Computational Studies on the Evolution of Metabolism

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

Dissertation

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM

im Fachgebiet
Informatik

vorgelegt
von Diplom Informatiker Alexander Ullrich

geboren am 31. März 1982 in Leipzig

Die Annahme der Dissertation haben empfohlen:

1. Prof. Dr. Peter F. Stadler
2. Prof. Dr. Daniel Merkle

Die Verleihung des akademischen Grades erfolgt mit Bestehen der
Verteidigung am 10.10.2011 mit dem Gesamtprädikat magna cum laude

Contents

Abstract

Living organisms throughout evolution have developed desired properties, such as the ability of maintaining functionality despite changes in the environment or their inner structure, the formation of functional modules, from metabolic pathways to organs, and most essentially the capacity to adapt and evolve in a process called natural selection. It can be observed in the metabolic networks of modern organisms that many key pathways such as the citric acid cycle, glycolysis, or the biosynthesis of most amino acids are common to all of them.

Understanding the evolutionary mechanisms behind this development of complex biological systems is an intriguing and important task of current research in biology as well as artificial life. Several competing hypotheses for the formation of metabolic pathways and the mechanisms that shape metabolic networks have been discussed in the literature, each of which finds support from comparative analysis of extant genomes. However, while being powerful tools for the investigation of metabolic evolution, these traditional methods do not allow to look back in evolution far enough to the time when metabolism had to emerge and evolve to the form we can observe today. To this end, simulation studies have been introduced to discover the principles of metabolic evolution and the sources for the emergence of metabolism properties. These approaches differ considerably in the realism and explicitness of the underlying models. A difficult trade-off between realism and computational feasibility has to be made and further modeling decisions on many scales have to be taken into account, requiring the combination of knowledge from different fields such as chemistry, physics, biology and last but not least also computer science.

In this thesis, a novel computational model for the *in silico* evolution of early metabolism is introduced. It comprises all the components on different scales to resemble a situation of evolving metabolic protocells in an RNA-world. Therefore, the model contains a minimal RNA-based genetics and an evolving metabolism of catalytic ribozymes that manipulate a rich underlying chemistry. To allow the metabolic organization to escape from the confines of the chemical space set by the initial conditions of the simulation and in general an open-ended evolution, an evolvable sequence-to-function map is used. At the heart of the metabolic subsystem is a graph-based artificial chemistry equipped with a built-in thermodynamics. The generation of the metabolic reaction network is realized as a rule-based stochastic simulation. The necessary reaction rates are calculated from the chemical graphs of the reactants on the fly. The selection procedure among the population of protocells is based on the optimal

metabolic yield of the protocells, which is computed using flux balance analysis.

The introduced computational model allows for profound investigations of the evolution of early metabolism and the underlying evolutionary mechanisms. One application in this thesis is the study of the formation of metabolic pathways. Therefore, four established hypotheses, namely the backwards evolution, forward evolution, patchwork evolution and the shell hypothesis, are discussed within the realms of this *in silico* evolution study. The metabolic pathways of the networks, evolved in various simulation runs, are determined and analyzed in terms of their evolutionary direction. The simulation results suggest that the seemingly mutually exclusive hypotheses may well be compatible when considering that different processes dominate different phases in the evolution of a metabolic system. Further, it is found that forward evolution shapes the metabolic network in the very early steps of evolution. In later and more complex stages, enzyme recruitment supersedes forward evolution, keeping a core set of pathways from the early phase. Backward evolution can only be observed under conditions of steady environmental change. Additionally, evolutionary history of enzymes and metabolites were studied on the network level as well as for single instances, showing a great variety of evolutionary mechanisms at work.

The second major focus of the *in silico* evolutionary study is the emergence of complex system properties, such as robustness and modularity. To this end several techniques to analyze the metabolic systems were used. The measures for complex properties stem from the fields of graph theory, steady state analysis and neutral network theory. Some are used in general network analysis and others were developed specifically for the purpose introduced in this work. To discover potential sources for the emergence of system properties, three different evolutionary scenarios were tested and compared. The first two scenarios are the same as for the first part of the investigation, one scenario of evolution under static conditions and one incorporating a steady change in the set of "food" molecules. A third scenario was added that also simulates a static evolution but with an increased mutation rate and regular events of horizontal gene transfer between protocells of the population. The comparison of all three scenarios with real world metabolic networks shows a significant similarity in structure and properties. Among the three scenarios, the two static evolutions yield the most robust metabolic networks, however, the networks evolved under environmental change exhibit their own strategy to a robustness more suited to their conditions. As expected from theory, horizontal gene transfer and changes in the environment seem to produce higher degrees of modularity in metabolism. Both scenarios develop rather different kinds of modularity, while horizontal gene transfer provides for more isolated modules, the modules of the second scenario are far more interconnected.

Acknowledgements

First of all, I would like to thank my supervisor Peter F. Stadler for his support throughout my young scientific career.

This work would not have been possible without the effort and help of Christoph Flamm, Martin Mann, Markus Rohrschneider, Petra Pregel and Jens Steuck. And this work would have been a lot less fun without my colleagues. In Leipzig, especially my dear mensa group David, Mario, Lydia and Joe, as well as my office mates Arli, Faktheh, Stefano and Rosalina. From my Vienna time, Christian, Ronny, Dill and ovviamente Ilenia.

Further, I want to thank my family and friends for being there for me whenever i needed them. Particular thanks go to my parents for their steady support and motivation.

Finally, a special thanks to Karen, for every day of help, motivation and love. With you, everything in life becomes easier.

This thesis is based on the following publications:

Ullrich A., Rohrschneider M., Scheuermann G., Stadler P.F. and Flamm C. (2011). In silico evolution of early metabolism. *Artificial Life*, Vol. 17:2 - Spring 2011

Rohrschneider M., **Ullrich A.**, Kerren A., Stadler P.F. and Scheuermann G. (2010). Visual Network Analysis of Dynamic Metabolic Pathways. Advances in Visual Computing, *Lecture Notes in Computer Science*, Vol. 6453

Ullrich A., Flamm C., Rohrschneider M. and Stadler P.F. (2010). In Silico Evolution of Early Metabolism. *Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems*, pp.57-64, MIT Press, 2010.

Flamm C., **Ullrich A.**, Ekker H., Mann M., Hgerl D., Rohrschneider M., Sauer S., Scheuermann G., Klemm K., Hofacker I.L. and Stadler P.F. (2010). Evolution of Metabolic Networks: A Computational Framework. *Journal of Systems Chemistry*, Vol. 1/1/4, 2010

Ullrich A. and Flamm C. (2009). A Sequence-to-Function Map for Ribozyme-catalyzed Metabolisms. *ECAL, Lecture Notes in Computer Science (LNAI)*, Vol. 5778

Ullrich A. and Flamm C. (2008). Functional Evolution of Ribozyme-Catalyzed Metabolisms in a Graph-Based Toy-Universe. *CMSB, Lecture Notes in Computer Science (LNBI)*, Vol. 5307

Ullrich A. Evolution of Metabolism in a Graph-Based Toy-Universe. *Diplomarbeit*, Universitt Leipzig, 2008.

Chapter 1

Introduction

> Give biologists a cell, and the'll give you the world. But beyond assuming the first cell must have somehow come into existence, how do biologists explain its emergence from the prebiotic world four billion years ago?
> Robinson [138]

This quote perfectly summarizes the situation of modern biology with its capabilities to describe today's organisms in great detail and at the same time its shortcomings in the explanation of how all of that emerged out of "dead" matter. Starting from Darwin's evolutionary theory [33] that explains the process of natural selection which shapes the evolutionary patterns of all species, to Mendel's laws [113], that tell about the inheritance of traits from parents to children. Later, Watson and Crick "found the secret of life" in the discovery of the double helix structure of DNA [171], a molecule made of simple chemicals carrying the genetic instructions necessary for the development of all organisms on earth. These great discoveries are the starting points of modern biology facilitating the advances of genetic technology making vasts amounts of data available, from sequence information of single compounds to interaction networks for entire organisms. Nevertheless, they do not solve the mystery of life.

While we can analyze the biological networks of many organisms and trace back evolution for billions of years, there is a barrier past which all the technology of modern biology is not able to look. This barrier is the last common ancestor of all living organisms and there is little to no data about the path to this first cell exhibiting all those features of modern organisms. Therefore, there can only be speculated about the evolution of this predecessor and the emergence of its properties. Besides the theories and experiments, that will be discussed briefly in the next chapter, trying to tackle this problem, simulation approaches have been an intriguing source of potential explanations beyond the scope of analyzing data. For the origin of life field, there are high demands on a simulation. The system of study, the protocell, has to be modeled on multiple scales, realistic to a certain extent but at the same time computationally feasible. Thus, making it a chemical, biological and computational challenge altogether, demanding the application and combination of knowledge and concepts

of all of these fields. **Providing a realistic and expressive model for the evolution of early metabolism will be a major goal of this thesis.**

All complex biological organisms have developed certain desired properties, such as the ability of maintaining functionality despite changes in the environment or their inner structure (e.g. through mutations), the formation of functional modules (e.g. metabolic pathways, organs) and most essentially the capacity to adapt and evolve in a process called natural selection. This observation becomes even more impressive considering that all of these systems are based on the interactions of chemical molecules on the most basic level.

Early metabolism appears to be a fitting field of research to tackle the questions of the emergence of complex properties and the evolution of biological systems. On the one hand, cellular metabolism is probably the best studied biological system, especially with the advance in genome sequencing technology and the corresponding reconstruction of whole-genome metabolic networks for many different organisms. On the other hand, the earliest steps in the evolution of metabolism, i.e. the formation of metabolic pathways from chemical reactions, the transition from uncatalyzed to catalyzed reaction systems and maybe even the origin of life itself, is still unknown and the evolutionary mechanisms are still to be discovered. The same could be said about the knowledge of the above mentioned properties that complex biological systems possess. While methods exist to measure these properties and theories trying to explain their existence, it is not whatsoever clear how they emerge and are preserved throughout evolution. Many of these problems have been discussed by generations of researchers providing insightful hypotheses and scenarios which will be used as anchors and guide lines for the research in this thesis. One popular thread in the research of evolution are theories on the formation of metabolic pathways, several hypotheses exist, from the backward evolution to the patchwork model. Some of these hypotheses exhibit some more evidence in the data than others, but all have some eligibility. Other important problems lie in the field of system properties, from their emergence to the reconciliation of intuitively opposing concepts, such as the findings that redundancy leads to robustness or the unification of robustness and evolvability in complex systems demonstrated on the example of the RNA sequence to structure map by Wagner et al. **Novel approaches to many of these problems will be introduced in this work, using the *in silico* evolution model and several new network analysis methods.**

Understanding the evolutionary mechanisms behind the development of biological systems and their properties benefits theoretical research on evolution and gives insights about the systems and their components itself. Beyond that, it provides valuable knowledge for the construction of complex artificial systems, from chemical reaction networks in metabolic engineering to abstract technological networks, that are desired to possess properties such as robustness and modularity.

In the following chapter, several problems in the field of metabolic evolution from the origin of life to the formation of metabolic pathways and the emergence of complex properties will

be introduced and historic as well as ongoing research will be discussed.

In Chapter 3, techniques for the modeling of chemical reactions systems, such as metabolism, are discussed with their applications, drawbacks and challenges. The following chapter will then provide a detailed account of a sophisticated multi-scale model for the evolution of early-metabolism.

The formation of metabolic pathways is investigated with the introduced *in silico* evolution model in Chapter 5, several existing hypotheses are tested and a new and unifying hypothesis is proposed. The emergence of complex properties in the simulated networks from this study are the focus of Chapter 6. Here existing and novel network analysis approaches are introduced and used to discover factors determining the evolution of metabolic networks and their properties.

Chapter 2

Origins of Life and early Evolution

> Life emerged, I suggest, not simple, but complex and whole, and has remained complex and whole ever since, not because of a mysterious elan vital, but thanks to the simple, profound transformation of dead molecules into an organization by which each molecule's formation is catalyzed by some other molecule in the organization. The secret of life, the wellspring of reproduction, is not to be found in the beauty of Watson-Crick pairing, but in the achievement of collective catalytic closure. The roots are deeper than the double helix and are based in chemistry itself. So, in another sense, life - complex, whole, emergent - is simple after all, a natural outgrowth of the world in which we live.
> Stuart Kauffman [93]

2.1 Origin of Life

The research on the origins of life, the transition from inanimate to animated matter, faces one major problem. While having some knowledge of the chemical processes and constitutions of the atmosphere before the alleged beginning of life on the early earth and more elaborate insights of the biochemical and biological organization of some ancient successors of early life forms, there is little known about the actual period of the transition to life, let alone representatives of the first living things. The core of the problem is that we can look back in the history of life only until a certain point of time, when there already exist fully functioning cells containing all the building blocks that can be observed in today's organisms. It is a consent in biological science that all living organisms on earth have one common ancestor. This last universal common ancestor (LUCA) used already a genetic code based on DNA, expressed proteins from these genes and performed many metabolic pathways that lie at the core of modern metabolism. As interesting as this observation may be, it also means that by looking at modern or ancient organisms, we can not know for certain how these

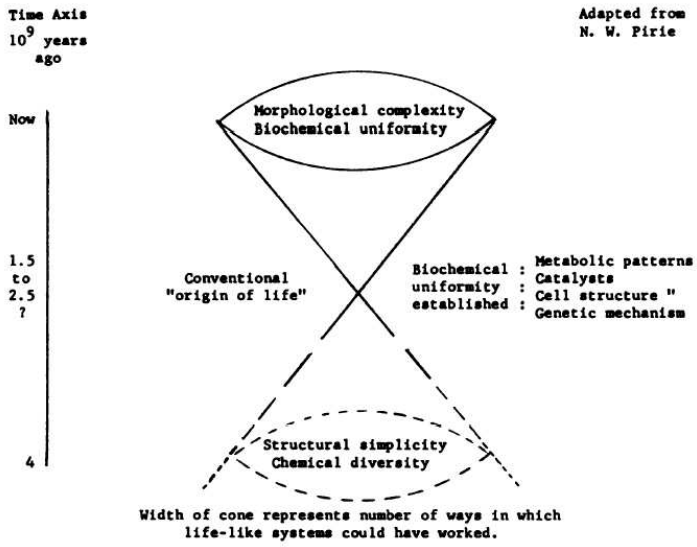


Figure 2.1: The dilemma of the origin of life research described by Eakin. It is difficult to look past the last universal common ancestor which possessed all biological features of modern cells. From [40].

essential components evolved and life itself began. However, many intriguing theories have been suggested and many experiments have been performed to shed some more light onto this mystery.

Before discussing some of these ideas, another question asks to be solved. What is life? Again, there is not just one answer and it is difficult to judge whether one is better than the other. The decision whether or not something is alive is gradual rather than clear-cut. Is a virus alive because it can evolve? Is a seed that has been dried for years alive because it has the potential of performing metabolism again under other conditions?

Whether or not something is alive depends on the processes going on in this something. Most definitions for life identify three essential processes that have to be active in a living organism. One such process is growth, in the sense of being able to use free energy and resources from the environment to maintain its functions and accumulate enough matter/stuff/molecules to reproduce. This process can be summarized as metabolism. Another essential process is replication, passing on heritable information which regulates the processes of the organism and is subject to evolution. In modern organisms this could be accounted to its genes and everything else belonging to its genome. The last functionality is keeping all these components and processes together, such as a membrane of a cell. The order of these processes in this paragraph is arbitrary and not based on their importance. One can argue that all three processes are mutually dependent or at least affect each other. For example, the chemical interactions of metabolism require a spatial concentration in order to work efficiently. Both, the compartment boundaries and the reproduction machinery need to be synthesized through metabolism. The heritable information in turn has to regulate these processes. Therefore, it

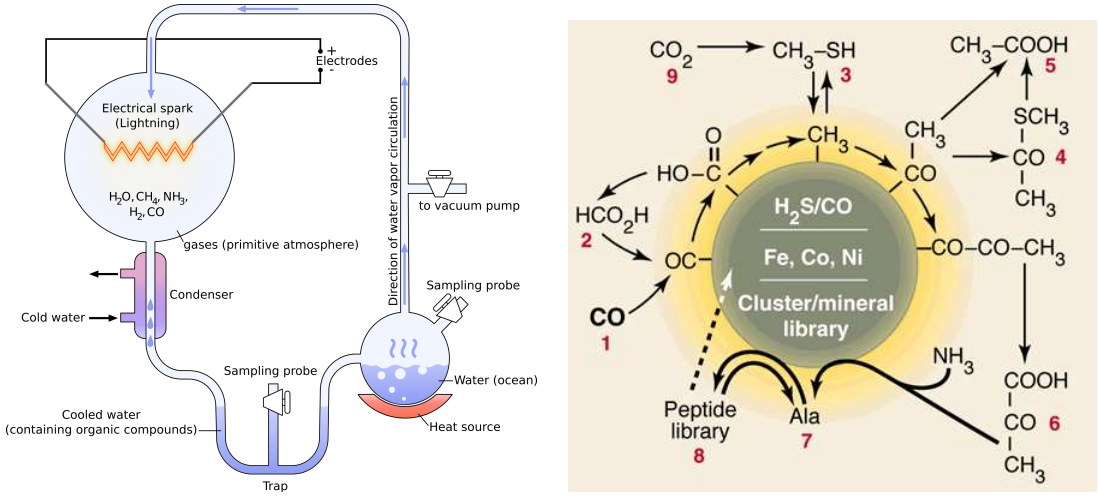


Figure 2.2: Arguments for the metabolism first scenario by a) Miller and Urey's experiment for producing amino acids out of inorganic compounds of a supposed early earth atmosphere (from [178]), b) Wächtershäuser theory on the evolution of metabolism around minerals (from [167]).

makes sense to view all three processes as fundamental basis of life.

However, for the question of the origin of life, matters are again more complicated, since the emergence of all three functionalities at the same time are less probable than all occurring separately. This poses another question. What came first? Theories and experiments for all three possibilities exist. The most famous experiment being that of Miller and Urey [115, 116], synthesizing amino acids from a mixture of anorganic molecules abundant in the atmosphere of the early earth, such as methane, water and ammonia. Since amino acids are the building blocks of proteins and thus the catalytic elements of all modern cells, these experiments would suggest that life started with the emergence of enzymes and thus metabolism. However, many questions remain. The first is about the assumptions of these experiments itself, it is not clear whether such a reducing atmosphere or similar environments were present at that time. Other questions concern the polymerization of amino acids and the replication of their polymers without the existence of ribosomes or any kind of RNAs. Another sort of theories, works with cofactors as catalytic elements instead of peptides [166, 110]. ATP could be such a catalyst, in modern metabolism it is a widely used coenzyme and an important carrier of chemical energy. Similarly, ATP could have played an important role in the early metabolism and exist in high concentrations. This would also mean that the related nucleotides ADP and AMP would be present in such metabolisms, therefore, one basic component of RNA and DNA. Starting as a parasitic side product RNA could have emerged from early metabolism and later develop to a catalyst and information carrier [109].

John von Neumann famously made the analogy between metabolism and replication in a living organism on the one hand with the hardware and software of a mechanic automata

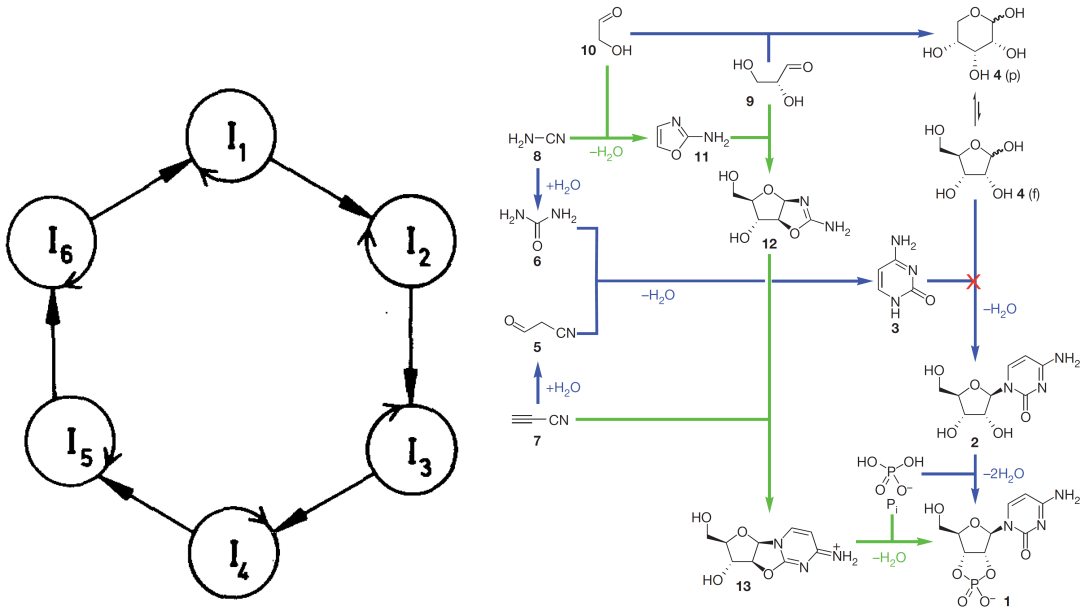


Figure 2.3: Arguments for the RNA world by a) Eigen and his hypercycle theory (from [42]), b) Sutherland providing a new route for the synthesis of RNA-nucleotides, blue- the old rout, green- new route (from [132]).

on the other hand [164]. In this analogy the chemical processes of metabolism correspond to the information processing of the automata's hardware, while the nucleic acids that hold the information correspond to the software. Accepting this assumption, it becomes clear that metabolism is more fundamental than nucleic acids and thus had to emerge first. However, this idea ignores one important molecule that can serve both as hardware and software. RNA in the form of mRNA as transcript for translation to a polypeptide or as genetic material in viruses can serve as simple information carrier, software. On the other hand, it can also act as hardware, performing chemical processes in the cell, for example as ribozymes or in its function as rRNA or tRNA. The first advocates of the RNA world [66] and their arguments, can be found in Eigen's theory of hypercycles [42] describing how sets of self-replicative RNAs could form higher-level autocatalytic cycles and in the experiments of Oro and Orgel showing how nucleotides could have been synthesized [124] and polymerize from a template [114], respectively, under simple conditions as on the early Earth. Recently, some of the weaknesses in the original synthesis routes were eliminated through a clever new idea for a nucleotide synthesis, making the emergence of nucleotides in an early earth atmosphere even more likely [132], see Figure 2.3. Further arguments for RNA's potential in the origin of life stem from *in vitro* evolution experiments made since the 1990s [43], showing that RNA can be evolved to bind all kinds of molecules, in some cases better than there protein enzyme counterparts [136]. Another theory of how life could have originated through the emergence of RNA, is described

in [23] introducing the idea that RNAs in protocells could have replicated by itself through thermal cycling. Protocells could have then be selected for faster replication in the form that protocells with faster replicating RNA molecules take away material from surrounding cells. In such a scenario nucleotides such as AMP would be used as chemical currency first and lead to their catalytic role in metabolism as ATP later.

There are also theories that focus on the importance of compartments in the development of the first living things such as the clay theory [15] proposing that organic molecules concentrated on the surface of the clay which was also able to catalyze their polymerization. Similar approaches exist accounting such functions to other minerals [165]. Another interesting idea that favors the compartment-centered scenario is the regosome model [121] where regoliths, porose dust grains, serve as compartments where chemicals would accumulate and react with each other. One current trend in the origin of life research follows this scenario. The abundance of many ancient bacteria around hot deep water vents brought up interest to these possible places for the origin of life [110]. The sulphide in the hot water of these vents forms membranes and spheres when going to the cold ocean water. Furthermore, these membranes can absorb organic molecules and some sulphide complexes could catalyze certain chemical reactions.

Although all of these theories do not solve the question of the origin of life indefinitely, they give some guidance in the further research and also for the development of a realistic model that incorporates all the important aspects of an early protocell.

2.2 Metabolic Evolution

Laying aside the troubles with the question of the origin of life, there are still paths to be discovered in the history of evolution. One of them, which will be a major focus of this work is metabolic evolution, including the evolution of enzymes and the formation of metabolic pathways from catalyzed chemical reactions. The goal is trying to understand the evolutionary mechanisms of complex biological systems, which has been an interest of research in biology as well as artificial life. In this section, some of the main theories on metabolic evolution will be discussed accompanied with their evidence from experimental data as well as other simulation studies.

2.2.1 Arguments from the Data

The mechanisms that governed the formation of metabolic pathways from chemical reactions has been discussed for decades and several hypotheses have been proposed since the 1940s. Research on the TIM β/α-barrel fold architecture [26], for instance, shows that the evolution of modern metabolism is mainly driven by enzyme recruitment, as suggested by the patchwork model [182, 89]). Enzymes with β/α-barrel fold architecture catalyzing similar chemical

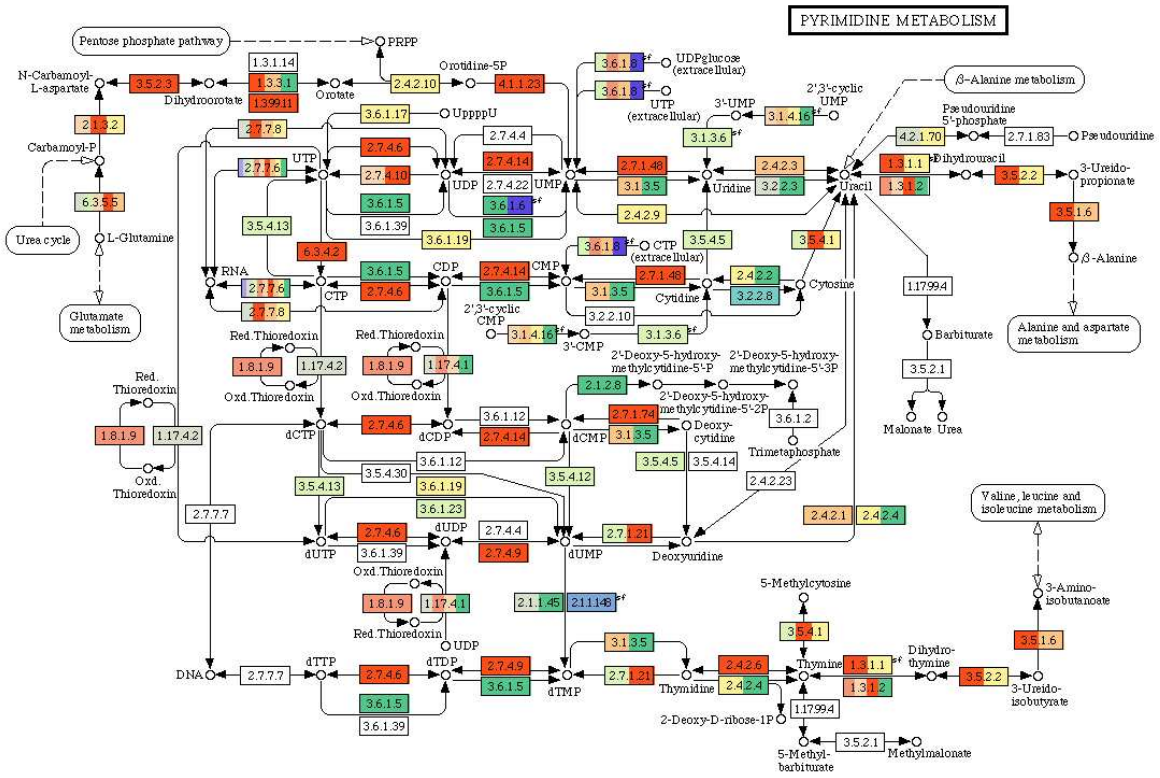


Figure 2.4: Metabolic network of the pyrimidine metabolism from the MANET database. Older enzymes are colored in red, evolutionary younger enzymes in blue. The mosaic-like constitution supporting the patchwork evolution scenario is apparent. [157]

reactions were found in many different metabolic pathways across the metabolism. This picture of metabolism as enzyme mosaic was shown for several enzymes of *E. coli*, through structural assignments and sequence comparison of their protein domains [156]. On the example of the pyrimidine metabolism in Figure 2.4, the mosaic-like constitution of metabolic networks concerning the evolutionary relationships of enzymes becomes apparent.

Gene duplications, on the other hand may facilitate the specialization of an originally multifunctional enzyme, such as the Carbamoyl-phosphate-synthetase, to diverse function in new pathways [82]. Similarly, entire metabolic pathway may duplicate and specialize, as it has been the case for the tryptophane and histidine biosynthesis [63, 89].

The ability of enzymes to catalyze additional reactions other than those for which they are physiologically specialized, dubbed "enzyme promiscuity" [96], forms an important evolutionary reservoir from which novel catalytic functions can be drawn. Promiscuous enzyme activities, although far less efficient than well-evolved ones, can be assembled into novel metabolic pathways [97], which can provide a selective advantage in particular environmental niches. The evolutionary potential of enzyme promiscuity thus extends far beyond mere

enzyme recruitment.

For an in depth discussion of these theories and further examples of pathway evolution it is referred to three interesting recent review articles [20, 45, 144].

2.2.2 Arguments from Simulations

Several simulation studies, modeling the evolution of complex networks in general or metabolic networks in specific, have provided some insights into the formation of complex systems and their properties. Most famously Barabasi and Albert [7] account of an computational model of network evolution provided an answer for the scale-free architecture of complex networks like metabolic networks. Their model grows a network through preferential attachment, i.e. new nodes are preferentially added to highly connected nodes, leading to networks with a scale-free node-degree distribution. Although this is a useful description of network evolution it does not actually explain the mechanisms behind it.

Pfeiffer [130] goes one step further in explaining the emergence of the scale-free architecture and existence of hub-metabolites in metabolic networks. The computational model starts off with a metabolism containing a few enzymes with broad specificity, i.e. catalyzing a great number of chemical reactions but at a low speed. Then duplication and specialization events occur throughout the evolution, leading to many specialized enzymes. The specialization of enzymes leads to the disappearance of chemical reactions and metabolites which then shapes the metabolic networks in the way we can observe them today.

Another approach by Hahndorf [71] identifies several metabolites and chemical reactions that are of great importance or even essential for the evolution of a metabolic network. Starting with small sets of molecules, they apply step by step known biochemical reactions from the KEGG database to grow a metabolic network. This procedure leads to so-called scopes, sets of compounds that are closed concerning the reaction application. Those metabolites that lead to scopes similar to the entire network can then be considered evolutionary relevant, thus, may have played a role in the early evolution of metabolism. Most of these important metabolites were cofactors such as ATP or Coenzyme A, which are present in many reactions of modern metabolic networks.

A computational model for the evolution of networks based on a simple artificial chemistry is presented in a work by Sanjay Jain [87]. The evolved networks exhibit autocatalytic structures and form a fixed core as well as peripheral modules [88]. The simulation starts from a random network graph where a node is a species and a connection between two species indicates that the one catalyzes the production of the other. The graph is then updated for several generations, by mutating (random rewiring) the least connected node, increasing the connectivity and complexity of the network.

2.2.3 Scenarios

Based on the evidence from modern metabolic networks, several hypotheses to explain the evolution of metabolism in general and the emergence of specific metabolic pathways have been suggested. The four most widely cited scenarios (see Figure 2.5) are briefly discussed in the following paragraphs.

The Backward Evolution Hypothesis was one of the first theories for the evolution of metabolic pathways, proposed by Horowitz [81]. It assumes that an organism is able to make use of certain molecules from the environment. However, individuals that can produce these beneficial molecules by themselves gain an advantage in selection in the case of depletion of the "food source". Therefore, new chemical reactions are added that produce beneficial molecules from precursors that are abundant in the environment or that are produced in turn by the organism's metabolism. As a consequence, one should observe more ancient enzymes downstream in present-day metabolic pathways. Towards the entry point of the pathway, younger and younger enzymes should be found (see Figure 2.5(a)). Backward evolution has been proposed for both the glycolytic pathway [57] and the mandelate pathway [129].

The Forward Evolution Hypothesis can be seen as an extension or counterpart of the backward evolution hypothesis, reversing the direction of pathway evolution. [69], and later [30], argue for a pathway evolution in forward direction, requiring that the intermediates are already beneficial to the organism. This is in particular plausible for catabolic pathways, where the organism can extract more energy by breaking food molecules down to simpler and simpler end products. Older enzymes are then expected to be upstream in the pathway, with younger enzymes appearing further downstream (see Figure 2.5(b)). The isoprene lipid pathway [126] is an example for the development of biosynthetic pathways in the forward direction.

The Patchwork Model [182, 89] explains the formation of pathways by recruiting enzymes from existing pathways. The recruited enzymes may change their reaction chemistry and metabolic function in the new pathways and specialize later through evolution. This introduction of new catalytic activities lead to a selective advantage. Looking at the constitution of a pathway formed by enzyme recruitment, we should observe a mosaic-like picture of older and younger enzymes mixed throughout the pathway (see Figure 2.5(c)). The observation that the TIM β/α-barrel fold architecture occurs in many different pathways corroborates widespread enzyme recruitment in modern metabolism [20]. Other examples are the pyrimidine metabolism and the histidine biosynthesis [129].

