uncertainties in ph, pmg, activity coefficients, metabolite concentrations, and other factors on the analysis of the thermodynamic feasibility of metabolic pathways. *Biotechnol Bioeng*, 103(4):780–795.

von Wettstein, D., Mikhaylenko, G., Froseth, J. A., and Kannangara, C. G. (2000). Improved barley broiler feed with transgenic malt containing heat-stable (1,3-1,4)-beta-glucanase. *Proc Natl Acad Sci U S A*, 97(25):13512–13517.

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

Wang, Y., Ruan, L., Lo, W.-H., Chua, H., and Yu, H.-F. (2006). Construction of recombinant bacillus subtilis for production of polyhydroxyalkanoates. *Appl Biochem Biotechnol*, 129-132:1015–1022.

Wang, Z., Chen, T., Ma, X., Shen, Z., and Zhao, X. (2011). Enhancement of riboflavin production with Bacillus subtilis by expression and site-directed mutagenesis of zwf and gnd gene from Corynebacterium glutamicum. *Bioresour Technol*, 102(4):3934–3940.

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

Williams, D. (1991). *Probability with martingales*. Cambridge mathematical textbooks. Cambridge University Press.

Wolf, Y. I., Karev, G., and Koonin, E. V. (2002). Scale-free networks in biology: new insights into the fundamentals of evolution? *Bioessays*, 24(2):105–109.

Yamada, T. and Bork, P. (2009). Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nat Rev Mol Cell Biol*, 10(11):791–803.

Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J. D., Osterhout, R. E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H. B., Andrae, S., Yang, T. H., Lee, S. Y., Burk, M. J., and Dien, S. V. (2011). Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat Chem Biol*.

Ying, X. and Wu, X. (2008). Randomizing social networks: a spectrum preserving approach. In *Proceedings of the SIAM International Conference on Data Mining*, pages 739–750. SIAM.

Ying, X. and Wu, X. (2009). Graph generation with prescribed feature constraints. *Information Systems Journal*, pages 966–977.

Zamocky, M., Furtmller, P. G., and Obinger, C. (2008). Evolution of catalases from bacteria to humans. *Antioxid Redox Signal*, 10(9):1527–1548.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. *Genes Dev*, 21(9):1010–1024.

Zlatic, V., Bianconi, G., Díaz-Guilera, A., Garlaschelli, D., Rao, F., and Caldarelli, G. (2009). On the rich-club effect in dense and weighted networks. *Eur. Phys. J. B*, 67(3):271–275.