The Shell Hypothesis was proposed by [117]. It argues for the case of the reductive citric acid cycle that in the beginning an auto-catalytic core is formed from which new catalytic activities and pathways could be recruited and fed. Thus a metabolic shell would form around this core. Enzymes in the core would likely be less prone to mutational changes because they are essential for the organism. Thus, one should still be able to observe a core of ancient enzymes (see Figure 2.5(d)). According to Morowitz [117] the reductive citric acid cycle

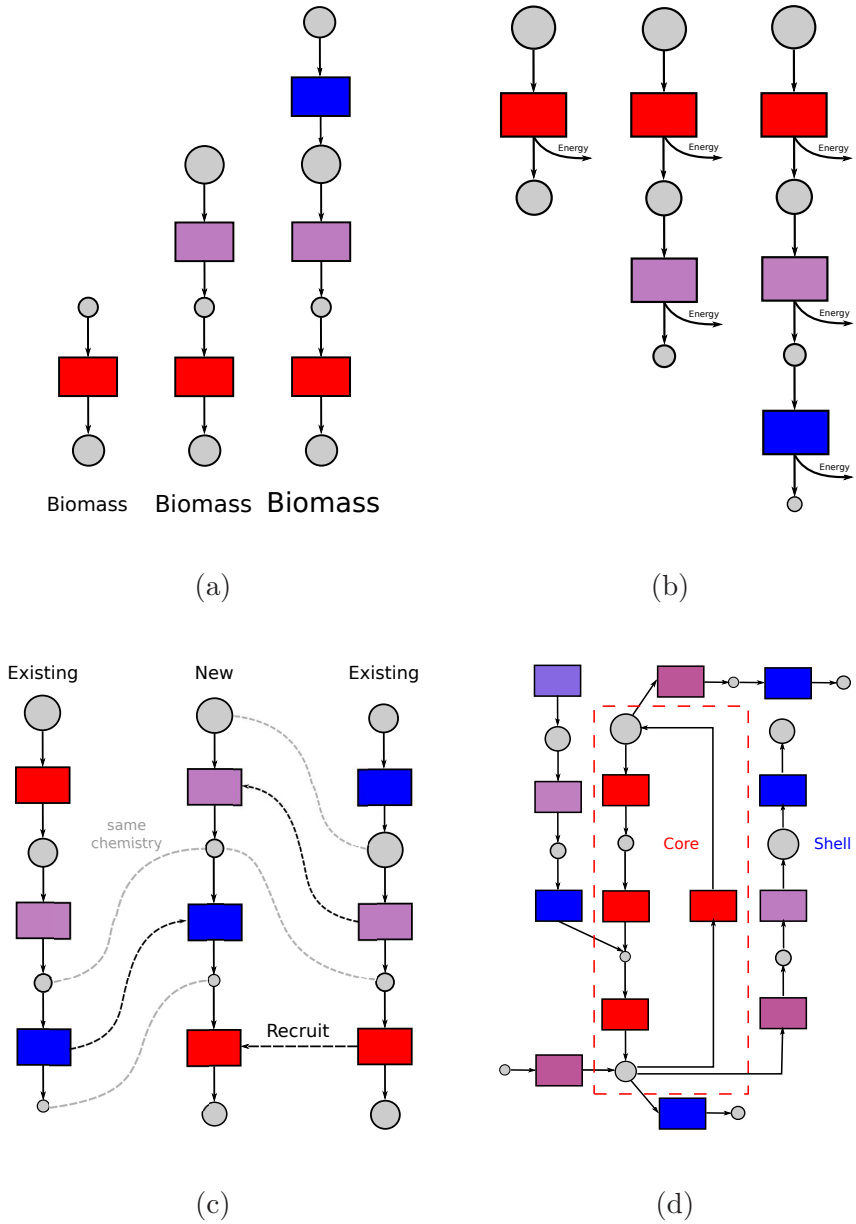


Figure 2.5: Hypotheses about the formation and evolution of metabolic pathways. (a) Backward evolution, (b) Forward evolution, (c) Patchwork model, (d) Shell hypothesis. Colored squares represent enzymes, gray circles are metabolites. Color encoding for enzymes stand for their age, red being older and blue being younger enzymes.

constitutes such an autotrophic synthetic system.

2.3 Complex Properties

Living organisms adopt to the environment by means of gradual change of their internal networks and regulations. Throughout the evolutionary process, biological systems developed certain desirable properties, such as robustness, modularity, flexibility and not least evolvability. Despite the profound knowledge of these properties and the processes within biochemical networks, the causes for the emergence of system properties are less well understood in most cases. In the next paragraphs, these properties will be defined and some of the challenges as well as findings concerning their origin and evolution will be elucidated. In Chapter 6 the emergence of complex properties will be investigated using the simulation framework introduced in Chapter 4 and several network analysis tools.

2.3.1 Robustness

Biological organisms are highly robust in the sense that they can maintain their basic functionality despite perturbations in their structure or dynamics, through genetic changes, such as mutations or epigenetic changes such fluctuations in metabolite concentrations. There are manifold sources of robustness in a system, such as structural and temporal modularity, functional redundancy and plasticity or dynamical stochasticity. Both experimental [16] and theoretical approaches [176] have tried to put a number on the impact of these sources.

While it is obvious that robustness is beneficial to a system, it is not as clear how that does not decrease its ability to evolve. If the system is less likely to change its functionality (phenotype) in the case of a change in its structure (genotype), then it should be more difficult to reach other potentially favorable (phenotypic) states. However, for complex biological systems the opposite is true. One explanation for this fact is the nature of the so-called genotype-to-phenotype maps which exist on many levels of biological organizations.

There are still open questions about the evolution of robustness. It has been discussed in the literature whether robustness is selected directly through natural selection or indirectly as an inherent property stemming from other properties of the system [58]. Another interesting problem is the relation between genetic and epigenetic robustness [168], e.g. are mutations necessary to develop epigenetic robustness and vice versa, is environmental noise necessary for genetic robustness or does it even decrease it.

Robustness is also under active research in the field of complex systems, where the focus lies on the analysis of the network structure and topology. A great variety of network types, such as social networks, road maps or the world wide web are investigated. Complex networks are often classified based on their connectivity distribution ($P(k)$), most prominently the classes of exponential and scale-free distributed networks. While both network types are highly robust,

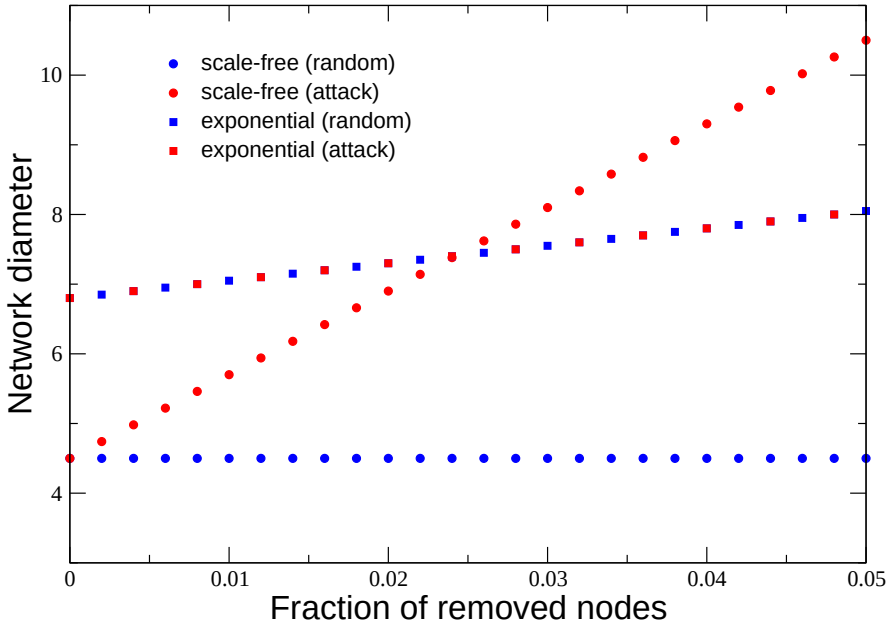


Figure 2.6: Error tolerance of scale-free and exponential networks. The scale-free networks (circles) have a higher tolerance against random errors (blue) than exponential networks (squares), but are more prone to targeted attacks (red) as can be seen by the increase in the network diameter. Adapted from [1].

scale-free are particularly robust against random errors. However, the scale-free architecture of these networks with the abundance of some highly connected nodes makes them also prone to controlled knockouts (attacks), see Figure 2.6.

2.3.2 Modularity

A system is modular if it contains subsystems (modules) that exhibit a distinct function. Studies on several biological systems [128, 102] have shown that the underlying networks are highly modular (more than random networks) and have even defined a specific structure that is abundant in most of these networks. This so-called core-periphery organization, consists of a densely connected core and a number of periphery clusters [32]. The structure of metabolic networks is further described as organized in highly connected modules that form larger modules in a hierarchical manner [134], combining the observations that metabolic networks have a scale-free connectivity distribution which usually does not occur in modular networks and at the same time they have a high clustering coefficient suggesting strong modularity.

Modularity in complex networks is often measured by the clustering coefficient of its nodes. Biological systems with a hierarchical modularity show high average clustering coefficients independent of the networks size. Further, the clustering coefficients of single nodes scale with a power law against the connectivity of these nodes [119]. Another approach is to find

modules through network decomposition [183]. Besides these methods focusing on structural and topological features of the underlying networks there has been proposed a different idea. The concept of metabolic pathways is considered a special functional subunit of metabolic networks [143]. Metabolic pathways are sets of reactions or enzymes which work at steady-state, i.e. metabolite concentrations are kept constant, and they fulfill certain thermodynamic constraints, such as irreversibility of reactions. Particularly interesting in this context is the set of extreme pathways, from which all possible pathways through the network can be generated.

Again the question about the emergence of this system property arises, and has been the subject of intensive research [170]. It is known that system that optimize toward a static goal tend not to evolve modular structures in their underlying networks. In fact, even already existing modularity is lost under these conditions [91]. If the optimization of fitness in a system does not let modularity emerge, then how does modularity emerge and what use does it have for the system. One crucial factor could be the ability of modularity to increase the evolvability of a system [168, 180], because small peripheral modules with a distinct function evolve faster than a large complex with an overall function. Is there then a selection for modularity because of the enhanced evolvability, or is modularity actually a side-product of other evolutionary processes. In the beginning it was stated that modularity does not emerge when optimizing toward a static goal, so change in the fitness landscape through adaptations to new environments [128] or new goals [108] can lead to the development of modularity. Another cause for the formation of modularity lies in the clustering of genes through genetic operations, in particular horizontal gene transfer [105]. Several comparative and simulation studies have been performed to confirm these hypotheses for the origin of modularity [106].

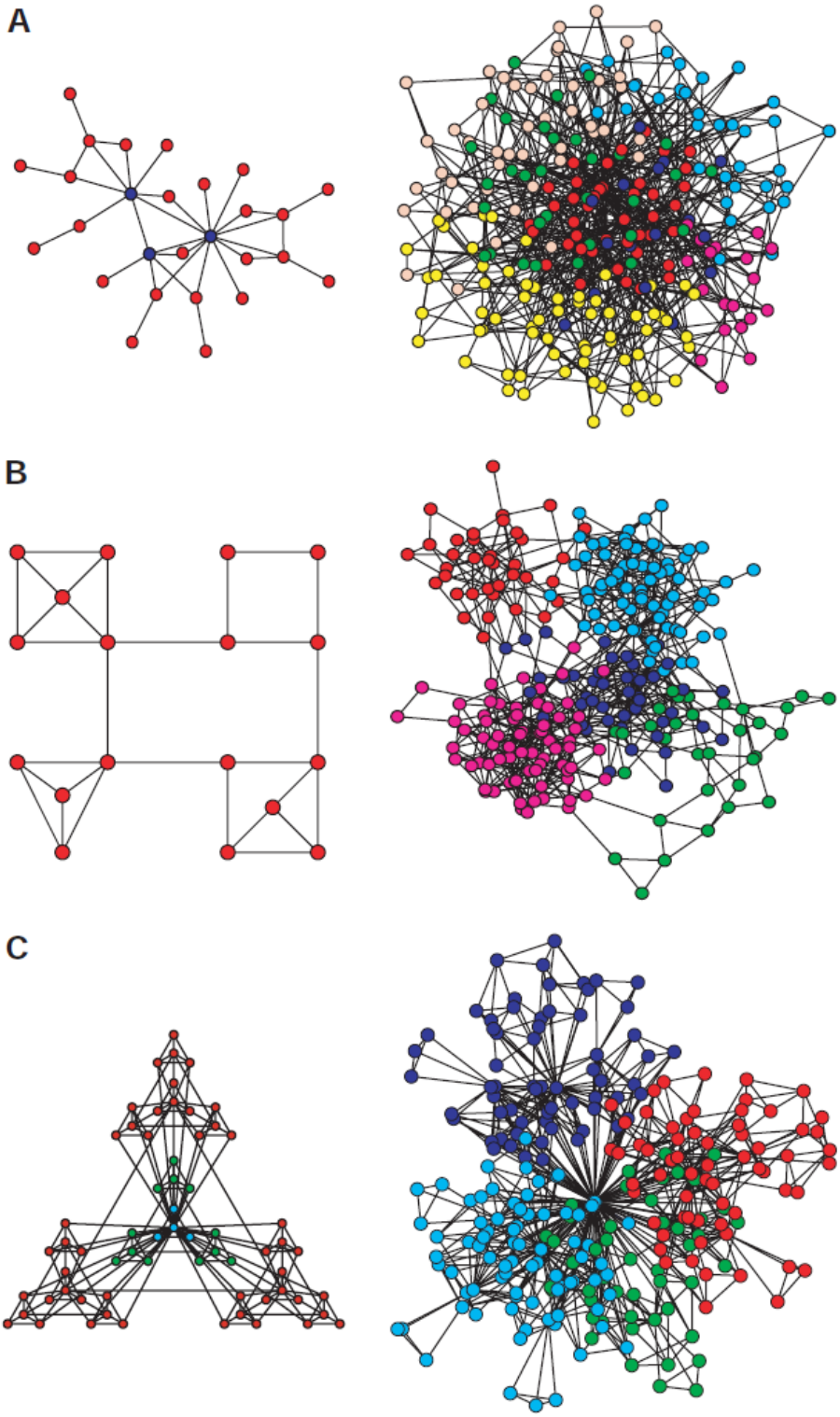


Figure 2.7: Three types of networks with different degrees of modularity. A scale free networks with very low modularity. B modular networks, with extremely high modularity. C networks with hierarchical modularity with a high degree of modularity and scale-free degree distribution. Taken from [134].

Chapter 3

Modeling Chemical Reaction Systems

> The mere formulation of a problem is far more essential than its solution, which may be merely a matter of mathematical or experimental skills. To raise new questions, new possibilities, to regard old problems from a new angle requires creative imagination and marks real advances in science.
> Albert Einstein

Chemical reaction systems occur in several forms all around us. They span from the atmosphere of the early earth or other planets, that formed the first forms of life, to the complex system of biochemical reactions in the cells of our bodies, the metabolism. Further, artificial chemical systems such as combustion or novel syntheses of products for the chemical industry and pharmacy. To understand the behavior and properties of these systems, the change of molecule concentrations and the essentiality of chemical reactions, they have to be modeled in an appropriate way. The knowledge that can be gained from the models of real world or artificial systems can be used to assess and manipulate these systems in a controlled direction.

In this chapter, several approaches to model and represent the different aspects of chemical reaction systems will be introduced and their applications and shortcomings will be discussed.

3.1 Chemical Reaction Systems

From complex biological systems such as metabolism to comparably simple artificial chemical systems like combustion, all chemical reaction systems have in common that they consist of molecules and chemical reactions, together forming chemical reaction networks.

3.1.1 Chemical Reactions

A chemical reaction is a process that transforms some initially existing molecules (reactants) to other molecules (reactants) of the same combined mass. Each chemical reaction has a reaction center which comprises all atoms and bonds which participate in the actual transformation, i.e. bonds that are broken or formed and the adjacent atoms with involved valence electrons. Depending on the Gibbs free (G) energy of the substrates and products, chemical reactions are either exergonic ($G_{substrate} > G_{product}$) or endergonic ($G_{substrate} < G_{product}$). However, in both cases the substrate molecules have to pass one or more transition state of maximal free energy which requires a certain activation energy (E_A). This activation energy further determines the speed of the chemical reaction or more precisely the reaction rate constant (k), which can be calculated from the Arrhenius equation (see Equation 3.1), where A is the frequency factor, e is the irrational number with a decimal approximation of 2.718281828, R is the gas constant and T the temperature.

$$k = Ae^{-\frac{E_A}{RT}} \tag{3.1}$$

However, to determine how fast the chemical reaction converts its substrates into its products in the actual system, one also needs the concentration of the initially participating molecules in the cell. The reaction rate (r) is computed from the rate constant and the molecule concentrations using the rate equation Equation 3.2, where $[A]$ is the concentration of a molecule A in the cell, the exponent in $[A]^a$ indicates the order of the reaction with respect to the molecule A.

$$r = k[A]^a[B]^b \tag{3.2}$$

Enzymes can catalyze chemical reactions by reducing the activation energy which increases the rate constant and reaction rate. The decrease of the activation energy of a reaction is achieved through interactions of the enzyme with the substrate molecules which changes the reaction path lowering the barrier of free energy that emerges by passing the transition state (see Figure 3.1). Not all possible chemical reactions actually occur in the biological systems we know here on earth, the biochemical reactions that we observe today are rather a small subset of them. Most molecules in the cell consist only of hydrogen, carbon, nitrogen or oxygen and in smaller quantities but also biologically important occur molecules containing phosphor and sulfur. Further, biomolecules have a limited number of linkage (connections between to atoms, e.g. C=C, N-H, P=O) types, a mere 16 types cover the majority of biomolecules. Therefore, it is possible to have almost complete databases for biochemical reactions [90], while for chemical reactions the respective databases do not hold such claims and the focus lies more on reaction mechanisms. There is a steady research for novel chemical reactions.

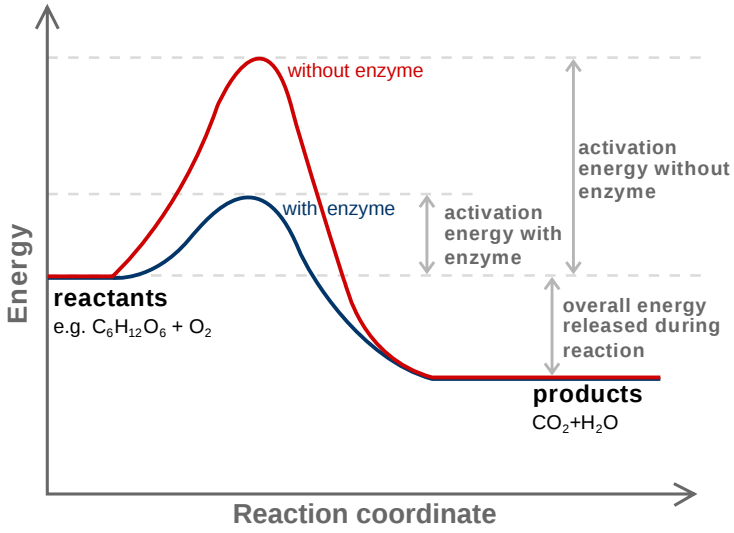


Figure 3.1: Reaction path of a chemical reaction showing the change in Gibbs free energy of the molecules along the reaction coordinate. The red curve illustrates the reaction path without the presence of an enzyme, while the blue curve is the path with enzyme. The blue curve has a lower activation energy, the free energies as well as the released energy are equal in both cases. Taken from [179].

3.1.2 Chemical Reaction Networks

A chemical reaction network is simply the composition of the molecules of the system and the reactions that are applied on them. It provides a unified representation of the interplay between the chemical reactions and, therefore, allows a more holistic analysis of the behavior and characteristics of a chemical reaction system.

3.2 Modeling

With the development of high-throughput technologies in biology and the accompanying vast amounts of data from different levels the need for models describing, analyzing and predicting biological systems and their processes arose. Away from traditional biology focusing on single parts of a system independently, the aim now is to view the system as a whole including processes at all scales, ideally.

Modeling is a powerful tool in many scientific fields from physics to biology, however, it is always an abstraction and simplification of the real system. Therefore, many different approaches, representations and models for the same system are possible, this is especially true for a complex biological system as metabolism. Further, different objectives command different strategies, thus, not all questions about one biological system will be solved with

one universal model. If the aim is to explain only individual aspects then minimal models covering limited components and making strong simplifications of the system may suffice. However, for a description and deeper analysis of processes encompassing different scales and parts, such as evolution or the emergence of complex properties the demands on the model become more challenging. In the following paragraphs and the entire work, the focus lies on models of metabolism able to predict the behavior of single components, like metabolites and enzymes on the one hand and the development of higher order constructs, such as pathways and the overlying network on the other.

3.2.1 Stoichiometric Matrix

A chemical reaction system can be represented by the chemical equations of its reactions. The stoichiometric information of these equations is summarized in a matrix. The stoichiometric matrix S of a metabolic network (Figure 3.2) comprises the stoichiometric coefficients s_{ij} for all metabolites ($m \in M$) and reactions ($r \in R$)

$$S = s_{ij}; \quad 0 \leq i < |M| \ \wedge \ 0 \leq j < |R| \tag{3.3}$$

where each row describes a metabolite and indicating participation in a reaction ($s_{ij} \neq 0$) and each column represents a reaction denoting which metabolites it uses ($s_{ij} < 0$) or produces ($s_{ij} > 0$). Furthermore, the stoichiometric matrix provides information about the dynamics of the molecule concentrations in the form of mass balance equations

$$\frac{dx_i}{dt} = \sum_r s_{ir} \ v_r \tag{3.4}$$

or in matrix form as

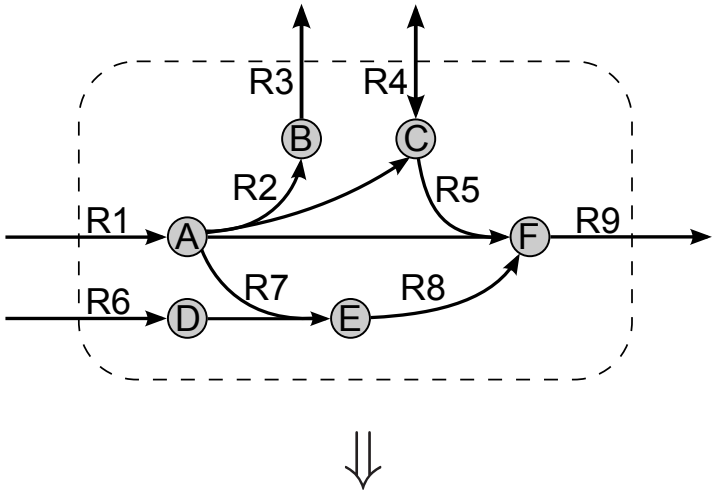
$$\frac{dx}{dt} = S\vec{v} \tag{3.5}$$

where, v are flux vectors, r the chemical reactions of the system and x_i the concentrations of a molecule i. This information about the time derivatives is used in the following modeling approaches. The information about the flux vectors, in particular the nullspace, is used later in the stoichiometric approaches.

3.2.2 Kinetic Modeling

Kinetic Modeling is the most common way to describe and predict the behavior of metabolic and other biological systems and the perfect tool to investigate change in metabolite concentrations and reaction activity. The metabolic network is thereby described simply by the set of biochemical reactions in the system and the dynamics of the individual reactions are investigated independently.

There are two major approaches to kinetic modeling, the classical deterministic modeling based on mass action kinetics and stochastic simulation considering stochastic effects in a



$$
\begin{array}{c}
 \\
A \\
B \\
C \\
D \\
E \\
F
\end{array}
\begin{array}{c}
\begin{array}{ccccccccc} R1 & R2 & R3 & R4 & R5 & R6 & R7 & R8 & R9 \end{array} \\
\left(\begin{array}{ccccccccc}
1 & -1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -1
\end{array}\right)
\end{array}
$$

Figure 3.2: Example network with corresponding stoichiometric matrix, including all inner metabolites and reactions as well as transport reactions. Reactions are represented as labeled hyperedges in the network graph and as columns in the stoichiometric matrix. Metabolites are circles in the graph and rows of the matrix.

system. Both ideas are closely related. Deterministic kinetic modeling provides general and exact numerical solutions through mathematical analysis of a system of ordinary differential equations representing the set of metabolic reactions. However, it is not applicable for more complex cases. Furthermore, if its assumption of continuous molecular concentrations does not hold as in small-scale processes with only limited numbers of molecules which are quiet common in real-world biological systems, deterministic modeling also cannot make correct predictions. In both situations, stochastic simulation is the preferable choice, because it can handle larger systems while also accounting for the stochasticity in systems with low molecule concentrations.

Deterministic Modeling

Deterministic kinetic modeling is based on the law of mass action which states that the rate v of a reaction is proportional to the product of the concentrations of the molecular species

involved in the reaction. For the simple reaction

$$A + B \rightarrow C, \tag{3.6}$$

the doubling of the concentration of A or B would double the rate of the reaction because the number of collisions between A and B that lead to the production of C would double. For the general reaction formula

$$nA + mB \rightarrow C, \tag{3.7}$$

applying the law of mass action gives us the following rate equation

$$v = \frac{d[C]}{dt} = k[A]^n[B]^m \tag{3.8}$$

as an ordinary differential equation. The rate of the reaction as well as the change of the molecule concentrations can be determined by solving the differential equation. For simple cases this can be done with mathematical analysis getting an analytic solution. However, in more complex situations only numerical solutions or simulations are possible. In the same fashion as described above for the simple reaction, many other types of reactions and processes in metabolism can be represented using the ordinary differential equations approach of deterministic kinetic modeling. For a reversible reaction

$$A \Leftrightarrow B, \tag{3.9}$$

the following differential equation describes its dynamics:

$$v = \frac{d[B]}{dt} = k_A[A] - k_B[B], \tag{3.10}$$

where k_A and k_B are the rate constants for the transition from A to B and back from B to A, respectively. Another example is the representation of Michaelis-Menten kinetics describing an irreversible enzymatic reaction with the assumption that the change of concentration in ES is neglectable in comparison with the product formation.

$$E + S \Leftrightarrow ES \rightarrow E + P \tag{3.11}$$

At this point only the differential equations for the ES-complex in quasi-steady state and the rate of product formation which is equal to the reaction rate is shown.

$$\frac{d[ES]}{dt} = k_1[E][S] - (k_{-1} + k_2)[ES] = 0 \tag{3.12}$$

$$v = \frac{d[P]}{dt} = k_2[ES] \tag{3.13}$$

Further applications for sequential, parallel and complex reactions as well as enzyme interactions, such as several types of enzyme inhibition, are described in [100].

Stochastic Simulation

One major drawback of deterministic modeling is the assumption of a continuous and deterministic dynamics of the chemical reactions and the metabolic system in the whole. While this approach in practice is useful for many applications the underlying assumption does not coincide with the physical reality of these systems. First of all, the concentrations of molecules exist and change in discrete not continuous amounts. Further, the dynamics of a population of molecules is not deterministic, certainly not in the sense that we are able to predict the exact state in which the system will be at the next point in time. Ignoring the stochasticity of the system can in some cases lead to significant deviations between prediction and actual behavior. Stochastic Modeling incorporates this stochasticity by implementing the observation that changes in the molecule population happen through collisions of a specific combination of molecules reacting together to form other molecules. Thus, the fundamental task in stochastic simulation is to find the probabilities for these collisions (chemical reactions) to occur within a certain time interval, given the present state of the system. In the following the exact stochastic simulation algorithm by Gillespie (direct method) will be outlined. Its detailed deduction, a comparison with the "master equation" approach and results on several chemical systems, can be found in [67].

Gillespie starts with the assumption of a well-mixed chemical system S in which non-reactive collisions occur more frequently then reactive collisions. The state of such a system is described by the discrete concentrations of all molecules.

$$S_t = ([A], [B], [C], [D], ...) \tag{3.14}$$

When a chemical reaction (e.g. $A + B \rightarrow C$) occurs after a certain time interval dt, the state at time t (S_t) undergoes a discrete change to state S_{t+dt}, where only the concentrations of the evolved molecules change.

$$S_{t+dt} = ([A] - 1, [B] - 1, [C] + 1, [D], ...) \tag{3.15}$$

The computation of the probability (a_μ)of a chemical reaction (μ) requires only some physical properties of the involved molecules, their concentrations and system properties, such as volume and temperature, using equations similar to 3.1 and 3.2. For pairs of states that can be transformed into each other through application of a chemical reaction, the transition probabilities are defined as follows

$$P(S_{t+dt}|S_t) = a_\mu dt. \tag{3.16}$$

Since these probabilities of the transitions from one system state to the next do not depend on any previous states S_{t-dt} it can be regarded as a Markov process. The probability of the system to be in a state S_t and its change over time can be determined from the differential equation

$$\frac{dP(S_t)}{dt} = \sum_s P(S_t|S_s)P(S_s). \tag{3.17}$$

The entire set of differential equations for all states is called the "master equation" of this system, from which the entire dynamics of the system can be calculated. However, the computation is infeasible for most complex systems. The idea of stochastic simulation is therefore to follow only one possible sequence of transitions rather than all. Starting from the state with the initial molecule population, repeatedly one chemical reaction (μ) is chosen and applied after a time interval (τ). Both μ and τ are random numbers chosen from the distribution of the reaction probabilities. Gillespie formulates the computation of μ and τ from two pseudo-random numbers r_1 and r_2 in the following way:

$$\tau = (1/a_0)\ln(1/r_1) \tag{3.18}$$

$$\sum_{v=1}^{\mu-1} a_v < r_2 a_0 \leq \sum_{v=1}^{\mu} a_v, \tag{3.19}$$

where a_0 is the sum of the probabilities of all reactions in the system. With each chemical reaction that is applied, the molecule concentrations and with that the reaction probabilities have to be updated and the time is set to $t = t + \tau$. The algorithm of Gillespie can then be described as a three step process.

1. Calculate the reaction probabilities a_μ given the system state S.
2. Generate random numbers r_1 and r_2. Compute μ and τ as in 3.19 and 3.18.
3. Update time t by τ and the molecule concentrations in S according to reaction μ.

Go back to Step 1.

The result of Gillespie's algorithm is one transition sequence which is chosen by exactly the probability that is given by the "master equation". This allows to make predictions on the dynamics of the system by averaging over several runs. Advances in computer technology and updates in Gillespie's algorithm [65] have made it feasible to produce large numbers of runs for complex systems.

3.2.3 Stoichiometric Approach

The modeling techniques introduced in the previous paragraphs focused on the dynamics of the reactions within a metabolic system. This section will describe approaches that regard the system as one connected entity allowing to investigate emergent properties beyond the reaction level. Stoichiometric modeling does not regard kinetic information, such as the rate constant, instead it incorporates topological information about the structure of the metabolic network and the reaction stoichiometry of a system in form of stoichiometric coefficients i.e.

the proportion of a metabolite in a reaction. The metabolic network which will be analyzed by the stoichiometric approaches comprises all internal metabolites and reactions as well as exchange reactions, such as influx of "food" metabolites and outflux of metabolites like biomass formation. Considering the law of mass conservation, the metabolites are here implemented to be in mass balance, i.e. metabolites can not appear out of nothing or disappear, they have to be produced or used in one of the chemical reactions. Consequently, for each metabolite a balance equation describing the time-evolution of its concentration can be formulated as

$$\frac{dm_i}{dt} = \sum_{j}^{|r|} s_{ij} v_j, \tag{3.20}$$

where s_{ij} is the stoichiometric coefficient of metabolite i in reaction j and v_j is the flux of reaction j. Furthermore, the investigated system is assumed to be in steady-state which means that all internal metabolites may not accumulate and metabolites that were input or created in a reaction either have to serve as substrate in another reaction or output. The differential equations from 3.20 can then be written as linear equations in the following form

$$\sum_{j}^{|r|} s_{ij} v_j = 0. \tag{3.21}$$

Some of the reactions in the metabolic system can be regarded as irreversible, i.e. flow in one specific direction, this can be modeled specifying inequalities as constraints.

$$v_j \geq 0 \tag{3.22}$$

Given the stoichiometric and topological information, one wants to compute functional subunits of the metabolic system, called metabolic pathways. Metabolic pathways are a sequence of reactions, which are either internal loops or paths connecting source metabolites ("food") with sink metabolites (biomass), which also have to fulfill the steady-state assumption. There are two types of stoichiometric analysis, one which focuses on one or few specific solutions from the distribution of all possible pathways obeying some additional constraints and a type of network-based analysis regarding the full distribution. To the former type belong for example flux balance analysis and metabolic flux analysis. The latter comprises mainly the elementary mode and the extreme pathway analysis. The next paragraph describes the steps from metabolic network to the stoichiometric model which is part of both methodologies. The succeeding two paragraphs underline their differences in computation and application.

Nullspace of S

Besides representing the structure of the network, the stoichiometric matrix S can be used to determine the dynamics of the metabolite concentrations and the steady-state component of the network. For the former, simply instead of using balance equations (Equation 3.20) the

$$
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}
\begin{array}{c}
\begin{array}{ccccccccc} R1 & R2 & R3 & R4 & R5 & R6 & R7 & R8 & R9 \end{array} \\
\left(\begin{array}{ccccccccc}
1 & -1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -1
\end{array} \right)
\end{array}
$$

$$\Downarrow$$

$$
\begin{array}{c} \\ R1 \\ R2 \\ R3 \\ R4 \\ R5 \\ R6 \\ R7 \\ R8 \\ R9 \end{array}
\begin{array}{c}
\begin{array}{ccccccccccccc} & & & & & & & & & & & & \end{array} \\
\left(\begin{array}{ccccccccccccc}
2 & -1 & -1 & 1 & 0 & 1 & & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & -1 & -1 & 0 & 0 & 0 & & -1 & -1 & 0 & -1 & 0 & -1 & -1 \\
1 & -1 & -1 & 0 & 0 & 0 & & -1 & -1 & 0 & -1 & 0 & -1 & -1 \\
0 & 1 & 0 & 0 & -1 & 1 & & 1 & 2 & 1 & 2 & 0 & 1 & 0 \\
1 & 0 & -1 & 0 & -1 & 1 & \ldots & 0 & 1 & 1 & 1 & 0 & 0 & -1 \\
0 & 0 & 1 & 1 & 1 & 0 & & 1 & 0 & 0 & 0 & 1 & 1 & 2 \\
0 & 0 & 1 & 1 & 1 & 0 & & 1 & 0 & 0 & 0 & 1 & 1 & 2 \\
0 & 0 & 1 & 1 & 1 & 0 & & 1 & 0 & 0 & 0 & 1 & 1 & 2 \\
1 & 0 & 0 & 1 & 0 & 1 & & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array} \right)
\end{array}
$$

Figure 3.3: The stoichiometric matrix (top) from example in Figure 3.2 and the respective kernel matrix (bottom). Rows in the kernel matrix correspond to the reactions of the network. Columns represent steady-state fluxes, not regarding inequality constraints yet.

stoichiometric matrix is multiplied with flux vector $\vec{v}$ to derive the vector of time derivatives for the metabolite concentration as shown in Equation 3.5.

Similarly, for determining the steady-state component which is the focus of this section, i.e. the space of possible fluxes trough the network satisfying the steady-state assumption, the system of linear equations from 3.21 is now written in matrix form as

$$S\vec{v} = 0, \tag{3.23}$$

where the solution for $\vec{v}$ is usually not a unique flux but a solution space spanned by linear independent flux vectors. This solution or null space can also be represented in matrix form, the kernel matrix K, where each row represents a reaction and its participation in a steady-state pathway. Accordingly, columns represent these pathways and the involved reactions. There exist efficient implementations of algorithms that can compute the kernel matrix, such as Gauss-Jordan elimination. Additionally to the use in the stoichiometric analysis that will be described in the following paragraphs, some information about the metabolic function can be extracted directly from the kernel matrix. For instance, by inspecting the kernel matrix for linear dependent rows (reactions) relationships among reactions, such as enzyme subsets,

can be discovered. Two reactions (k, l) belong to the same enzyme subset if their row vectors $(\vec{r^k}, \vec{r^l})$ in the kernel matrix are equal up to a scalar, i.e. if it holds that

$$\forall i : r_i^k = \alpha \times r_i^l \tag{3.24}$$

where a positive α indicates parallel reactions and a negative value opposing reactions. Those opposing reactions could be excluded from further inspection since they will not be involved in a valid pathway. Parallel reactions can be lumped together and regarded as one component, which simplifies the further pathway analysis. Several other reductions exist, such as discarding strictly detailed balanced reactions and conservation relations. In most cases the solution space, which can be regarded as a convex polyhedral cone in the flux space (see Figure 3.4b), is further cut down by introducing thermodynamic constraints, specifying some reactions as irreversible (see Equation 3.22) or limiting the flux of a reaction to a maximum value (k_M). In flux balance analysis, the polyhedral cone of solution space also has to be closed through capacity limits on the fluxes of the reactions,

$$v_i \leq v_{max} \tag{3.25}$$

so that the optimization problem and consequently the polyhedral cone become bounded (see Figure 3.4c).

Flux Balance Analysis

The aim of flux balance analysis is to find an optimal solution within the solution space described above. The assumption is that the metabolic system functions optimally and evolved to maximize a certain function, such as growth rate or biomass formation. Linear optimization is applied to maximize or minimize a certain objective function, satisfying the set of linear equations (Equation 3.23) and inequality constraints (see 3.22). In most cases biomass production is chosen as the objective function Z, which would then comprise of a certain number of exchange reactions v^e and specific proportions of them (c^e), depending on the respective metabolic system.

$$max: \ Z = \sum c_i^e v_i^e \tag{3.26}$$

The solutions of linear optimizations lie at the edges of the polyhedral cone. The solution for a specific objective function is a single intersection point between the linear objective function and the edges of the cone. Since the solution space in flux balance is constrained in a way that it represents a bounded convex polyhedral cone, the solution to the linear optimization is either a line, when Z coincides with an edge of the cone, or optimally a single point in flux space if Z cuts the cone in an extreme point (see Figure 3.5). A solution represents one specific flux distribution that leads to the optimal value for the objective function.

Flux balance analysis has been shown to be consistent with experimental studies in predicting the correct phenotype (flux distribution) that will be expressed in the real metabolic system

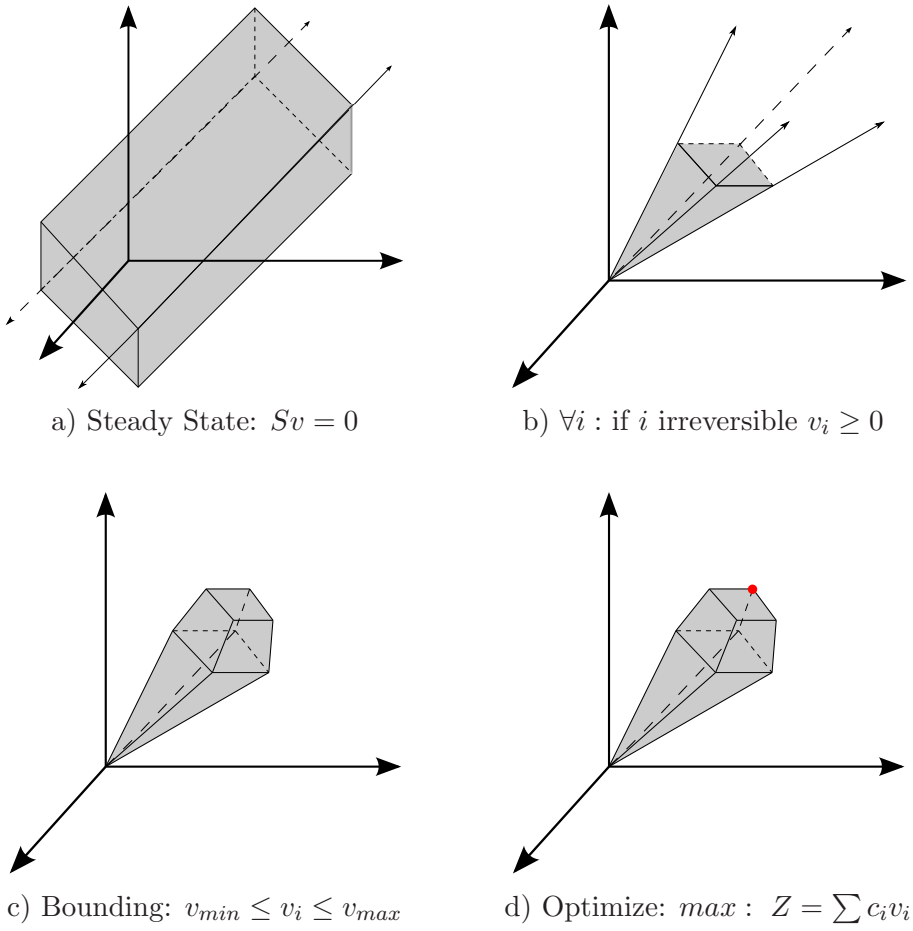


Figure 3.4: Four steps in the analysis of the nullspace of the stoichiometric matrix in terms of the solution space (gray shape). a) Fulfilling the steady-state assumption gives the nullspace of the stoichiometric matrix S as solution space. b) Satisfying the inequality constraints of the reactions shapes a polyhedral cone as solution space, with the extreme pathways as boundaries. c) Additional maximum flow constraints bound the cone. d) Linear optimization of an objective function gives one point in the solution space as optimal solution (red point).

[123, 92]. Further it can be used for sensitivity analysis, where either certain parameters in the model are changed, enzymes are knocked out to simulate gene deletion or capacities of influxes are changed (change in growth medium). The differences between the optimal solutions in the wildtype model and the changed models give insights about the sensitivity or robustness of the modeled system concerning the respective perturbation.

Metabolic Pathway Analysis

Instead of finding one optimal solution, in metabolic pathway analysis the entire space of possible pathways through the network is studied. It is therefore able to shed more light on the systematic properties, such as flexibility or modularity.

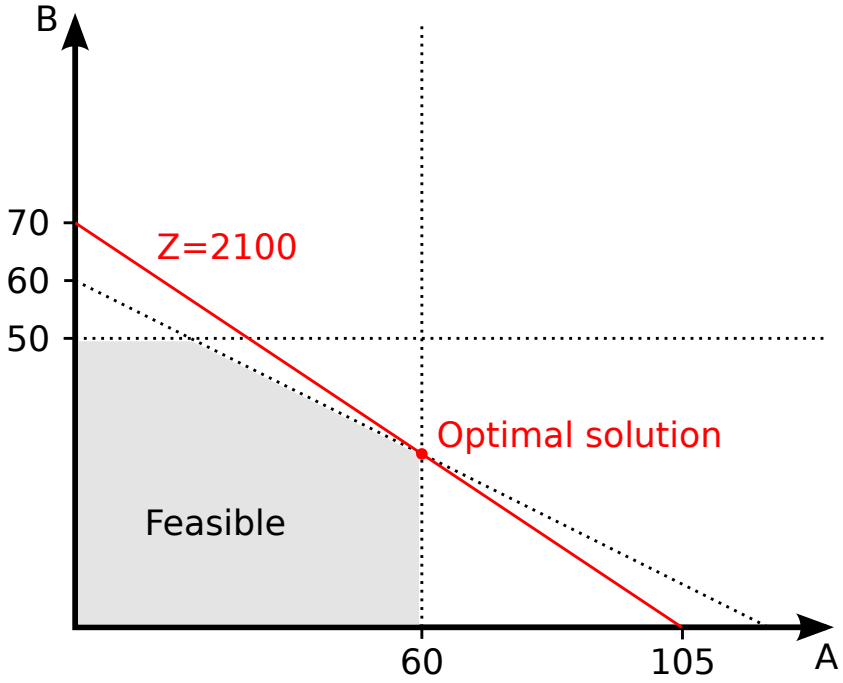


Figure 3.5: Simple example of linear optimization in a two-dimensional solution space (gray shape). The inequality constraints $0 \geq A \geq 60$ and $0 \geq B \geq 50$ as well as the flow constraint $A + 2B \leq 120$ are drawn as dotted lines. The objective function $Z = 20A + 30B$ is the red line. The optimal solution is the red point in the intersection of the objective function and the constraints. Adapted from [127]

The first step in metabolic pathway analysis is to determine a set of descriptive pathways from which the full set of possible pathways can be described and gained, so to say basis vectors spanning the space of admissible fluxes through the network, i.e. every allowed flux can be generated by non-negative convex combination of the basis flux vectors. There are two very related concepts that implement this idea, the approach of elementary modes and that of the extreme pathways. Both approaches are solved using convex analysis with a convex polyhedral cone P as the solution to the system of linear equations of the steady state assumption and the inequality constraints representing thermodynamical irreversibility of reactions.

$$P = \{\vec{v} \in R^n : S\vec{v} = 0 \ \wedge \ v_i \geq 0 \ if \ i \ is \ irreversible\} \tag{3.27}$$

The difference between elementary modes and extreme pathways lies in the fact that in the concept of extreme pathways all reactions underlie an irreversibility constraint, whereas elementary modes allow for the inclusion of reversible reactions. In terms of convex analysis this means that extreme pathways represent the outer edges of the cone which are linear independent of each other, while elementary modes do not necessarily fulfill this requirement but have to be minimal in the sense that there may not be any other flux vector in the cone which is a proper subset concerning the involved reactions. Biologically, this means that the set of elementary modes comprise all minimal routes through the metabolic network that are

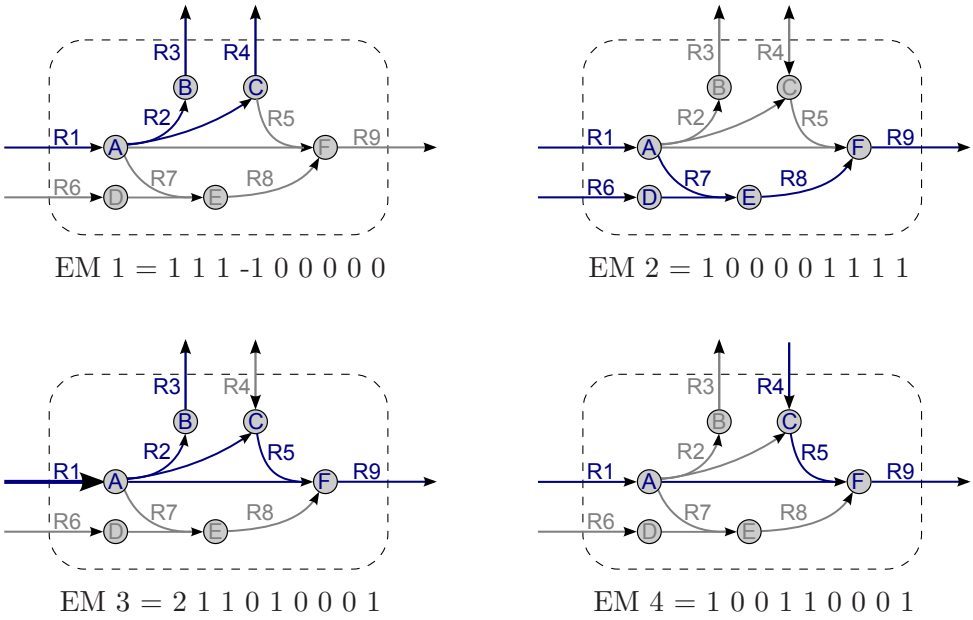


Figure 3.6: The full set of elementary modes of the example network in Figure 3.2. The blue reactions are part of the respective elementary mode. A thick blue line indicates a flow of two, while a thin line represents a flow of one. Below each graph are given the flux vectors of the modes.

stoichiometrically and thermodynamically feasible. The extreme pathways are just a subset of the elementary modes, which can pose problems in the metabolic analysis. However, if reversible reactions are modeled as two irreversible reactions of opposite direction the set of extreme pathways coincides with that of the elementary modes.

Minimal Knockout Sets

Elementary modes are a concept for functional subunits of the network, in this paragraph now the focus will lie on a related but opposing concept. Minimal knockout sets are so to say dysfunctional subunits of the network. They are sets of reactions which block a certain target function if they are knocked out simultaneously, see Figure 3.7. The knockout of the reactions of one knockout set has to block the target function in all situations, i.e. in all elementary modes where the target reaction is active, see the sets on the top in Figure 3.8. Knockout sets could , thus, also be described as cutting or hitting sets of elementary modes. This means that for all elementary modes in which the target reaction is active (non-zero), there has to be one reaction in the knockout set that is also active in this elementary mode [99]. Further, the knockout sets have to be minimal in the sense that there is no subset of the knockout set that is itself a knockout set for the specified target function, see some not minimal knockout sets on the bottom of Figure 3.8.

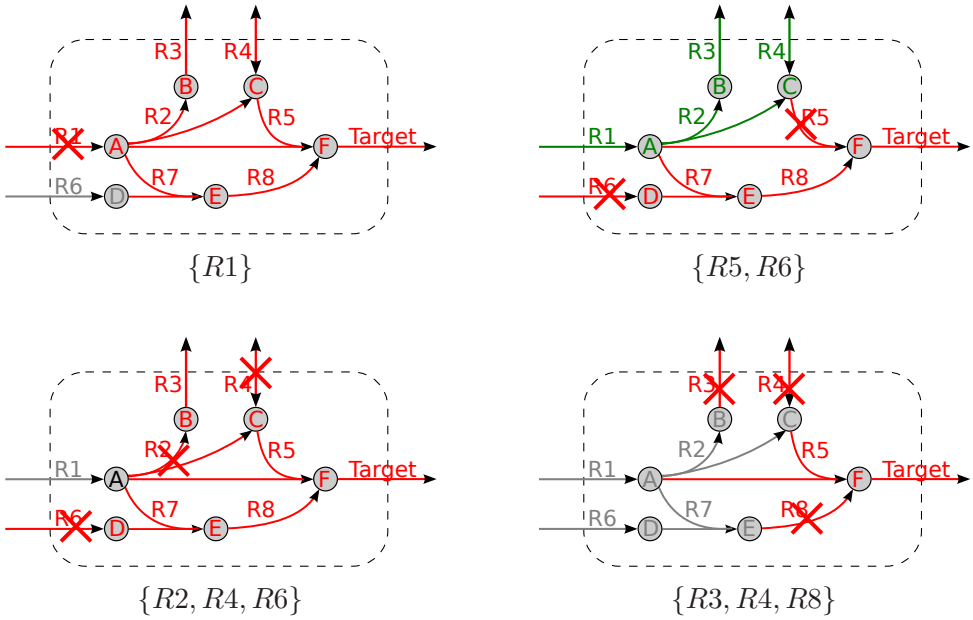


Figure 3.7: Four of the ten minimal knockout sets of the example network in Figure 3.2. The target function here is the transport reaction of metabolite F. A crossed out reaction is knocked out and part of the respective minimal knockout set. Red lines represent reactions that are blocked through the knockouts. In all knockout sets the target reaction is knocked out. Green lines indicate reactions through which there is still flow. Below the graphs the reactions of the minimal knockout sets are given.

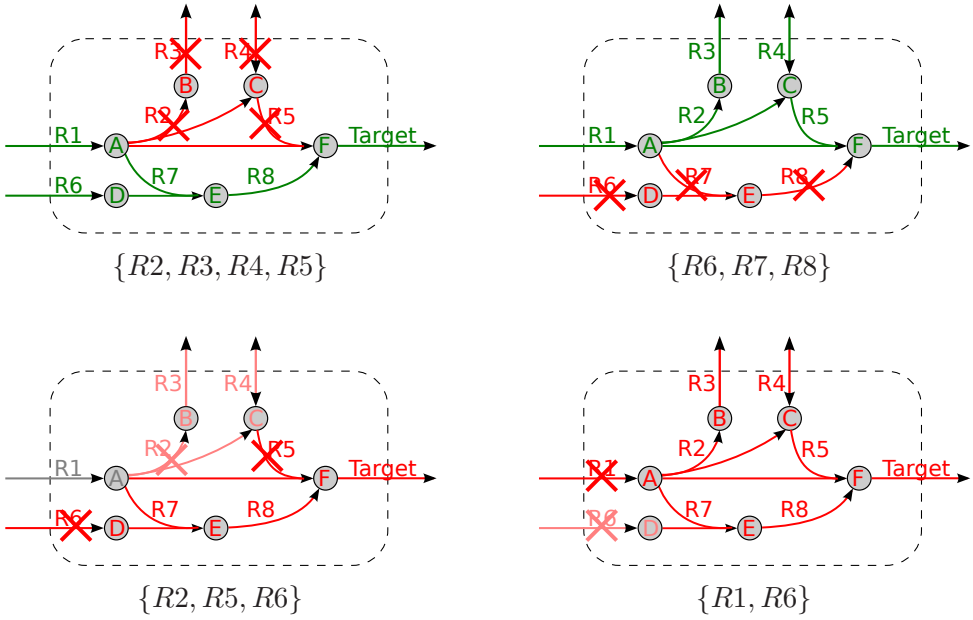


Figure 3.8: Four sets of knocked-out reactions that are not minimal knockout sets. The two sets on the top are not MKS because the target function is not blocked, even though a large number of reactions is knocked out. The two sets on the bottom block the target function but are not minimal, not all reactions (light red cross) of these sets are needed for the blocking.

3.2.4 Chemical Organizations

Another way to view chemical reaction systems is through so called chemical organizations, sets of molecules and reactions of the chemical reaction system that are closed and self-maintaining. This means on the one hand that the application of the included reactions to the molecules of the set does not produce new molecules outside this set (closeness). On the other hand, all molecules of the organization has to be produced from molecules and through reactions within the organization (self-maintenance). Thus, chemical organizations are steady-state subsystems of a chemical reaction system, but of the form that they do not evolve new behavior. In fact, all steady states of a reaction system are included in the full set of chemical organizations [38]. Therefore, they can tell us something about the dynamics of the system. Since the set of chemical organizations forms a partially ordered set with the inclusion operation [22], it can be represented through a Hasse-diagram, which will be called organization hierarchy, throughout this work. See Figure 3.9 for an example reaction system with the corresponding subsets and the organization hierarchy.

3.2.5 Computational Representations

A model of metabolism comprises several objects on different levels, in the following paragraphs possible computational representations of these parts are introduced. The focus will lie on representation in graph form which is an easy to grasp concept and the most intuitive way to view the objects of this study, such as molecules, reactions and networks. Further, graph theory provides many useful tools for their analysis and manipulation.
A graph G is a tuple of a set of vertices V and a set of edges E connecting pairs of vertices.

$$G = (V, E); \quad V = v_1, \ldots, v_n; \quad E \subseteq V \times V \tag{3.28}$$

A subgraph S of a graph G is the subset of the vertices and edges of G.

$$S = (V_S, E_S); \quad V_S \subseteq V; \quad E_S \subseteq E \tag{3.29}$$

Molecules

The most common computational representation of a molecule is as molecular graph, in which the vertices represent the atoms and the edges correspond to the bonds of the molecule. The molecular graph representation provides storage of important chemical information such as atom and bond type as well as charges and aromaticity through vertex and edge labels. Furthermore, it allows to compare molecules and search for molecules which contain or are contained in another molecule, using well known graph theoretic approaches for graph and subgraph isomorphism. A graph G_1 is a subgraph of another graph G_2 if there is a mapping of the vertices and edges of G_1 to the vertices and edges of G_2.

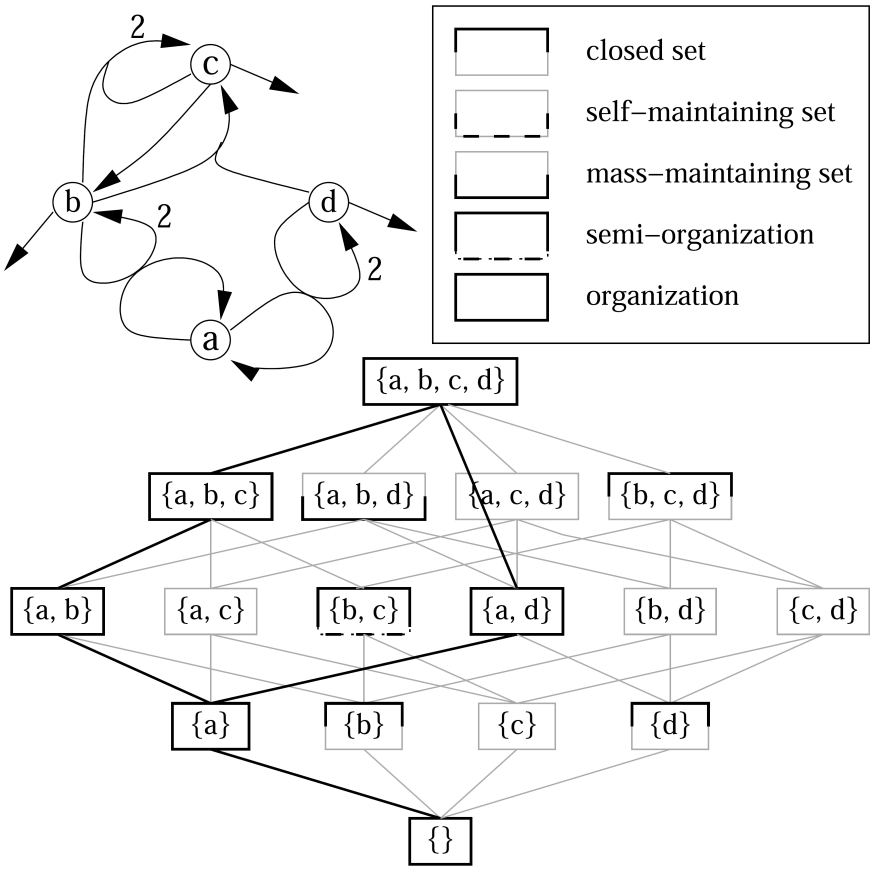


Figure 3.9: Example reaction system with its chemical organization hierarchy. Top left: The reaction system, with molecules (circles: a,b,c,d), reactions (arrows) and stoichiometry (caption of arrows). Top right: The different categories of subsets. Bottom: Hierarchy of all subsets, including the organization hierarchy (solid lines, solid boxes). The top organization of the hierarchy is the full reaction network, the bottom organization is the empty set. From [37]

Two graphs are isomorphic if both graphs are a subgraph of the other, i.e. if there exists such a mapping for both directions. There exist several efficient algorithms for subgraph and graph isomorphism checking [162]. However, performing this check for a huge number of graph pairs it becomes a computationally costly factor. It is sometimes important to know whether a certain graph has already appeared in the course of an algorithm, which means that every graph has to be checked against all previous graphs. For molecular graphs representing valid chemical molecules there exist more efficient tools than graph isomorphism checks. It is possible to represent molecules in a canonical line or string format such as `SMILES` [175]. In chemical textbooks the structural formula is the commonly use string representation, a short and comprehensible format, however, it is not unique which makes it uninteresting for the use in search strategies.

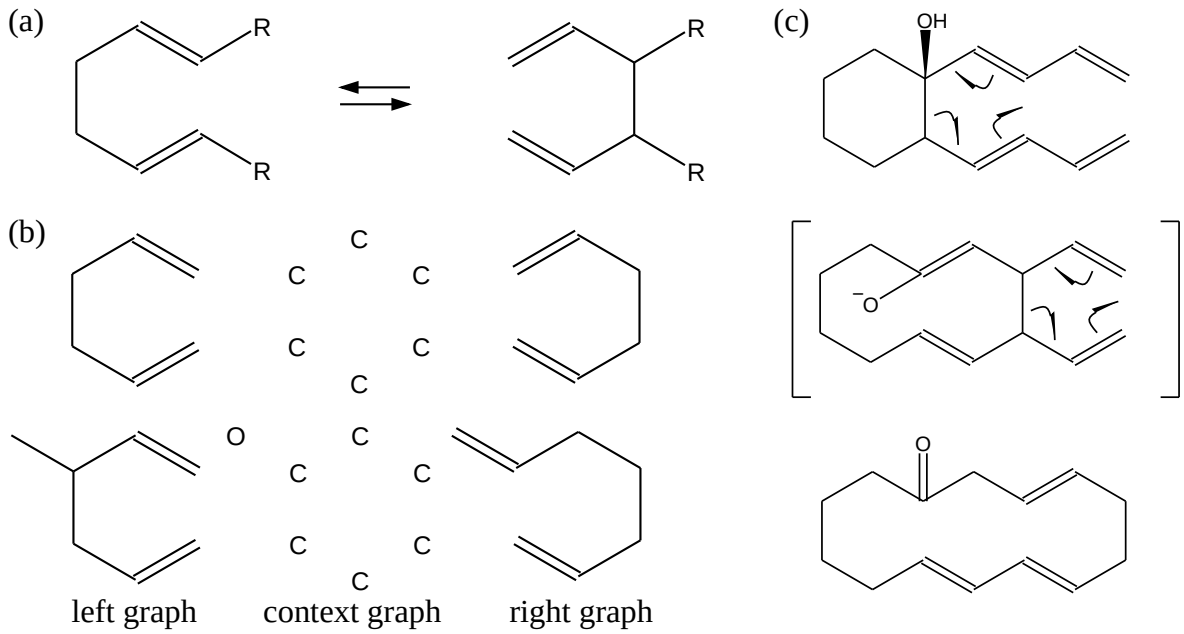


Figure 3.10: Reactions as graph grammars: Chemical transformations very naturally translates into graph transformation rules. As an example the Cope rearrangement, a concerted pericyclic [3,3]-sigmatrope rearrangement, is shown (1st row on the left). A graph rewrite rule consists of 3 graphs: (i) the left graph which is composed of all the atoms and bonds which vanish during the reaction (ii) the context graph comprised of all atoms and bonds which do not change (iii) the right graph consists of all the atoms and bonds which are formed during the reaction. The conjunction of left and context graph forms the pre-condition for the applicability of the rewrite rule. the rules for the Cope and oxy Cope rearrangement are shown (2nd and 3rd rows on the left). The context sensitivity of graph rewrite rules is illustrated by Wender's methatese, a tandem reaction (oxy Cope rearrangement followed by Cope rearrangement). While the Cope rule applies to both steps, the oxy Cope rule is only applicable to the first step of the tandem reaction.

Chemical Reactions

Once molecules are represented as (labeled) graphs it becomes natural to view reactions as graph transformations. Again, this matches the intuition. After all, a chemical reaction mechanisms is taught and understood as a sequence of events that break and/or form chemical bonds among the atoms (vertices) of small assembly of molecules (graphs). From a computer scientist's point of view, chemical reactions are thus just graph-rewriting rules, see Figure 3.10 for an example.

A graph rewriting rule is specified as a triple consisting of *left graph*, *context*, and *right graph*. Left and right graphs consist of all atoms and bonds that vanish or are newly formed in the transformation, respectively. The context specifies the necessary prerequisites for the applicability of the rule beyond the atoms that are actually affected by the reaction itself. Note that in proper chemical reactions all vertices (atoms) involved in the reaction are part

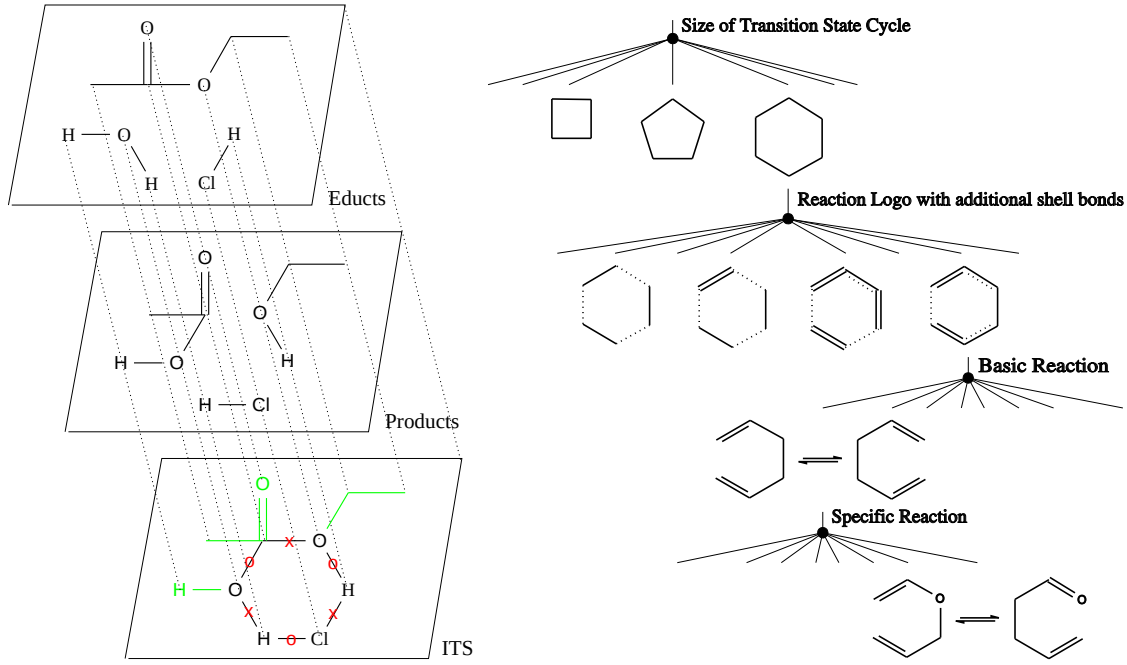


Figure 3.11: Imaginary Transition State (ITS) and their hierarchical organization: Superimposition of educt and product molecular graphs and subsequent removal of all atoms and bonds which do not directly participate in the chemical reaction (marked in green) yields a cyclic ITS for a chemical reaction (e.g. acidic hydrolysis of ethylacetate). Bonds which are broken/formed during the reaction are marked with a red x/o. The ITS can be organized in a hierarchical structure where each tree level adds additional information to the base cycle of the ITS such as bonds or atom labels. Specific instances of reactions are found as leafs of the tree.

of the context of the rewrite rule because they neither disappear nor are newly created.

Another representation describing chemical reactions is the imaginary transition state structure, ITS. The ITS of the reaction is intimately connected to the left and right graphs. It is obtained from the superposition of educt and product molecular graphs and subsequent removal of all atoms and bonds which do not directly participate in the reaction (see Figure 3.11). The form of monocyclic ITSs that are used in this work, account for over 90% of all known chemical reactions [73].

Chemical Reaction Networks

Mathematically speaking, a chemical reaction network consists of a set of nodes (Molecules) and a set of subsets of these nodes (Reactions). This corresponds perfectly to the notion of a hypergraph $H = (N, H)$, where H are the set of hyperedges. Hyperedges, just as chemical reactions, are non-empty subsets of the set of nodes N. Hypergraphs are an intuitive way to view chemical reaction networks, see Figure 3.12a) for the above used example network as

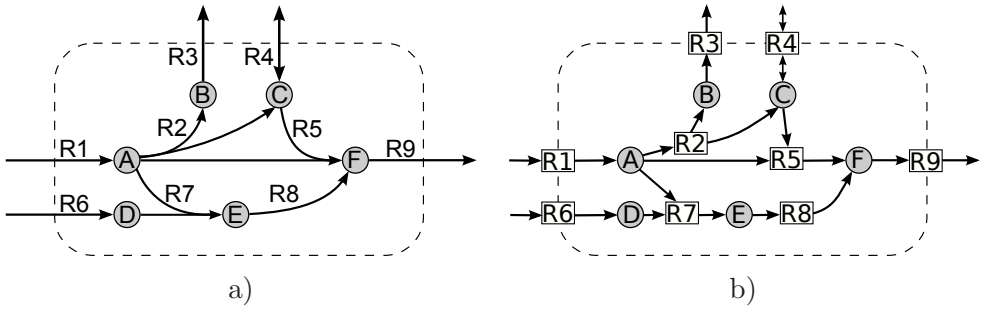


Figure 3.12: Example network in a) hypergraph representation and b) as bipartite graph. In both representations gray circles are metabolites. In the hypergraph, reactions are represented as hyperedges (arrows). In the bipartite graph, reactions are the white boxes and participation is indicated through connection with arrows.

hypergraph. Often in biology textbooks metabolic networks or chemical routes will be shown as hypergraphs.

Another appropriate graph representation of chemical reaction networks is the bipartite graph. As for the hypergraph, no connectivity information is lost as would be for simpler reaction or substrate graphs. A bipartite graph $B = (N1, N2, E)$ is a tuple of two independent sets of nodes $N1, N2$ and a set of edges E, where an edge e always contains one node $n1$ of the nodeset $N1$ and one node $n2$ of the other nodeset $N2$. In case of a chemical reaction network, the two independent node sets are the sets of metabolites and reactions, respectively. A connection between a metabolite node and a reaction node in the bipartite graph indicates the participation of the metabolite in the respective chemical reaction.

3.3 Artificial Chemistry

Artificial chemistries are models of real chemical systems or systems that behave like one, that can deliver insights about the evolution of complex systems and possibly some of their properties. An artificial chemistry model is defined through molecules, reactions specifying the interactions between molecules and a dynamic on the reactions, where the notion of molecule and reactions can be understood as metaphors for objects and the rules for changing objects, respectively. There exist several approaches for such models, which differ in the level of abstraction and the definitions of the set of molecules and reactions. An interesting property of an artificial chemistry is to be constructive in the sense that new molecules outside the set of initial molecules are created, thus one can not enumerate and define the set of possible molecules explicitly and possibly not even implicitly, they just emerge through the interaction of the existing molecules.

Based on the structure of molecules and definition of the molecule interactions, artificial chemistry approaches can be ascribed to specific classes, such as rewrite, arithmetic or turing-

machine like systems. Many models of artificial chemistries have been developed in recent years. Most famously, Walter Fontana's `AlChemy` [49, 50], representing molecules as λ-calculus expressions and defining reactions by the application of one λ-term to its reaction partner. The result is a new λ-term. Related models are based on a wide variety of different computational paradigms from strings and matrices to Turing machines and graphs [3, 6, 36, 161, 158, 112, 141], for a broad review it is referred to [39, 153].

3.3.1 Molecules

The molecules in artificial chemistry models are sometimes sets of strings, bit-strings, graphs, lambda-expressions, numbers or other abstract symbols and objects. The set of molecules M can be defined either explicitly,

$$M = \{m_1, \ldots, m_n\}; \quad n \in N, \tag{3.30}$$

where all molecules can be enumerated, or it is defined implicitly through a grammar or another form of construction definition

$$M = \{m_i : m_i \in L(G)\}, \tag{3.31}$$

where $L(G)$ is the set of all words that can be derived from grammar G. Sometimes, it is sufficient to define an initial set of molecules

$$M_{init} = m_1, m_2 \tag{3.32}$$

and additional molecules emerge from the interactions between the molecules, i.e. the definition of the reactions. In this work the graph representation of real chemical molecules is used an thus will be the focus.

Historically, the description of molecular structures was one of the roots of graph theory [21, 154]. Graphs with vertex labels denoting atom types and edges indicating bond orders are ubiquitous in every book and journal article on Organic Chemistry and in practice convey enough information to provide chemists with a good idea of the molecules behavior in particular chemical reactions.

By construction, the graph representation abstracts spatial information to mere adjacency. Thereby avoiding the most time-consuming computation step: embedding the atoms in 3D by means of finding the minima on a potential energy surface [72]. On the other hand, the restriction to graphs implies that several features of real molecules cannot even be defined within the model: (1) There is no distinction between different conformers and, in particular, between *cis* and *trans* isomers at a $\mathsf{C} = \mathsf{C}$ double bond. (2) there is no notion of asymmetric atoms and chirality.

3.3.2 Rules

The rules of an artificial chemistry define interactions between molecules, the chemical reactions. A rule r is thus a function from M to M ($r : M \to M$) and can be written similarly to a chemical reaction as

$$m_1 + m_2 \xrightarrow{r} m_3 + m_4; \quad m_i \in M \tag{3.33}$$

If all molecules of the left side of a rule are existent, then a rule is applied by replacing those molecules with the molecules of the right side.

If molecules are represented as graphs the set of rules are modeled as graph-rewrite rules and can be understood as a graph grammar. A graph rewrite system [120] interprets the graph rewrite rule and performs the graph rewriting step if the graphical pre-condition is matched in a host graph. The applicability of rewriting-based approaches to metabolic network data was demonstrated recently in an analysis of KEGG data [47].

3.3.3 Dynamics and Energy

The dynamics of an artificial chemistry determine how its rules are applied to the set of molecules. The two major approaches differ in the way they treat molecules, either as single entities or as frequencies of molecule types. The former view usually leads to stochastic-simulation-like dynamics, where molecules are randomly drawn and, if possible, rules are applied, creating new molecules. The latter approach is best described by a system of differential equations, based on the rules of the chemistry, that describe the chemical evolution of the molecules frequency or concentration. A discussion of both modeling approaches can be found above in Section 3.2.2.

Reaction energies can introduce important constraints on the dynamics and the development of artificial chemistries, by selecting one or a few preferred reaction pathways from the entire space of possible reaction channels. Therefore, an energy function is indispensable for a realistic model of chemical reaction systems. Despite substantial progress in theoretical chemistry, detailed quantum chemical computations are in many cases still too expensive to be employed in large scale computer simulations. Comprehensive reaction databases used e.g. in synthesis planning, on the other hand, are mostly commercial products which come a substantial access costs. It also remains unclear to what extent the network of the millions of reactions performed and compounds synthesized by organic chemists over the past two centuries [70] provided a view biased by the history of chemical research. Knowledge-based approaches hence appear less attractive for this purpose.

ToyChem model

The `ToyChem` model [13] utilizes a caricature version of quantum chemistry to compute total binding energies directly from the labeled graphs. In particular, the chemical structure

graph is decomposed in an unambiguous way into hybrid orbitals using the VSERP rules [68]. Application of a simplified version of the Extended Hückel Theory (EHT) [80] yields a Schrödinger type secular equation which is parametrized in terms of the *atomic valence state ionization potentials* and the overlap integrals between any two orbitals. The physical properties of a molecule are determined by the eigenvalues of the Hamilton matrix and their associated eigenvectors as well as by the number of valence electrons and the electrons in the various molecular orbitals. For details it is referred to [13].

The `ToyChem` model was used to study the generic graph-theoretic properties [12] of chemical reaction networks under thermodynamic constraints. A straightforward extension of the `ToyChem` model to solvation energy made it possible to study chemical reaction networks in a multiple phases setting [14].

More realistic estimates of reaction rates require the use of state-of-the-art methods from well established quantum mechanical program packages such as `GAUSSIAN` or `Schrödinger Soft`. Unfortunately, many of these sophisticated quantum mechanical methods are very expensive in terms of computer time. Semi-empirical methods like PM3 (implemented for example in `Mopac` and `GAUSSIAN`) are computationally less costly but also provide less reliable results. Another popular choice nowadays is DFT on the B3LYP level of theory, which works well for certain organic molecules, but not across board for the whole organic chemistry subset [181, 19].

3.4 Metabolism

> We are seeing the cells of plants and animals more and more clearly as chemical factories, where the various products are manufactured in separate workshops. The enzymes act as the overseers. Our acquaintance with these most important agents of living things is constantly increasing.
> Eduard Buchner in 1907

Metabolism comprises a set of catalyzed chemical reactions, responsible for the uptake of food molecules and building new structures essential for the survival of the cell, using some form of energy. Thus metabolism can be seen as a chemical reaction system. However, it has some characteristics that make it a specialized form. One important difference to general chemical reaction systems, is the fact that metabolism contains enzymes that catalyze the chemical reactions and give them direction. These enzymes can form subsets or entire metabolic pathways shaping the metabolic network. The way enzymes evolved through series of duplications and specialization, as parts of clusters and pathways that also duplicate and specialize makes metabolism a unique chemical reaction system. This of course also true for the evolutionary history of the entire system, in which it adapted to environmental changes, overcame barriers and was subject to natural selection.

3.4.1 Enzymes

The vast majority of chemical reactions in metabolism are performed and catalyzed either by protein enzymes or ribozymes. Enzymes differ in their reaction mechanisms and substrate specificity, some react only with a very restricted set of metabolites while others can react with entire classes of metabolites. Also the way in which reactions are catalyzed, i.e. how the activation energy of the chemical reaction is lowered, differs among enzymes, from the stabilization of the transitions state to spatial orientation of the substrates. The activity of the enzymes activity can be inhibited or activate through interactions with specific substrates or other enzymes.

3.4.2 Metabolic Pathways

A metabolic pathway is an organized sequence of chemical reactions where the products of one reaction are the substrates of the subsequent reaction. The orchestrated procedure in a pathway, which is achieved for instance through enzyme complexes or spatial proximity of enzymes catalyzing subsequent reactions, ensures a high rate of throughput. The products of a pathway can accumulate and then be used in biomass formation, as input for another metabolic pathway or released out of the cell or compartment. Intermediary products, however, are never accumulated because it is constantly consumed by other reactions, this means that metabolic pathways are in steady-state. The steady-state condition does not imply an energy equilibrium, in fact, metabolic pathways constantly release (catabolic) or use (anabolic) energy. Catabolic and anabolic pathways, thus, are connected with each other and two parallel but opposing pathways are usually regulated reciprocally, i.e. if the catabolic pathway is up-regulated, the according anabolic pathway will be down-regulated. Therefore, never all metabolic pathways will be active at the same time, the cell rather switches between specific patterns or modes of pathways. The union of all metabolic pathways build one overlying metabolic network.

3.4.3 Metabolic Network

The metabolic network of a cell comprises all its metabolic pathways and chemical reactions serving the function to produce energy and biomass molecules from the steadily imported food metabolites. Since the metabolism of a cell is in steady state only certain combinations of pathways and chemical reactions can act simultaneously. The organization of metabolic networks evolved in a way to efficiently reorganize these pathway combinations according to the needs of the cell. The structure of the networks is hierarchical and modular, i.e. small modules group into larger modules which itself build even larger modules. These modules are never fully separated, in most cases some highly connected metabolites (e.g. pyruvate, coA)ensure their connectedness. These so called hub-metabolites account for the majority of the overall connectivity, the remaining metabolites participate only in one or very few

reactions. The benefit of such a scale-free topology for the network are short paths between pairs of metabolites and high robustness against random knockouts [1], which allows the cell to respond quickly to environmental (e.g. depletion of food sources) and internal changes (e.g. mutations, different energy levels).

Chapter 4

Computational Framework

In the previous chapter, the possibilities and potential benefits for modeling chemical reaction networks that underlie metabolism were discussed. In this chapter, a novel multi-scale computational framework for the simulation of the evolution of early metabolism is presented in its details, including the single components as well as the big picture.

The particular Chemical Universe (see Figure 4.1) that underlies the simulation framework is motivated by the way how chemical reactions are explained in introductory Organic Chemistry classes: in terms of structural formula (labeled graphs) and reactions mechanisms (rules for modifying graphs). It further contains a minimal, RNA-World style, genetics and a simple fitness function linked to metabolic efficiency.

4.1 Protocells

The "players" in the Simulation Universe are modeled as complex agents, referred here to as protocells. These protocellular entities are characterized by individual genomes which encode for its catalytic elements. These ribozymes represent the individual's metabolic capability, which is internally realized through an algebraic chemistry. Besides the catalyzed chemical reactions, exchange of metabolites with the environment is performed. There is also selection process on the protocell population. The selected individuals will then multiply and the resulting child protocells incorporated in the population, while their genomes undergo some genetic operations.

4.2 Genome

The interest of this work lies primarily in the earliest stages of metabolic evolution, which arguably took place in the setting of the Early RNA World [64]. In this setting, RNA has the double role of genetic material and serves as catalysts. Both the analysis of naturally

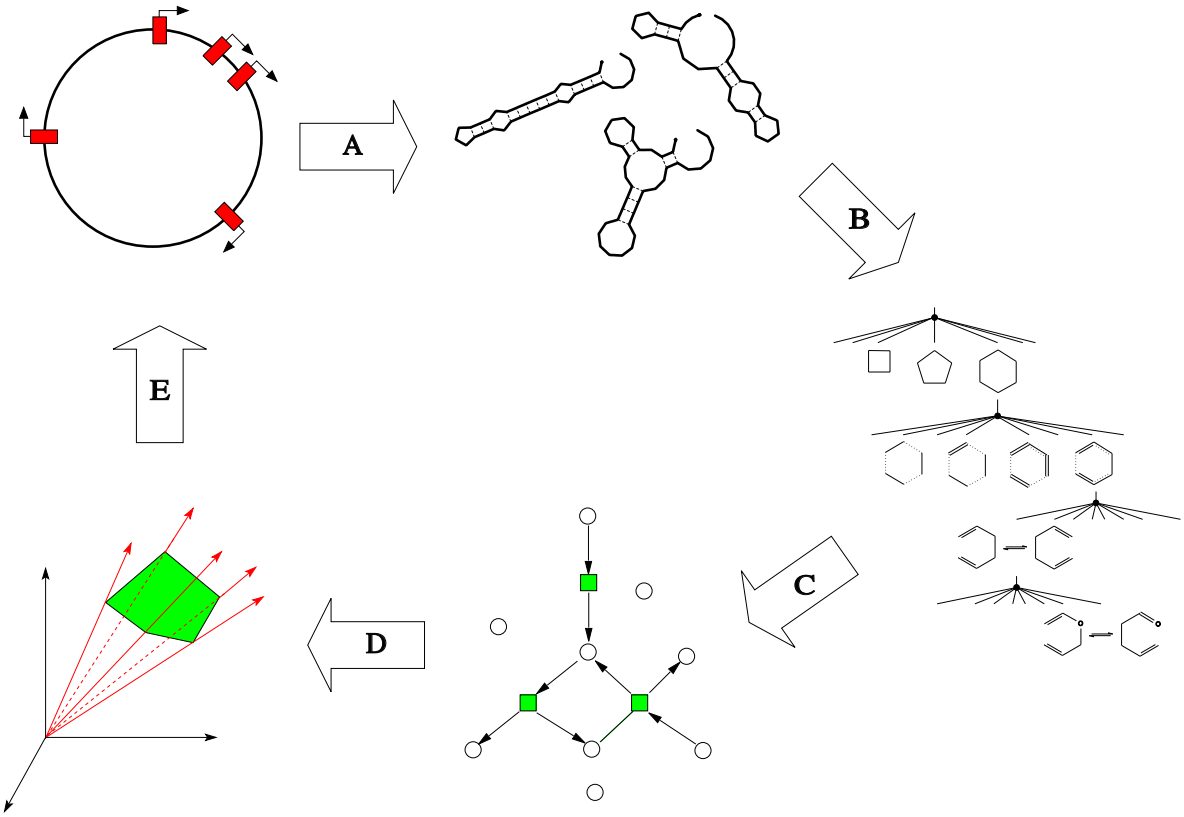


Figure 4.1: Overview of the simulation system: (A) Transcription of genes into catalytic ribozymes; (B) Assignment of catalytic functions to each ribozyme; (C) Estimation of reaction rates; (D) Construction and stochastic simulation of the metabolic network; (E) Metabolic Flux analysis and fitness evaluation; (F) Application of genetic variation operators.

occurring ribozymes and a wide variety of artificial selection experiments have shown that RNA molecules of about 100nt are capable of catalyzing most important types of chemical transformations that occur in a modern organism, see [118, 24, 155] for recent reviews. Thus it makes good sense from a prebiotic evolution point of view to implement "enzymes" as structured RNAs of approximately tRNA-size. For simplicity, a very simple genomic organization is used: A single RNA sequence serves as genome carrying a collection of non-overlapping "genes" encoding ribozymes. Start and stop positions of genes are marked by special sequence motifs.

The organisms in this model are thought to be haploid. As genetic operators, currently point mutations, deletions as well as gene duplication and horizontal gene transfer are used. More sophisticated modes of genome evolution such as rearrangements or recombination are excluded at present but could easily be incorporated into the computational framework.

The detailed modeling of any form of gene regulation is discarded in the present model to reduce the computational efforts. Again, such refinements could be included e.g. along the

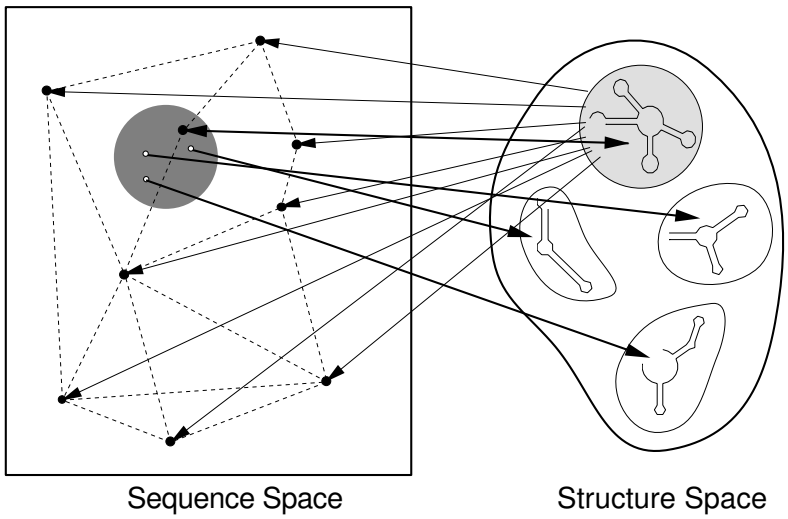


Figure 4.2: The RNA sequence-to-structure map: There are many more sequences than structures which brings redundancy into the map. Sequences which fold into the same secondary structure form extended neutral networks in sequence space. The strong interweavement of the neutral networks implies that the sequences in a small volume around an arbitrary sequence realize all possible secondary structures.

lines of [152, 48]. The minimal organisms of this model thus exhibit constant metabolic characteristics throughout their life-time, thus dispensing with the need to explicitly model any aspects of growth or development at the level of individuals.

4.3 Ribozymes

The catalytic activity of ribozyme as well as a polypeptide enzyme is dependent on the three-dimensional structure of the catalytic heteropolymer. The map from sequence to catalytic activity can be understood in two steps: sequence $\mapsto$ structure $\mapsto$ function. In the case of protein-enzymes, translation of the genomic nucleic acids sequence into the polypeptide sequence forms an additional mapping step.

The first step, the sequence-to-structure map [51] (Figure 4.2), is well approximated by the usual RNA folding algorithms. RNA molecules form secondary structure by folding back onto itself to form double helical regions interspersed with unpaired regions termed “loops”. The resulting secondary structure can be represented by an outer planar graph with nucleotides as vertices and base pairs as edges. A well established energy model [111], with parameters derived from melting experiments, assigns a free energy to every possible secondary structure. The simplest approach to RNA folding consists then of selecting the structure with minimal free energy from the combinatorial set of all possible structures. Fortunately, this task can be solved efficiently by dynamic programming algorithms that run in time proportional to the

cube of the sequence length. Here the folding routines as implemented in the `Vienna RNA package` [79] are used.

The statistical architecture of the RNA sequence-to-structure map and it's implications for the evolutionary dynamics [53, 54] has been extensively studied over the past decade. In particular, the map possesses a high degree of neutrality, i.e. sequences which fold into the same secondary structure are organized into extended mutationally connected networks reaching through sequence space. A travel along such a "neutral network" leaves the structure unchanged while the sequence randomizes. The existence of neutral networks in sequence space has been demonstrated in a recent experiment [145]. Due to the fact, that the neutral networks are strongly interwoven, the sequence-to-structure map shows another interesting property called "shape space covering" [146]. Meaning that within a relatively small volume of sequence space around an arbitrary sequence any possible secondary structure is realized. Both features of the RNA sequence-to-structure map account for directionality and the partially punctuated nature of evolutionary change.

For the structure-to-function mapping, unfortunately, there does not exist any well-understood physically realistic model. Instead, a simple purely computational model based on structural features motivated by early models of RNA evolution [53] is used. Catalytic structures typically depend on the molecular details of an active center, which is abstracted here to a local motif contained in a secondary structure. Here the longest "loop" (cycle) of the secondary structure is used as a computationally easily accessible feature of this type.

Without any claim of physical realism, this cycle can be interpreted as an encoding of the imaginary transition state of the catalyzed reaction. This type of mapping was inspired by the fact that many enzymes catalyze a chemical reaction by stabilizing its transition state and the work on reaction classification systems, in particular Fujita's imaginary transition structures (ITS) approach [59], in which cycles also play a central role. All common homo- and ambivalent reactions, which account for over 90% of all known reactions [77], can be described by a mono-cyclic ITS [73]. The rest of the reactions are usually composites of successive mono-cyclic reactions in sequence (rarely more than two [74]) with unstable intermediates like carbene or nitrene.

In order to construct and evaluate the structure-to-function map a hierarchical classification of imaginary transition states [76] is utilized here. The size of the ITS, i.e. the number of atoms involved in the electron re-ordering in course of the chemical reaction, corresponds to the length of the loop and constitutes level 1 of classification hierarchy (see rhs of Figure 4.3). The "reaction logo" specifies in addition the bond types in the transition state. The length and the type of the enclosing base pairs of the adjacent stems is further used to determine the bond types of the transition state. The absolute positions of the stems within the loop determine the arrangement of the electron re-ordering corresponding to level 3, the basic reaction. The information that leads from the basic reaction to the specific reaction (level 4), the atom-types, stems from the sequence within the loop. Again, each of the different loop

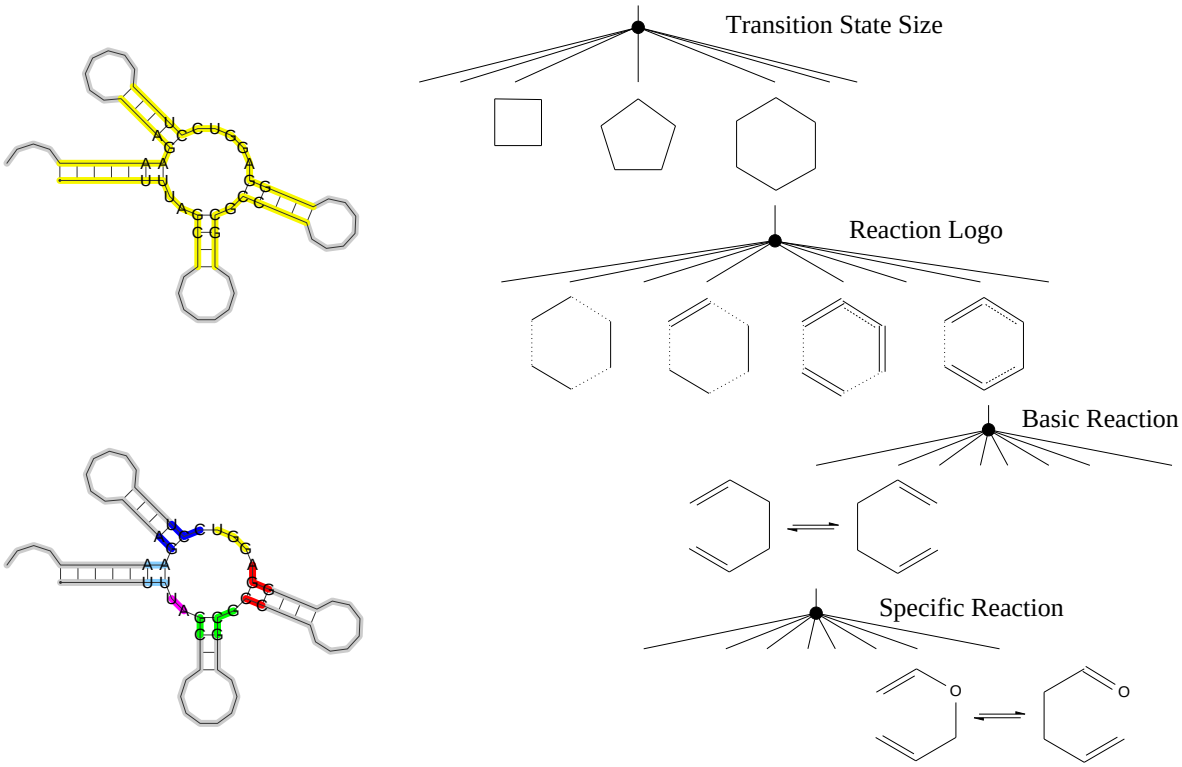


Figure 4.3: The structure-to-function map: (left) The colored regions of the ribozyme fold determine the catalytic function i.e. which leaf in ITS-tree is picked; (right) Along the levels of the ITS-tree the amount of chemical detail increases.

regions stands for one part in the transition state, here the atoms.

Since the structure-to-function map is not based on an approximation of physico-chemical principles but on an *ad hoc* model, it is necessary to investigate its statistical properties. To this end, the autocorrelation function of the sequence-to-function map is considered and compared to the autocorrelation function of the sequence-to-structure map of RNA folding [55]. For this, a distance measures on the spaces of RNA structures and transition states is needed, respectively.

For the structure space, an existing tree edit distance is used that is obtained through a sequence alignment procedure and the minimization of the cost for transforming one tree into the other, allowing deletions, insertions and relabeling of nodes as edit operations [51]. Similarly, the distance measure for the transition states starts with an alignment procedure. This can either be done on the graph representation or a unique string form of the transition state [163]. Edit operations include substitution of atoms, rearrangement of electron reordering, substitution of bonds and increase/decrease of transition state size. The cost of the edit operations rises in this order, atom substitution thus being the cheapest operation. The total cost for transforming one transition state to the other is then minimized.

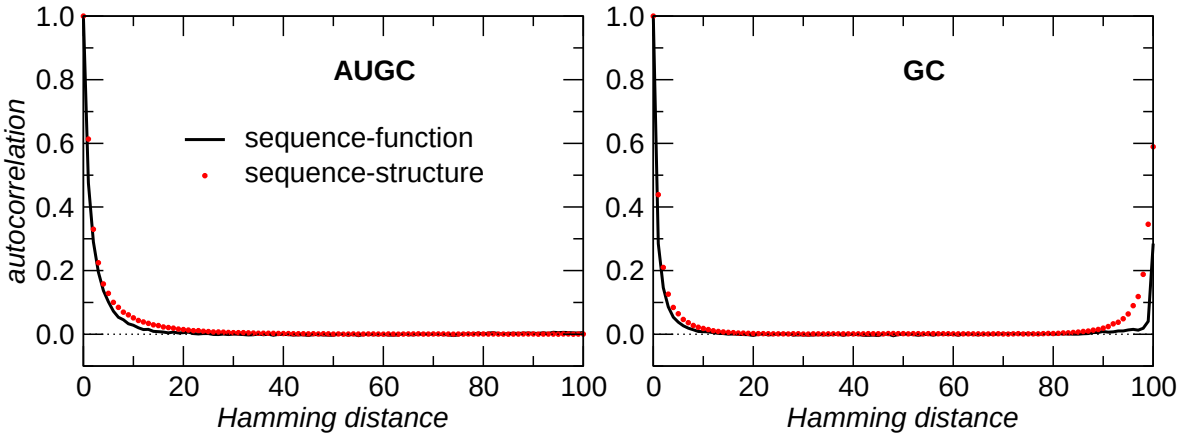


Figure 4.4: Autocorrelation functions for sequences of length $n = 100$ for secondary structure landscape and transition-state landscape, with alphabets AUGC (left) and GC (right). For each of 1000 randomly generated reference sequences 1000 mutants were produced for each of the 100 Hamming distance classes.

The autocorrelation function of a map $\varphi : (X, d) \to (Y, D)$ between metric space X with distance d and Y with distance D can be defined as

$$\rho(d) = 1 - \frac{\langle D(\varphi(x), \varphi(y)) \rangle_{d(x,y)=d}}{\langle D^2 \rangle}, \tag{4.1}$$

where $\langle D^2 \rangle$ denotes the expected distance between the images $\varphi(x)$ and $\varphi(y)$ of two independent elements $x, y \in X$ [51]. Figure 4.4 shows that the composite sequence-to-function map behaves much like the underlying sequence-to-structure map. This is not surprising: if the sequence-to-structure map is dominated by neutral and essentially randomized structures, as in the case of RNA folding, then the second component, the structure-to-function map, has little influence on the overall behavior of the composite sequence-to-function map [150]. This observation in particular justifies the use of an *ad hoc* artificial structure-to-function map in this simulation setting.

It can also be shown that the composite map, of RNA sequence to structure map and the novel structure to function map described above, performs superior against other artificial genotype-phenotype mappings, as well as other maps based on RNA folding, in terms of evolvability, connectivity and extension of the underlying neutral network (see Section 6.5.1). Thus, making it the preferable choice for the present model and possibly other similar optimization tasks.

4.4 Reaction Network Generation

To apply the system of graph-rewrite rules, here a generic graph rewrite engine is utilized. The computationally most difficult step is the identification of all occurrences of the left graph of

the rule in an input graph. To solve this subgraph isomorphism problem the dedicated state-of-the-art `VF`-algorithm (freely available in the C++ `VFlib-2.0` library [28, 29]) is applied. For each match, the input molecule is then rewritten according to the current graph rewrite rule. The resulting molecule graphs are converted into unique `SMILES` [174] to test for duplicates. The initial molecule(s) and the resulting ones are needed to calculate the transition rate for the applied reaction.

The energy calculation is performed through the ToyChem model as discussed earlier in Section 3.3.3. For situations in which a faster rate calculations are needed, an estimation using quantitative structure-property relationship (QSPR) and the Wiener numbers of reactants and products can be used. Here, the QSPR and the approach for activation energy computation from Faulon is applied, delivering still realistic enough results [46], for the calculation of the rate constants. The final reaction rate is gained, by multiplying the rate constant with the reactant concentrations divided by the volume (here, the sum of concentrations of all molecules in the particular cell).

The complete chemical reaction network can be constructed by exhaustive enumeration. In practice, however, this is not feasible due to the combinatorial explosion that would result from iteratively applying all possible reactions to all combinations of molecules. It is imperative therefore to prune the growing network at each step by removing energetically unfavorable products and by ignoring highly unlikely reaction channels [135, 46], Figure 4.5.

Suppose a given list of reaction mechanisms and an initial list $\mathfrak{L}_0$. Performing all unimolecular reactions on each molecule $\mathsf{M} \in \mathfrak{L}_0$ and all bimolecular reactions with each pair of molecules $(\mathsf{M}_1, \mathsf{M}_2) \in \mathfrak{L}_0 \times \mathfrak{L}_0$ we obtain a new list $\mathfrak{L}_1'$ and a list of new molecules $\mathfrak{L}_1 = \mathfrak{L}_1' \setminus \mathfrak{L}_0$. The recursion then proceeds in the obvious way:

$$\mathfrak{L}_{k+1}' = \left(\bigcup_{j=0}^{k-1} \mathfrak{L}_j \right) \times \mathfrak{L}_k \cup (\mathfrak{L}_k \times \mathfrak{L}_k) \tag{4.2}$$

and $\mathfrak{L}_{k+1} = \mathfrak{L}_{k+1}' \setminus \bigcup \mathfrak{L}_k$. This type of strategy [46] was applied in practice e.g. to predicting product distributions from simulations of chemical cracking and combustion processes, which have notoriously large reaction networks.

In addition to kinetically inaccessible reaction products also molecules with more than 30 atoms are excluded in order to keep the efforts of computing molecular properties within manageable bounds. Note that the resulting reaction networks could contain autocatalytic compounds whose production would have to be kick-started by external addition of a small amount of that compound. Evidence for such autocatalytic compounds (notably ATP) has been reported by Kun and collaborators [104] in the metabolic networks of several species.

In order to check whether a newly generated molecule m is already contained in a previous list a comparison of the structural formulae must be performed. This is done by transforming the molecular graphs into their *canonical SMILES* representation [175], which then are compared as strings.

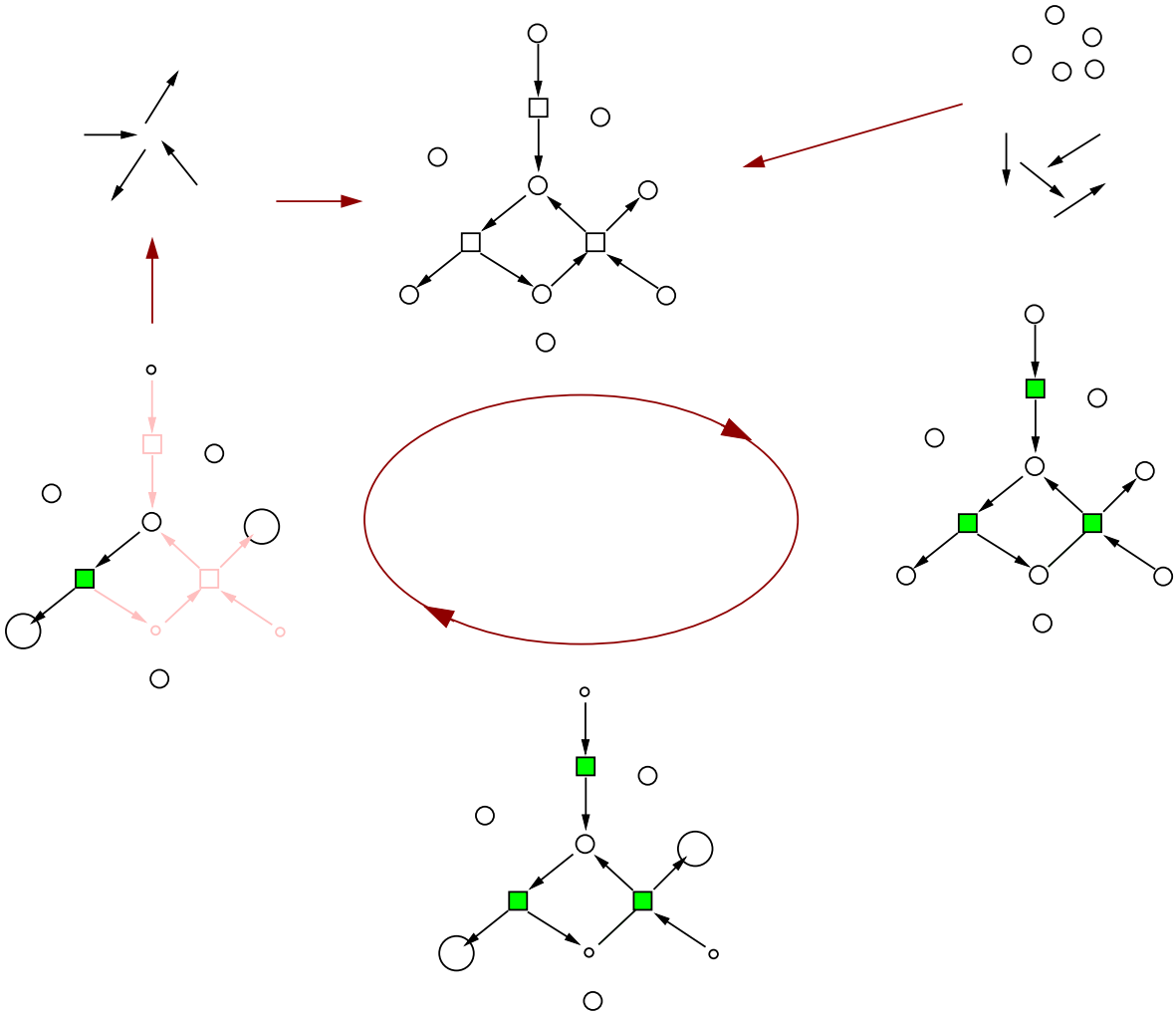


Figure 4.5: Generating reaction network: To avoid combinatorial explosion during reaction network generation a filtering step, which prunes unproductive parts from the reaction network, is needed after each application of the reaction set (arrows) to the (current) set of molecules (circles). The network usually quickly converges in size if the filtering is done on a kinetic basis. In particular, after the estimation of reaction rates (green squares), the dynamics of the reaction network is simulated by a Gillespie type stochastic method, followed by removing nodes from the reaction network which have not accumulated enough particles, due to small reaction rates. This type of strategy [46] has been used to predict the right product distributions in simulations of chemical cracking and combustion processes, which are notoriously large reaction networks.

4.5 Fitness and Selection

The final ingredient in this minimal model of evolutionary processes is the choice of fitness function and a scheme for selection.

The fitness of the protocells is derived directly from their metabolic yield, more precisely, the amount of "desirable end products" that can be produced from a defined quantity and composition of input material. Its explicit computation is again a computationally nontrivial task. First the pathway distribution of the metabolic network under the steady-state assumption is determined using metabolic pathway analysis (MPA) [127]. This approach starts from the stoichiometric matrix S of a metabolic network which is extracted from the structural information encoded in its graph representation. (Internally, the simulations represent metabolic networks as bipartite graphs composed of metabolite and reaction nodes.) The steady state assumption implies that the interest lies in non-negative flux vectors $\vec{v}$ in the null-space of S, i.e., $S\vec{v} = 0$. The assumption is that catalyzed reactions have a non-zero flux only in one direction. The implementation of MPA in this computational framework delivers the set of extreme pathways from which all other admissible pathways through the metabolic network can be derived as linear combination. The optimal yield of the entire network is therefore realized by one of the extreme pathways[60]. The fitness is consequently computed as the maximum of the (linear) yield function over all extreme pathways.

This fitness function depends on the definition of a set of metabolites that need to be produced as "desirable end products". This set can be either chosen explicitly by the user (entering a set of target molecules and a graph-similarity measure), or by defining an order on the produced metabolites with the help of molecular descriptors. Here, several different topological indices such as Balaban-Index [4] or Wiener-Number [177] are offered. A certain number of produced metabolites with maximal/minimal (user's choice) values (graph-similarity or topological index) then constitutes the set of "desirable end products".

The selection process is modeled here as adaptive walk, which applies in the limit of strong selection, weak mutation, negligible interactions between individuals, and constant environment [125]. An adaptive walk amounts to accepting a genomic mutation if and only if it increased this yield function. A similar setup is used e.g. in simulations of metabolic evolution based on group-transfer reactions [130] that explain the emergence of hub metabolites.

4.6 Visualization

For the analysis of complex simulations an efficient visualization tool is of particular interest. On the one hand it is used to identify simulations with worthwhile properties for further inspection and on the other hand it supports the choice of appropriate statistics and measures to summarize the simulation results.

Several tools have been developed for the visualization of large-scale biological networks [159, 101, 139] and there exist methods for dynamic graph drawing [94], however, there is no tool which combines these two fields to handle dynamic biological systems.

The visualization tool used here was specifically build for the use with the evolving metabolic networks produced by the computational framework introduced in this chapter. Almost all aspects of single simulation runs can be investigated through the use of versatile points of view and the high scalability that is provided. This supports a profound and efficient analysis of the structure and properties of the generated metabolic networks and its underlying components, while giving the user a vivid impression of the dynamics of the system. The analysis process of the visualization tool is inspired by Ben Shneiderman's mantra of information visualization [148]. For the overview, user-defined diagrams give insight into topological changes of the graph as well as changes in the attribute set associated with the participating enzymes, substances and reactions. This way, "interesting features" in time as well as in space can be recognized. A linked view implementation enables the navigation into more detailed layers of perspective for in-depth analysis of individual network configurations.

4.6.1 Data

The most prominent data that is visualized are the metabolic networks representing the molecular interactions of the simulation's protocells. A directed bipartite labeled graph is used to visualize the metabolic networks, which is a commonly used representation. Metabolites and enzymes (reaction) build the nodes (subnodes) of this graph. While its edges indicate participation in the same metabolic reaction, which is further specified through edge labeling (see Figure 4.6 for a reaction schema). Several attributes of the network's components, such as metabolite concentration and enzyme activity, are also visualized as will be explained in the following paragraphs.

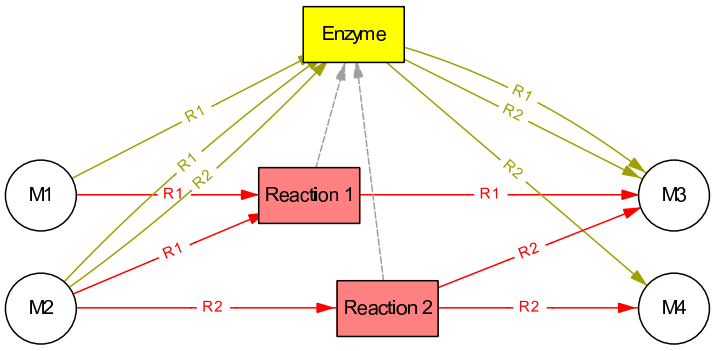


Figure 4.6: One enzyme node can have an arbitrary number of reaction nodes as children. Metabolites (circles) are un-grouped. Either the red elements (Reaction View) or the yellow elements (Enzyme View) are visible. Metabolites are always shown. The dashed lines indicate the child-parent-relationship. From [140].

The second type of information which is visualized is the set of extreme pathways of the metabolic networks in from of a matrix, where each row represents one pathway and columns are reactions. The matrix entries indicate the participation of a reaction in the corresponding pathway.

Finally, the information about the sequence of the metabolic networks is used to build a coherent picture of the evolution of the metabolic system over all generations of the simulation.

4.6.2 Framework

In this paragraph, the possibilities of the visualization tool and the responsible methods are given, for a more detailed description see [140].

The central object of interest and starting point of all visualization tasks here is the network graph depicted in Figure 4.7. The constitution of this graph with enzyme nodes and corresponding reaction child-nodes (see Figure 4.6) allows switching between views, the enzyme view (Fig 4.7 top) with enzymes as nodes and reactions encoded only in the edge label, and the reaction view (Fig 4.7 bottom) discarding enzymes and representing each reaction as a single node.

The network graph can be augmented with further visualizations of information about the network's dynamics and some of its components. The set of extreme pathways is used here in two ways. Firstly, single pathways or combinations of pathways can be selected, highlighting all of its components. Secondly, the information about which components (metabolites / reactions / enzymes) participate to which extent in the extreme pathways is used to define node and edge weights. This is particularly useful in combination with the dynamic graph drawing capabilities of this visualization tool, allowing it to view the change in material flow through the network (see Figure 4.8). Another information that is encoded in the augmented network graph is the metabolite concentration, which is represented by the fill level within a metabolite node.

The data from one simulation can already produce large network graphs, with potentially important information in small sets or even single components. The visualization framework therefore provides the ability of overviews and detailed views. As a simple example, the labeling for metabolite nodes changes from a mere index in the broad overview, to a structure formula, to a `SMILES` notation (see Appendix) and finally in the most detailed view to the molecule graph of the metabolite, enabling to view the exact reaction educts and products (see Figure 4.9).

Another part of the visualization framework are secondary visualizations, e.g. overviews that span over all generations of the simulation, such as the life-time diagram and time-series charts tracking certain network properties such as average or maximum node degree, and detailed attribute diagrams for single networks from a specific generation, giving a more precise view

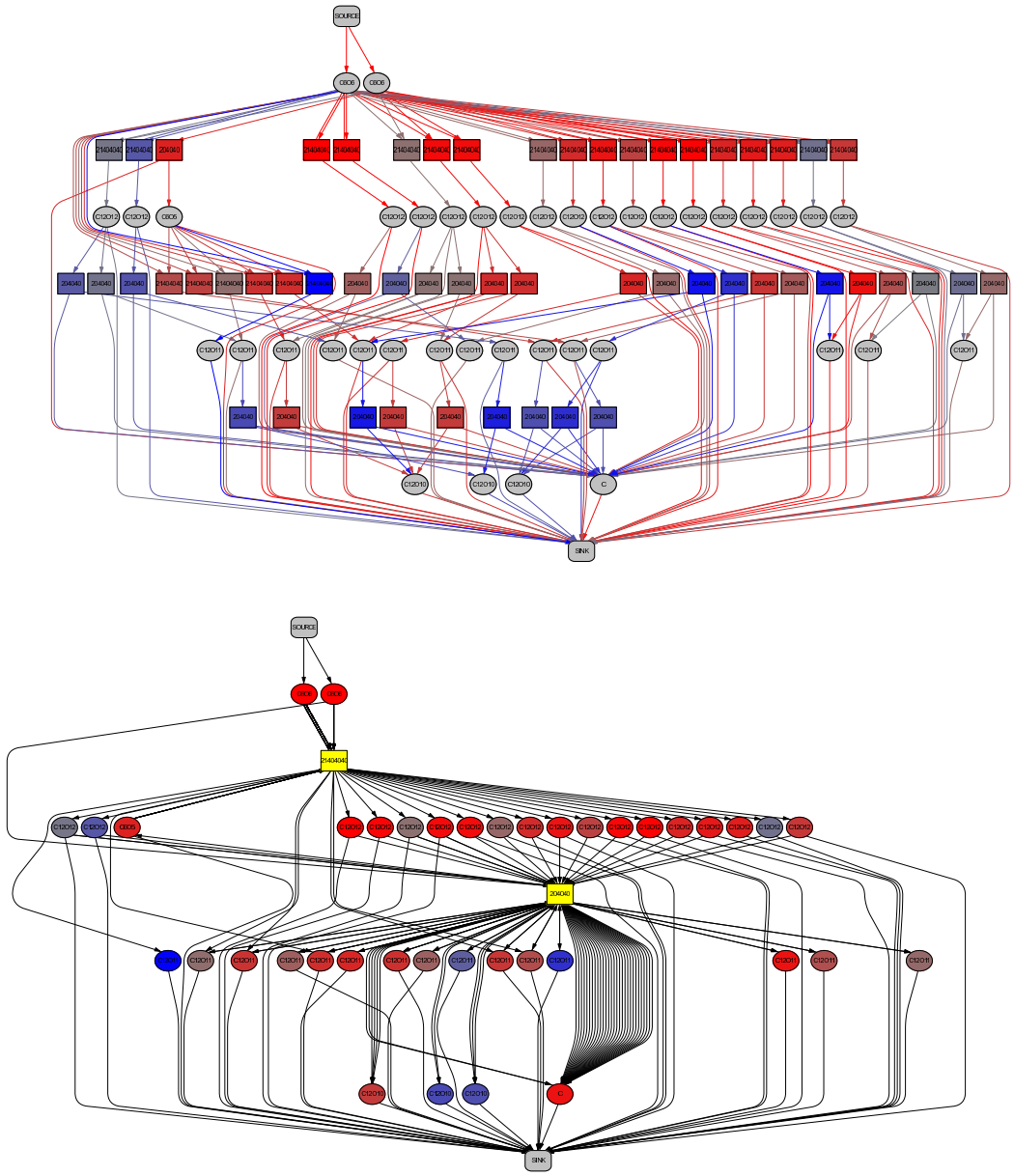


Figure 4.7: Union graph laid out using Sugiyama layout algorithm. The reaction nodes (rectangles) are colored according to their first appearance (red: earlier, blue: later). Note that the positions of metabolite nodes (ellipses) remain the same. From [140].

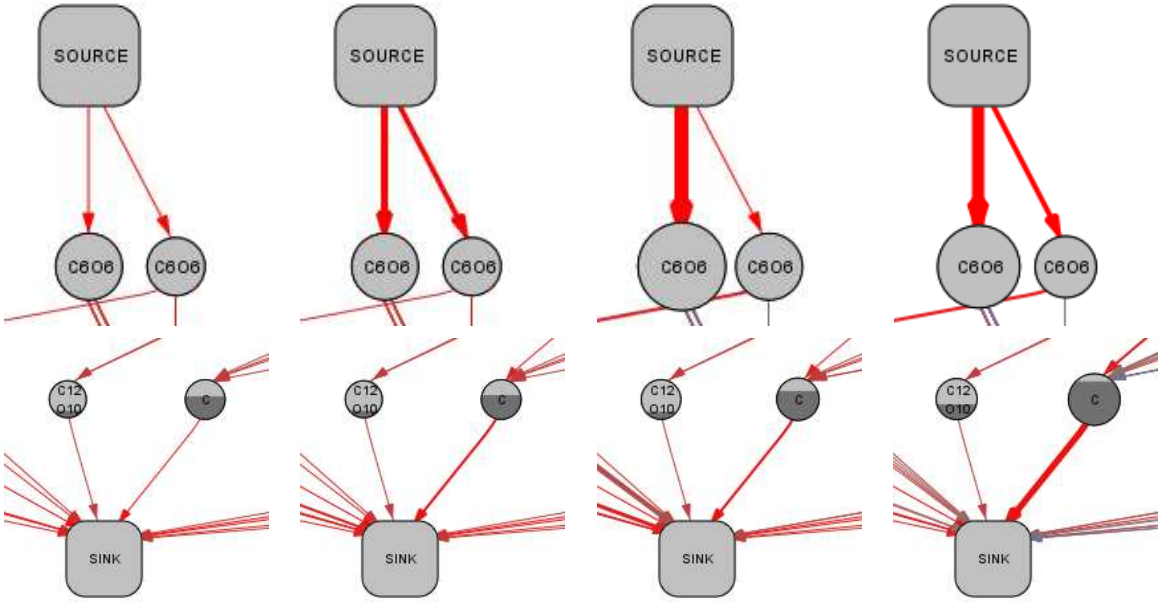


Figure 4.8: Scaling of the nodes and edges in the network graphs of several generations. The size of the nodes and the thickness of the edges decodes the strength of the flow through the metabolites and reactions, respectively. The filling of the metabolite nodes represents their concentration in the network at a specific time.

on the network properties, e.g. the node-degree distribution is the detailed counterpart to the average node degree time-series.

The life-time diagrams depict the inclusion of all nodes (metabolites / reactions / enzymes) in the network for each generation. Every row of the diagram stands for one node. If a node was included in the network at a specific generation then the respective row is colored at this time point. The color of the row encodes the 'age' of the node, i.e. the generation in which the node entered the network for the first time. Nodes that entered early ('old') are colored red, nodes from later generations ('young') are colored blue. Inside the row, attributes specific for each node type are depicted as dark columns. For metabolites the height of the columns indicate the concentration, for enzymes it shows the number of catalyzed reactions and for reactions the amount of material flow through this node.

All of these visualizations are closely linked to the network graph, on the information level as well as in terms of user interaction, which is reached through a linked view implementation [137]. This means that all interactions in the network graph are updated and immediately shown on all other views, which allows a fast navigation in space (nodes and edges) and time (generation of the simulation). The same is true for the reverse direction, if interesting generations or nodes are discovered in overviews, the respective network is immediately shown and the corresponding parts of the graph are highlighted.

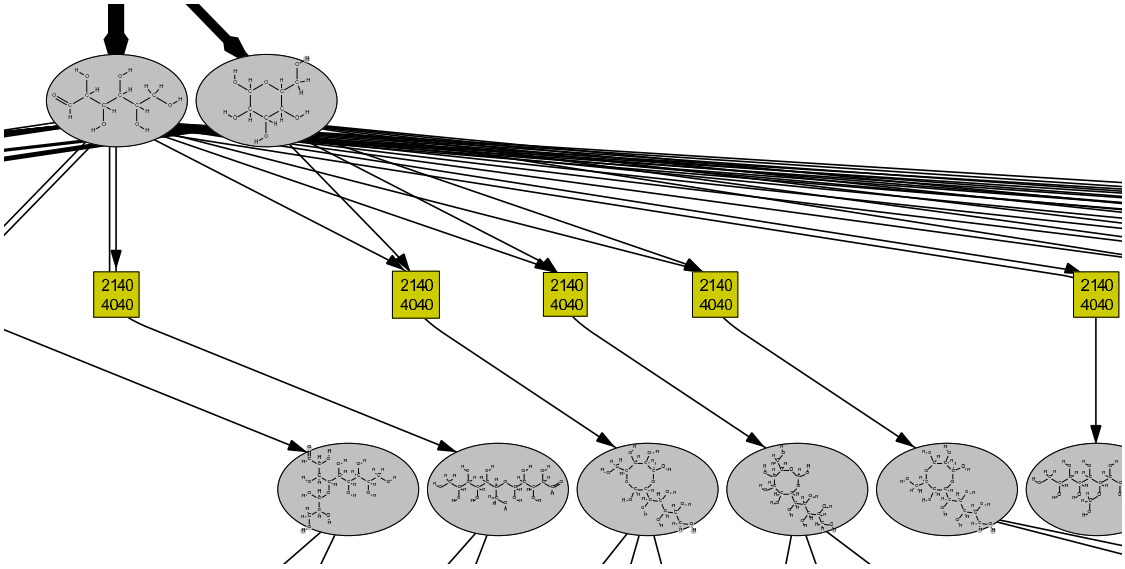


Figure 4.9: Semantic Zoom: Below a certain level-of-detail threshold, the chemical structure of the molecule is shown instead of the totals formula. From [140].

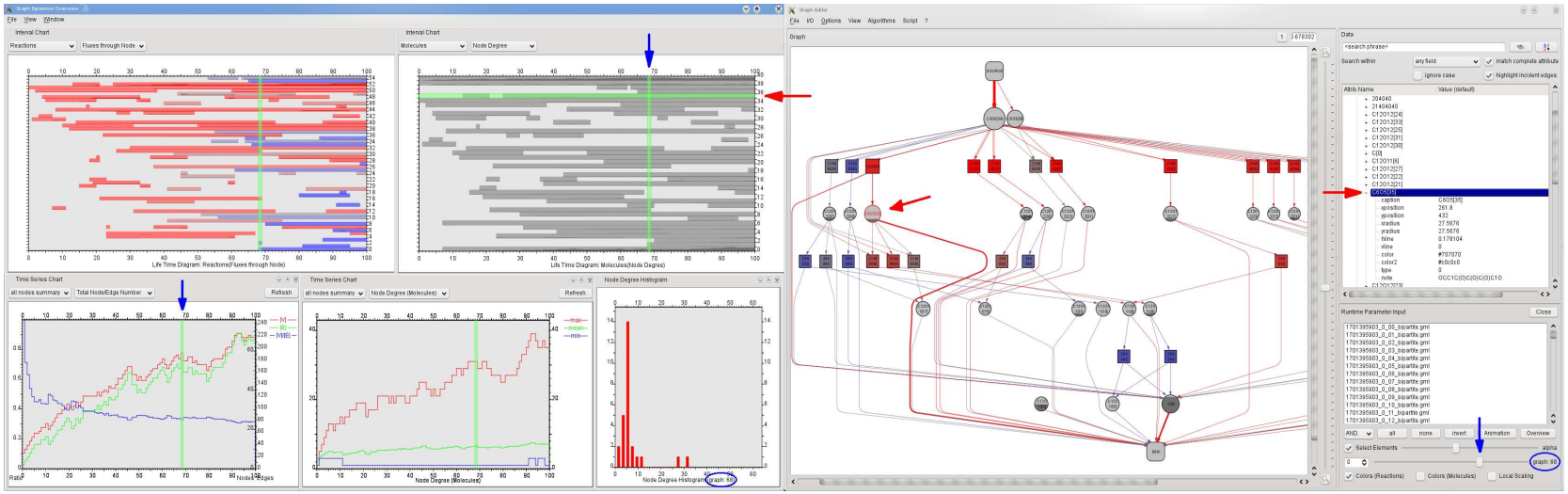


Figure 4.10: Linked View realization facilitates browsing different graph snapshots in time. The blue arrows indicate the current position in time, red arrows indicate the selected node in the current generation. These components of the graphical user interface are also sensitive to user input and can be used for navigation. Selecting a node in the Graph Scene (r.h.s.) highlights the associated row in the appropriate interval chart as well as the associated point in time in all charts. The five diagrams given on the r.h.s. display the following data. Top: Life time diagram of reactions overlaid with the number of pathways through each reaction node. Life time diagram of metabolites overlaid with each node's degree. Bottom: Time series chart giving number of nodes, edges, and nodes-to-edges-ratio. Time series chart of summarized node degree (minimum, maximum, average) over all metabolites. Node degree histogram of the currently displayed graph generation. From [140].

Chapter 5

In silico Evolution of early Metabolism

In Chapter 2 the problem of metabolic evolution and in particular the formation of metabolic pathways was elucidated. Some of the most popular scenarios and evolutionary mechanisms were introduced. Each of these mechanisms was shown to be supported by evidence from certain pathways, so that none of these mechanisms is exclusive. As discussed, studies on hypotheses of pathway evolution [20, 117] suggest that metabolism has evolved differently in different phases. Furthermore, only traces, or "shadows", are still observable from the events in the very distant past of terrestrial life. Many aspects of the evolutionary history are therefore still not well understood. In particular, the first steps that lead to the emergence of the earliest forms of metabolism evade observation by conventional approaches.

Thus, there is an urgent need for detailed and realistic models of early metabolism that consider all its components and scales. As shown, simulation approaches have proven to be useful in finding and challenging explanations for the evolution of biological networks. In Chapter 4, a computational framework for the early evolution of metabolism was introduced, modeling all its significant components in a realistic way. In this chapter, the focus lies on the detailed analysis of evolutionary transitions, aiming in particular at an understanding of the processes underlying metabolic innovation.

5.1 Computational Approach

Innovation is hard to model. In contrast to population dynamics or quantitative genetics, where the dynamics is governed by Darwinian selection and the generation of variability can be described by simple statistical models, we need a way of judging whether an innovation has been selected, or whether an observed fitness increase is the result of an incremental adaptation. This implies that phenotypes must be represented explicitly as objects whose fitness can be evaluated. This paradigm has been explored already two decades ago in the

context of evolving RNA molecules [52], where phenotypes are modeled as RNA secondary structures. Subsequent investigations have demonstrated that the sequence-structure map or, in a more general setting, the genotype-phenotype map [146, 85] plays a crucial role. More recently, neutral networks were studied in the context of gene regulation [25] and metabolic networks [142].

In particular, the accessibility of potential novelties plays a crucial role: in a realistic setting the search spaces are so large that evolutionary trajectories are determined to large extent by the ease with which advantageous mutants can be generated from extant populations [53, 149]. It is crucial, therefore, to devise models of phenotypes that are as biophysically realistic as possible and computationally feasible. While in the case of RNAs the fairly simple and well-understood relation of RNA sequences and RNA secondary structures, i.e., the map defined by RNA folding, could be employed, it is a highly complex task to devise realistic genotype→phenotype→fitness mappings for even minimal organisms. A very primitive "ribo-organism" was devised, for instance, to study the evolution of primitive genetic codes [173].

5.2 Results

In this section, the computational model described in this work is used to simulate the evolution of metabolic networks and analyze the change of its structure and components over several generations. All simulation runs performed for this section were initialized with the full set of chemical reactions to chose from, the same configurations for genome length (5000 bases), and the same TATA-box constitution ("UAUA") and gene length (100 bases). They differ in initial conditions, population size, environmental conditions, selection criteria, and simulation time (number of generations).

5.2.1 Quantitative Analysis

To gain some quantitative insights into the general principles of metabolic evolution, a series of simulation runs was performed to investigate certain measures that give a picture of the evolutionary constitution of the metabolic networks throughout the evolution process.

As a starting point the connectivity of enzymes and metabolites throughout the evolutionary process is investigated. The assumption from biological observations and simulation studies [130] is that enzymes from early stages show a higher connectivity than those from later stages, which are more specialized in the sense that they catalyze only few reactions. Similarly, highly connected metabolites, so called hub metabolites usually are ancient components of metabolic networks. Here, these findings can be confirmed by the analysis of the networks from a sample of 100 simulation runs starting from a simple set of initial metabolites (cyclobutadiene, ethenol, phthalic anhydride, methylbutadiene, and cyclohexa-1,3-diene). For every generation the average contribution of enzymes and metabolites that originated at the particular time

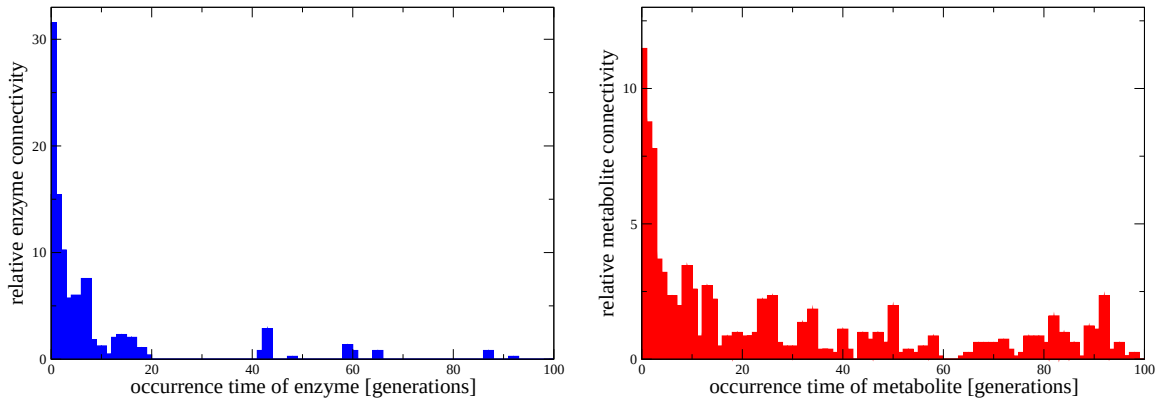


Figure 5.1: Average relative connectivity of a) enzymes and b) metabolites introduced in the same generation, for 100 generations. The height of the bars shows the fraction of the overall connections that are accounted by enzymes/metabolites from a particular generation. All values are averages over 100 simulation runs. Input molecules are not considered in the statistic, they account for nearly 50 percent of metabolite connectivity.

is tracked. Figure 5.1(a) shows a clear trend for enzymes from the first generations to be responsible for the major part of connections in the metabolic network. On the one hand, this can be explained simply due to the fact that enzymes that enter the system earlier have more time to form connections. On the other hand, this observation could also indicate that enzymes with higher and higher specificity evolve in the later stages. It could be anticipated, that enzymes with all specificities still appear in later generations but only specific enzymes catalyzing few reactions are taken to the next generation, while multi-functional enzymes are discarded because they would change the structure of the network too rigorously. Considering the connectivities of metabolites (see Figure 5.1(b)), the highly connected nodes can still be found in the early steps, especially if considering environment metabolites that are always abundant which account for about 50 percent of connectivity (not shown here). However, there is a constant production of metabolites potentially becoming highly connected. The evolution of enzymes and metabolites show similarities in that in both cases highly connected components mostly stem from the early phase of evolution. However, the extent to which that happens is different. The reason for this observation might be that evolution selects for enzymes but not for metabolites.

5.2.2 Evolution of metabolic Pathways

In order to find arguments for some of the evolution hypotheses, the occurrence time (age) of reactions and metabolites along pathways is observed and studied. It is of particular interest to determine in which direction (downwards – with the flow of mass, or upwards – against mass-flow) pathways are formed by addition of chemical reactions that recruit or produce

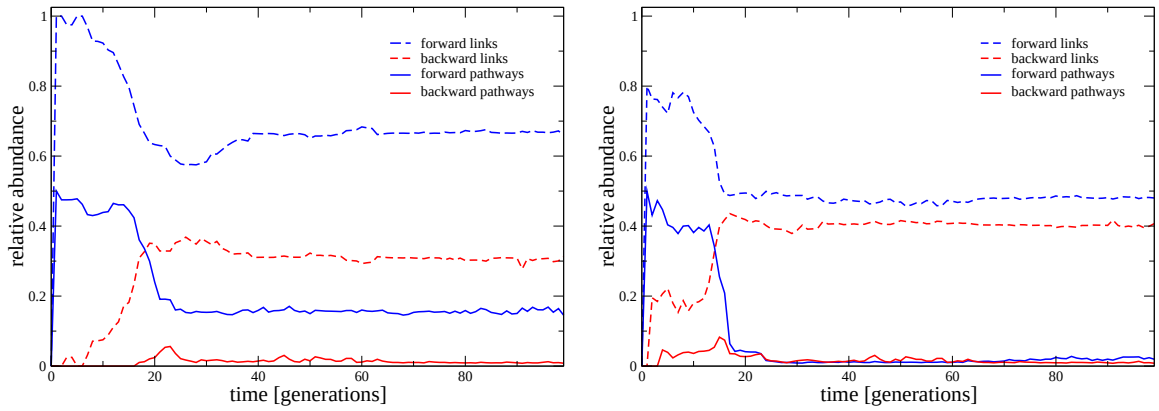


Figure 5.2: Evolutionary history of simulated metabolic networks. For the first 100 generations, we show the number of links and pathways that conform to the forward and backward evolution scenarios, respectively. Links are pairs of a) consecutive reactions or b) consecutive metabolites along a pathway. A pathway is identified as "forward-evolved" if at least one of its links is forward and none backward. In the first generations, the network consists predominantly of forward (reaction) links and pathways. After about 20 generations, the relative abundance of forward pathways decreases drastically but quickly reaches a persistent plateau value.

new metabolites. For the investigation of reactions, the term forward (backward) link will be used if, in a pair of reactions in a pathway, the successor is evolutionary older (younger), i.e. the catalyzed reaction that lies more downstream in the pathway occurred later (earlier) in the evolution process. In the same vein, a forward (backward) link between metabolites refers to a situation in which the products of a reaction are evolutionarily older (younger) than the educts. Accordingly, forward (backward) pathways are defined as pathways in which there is at least one forward (backward) link and no backward (forward) link. Given these definitions, the set of extreme pathways is computed for every generation and all cells. For each pathway the percentage of forward and backward links and pathways is determined, for both reactions and metabolites.

For this study, 100 simulation runs were performed with the following settings: a population size of 100 cells running for 100 generations and performing 100 network expansion (stochastic simulation) steps per generation. The always abundant input molecules were cyclobutadiene, ethenol, phthalic anhydride, methylbutadiene, and cyclohexa-1,3-diene. In Figure 5.2, one can see the change from generation to generation in the constitution of the metabolic networks regarding the above mentioned measures of forward/backward links and pathways. Considering the reactions of the networks, it can observed that in the first generations, the networks consist mainly of links and pathways conforming to the forward evolution scenario. However, in later generations a much more mixed mosaic-like picture appears, arguing in favor of the patchwork model. This trend becomes even more evident from the metabolite's point of view:

almost all pathways consist of forward and backward links in equal numbers. Another observation from the reaction's point of view is that most forward pathways from the early stages remain even until the last stages, which could mean that they form a core of pathways that are not subject to evolutionary change. This would support the shell hypothesis.

For the gene history analysis in the next section, longer simulation runs of 2000 generations were performed, while keeping the same initial conditions as in the previous simulations, however, with fewer network expansion steps per generation. The statistics for these runs is summarized in Figure 5.3. The main observation is the ongoing trend of a mix of forward and backward links while retaining a certain fraction of forward pathways throughout evolution, giving further support to the observations from the previous shorter simulation runs. The situation as studied so far can be interpreted as follows. While the first generations of network evolution are dominated by the forward evolution scenario, patchwork evolution takes over after a sufficiently diverse repertoire of enzymes has built up from which enzymes can be recruited. An evolution in layers, as proposed in the shell hypothesis, so far has not been observed. However, the maintenance of the set of pathways that originated by forward evolution in the earliest generations at least suggests the possibility of an ancient metabolic core from which later pathways are build by enzyme recruitment or other strategies, such as enzyme or pathway duplication.

Until now, the simulation results do not provide any support for the backward evolution scenario. However, the simulations investigated so far have not incorporated an environment with temporary depletion of "food" metabolites, which is one of the major assumption of this theory. To this end, the following study is investigating exactly this impact of variations in resource abundances to metabolic evolution. For this purpose, now 100 simulations over 900 generations are run and analyzed. The initial conditions are the same as in the previously described simulations. However, the set of "food" metabolites is changed for certain time periods. For the first 100 generations the system is left unperturbed and thus under the same conditions as in the previous simulations. Starting from generation 100, one of the five "food" metabolites will be removed from the input set. This means that the particular metabolite can still be produced by other metabolites but is not permanently imported from the environment and might deplete in the same way as other internal metabolites. After a time interval of 50 generations the removed metabolite is added back to the set of "food" metabolites and the next "food" metabolite is removed. After every "food" metabolite has been removed once (after 350 generations) the initial "food" set is reintroduced for the next 50 generations. The upper plot in Figure 5.4 shows the metabolic pathway evolution statistics for these first 400 generations. In the next phase which is depicted in the lower plot in Figure 5.4 , pairs of "food" metabolites are removed for intervals of 50 generations in the same procedure as described above for the removal of single metabolites. After 900 generations every possible combination of two metabolites from the full set of "food" metabolites was removed once.

In the first 100 generations, no fundamental differences to the previous simulations can be observed. This is to be expected since the conditions up to this point are the same. In the

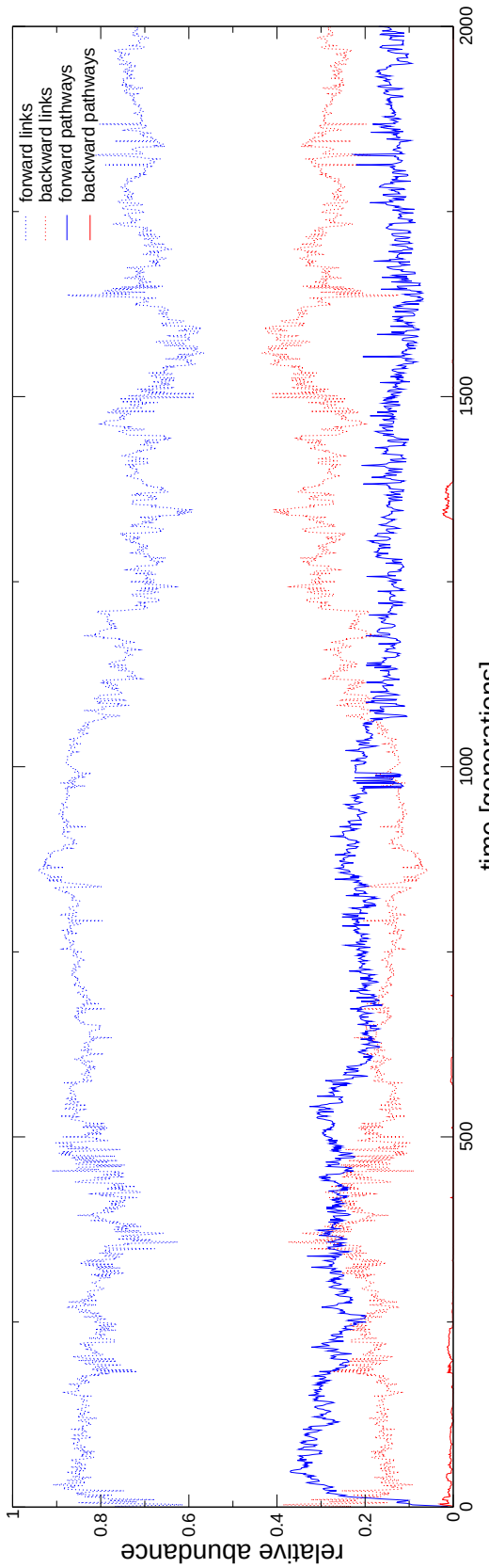


Figure 5.3: Evolutionary history of longer (2000 generations) metabolic network simulations. The statistics, as explained above, show the number of links and pathways that conform to the forward and backward evolution scenarios, respectively. The trend is similar to the shorter simulations in that forward evolution dominates the starting phase, converging in a mixed situation typical for the patchwork model.

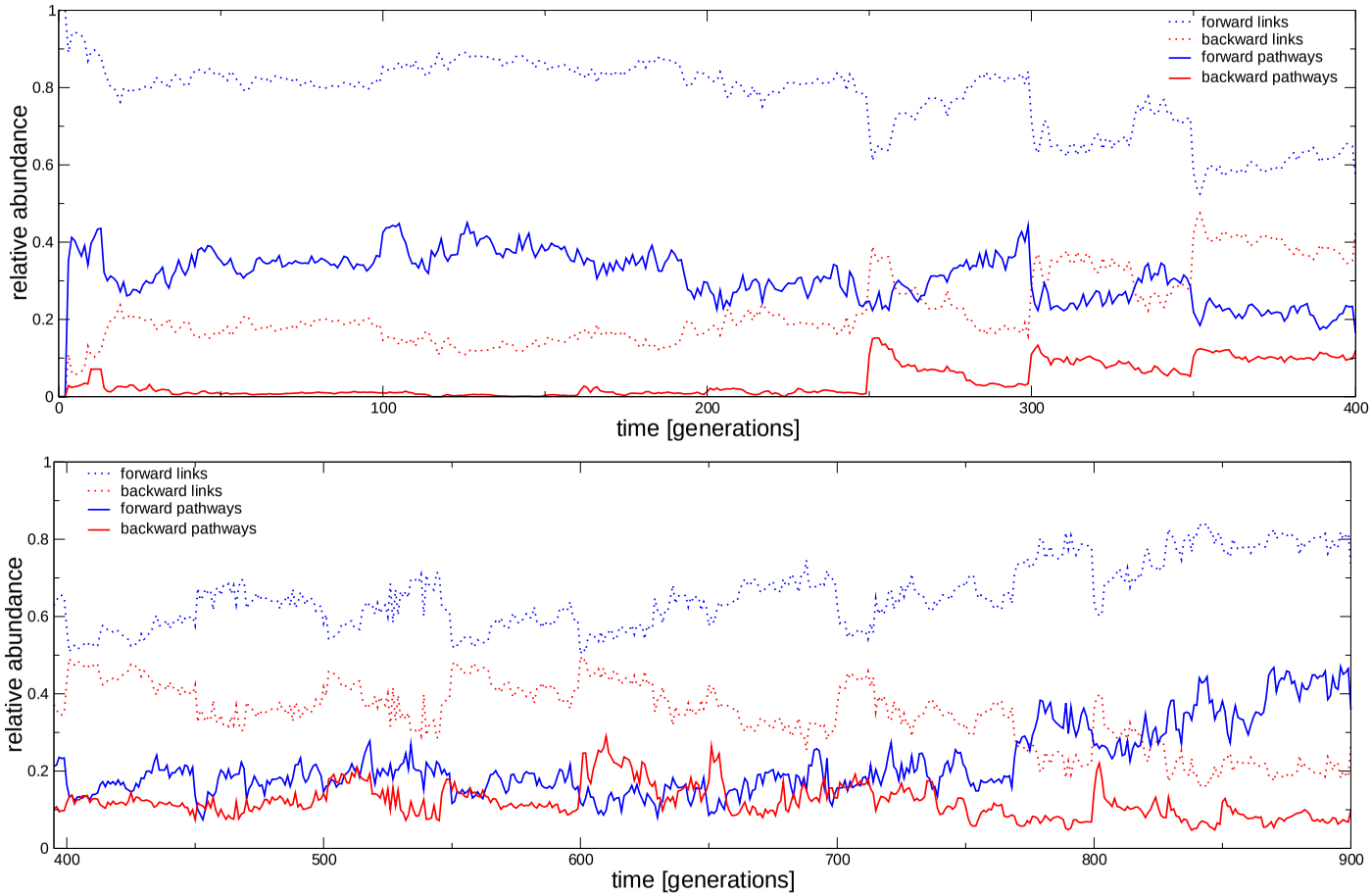


Figure 5.4: Simulated evolutionary history of networks in an environment with perturbations in the set of "food" metabolites. The same statistics as for the other simulations is used, i.e., the number of links and pathways that conform to the forward and backward evolution scenarios, respectively. For the first 100 generations, the food set is the same as for the previous simulations. In the next five intervals of 50 generations each, the "food" metabolite set is perturbed by removing one of the five initial metabolites at a time. From generation 350-400, the original full set is reintroduced. The remaining simulation (400-900) is divided in intervals of 50 generations. In each interval, exactly two of the "food" metabolites are withdrawn. The system reacts to the largest perturbations with a sudden increase in backward links and pathways, followed by a slower decrease, suggesting that metabolite depletion can indeed induce backward evolution.

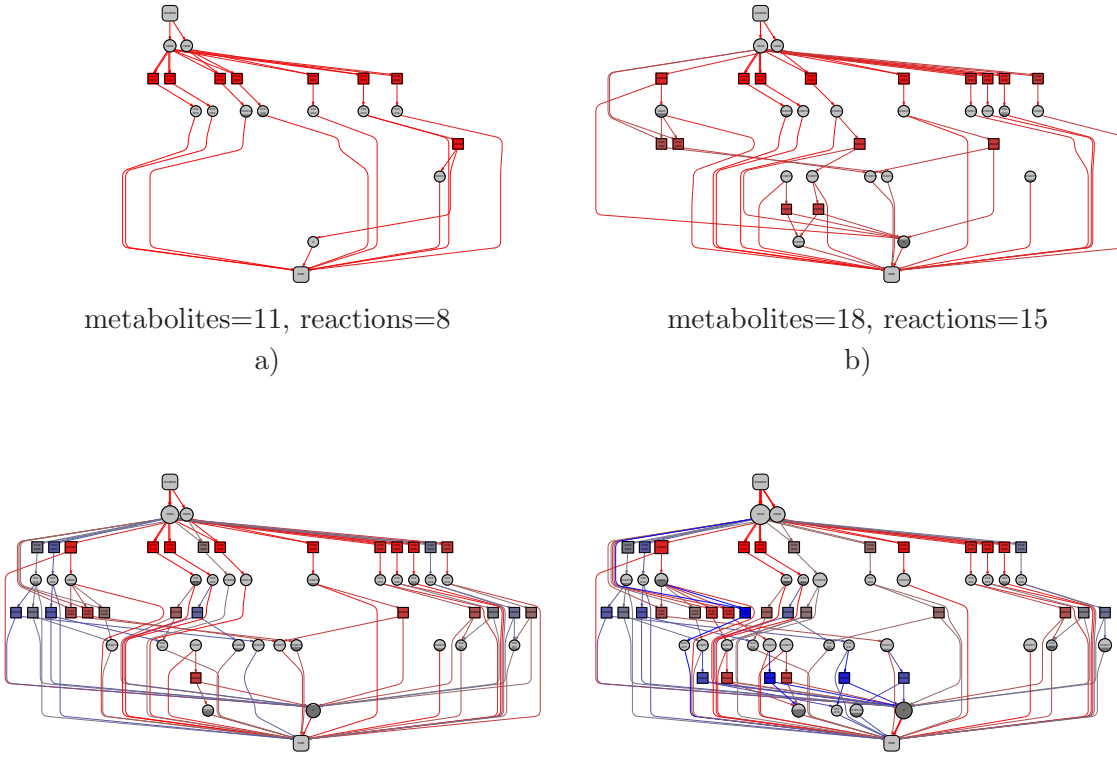
following three intervals of 50 generations, from generation 100 to 250, still no significant changes in the network evolution are apparent despite the perturbation in the set of "food" metabolites. For the next 100 generations (250-350), however, an increase in backward links and the emergence of backward pathways can be observed. More specifically, at the beginning of the two intervals there is a sudden and drastic increase of backward evolved pathways followed by a slow decline for the rest of the interval. The explanation for the increase in backward pathways is that of the retrograde evolution hypothesis: due to the depletion of an important "food" metabolite, there is a strong selective advantage for pathways that can produce this metabolite from other sources. The sudden increase suggests that the reactions or at least the enzymes of these backward pathways must have been already present in the network rather than invented through the emergence of entirely new enzymes or catalytic mechanisms, i.e., capitalizing on enzyme promiscuity. The other factor explaining the sharp increase in the beginning and the slow decline afterwards is that the network complexity and thus the overall number of metabolic pathways decreases after depletion of the "food" metabolite. Therefore, pathways that depend upon the depleted metabolite, mostly forward pathways, disappear since they no longer contribute significantly to the organism's fitness. When the network later evolves new (forward or mixed) pathways based on the new conditions, the fraction of backward pathways among all pathways decreases. The five intervals of perturbation are followed by an interval (350-400) with the full "food" metabolite set. Surprisingly, this interval also starts with a sudden increase in backward links as well as a slightly smaller increase in backward pathways and is not followed by a decrease later in the interval. No satisfactory explanation can be provided for this observation so far.

In the second phase (400-900), in which pairs of "food" metabolites are withheld, similar observations to the first phase of perturbations can be made. However, the modifications to the network are not as significant and less clearly visible. The depletion of "food" metabolites is again mostly followed by an increase in backward links and pathways, but less dramatic as before. Further, the slow decrease after the sharp rise is not as clear as in the previous phase and sometimes interrupted by smaller increases. In the later intervals (750-900), one can even observe an overall decrease in backward links and pathways compared to forward pathways. To conclude, synchronous depletion of multiple metabolites can induce some form of backward evolution. At the same time, however, too invasive perturbation can disrupt the entire system. Furthermore, it can be suggested that the backward evolution observed within the simulations is driven by an enzyme recruitment process rather than a formation of completely novel reaction pathways.

5.2.3 Detailed Analysis of simulated evolutionary Histories

In the following, some of the findings from the previous study are illustrated in more detail for a single simulation, starting with only two input molecules and developing only few enzymes. For the visualization of an evolutionary time series see Figure 5.5. Figure 5.6 gives an overview

of reaction- and metabolite-lifetimes. The genome, and hence the set of enzymes, is chosen at random in the beginning. The two input molecules of this simulation are the cyclic and sequential forms of glucose. The simulation run is terminated after 100 generations.



metabolites=11, reactions=8
a)

metabolites=18, reactions=15
b)

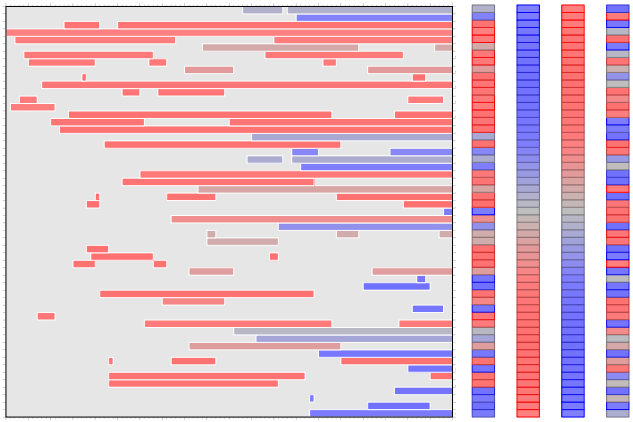
metabolites=27, reactions=27
c)

metabolites=30, reactions=30
d)

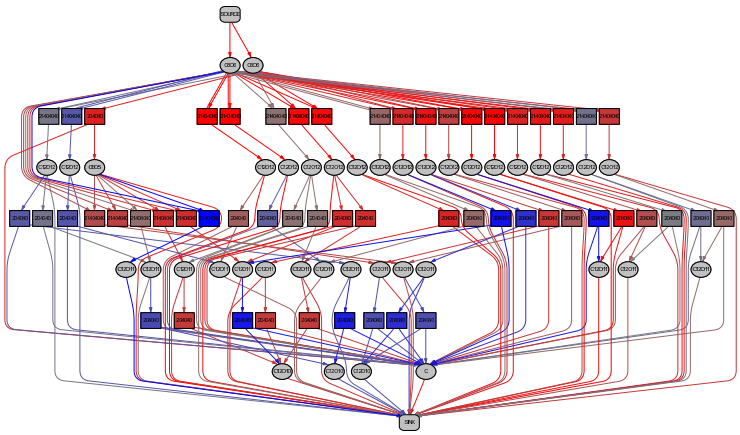
Figure 5.5: A series of simulated metabolic networks after a) 10, b) 30, c) 66, and d) 100 generations. Colored squares represent chemical reactions, gray circles represent metabolites. Metabolites involved in a reaction are connected to it in the network graph. The size of the nodes and the width of the edges encode for the number of extreme pathways in which the respective object is involved. The coloring for the reactions encode their age, where red stands for older (occurrence in early generation) and blue for newer (later generation) reactions.

The focus lies again on the evolutionary constitution of the metabolic network, i.e. investigating the relation between the occurrence time (age) of chemical reactions and their position in the network (downstream vs upstream) to draw conclusions about one of the evolution scenarios being at work. The four snapshots in Figure 5.5 showing the metabolic network in different stages are aligned to a union graph over all generations [140]. From this view, it is easy to see that in the first generations the reactions upward in the network are added. The pathways are formed further in this forward direction. Looking at the last generation, basically all pathways from source to sink follow the forward evolution scenario.

a)



b)



c)

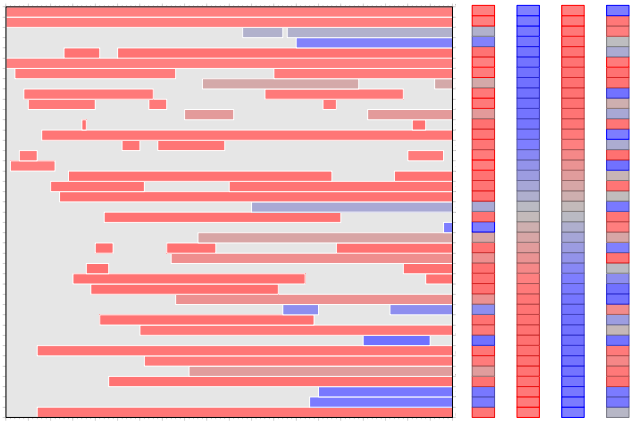


Figure 5.6: Life-time diagrams for reactions and metabolites. a) Life-time of reactions, b) union network graph over all 100 generations, c) life-time of metabolites. The reactions and metabolites (rows) in the life-time diagrams are positioned corresponding to their position in the union network graph, i.e. reactions/metabolites close to the source metabolites are in upper positions, reactions/metabolites close to the sink metabolites are placed at the bottom. The rows have colored entries if the corresponding reaction/metabolite was present at a certain generation (columns 1-100). We use the same coloring scheme as above, older reactions/metabolites are red, newer blue. The colored bars show the age distribution of reactions in the network in the same order as in the lifetime overview. The first bar represents our results, following the pattern for backward evolution, forward evolution and the patchwork model.

This observation is further supported by the life-time diagram for all chemical reactions in Figure 5.6. The reactions are here ordered according to their position in the union graph, combining all components that occurred throughout the simulation. There is a clear trend of older reactions being on the top of the metabolic network (upstream) and younger ones following more downstream. The colored bar next to the life-time diagram shows the pattern of the relation between age and position of reactions and metabolites for the example simulation run. The other three bars show the patterns for backward, forward evolution and the patchwork model, respectively. The forward evolution pattern shows the highest similarity to the simulated pattern. This illustrates again the speculation from the general analysis that in the early phase of metabolic evolution, forward evolution seems to be dominant. However, for metabolites there appears to be not a clear relation between the position along pathways or the network and their first appearance in the system. Similar to the general results, a much more mixed picture is observed for the metabolites. Therefore, no clear explanation can be made for the metabolite constitution.

Another, more complex, setting is used in a simulation run in which the evolutionary history of the involved genes/enzymes is investigated, depicted here in the catalytic function genealogy for all generations (Figure 5.7). The simulation takes the same five input molecules from the above general study, but with a higher mutation and duplication rate and runs for a total of 2000 generations. The simulation frameworks allows to study the emergence of divergence and convergence of catalytic functions [2] since it can record the genealogy of each gene (reaction catalyst) throughout a simulation run, and it can utilize the ITS classification of the catalyzed reaction as a representation of the enzymatic function. Divergence of function is caused by gene duplication followed by sequence mutations, creating functionally different but structurally related catalysts. Convergence of function occurs when catalysts from genealogically unrelated genes independently accumulate mutations resulting in the catalysis of the same reaction (or class of reactions). In Figure 5.7 convergence events are marked by circles. A small selection of divergence events, which were very frequent in all simulations, are marked by broken circles.

In Figure 5.8a) the history for one specific gene/enzyme (ITS code: 404040) is shown, together with the ITS-structures of the enzymes leading to the formation of this enzyme and those that originated from it. Interestingly, the divergence of the enzyme is always preceded by an increase in the number of genes, either through duplication or convergence of other genes/enzymes. That corresponds to the duplication and divergence scenario proposed in several biological studies on the evolution of enzymes. Furthermore, the analysis of the functional transitions on the basis of the ITS graphs reveals that catalysts can alter their substrate specificity by small changes in the context of the graph rewrite rule, i.e. the necessary precondition for the applicability of the graph transformation rule. In this example, most of the adjacent enzymes have a similar ITS structure and consensus in most parts of the substrate structure as well as in the reaction mechanism. However, the first as well as the last adjacent enzyme do not show any significant similarity to the ITS structure of the studied enzyme.

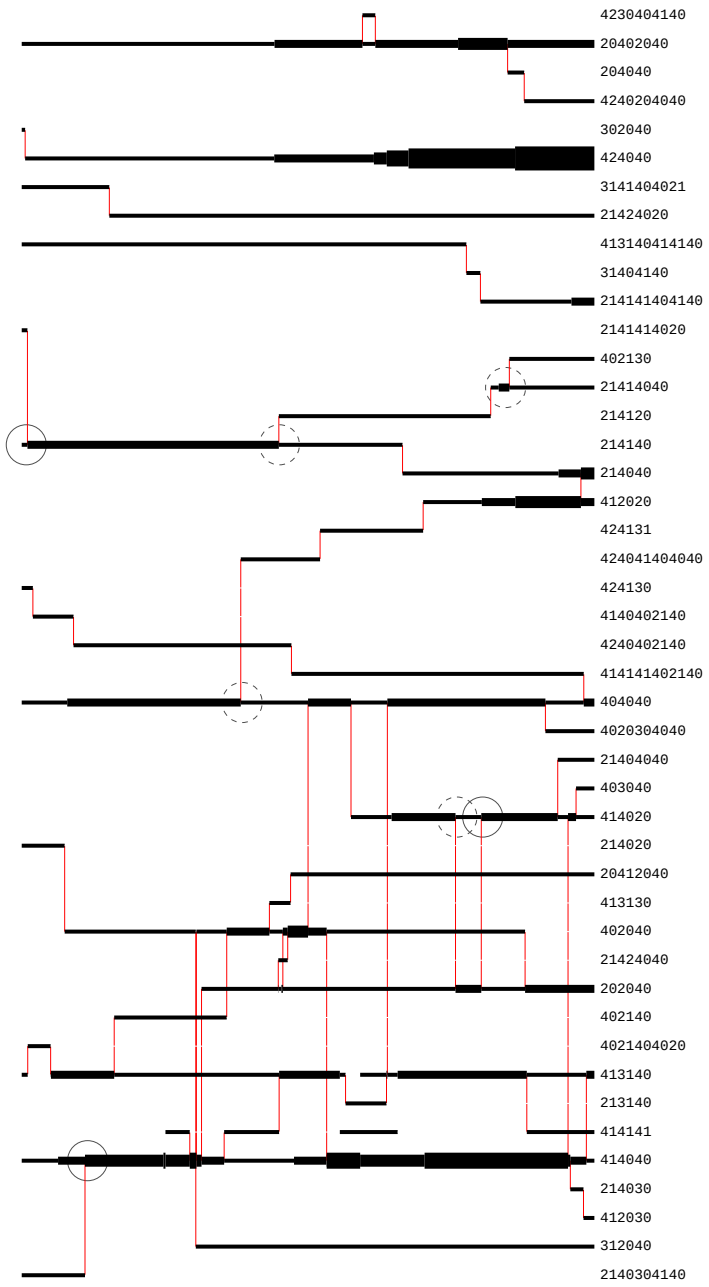


Figure 5.7: Genealogy of catalytic functions and gene dosage over 2000 generations. Each row represents an observed catalytic function. Black horizontal lines indicate time intervals in which genes coding for that catalytic function were present in the genome (0-200: from left to right). The thickness of the black lines indicates the number genes with a given function. Thin vertical red lines indicate points where the accumulation of mutations caused a transitions between catalytic functions. If the number of genes copies in a function class increases without a transition from another gene, then the increase is due to a gene duplication. A new gene can be created in the genome through the fortuitous formation of a TATA-box. Conversely, a gene can vanish if its TATA-box is destroyed by mutation. On the left of the chart a numerical encoding of the graph transformations performed by the "enzyme" is plotted.

The four transitions which show more similarity in terms of their ITS are of particular interest and thus only these transitions will be discussed here.

The first of these transitions is described in Figure 5.8b, with the ITS codes, the ITS structures and the corresponding reaction mechanisms of the enzymes, as well as a sample reaction using one of the five “food” metabolites catalyzed by them. Here, only one atom in the context of the ITS is changed (from O to C) while all other parts remain the same. This preserved the exact same reaction mechanism. Nevertheless, both enzymes react with different metabolites from the original set of “food” molecules. The enzyme with the ITS code 402040 reacts with phthalic anhydride, while 404040 is able to make use of methylbutadiene (see example in (b) and (c)) and cyclohexa-1,3-diene (see example in (d) and (e)).

In the next transition (Figure 5.8c) one additional bond is introduced to the context of the reaction, increasing the substrate specificity of the new enzyme and also resulting in a significant change in the product structure, despite keeping the overall reaction mechanism. The new enzyme uses ethenol, which is also available among the five “food” metabolites. It is an interesting observation that although all of the three enzymes discussed until now posses the same reaction mechanism and differ only slightly in the stabilized transition state structure, they are able to make use of different starting molecules and consequently also introduce metabolites not present in the metabolism before.

The transition in Figure 5.8d causes a loss in substrate specificity through removal of two bonds from the reaction context. As before, the change of the product structure is more significant than the comparably moderate change in the substrate molecule. However, this enzymes is not able to use any of the original “food” metabolites. This might explain the accepted specificity loss of this transition which is more rare in the studied simulations than transitions with a gain in specificity. However, since the new enzymes is able to use some of the constantly abundant ”food” molecules, the transition is beneficial.

A more interesting case is the last of the four transitions (Figure 5.8e). Although the ITS structures of the enzymes before and after the transition seem very different, a large part of the reaction mechanism is retained and the change in the substrate binding can be described as an increase of substrate specificity by adding two atoms (N and C) to the context of the reaction. The upper parts of the substrate, product, and reaction are equal in both enzymes, only the lower part of the new enzyme differs from the original enzyme. The mutated enzyme is able to make use of the same “food” metabolite (phthalic anhydride) as the enzyme in (b) and yields similar product metabolites but has a more restrictive context.

These examples demonstrate that, in the simulation universe, relatively small changes in the gene sequence can lead to new enzymes with typically similar reaction mechanisms but different substrate specificity. This in turn sometimes causes quite drastic changes in the reactions that are actually catalyzed. This also shows the power of the introduced structure-to-function mapping.

The impact of enzyme promiscuity on enzyme evolution, enzyme engineering and biocatalysis

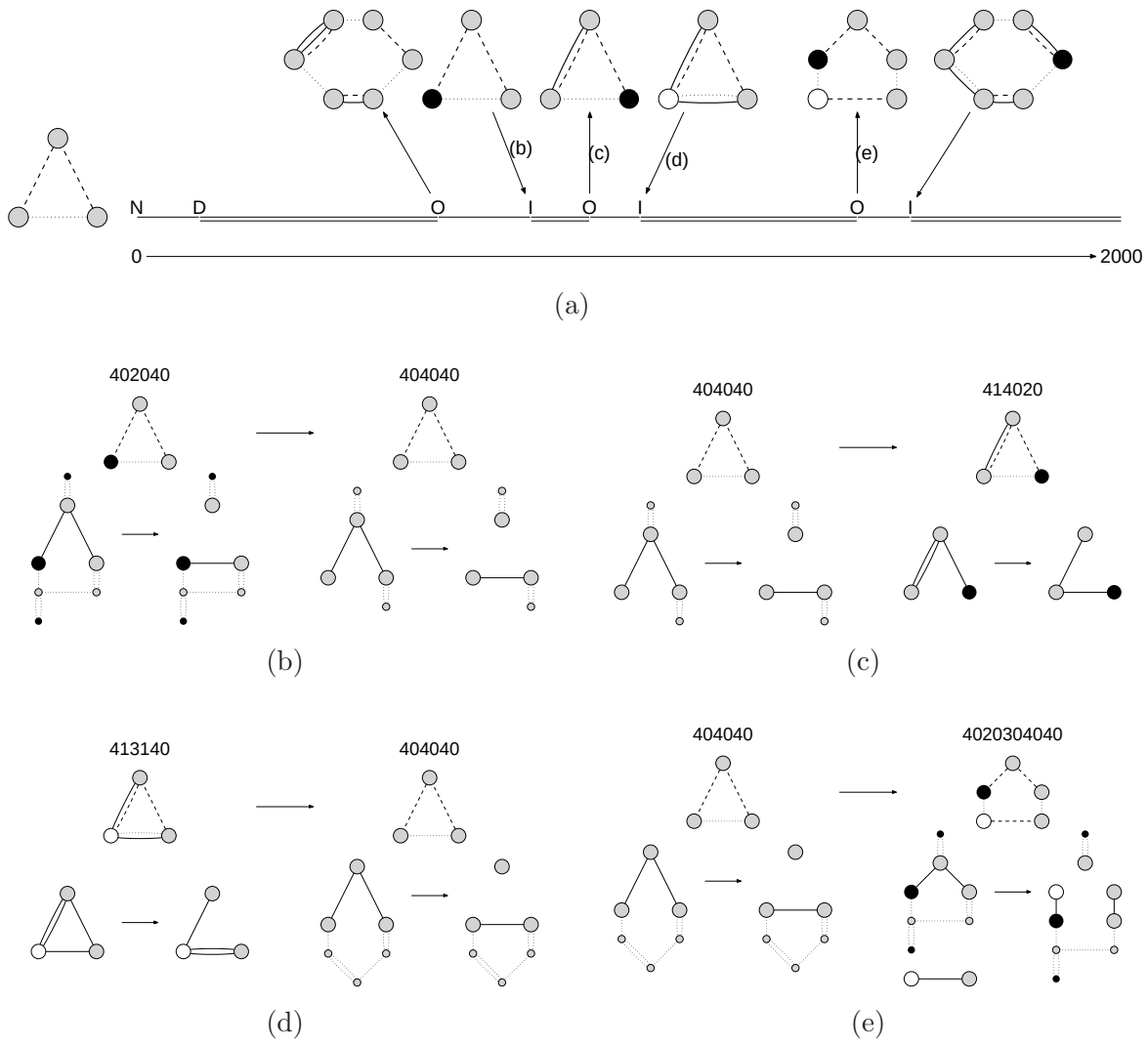


Figure 5.8: (a) History for one enzyme (ITS: 404040) of the 2000 generation simulation run, with ITS structure of this enzyme and all adjacent enzymes. ITS structure - Lines: solid = reaction context, dashed = bonds that are broken by reaction, dotted = bonds that are created; Circles: black = oxygen, gray = carbon, white = nitrogen. Evolutionary events are marked in the timeline when they occur, N = New occurrence, D = Duplication, I = (In) convergent event, O = (Out) divergent event. The number of lines parallel to the timeline indicate the number of gene copies for that enzyme. Four of the six functional transitions are depicted (b)-(e) with the ITS codes (top), the ITS structures (middle) and the reaction mechanisms (bottom) of the two adjacent enzymes. The actual reaction mechanism is represented by the big circles and solid lines only. Including small circles and dotted lines gives a sample reaction using one of the original "food" metabolites. In (b) substrate changes in only one atom position (O to C), in (c) substrate specificity increases through addition of a bond to the context, in (d) substrate specificity decreases due to removal of two bonds from the context and in (e) substrate specificity increases by adding two atoms (N and C) to the context.

[83, 95, 18, 27, 122] has been discussed widely throughout the literature for quite a long time. For instance many enzymes exhibit so called substrate promiscuity i.e. they perform the same chemical transformation on a wide range of substrates. An impressive natural example for this type of enzyme promiscuity is methane monooxygenase (EC 1.14.13.25), which hydroxylates approximately 150 different alkane substrates in addition to its major substrate methane. This characteristic of natural enzyme function is fully represented in the presented model and is as well a target of evolutionary change. Other aspects of enzyme promiscuity such as two different reaction mechanisms are either implemented by the same residues in the active site (thymine hydroxylase EC 1.14.11.6), or the residues in the active site are used in different mechanistic context, do not have a representation within the model.

Noteworthy, the sample enzyme and its six neighboring enzymes (connected trough functional transitions in this sample simulation) are already able to catalyze chemical reactions using four of the five originally added "food" metabolites. While most functional transitions from one enzyme to another introduce only little innovation to the reaction repertoire of the metabolic system, some give access to previously unreachable parts of the existing chemistry.

5.2.4 Summary of Results

Using both simple examples and a series of more complex simulation runs, the evolution of the components on the small scale (metabolites, enzymes) as well as on the level of systems (pathways, networks) was investigated. The analysis of the genes history, showed all different kinds of evolutionary events, such as convergence, duplication and divergence, and many different functional transitions from one gene/enzyme to another, increasing the substrate specificity or changing the reaction chemistry. The simulations further allow to discriminate between different scenarios for the evolution of metabolic pathways. Based on the observations from this study, it can be argued that the different evolutionary hypotheses can be reconciled, in that they act in different phases of evolution, i.e., in different scenarios one might observe another strategy at work. Here, it is suggest that forward evolution dominates in the earliest steps and is then superseded by a phase of enzyme recruitment, however, leaving behind a trace in form of a core set of forward evolved pathways. Another noteworthy finding of this study is that the depletion of important "food" metabolites introduces backward evolved pathways. However, the formation is rather driven by enzyme recruitment than a formation from scratch according to the retrograde hypothesis.

Chapter 6

Emergence of complex Properties

As discussed in Chapter 2, the emergence and evolution of system properties in complex biological systems is an intriguing field of research in biology with many unresolved questions. The knowledge that can be gained from it is not only beneficial for the understanding and optimization of existing systems but also for constructing many kinds of novel artificial systems.

In this chapter, the emergence and evolution of complex properties in biological systems is investigated by studying the *in silico* evolution of early metabolism and observing the structure and dynamic behavior of the underlying metabolic networks. For the former, the already introduced multi-level computational model for the evolution of catalyzed reaction-networks is used to simulate different evolutionary scenarios and thus provide appropriate network data from an evolving biological system. The latter, is achieved with the help of conventional network measures as well as measures suited for metabolic reaction networks. The goal is to gain insights about the complex properties of the investigated networks and their evolution throughout the simulation.

In the following, several measures of system properties, such as robustness and modularity are introduced and the results of their application to the simulated networks are discussed.

6.1 Simulation Data

For the investigation of this chapter, several extensive simulation runs were performed. Three different evolutionary scenarios were considered. For each scenario 100 simulation runs are recorded. One simulation consists of 500 generations (1000 generations for one of the three scenarios) and 100 network expansion steps per generation. The set of initial input metabolites, the "food" metabolites, contains 17 molecules from the citric acid cycle (see Table 6.1), as it can be found in the KEGG [90] database. Further, all simulation runs are initialized in the same way as for the studies in the previous chapter, i.e. the full set of chemical reactions

KEGG ID	SMILES/Formula	NAME
C00022	CC(=O)C(O)=O	Pyruvate
C00122	OC(=O)C=CC(O)=O	Fumarate
C00036	OC(=O)CC(=O)C(O)=O	Oxaloacetate
C05379	OC(=O)CC(C(O)=O)C(=O)C(O)=O	Oxalosuccinate
C00024	C23H38N7O17P3S (SF)	Acetyl-CoA
C00149	OC(CC(O)=O)C(O)=O	(S)-Malate
C00311	OC(C(CC(O)=O)C(O)=O)C(O)=O	Isocitrate
C00417	OC(=O)CC(=CC(O)=O)C(O)=O	cis-Aconitate
C00042	OC(=O)CCC(O)=O	Succinate
C00158	OC(=O)CC(O)(CC(O)=O)C(O)=O	Citrate
C00068	C12H19N4O7P2S (SF)	Thiamin diphosphate
C00091	C25H40N7O19P3S (SF)	Succinyl-CoA
C00026	OC(=O)CCC(=O)C(O)=O	2-Oxoglutarate
C05381	C16H25N4O10P2S (SF)	3-Carboxy-1-hp-ThPP
C05125	C14H23N4O8P2S (SF)	2-Hydroxyethyl-ThPP
C00068	C12H19N4O7P2S (SF)	Thiamin diphosphate
C00074	OC(=O)C(=C)O[P](O)(O)=O	Phosphoenolpyruvate

Table 6.1: Metabolites of the citric acid cycle used as set of "food" molecules in the simulations of this chapter. Given are the KEGG ID, the SMILES notation and the name for all 17 metabolites. For the six largest molecules, their structural formula (SF) are shown instead of the SMILES notations, due to their length.

is available, the random genome length is 5000 bases, the TATA-box sequence is "UAUA" and the gene length is 100 bases.

The first scenario will be referred here to as the static scenario, because the set of "food" metabolites stays the same for the entire time of the simulation. The multiplication and duplication rates in this scenario are the same as in the previous simulations in Chapter 5. This scenario should allow for a steady evolution of chemical reaction networks with few perturbations. It will be interesting to see, whether both genetic and non-genetic robustness will evolve in this scenario. Modularity is not expected to arise under these conditions.

In the second scenario the set of input metabolites is not static anymore but will change after periods of 50 generations. It starts with 50 generations with the full "food" set, followed by eight periods in which a small number (3-4) of metabolites is taken away from the full set. The eliminated metabolites are reintroduced at the end of a 50 generation period and no metabolite is discarded for more than one period. This first phase of moderate changes (50-450) is followed by three periods (450-600) of strong perturbation. In these periods, only rather small "food" sets of six to eight metabolites remain. For the next period (600-650) the full "food" set is reintroduced, while in the subsequent period (650-700) all input molecules are

discarded at the same time. All of the described periods are repeated again (700-1400), but in this study mostly only the first 1000 generations will be considered and compared with the other two scenarios. The idea behind this scenario stems from the hypothesis that a biological system under varying environmental conditions will develop modularity. Furthermore, the effect of non-genetic perturbations on genetic robustness will be investigated.

The third and last scenario will incorporate another type of genetic operation besides mutation and duplication as in the first two scenarios, the horizontal gene transfer. Additionally, the mutation and duplication rate will be increased ten fold to the original rates in the previous simulations. The "food" set is again the full set as in the first scenario. Horizontal gene transfer is believed to be another source of modularity beside environmental change. Therefore, one interest in this scenario will be whether or not modularity emerges. Another uncertainty in the outcome of this scenario is the degree of genetic and non-genetic robustness. Will the increased rate of mutations make the system adapt to coping with mutational errors or will it weaken it to be overall less robust then systems evolved under steady conditions?

The analysis and comparison of all three scenarios can shed some light on the sources of robustness, modularity and its relatives such as flexibility and evolvability of biological systems, as well as the structure and evolution of the underlying reaction networks itself.

6.2 Network Analysis

The goal of general network analysis is to discover certain topological features in complex networks that can give valuable insights about their structure and dynamics. Many different measures exist that approach this goal from various angles [31]. Most of the work in network analysis was developed for random graphs [44], however, recent research focuses more and more on the application of real world networks, such as the world wide web [8], biological networks [9] or social networks. These networks usually diverge from random graphs in one or the other topological feature. As was shown by [172] most of the real world networks are also so-called small-world networks, they have a comparably small diameter such that the minimal distance between any two nodes is usually very low. Later, [7] analyzed the node-degree distribution and found that complex networks belong to the class of scale-free networks with the majority of nodes connected only to few other nodes and some highly connected nodes called hubs.

Network analysis is able to give a better picture of complex networks, regarding their structure and properties. In the following paragraphs some network measures, such as the connectivity distribution, spectral graphs and the clustering coefficient will be introduced and applied to the evolved networks from the simulations described above. The results will be discussed in terms of their expressiveness about system properties, with the focus on robustness and modularity.

6.2.1 Connectivity Distribution

As mentioned above, [7] used the connectivity distribution to classify complex real world networks and compare them with random networks. It was suggested that these biological or social networks belong to the class of scale-free networks differing from random networks that show normally or exponentially distributed connectivities. Interestingly for this study, scale-free networks exhibit a much higher robustness than the normally distributed or exponential random networks. The high degree of robustness can be accounted to the specific structure of scale-free networks, where most nodes of the network are only connected to one or few other nodes and only a few highly connected nodes connect the entire network. Consequently, in scale-free networks the random removal of one node will have only a minute impact. Although, the knock-out of one of the hubs is likely to disturb large parts of the networks structure and dynamics, this is more than compensated with the very low probability of this event due to the low frequency of hubs compared to the much higher probability of removing a node connected only to few other nodes and thus having almost no negative effect.

In Figure 6.1 the connectivity distribution for the three described scenarios averaged over all simulations are shown and compared with the average connectivity distribution of full pathway maps from the KEGG database from ten different organisms (see Table 6.2.2). Scale-free distributions can be represented by power law graphs defined through the polynomial of the form $k^{-\lambda}$. For this purpose, k is the node degree, λ is a constant and $k^{-\lambda}$ is the frequency of nodes with node degree k. Here, two power law graphs (with $\lambda = 1,\ 2$)are depicted for further orientation, as well as an exponential function (e^{-k}). For the real data (KEGG pathway maps), the connectivity distribution of the networks follows approximately the power law with exponent $\lambda = 2$, as expected. The networks from the first (Static) and third scenario (HGT/Mutation) show a very similar distribution, while those for the second scenario (Changing) also approximate a power law but more closely one with exponent $\lambda = 1$.

The first conclusion that can be drawn from these results is that the simulations under certain conditions (Static and HGT) lead to network structures similar to real-world metabolic networks. Interestingly, the mutation rate per generation does not affect the overall connectivity of the network, since the connectivity distribution from both scenarios are highly similar for all node degrees. On the other hand, the change in the constitution of the environment or "food" set appears to have some (negative) impact on the connectivity of a network, while still resulting in a realistic network structure, in the sense that it is scale-free. The main difference in the second scenario is the higher number of highly connected nodes (node degree > 80) but also nodes with ten to twenty connections. A possible explanation could be the emergence of modularity and the subsequent building of interconnections between the modules. While these interconnections might be a necessity in order to switch more flexibly between modules, it decreases the robustness of the overall structure.

In the Figures 6.2-6.4, the change of the connectivity distribution throughout the simulation

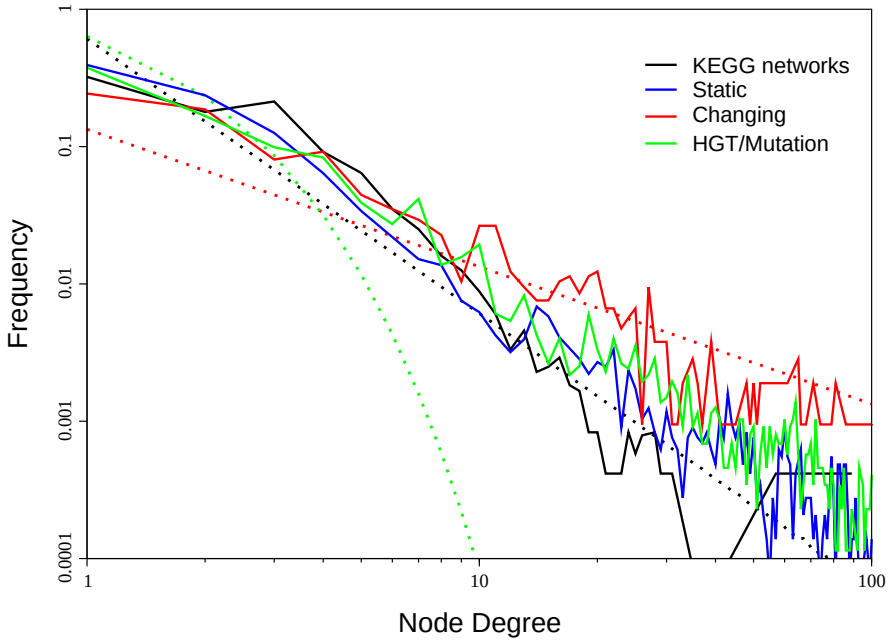


Figure 6.1: Connectivity distribution for all three scenarios compared with real-world metabolic networks from the KEGG database. Shown is the frequency of metabolites with a node-degree up to one hundred. The black line here represents the expected distribution for real metabolic networks. The static scenario (blue line) comes closest, while the changing scenario (red line) diverges most of all three scenarios. The scenario with increased mutation rate and HGT (green line) is similar to the static scenario. The dotted lines represent power law (black, red) and exponential distributions (green). The changing scenario is more similar to the power law with exponent one (red dotted line) while the other distributions are similar to the power law with exponent two (black dotted line).

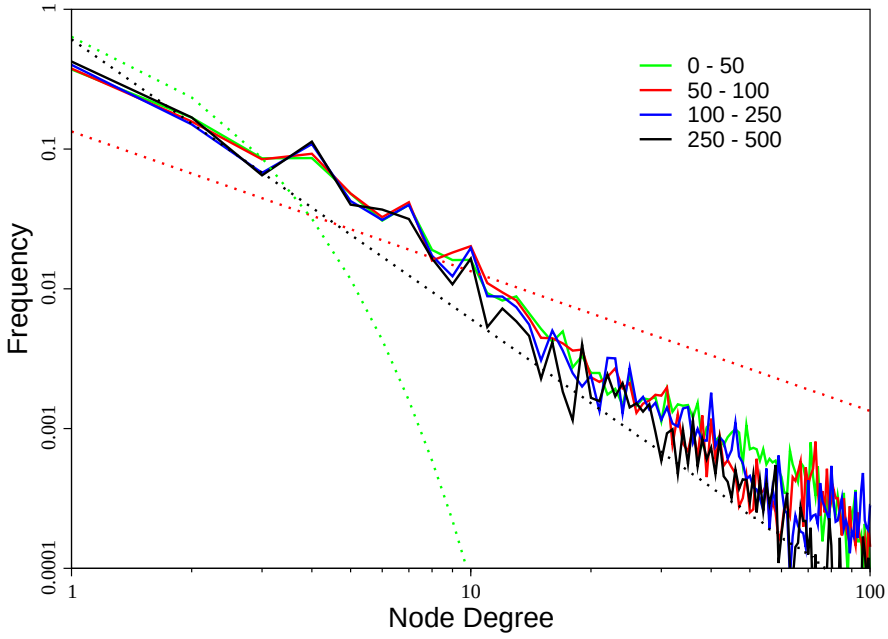


Figure 6.2: Connectivity distribution for the networks of the static scenario in four phases of the simulations. The distributions for all phases are highly similar and similar to a power law distribution with exponent two (black dotted line).

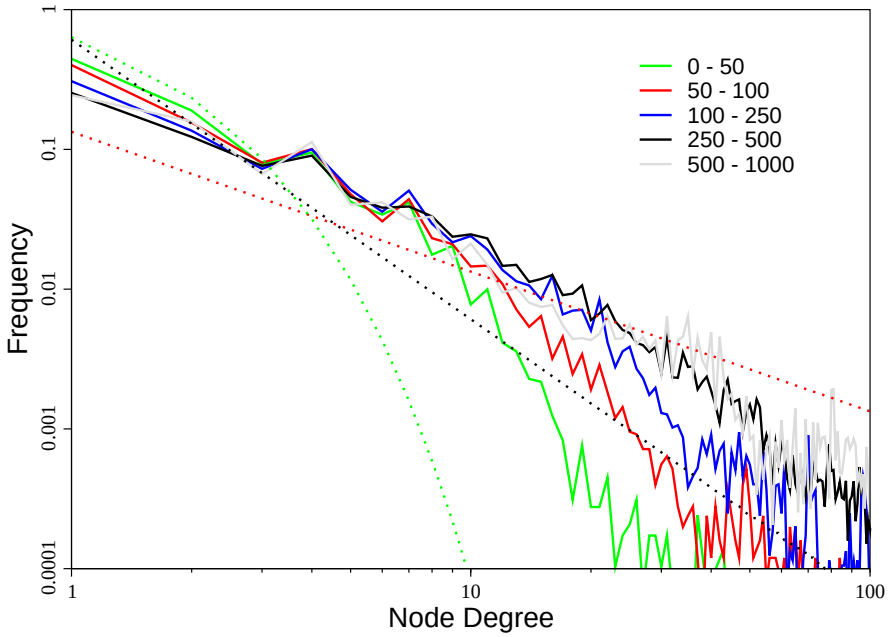


Figure 6.3: Connectivity distribution for the networks of the changing scenario and five phases of the simulations. The distribution of the earliest phase (green line) corresponds more to an exponential distribution. The distributions from the later phases more and more resemble power law distributions.

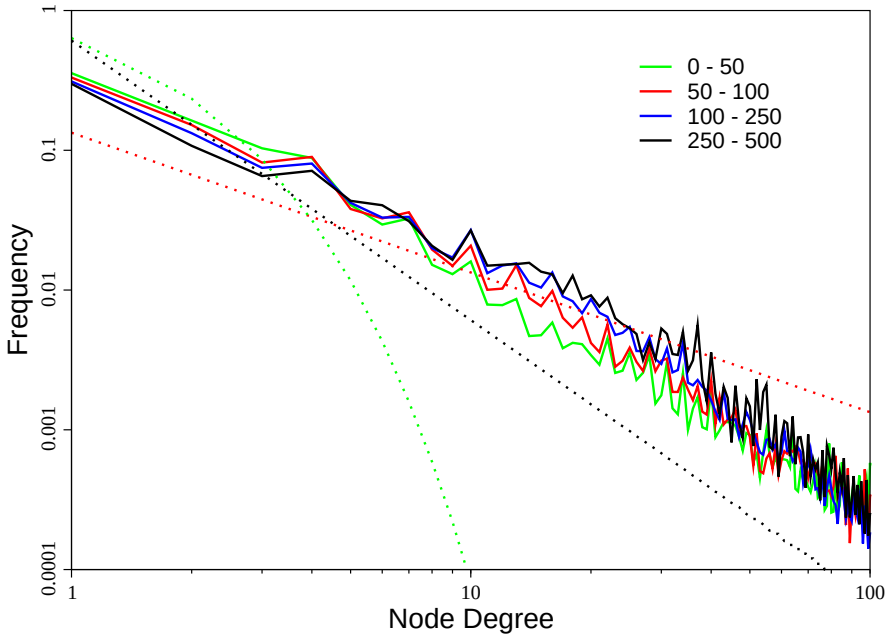


Figure 6.4: Connectivity distribution for the networks of the HGT/MUT scenario and four phases of the simulations. Similarly to the static scenario, the distributions of all phases are very similar and there is no significant change throughout the network evolution.

is shown on snapshots from four generations (50, 100, 250, 500). The development of the first and third scenario are rather minute. In both cases, the overall distribution is reached already in the earliest generations and does not change notably in the further evolution. In contrast, the networks in the second scenario undergo considerable change in terms of their connectivity distribution. In the earlier generations, highly connected hub metabolites are significantly underrepresented. In the course of evolution this number is steadily increasing and at the last stages even at a higher level than in the other two scenarios. Possibly, in this scenario modules are formed first, followed only later by a drive to connect these modules among each other. This would consequently lead to the increased number of highly connected nodes.

All three scenarios evolve scale-free networks, resembling real world complex networks. Therefore, they should also exhibit similar behavior in case of perturbations in the network structure. Since scale-free networks, as mentioned above, are particular robust this is then also true for the networks of the three scenarios. For the second scenario, one distinction has to be made. Although, the networks evolve to a more favorable constitution in terms of their connectivity distribution, due to the high number of highly connected nodes these networks are less robust against knockouts then networks from the other two scenarios or real-world networks.

6.2.2 Clustering Coefficient

It has been shown that biological networks contain many small tightly connected groups. The number of these clusters is higher than in random networks [172] and even more interestingly scale free networks. This observation is explained by the hierarchical modularity that underlies most biological systems. The clustering coefficient that will be discussed here is a measure for the extent of clustering a network possesses. The average clustering coefficient (see Equation 6.2) can be used as and indicator of modularity in general, while the local clustering coefficient (see Equation 6.1) may serve as proof for the real-world typical hierarchical modularity. Biological networks have a high average clustering coefficient independent of the network size and the local clustering coefficient scales against the node degree with approximately $d(n)^{-1}$ [119]. In this section, both measures will be applied to the simulated networks of all three scenarios to look for signs of modularity and hierarchically modular organization.

In a graph or network $G = \{N, E\}$, the local clustering coefficient cc_i of a node n_i is the ratio between the actual number of connections (C_{A_i}) between all nodes that are adjacent to this node (A_i) and the possible number of connections between these adjacent nodes.

$$\begin{aligned} cc_i &= \frac{|C_{A_i}|}{|A_i|(|A_i|-1)} \\ A_i &= \{n_j : (i,j) \in E\} \\ C_{A_i} &= \{(j,k) : (j,k) \in E \wedge j,k \in A_i\} \end{aligned} \tag{6.1}$$

KEGG Code	Organism	Size $\|N\|$	CC
ath	Arabidopsis thaliana	2068	0.273229
cel	Caenorhabditis elegans	1748	0.270794
cyt	Cyanothece sp.	2003	0.277287
dme	Drosophila melanogaster	1738	0.271532
dre	Danio rerio	1874	0.275166
eco	Escherichia coli	2009	0.272218
eok	Escherichia coli (EPEC)	1991	0.271731
hsa	Homo sapiens	1835	0.277645
mmu	Mus musculus	1835	0.277645
sce	Saccharomyces cerevisiae	1578	0.26941

Table 6.2: Ten organisms that serve as real-world metabolisms for the comparison with the simulated networks. Here their KEGG Organism Code, name, number of metabolites ($|N|$) and the clustering coefficient (CC) of their metabolic networks are given.

The average clustering coefficient CC is the sum of the local clustering coefficients of all nodes in the network, divided by the number of nodes.

$$CC = \frac{1}{|N|} \sum_{i \in N} cc_i \tag{6.2}$$

In Figure 6.5 the average clustering coefficients of the networks from the simulations is depicted in dependency to the network size, compared with metabolic networks from the KEGG database (see Table 6.2.2). Those real-world networks are much bigger than the networks from the simulations. However, the assumption was that for biological networks the average clustering coefficient is independent of the network size. In fact, the majority of the simulated networks has a clustering coefficient similar to the value (≈ 0.27) of the KEGG networks. Therefore, the degree of modularity and the structure of the networks from all three scenarios appears to correlate with that of actual metabolic networks, rather than random networks or single metabolic pathways.

When comparing the distributions of the three evolutionary scenarios, two main observations should be noted. First, the clustering coefficient is mostly independent of the network size, ignoring the divergence in some scenarios, but for the largest networks the clustering coefficient is dramatically plunged in all scenarios. A possible explanation is that these networks stem from the last evolutionary stages and thus are not yet adapted in the same way as those from previous stages. The second observation is that only the first scenario (Static) is really constant over all network sizes, while the other two show some peculiarities. For the second scenario (Changing) the clustering coefficient increases with network size until the sharp decline. In the third scenario some small but highly clustered ($CC > 0.5$) networks are

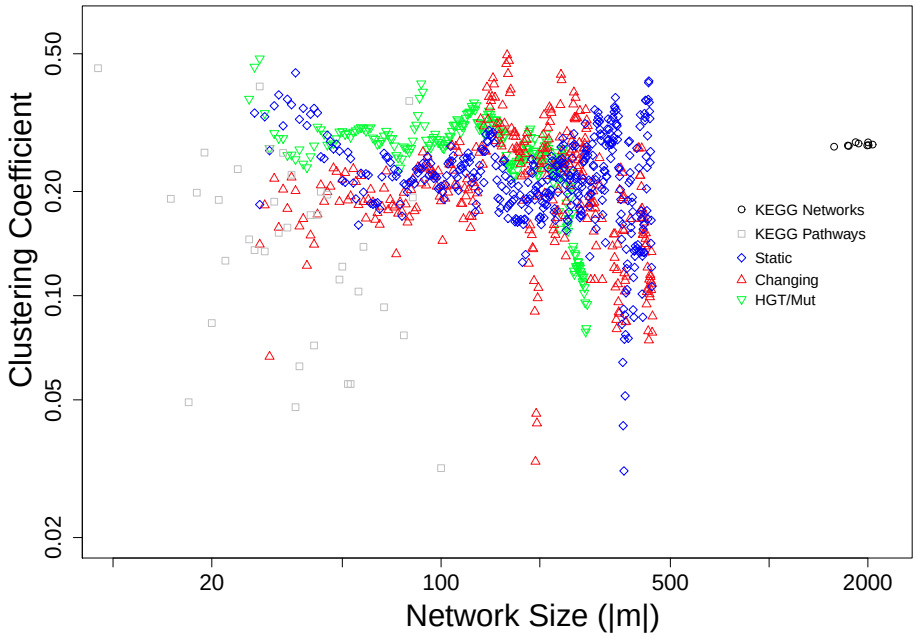


Figure 6.5: Clustering coefficient vs network size, for all three scenarios compared with real-world metabolic networks and single pathways. The average clustering coefficients of networks with the same number of metabolites are shown for the static (blue diamonds), changing (red upwards triangle) and HGT/MUT scenario (green downward triangle) as well as for entire pathway map (black circles) and single pathways (gray squares) from the KEGG database.

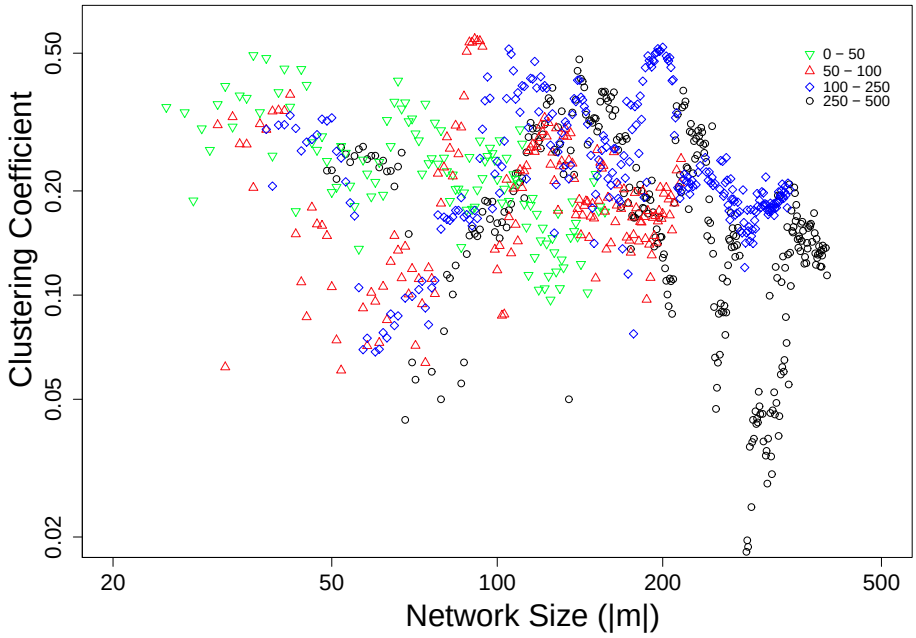


Figure 6.6: Clustering coefficient vs network size, for networks of the static scenario and four time intervals. Shown here are the clustering coefficients for the intervals of generation 0 to 50 (green), 50 to 100 (red), 100 to 250 (blue) and 250 to 500 (black).

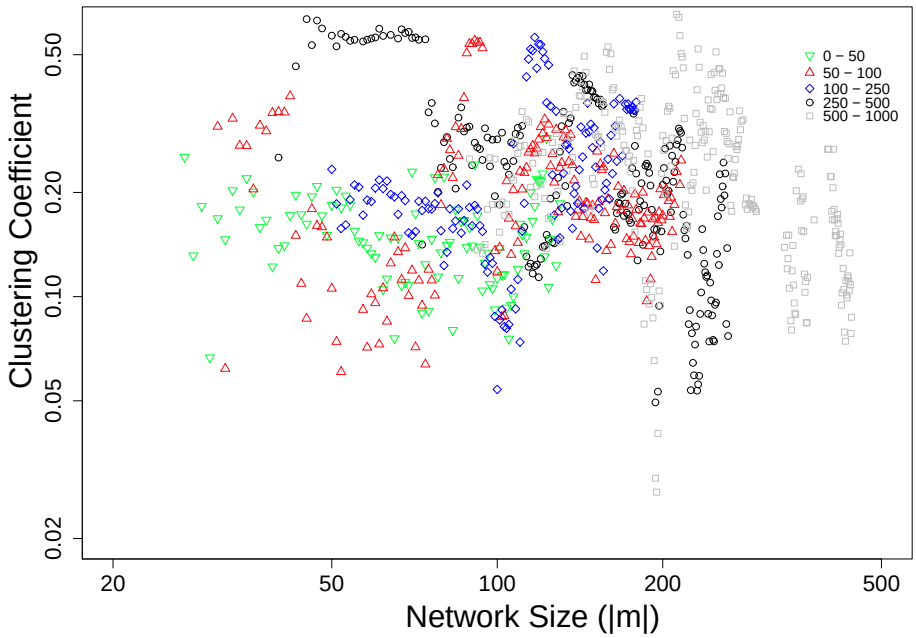


Figure 6.7: Clustering coefficient vs network size, for networks of the changing scenario and five time intervals. The clustering coefficients for the same intervals as above are shown plus the interval from generation 500 to 1000 (gray).

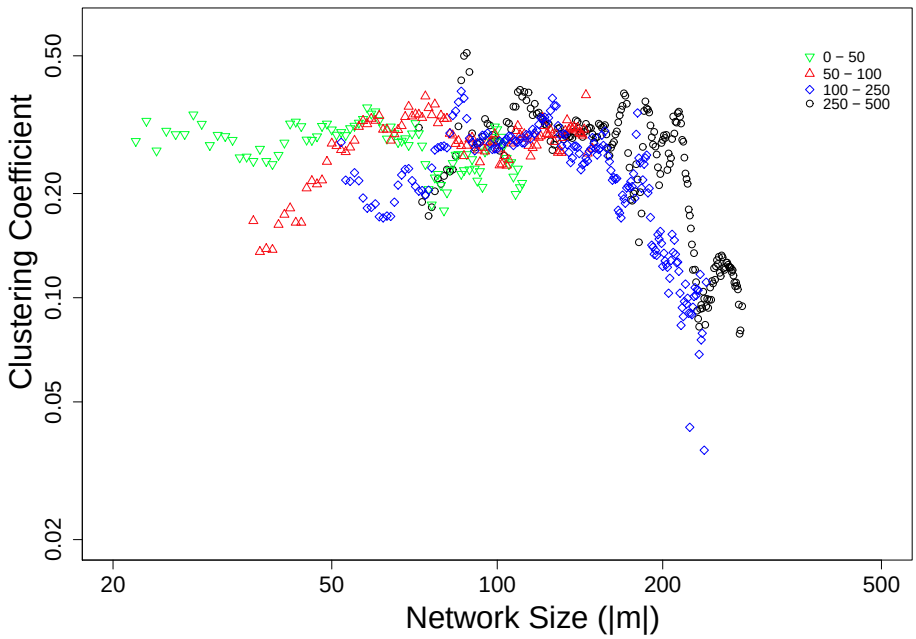


Figure 6.8: Clustering coefficient vs network size, for networks of the HGT/MUT scenario and four time intervals. Shown here are the clustering coefficients for the intervals of generation 0 to 50 (green), 50 to 100 (red), 100 to 250 (blue) and 250 to 500 (black).

formed. To explain these observation, the evolutionary steps of the simulations has to be inspected in more detail.

To this end, Figures 6.6-6.8 illustrate this evolutionary transition for the three scenarios, by depicting the clustering coefficient for several time steps of the simulation (generations 0 - 50, 50 - 100, 100 - 250, 250 - 500 for all and 500 - 1000 for the second scenario). In the first scenario (Static), in all phases there is a similar trend of slightly decreasing clustering coefficients, which indicates a slightly lower modularity and divergence from the structure of real-world networks. However, this trend becomes smaller in the later phases and might vanish for longer time spans. As discussed above, the large networks ($|N| > 250$) and significantly lower clustering coefficient ($CC < 0.1$) all stem from the latest phase. Figure 6.7 depicting the evolution of the second scenario (Changing), provides a clear explanation for the increasing clustering coefficient noted above. The average clustering coefficients in the earliest phase (0 - 50) are around 0.15 and thus lower than in the next phase (50 - 100; $CC \approx 0.2$) and much lower than in the latest phases with clustering coefficients around 0.3. Thus, evidently the increased clustering can be accounted to the evolution in this scenario. The main driving force in this scenario was the environmental change in the set of "food" metabolites. So it could be concluded that this change of the environment leads to an increased clustering in the underlying networks and a higher modularity in systems that evolved under such conditions. In the third scenario, even the early phases (0 - 50 and 50 - 100) exhibit high clustering coefficients, suggesting that an increased mutation rate and/or horizontal gene transfer can lead to a rather spontaneous increase in modularity. Nevertheless, there also appear to be limitations on the impact of these two factors. First of all, the clustering coefficient does not increase further with more mutations and gene transfer taking place throughout the evolution. Secondly, the proposed modular structure does not hold for many of the large networks. It is possible that too many mutations disrupt the overall structure to a less clustered organization. To investigate this and identify further possibilities, it is necessary to take a closer look at the structural organization of these networks in terms of their clustering/modularity.

Figure 6.9 provides this deeper look by showing the local clustering coefficients of nodes with the same node-degree. Again, in this overview the measure is shown for all three scenarios. For a comparison with real-world biological networks the clustering coefficient of full metabolic networks maps and single pathways from ten organisms out of the KEGG database is shown too. All network types, except the single pathways which are not full networks in that sense, share a similar distribution. Until a node degree of around 20, the local clustering coefficients increases with its connectivity, while after this point it decreases. Beginning from the turning point the clustering coefficient of the KEGG networks scales against the node degree as a function of the form $k^{-\lambda}$ as proposed above. The clustering coefficients of the second and third scenario behave in a similar way. Only the first scenario deviates more strongly from this scaling. Due to the lack of environmental change or horizontal gene transfer, the two most prominent potential sources of modularity, the extent of hierarchical modularity observed in real biological networks may not be reachable. In the second scenario, even nodes with higher

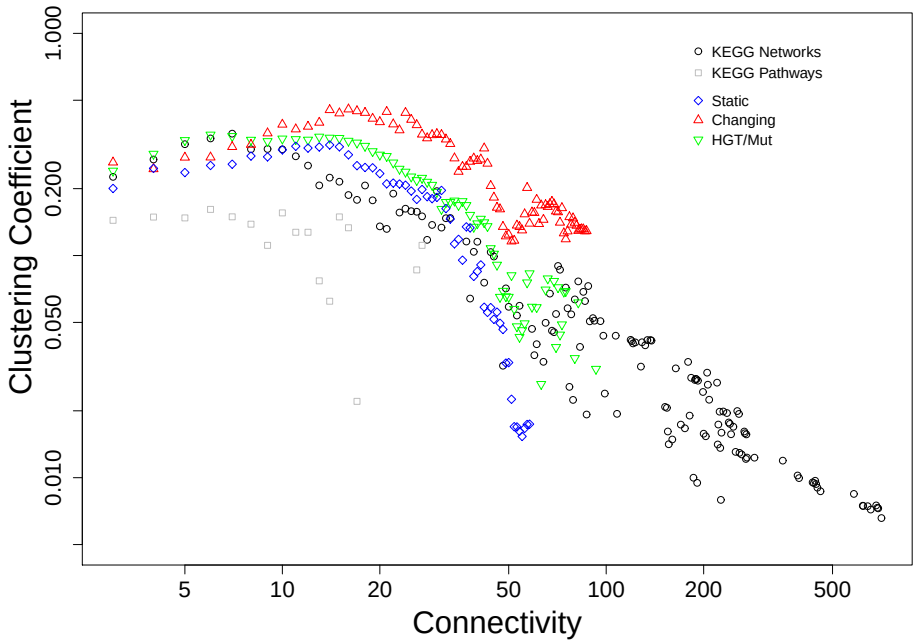


Figure 6.9: Local clustering coefficient vs node-degree for all three scenarios compared with metabolic networks and single pathways from KEGG. Real-world metabolic networks are known to exhibit hierarchical modularity which shows in the scaling of their metabolites' clustering coefficients against their node degree (black circles). The nodes of the simulated networks show a similar but slightly divergent scaling, in particular the static scenario (blue diamonds) misses highly connected nodes with high clustering coefficients.

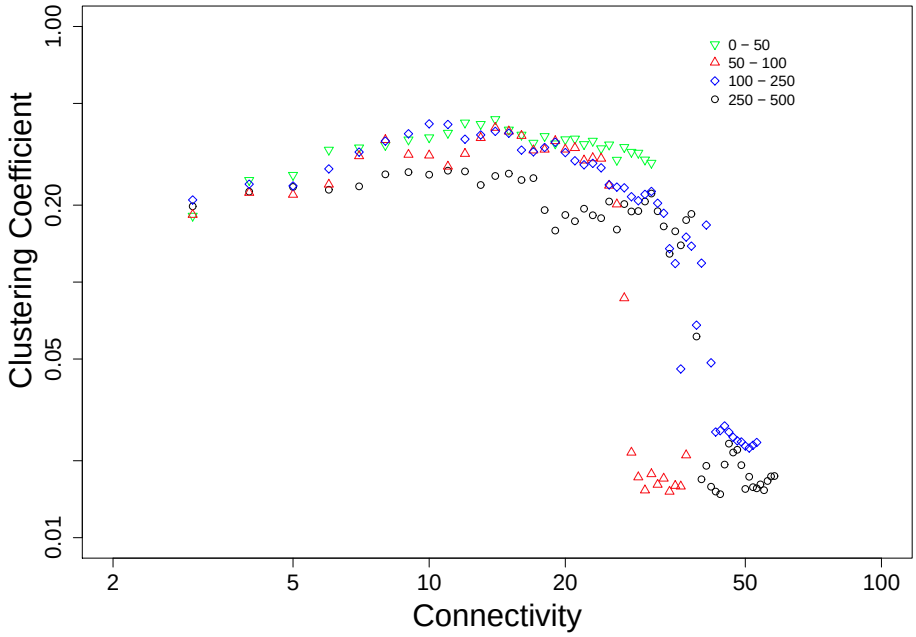


Figure 6.10: Local clustering coefficient vs node-degree for the metabolites of the simulated networks from the static scenario. Results are presented for the intervals from generation 0 to 50 (green), 50 to 100 (red), 100 to 250 (blue) and 250 to 500 (black).

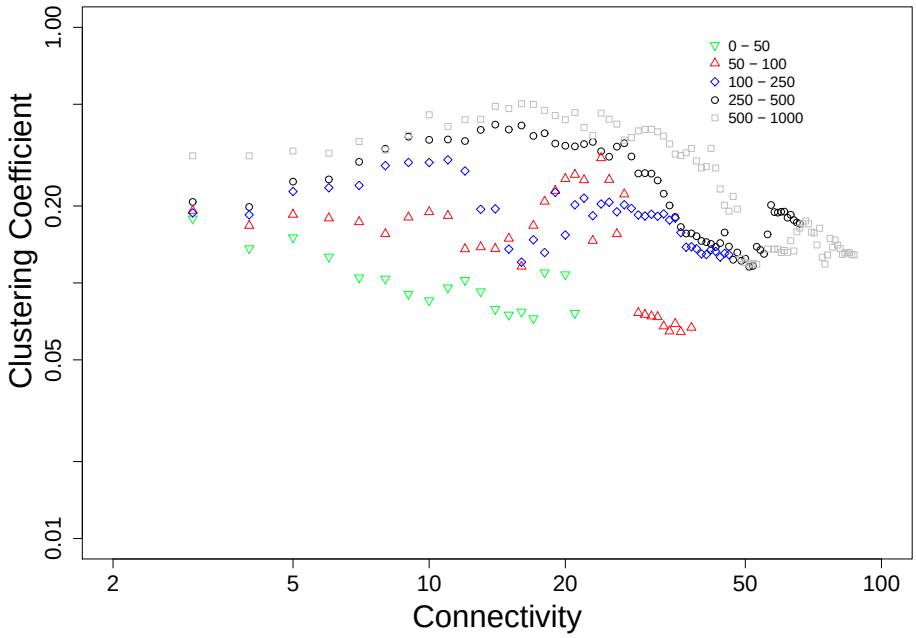


Figure 6.11: Local clustering coefficient vs node-degree for the metabolites of the simulated networks from the changing scenario. Results are presented for the intervals from generation 0 to 50 (green), 50 to 100 (red), 100 to 250 (blue), 250 to 500 (black) and 500 to 1000 (gray).

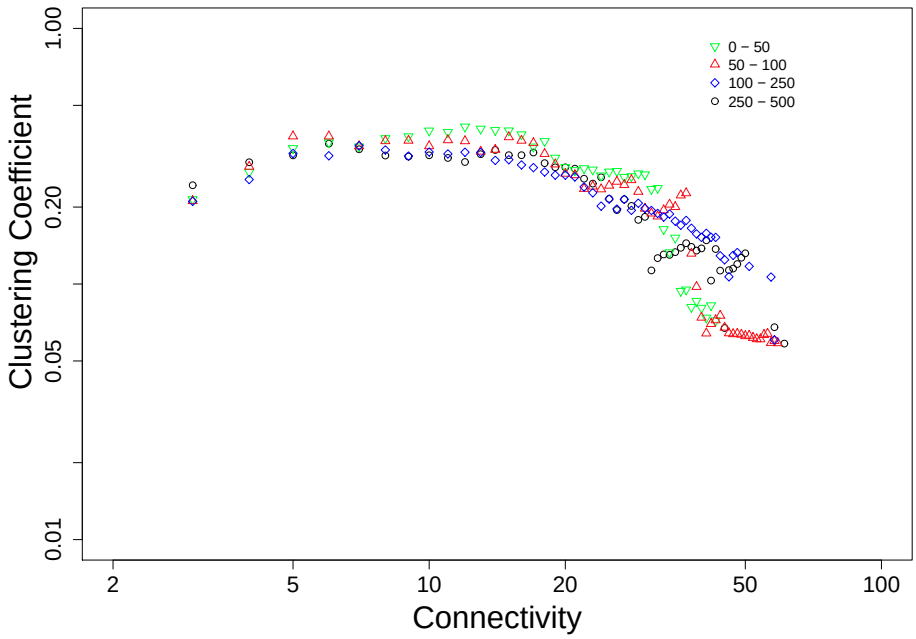


Figure 6.12: Local clustering coefficient vs node-degree for the metabolites of the simulated networks from the HGT/MUT scenario. Results are presented for the intervals from generation 0 to 50 (green), 50 to 100 (red), 100 to 250 (blue) and 250 to 500 (black).

connectivity have a relatively high clustering coefficient compared to the other scenarios and the KEGG networks. If the environmental change in this scenario causes its modularity then this might also have the side effect of adding many interconnections between the various modules in order to be able to switch between one another. This would not be necessary for modularity from horizontal gene transfer, here modules can be more independent of each other.

For a detailed evolutionary study of the relation between local clustering coefficient and node degree, the Figures 6.10-6.12 depict this relation for several time steps in the evolution of all three scenarios. All scenarios show remarkably different developments, while the first scenario starts off (0 - 50) with many highly clustered nodes ($cc_i > 0.35$) evolving to less clustered nodes ($cc_i < 0.25$), the opposite development can be observed in the second scenario, where highly clustered nodes only form late in the simulation history. Further, there is almost no change in the constitution of network nodes, with respect to the clustering coefficient, in the third scenario. Again, these differences have to be viewed in consideration of the forces that shape the evolution of the different scenarios. As above, horizontal gene transfer leads to a certain level of clustering and modularity in short time but does not increase significantly. Under environmental change, on the other hand, modularity increases steadily but slower. In the absence of any of the two factors, modularity is stagnating if not even decreasing.

6.2.3 Graph Spectrum

Another graph measure that can be used to analyze the properties of the studied metabolic networks is the graph spectrum, or better the spectrum of the Laplace operator δ of the network. Here the network is assumed to be a directed graph with nonnegative edge weights. This measure has been used to distinguish different types of biological networks and identify some of their structural characteristics and even evolutionary processes that lead to their formation [10]. It was shown for instance that in metabolic networks are more strongly connected components than for example gene regulatory networks, and some of these components could be defined (see Figure 6.13). Furthermore, the high number of non-cyclic components could be explained to a considerable extent by the duplication of nodes [5]. The main use in this work is the comparison of the overall structure between the simulated networks and real metabolic networks as well as the identification of missing (or additional) complexity and network components in those simulated networks.

The Laplacian L is a representation of a graph ($G(V, E)$) in matrix form, combining the degree and adjacency information of this graph. It can, thus be written as $L = D - A$, where D is the degree matrix and A the adjacency matrix of the graph. The degree matrix contains all node degrees in its diagonal and is otherwise empty. The adjacency matrix holds the information about connections between nodes, a connection between a pair of nodes (i, j) is represented by a nonzero entry in the adjacency matrix at row i and column j (for undirected

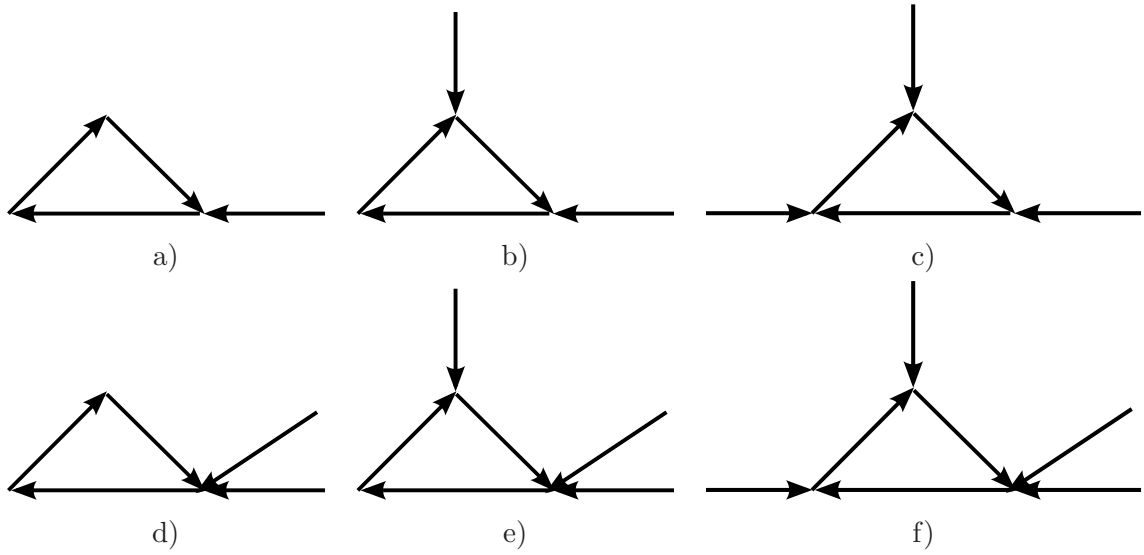


Figure 6.13: Examples for strongly connected components that contribute to the spectrum with eigenvalues other than 0 or 1. Arrows are reactions and point to metabolite nodes. The corresponding eigenvalues to the subnetworks are a) $0.21, 1.40 \pm 0.87i$; b) $0.37, 1.32 \pm 0.55i$; c) $0.5, 1.25 \pm 0.43i$; d) $0.31, 1.35 \pm 0.60i$; e) $0.45, 1.28 \pm 0.47i$; f) $0.56, 1.22 \pm 0.38i$. Adapted from [10]

graphs also the reverse, row j and column i).

$$L(i,j) = \begin{cases} deg(i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } (i,j) \in E \\ 0 & \text{else} \end{cases} \tag{6.3}$$

The spectrum is now the set of eigenvalues of the network's laplacian matrix. Here three types of eigenvalues are of interest, 0 eigenvalues, 1 eigenvalues and eigenvalues different from 0 and 1. All types represent different network structures. O eigenvalues indicate isolated strongly connected components. 1 eigenvalues in the spectrum stand for non-isolated non-cyclic nodes in the network. Eigenvalues that are unequal to 0 and 1 result from non-isolated but cyclic nodes, some instances are strongly connected components such as in Figure 6.13. For a more detailed description of the analysis spectra from directed and undirected networks in general and its application to biological networks in particular see [10].

In Figure 6.14 spectra of four networks are depicted. The first three spectra stem from example networks of the three evolutionary scenarios after 500 generations, the last spectrum is that of the *E. coli* network from the KEGG database. Several full network maps of different organisms were tested, but only one is shown here because all spectra were highly similar. At first glance, also the four spectra in Figure 6.14 appear to be very similar. While this is

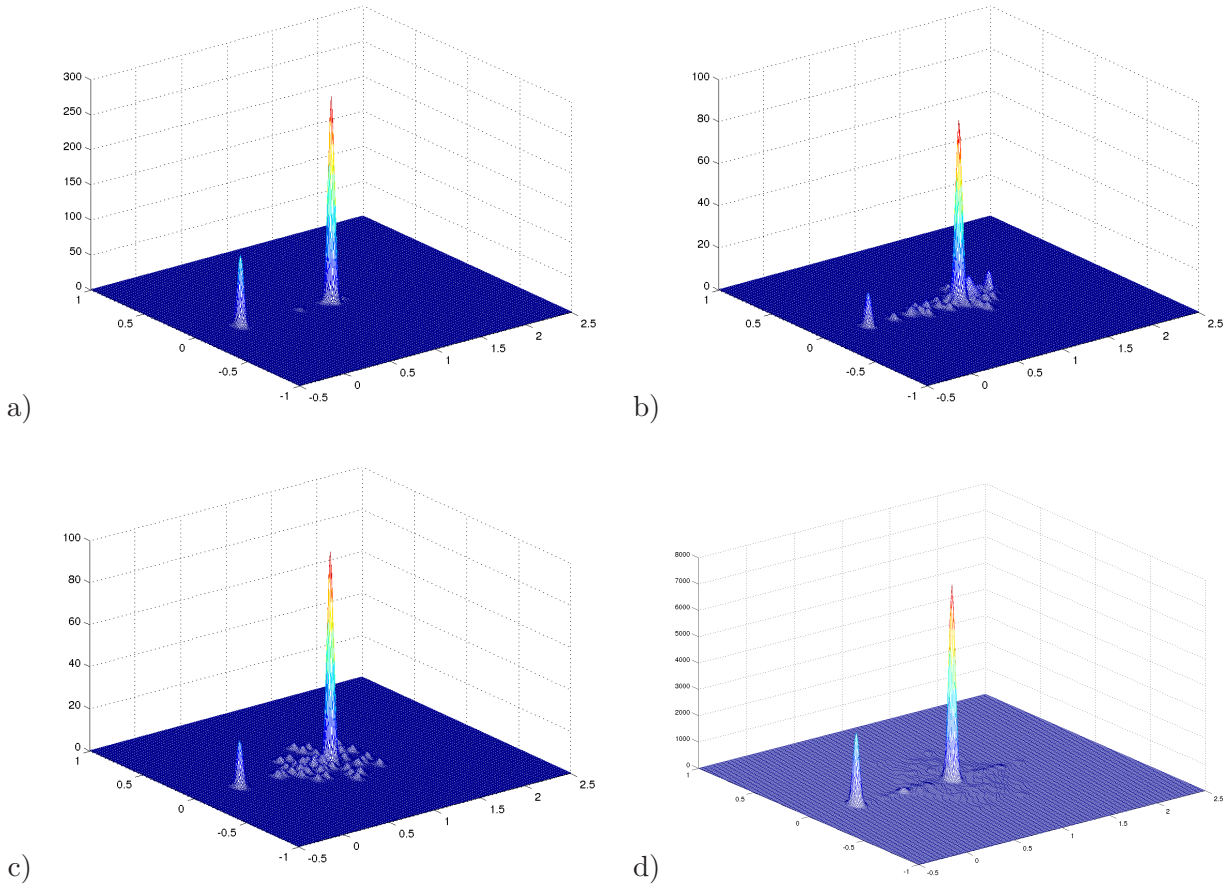


Figure 6.14: Spectra of example network graphs from all three scenarios compared with a real-world metabolic network. The spectra are three-dimensional eigenvalue distributions, where the x-axis is the real-part of the eigenvalue, the y-axis the imaginary part and the z-axis is the frequency of this eigenvalue in the network. d) The *E. coli* network shows the most diverse eigenvalues, whereas the network from a) the static scenario shows a restricted repertoire of eigenvalues, those of the b) changing and c) HGT/MUT show parts of the eigenvalues.

true for the multiplicity of 0 and 1 eigenvalues, the distributions of the eigenvalues different from 0 and 1 are rather divergent in all four cases. The spectrum of the metabolic network of *E. coli* contains several eigenvalues widely spread around the 1 eigenvalue and an increased number of eigenvalues with an imaginary part of 0. The spectrum of the example network of the third scenario shares the distribution of eigenvalues around the one eigenvalue, while the network from the second scenario has a similar accumulation of eigenvalues with imaginary part 0 as the *E. coli* network. The network from the simulation of the first scenario does not show a considerable multiplicity of eigenvalues other than 0 and 1 eigenvalues. It appears that in all three scenarios certain network structures are missing or are underrepresented compared with the real-world metabolic network, however, this might also be due to the

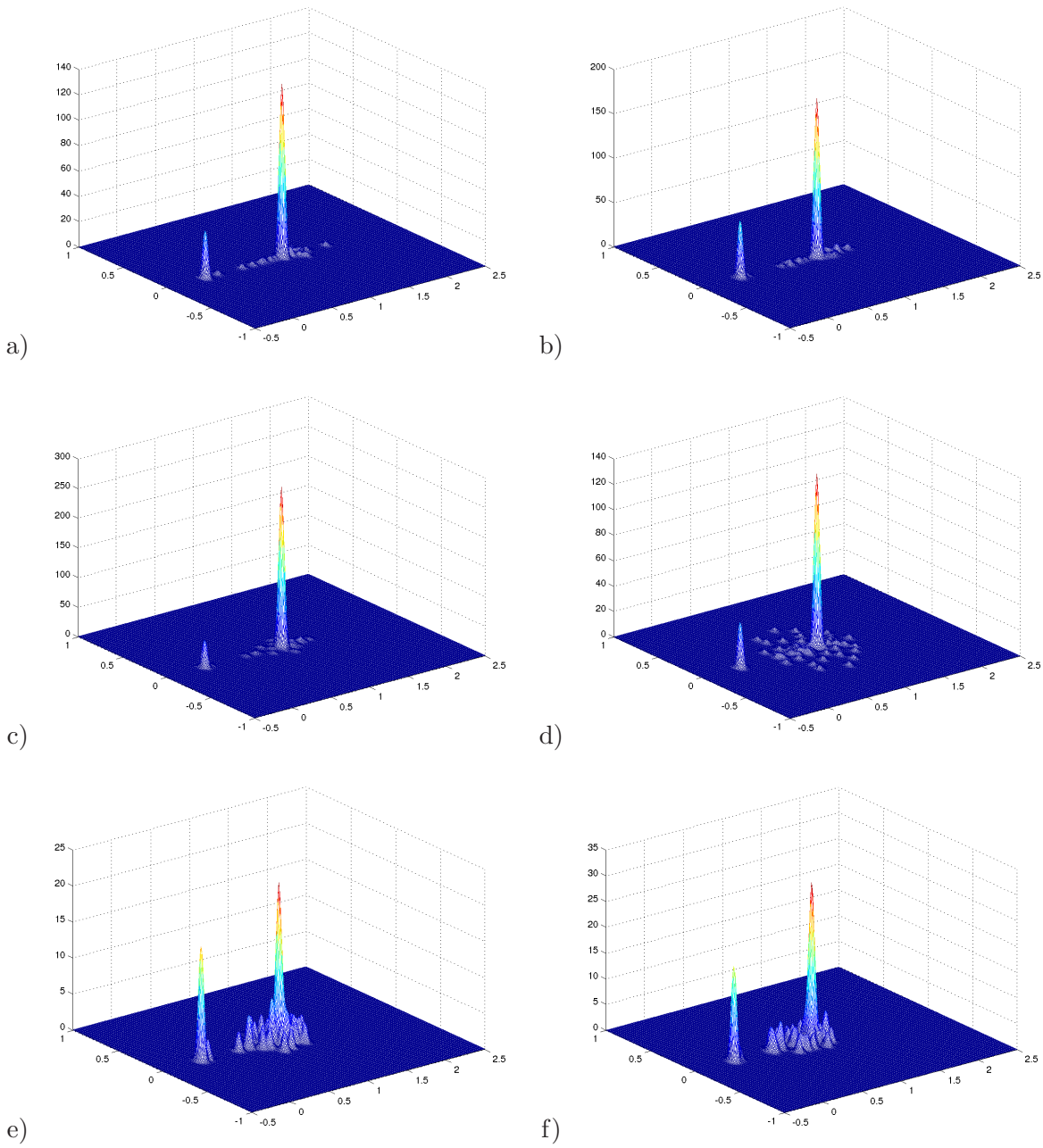


Figure 6.15: Spectra of example network graphs from the early and late phases of all three scenarios. The spectra are three-dimensional eigenvalue distributions, where the x-axis is the real-part of the eigenvalue, the y-axis the imaginary part and the z-axis is the frequency of this eigenvalue in the network. On the left side, networks after 100 generations are shown, for a) the static, c) changing and e) HGT/MUT scenario. Compared to the right side, with networks after 250 generations, for b) the static, d) changing and f) HGT/MUT scenario. In all three scenarios, the peaks for the 0 and 1 eigenvalues get more dominant. The change of the eigenvalues other than 0 or 1 is differs in the scenarios.

smaller size of these networks. Nevertheless, it can be speculated about the types of network components that are more or less abundant. The higher mutation/duplication rate and horizontal gene transfer possibly leads to a multiplicity of duplications of nodes or bigger structures in the network that is higher than in the other two scenarios. Consequently, those eigenvalues that are more abundantly present in the third scenario might be causes of such duplication and gene transfer processes. In the previous paragraphs on the connectivity and clustering of the simulated metabolic networks, differences in the structure of networks from the second scenario were discussed. The overrepresentation of highly connected nodes and highly interconnected modules then possibly explains the increased number of eigenvalues with imaginary part 0, that were observed specifically in the spectrum of the second scenario.

For all three scenarios one network from its early phase (≈ 100) and one from a later phase (≈ 250) is shown with its spectrum in Figure 6.15 and one from the latest stage (500) in Figure 6.14. The change of the spectra of the three scenarios differ on the first glance, but have one similarity, the dominance of the 0 and 1 eigenvalues becomes stronger in later generations, which can be seen by the increase of the peaks at these positions, while the frequencies for the other eigenvalues decreases. The differences in the changes of the spectra, are the occurrences of eigenvalues other than 0 and 1. In the early phase of the first scenario, still several eigenvalues have considerably high frequencies, while almost disappearing in the later stage. In the beginning of the second scenario the number and variety of eigenvalues other than 0 and 1 is increasing, while later loosing in variety. Some of the eigenvalues remain from the middle stage, but their frequency is less equally distributed then before. The trend is that a few eigenvalues are established suppressing the others and thus decreasing the variety of eigenvalues. This indicates that certain modules become dominant through duplication or convergent evolution. An almost opposite effect can be observed for the third scenario. The variety of eigenvalues other than 0 and 1 is actually increasing in the later stage to the cost of a few frequent eigenvalues from the earlier stage. The high mutation rate disrupting also the network structure and the steady insertion of new modules through the horizontal gene transfer in this scenario might be the cause for this development.

6.3 Steady-State-based Approach

In the paragraphs of the previous section, predictions about the system properties of metabolic networks were made based on the structural and topological features derived from general network analysis methods. Although this has revealed interesting characteristics of the evolution of the networks, a more-in-depth study has to consider also the dynamics and functionality of the studied systems. To this end, concepts specific for chemical reaction systems will be applied for the further analysis. The concepts that are used in the next paragraphs are all based on the steady-states of a system. The focus here is on metabolic pathways and chemical organizations.

6.3.1 Elementary Modes

Metabolic pathways are concepts special to metabolic systems, that can be seen as functional subunits, i.e. parts of the overall system that can perform independently in steady-state, exhibiting some functionality such as the conversion of metabolites from the environment to vital parts of biomass, or the provision and preservation of important inner metabolites. From that, the importance of this concept can be seen and the need for analyzing the space of potential states of the cell, the set of all possible pathways through its metabolic network, becomes apparent. For instance, if two metabolic systems of equal size and with similar structural features are compared in terms of their flexibility, it is fair to say that the system with the higher number of metabolic pathways will in average be the more flexible one, while this will not be true for all cases, though. Nevertheless, the idea is that a system with many metabolic pathways has also many alternative routes allowing a better response to any perturbations. Furthermore, predictions about the robustness of a system can be made by observing the change of the set of pathways from the original unperturbed network to networks in which single or multiple enzymes are knocked out at the same time or networks in which some of the metabolites are depleted. If the system can maintain the majority of pathways despite perturbations then this indicates a high robustness.

Since it is impossible to enumerate all metabolic pathways and analyze the entire space of possible routes through the network, it is better to focus on a specific set of pathways that can give valuable insights to the overall flux distribution. Elementary modes, as discussed in Section 3.2.3 are such a set of pathways that can generate all other pathways through linear combination. The set of elementary modes has been computed for many metabolic networks and used for their analysis [147, 60, 35]. In particular, the work of Behre [11] has addressed the above mentioned suggestion that the number of pathways can tell us something about the properties of a metabolic system. The main goal in this work is to define a measure of robustness based on the change in the set of elementary modes in cases of network changes. They introduce several such measures, four of which will be explained briefly in the following paragraphs (Equation 6.4-6.7) and applied on the networks of the three simulation scenarios.

The first measure considers the robustness towards single knockouts of enzymes. Therefore, the ratio of the number of elementary modes of a network with one removed enzyme (z^i) and the number of pathways before the knockout (z) is calculated. The ratios for all enzyme knockout events are then summed up and averaged over the number of enzymes ($|E|$).

$$R_1 = \frac{1}{|E|z} \sum_{i=1}^{|E|} z^i \tag{6.4}$$

Similarly, the second measure defines the robustness against metabolite depletion as the sum of the ratios for all possible metabolite depletions divided by the number of metabolites ($|M|$).

$$R_2 = \frac{1}{|M|z} \sum_{i=1}^{|M|} z^i \tag{6.5}$$

The next measure goes a bit further then the first by considering multiple knockouts. The ratios therefore have to be calculated for all possible combinations of knockouts ($s(k)$), given a certain number of simultaneous knockouts (k).

$$R_3(k) = \frac{1}{s(k)z} \sum_{i=1}^{s(k)} z^i \tag{6.6}$$

The last measure is one for the overall knockout robustness, simply the combination of the multiple knockout robustness for all sizes below and equal to a certain threshold (K).

$$R_4(\leq K) = \sum_{k=1}^{K} R_3(k) p_k \tag{6.7}$$

In Table 6.3.1 the robustness measures are shown for some example networks to illustrate their differences. For the simple robustness measure $R1$ the number of elementary modes is still a good indicator, networks with a higher number are more robust and mostly those with the same number have a similar robustness. This must not be true at all anymore for double knockouts or robustness against depletion.

The development of the robustness measures (R1, R2, R3(2)) for the entire history of the three evolutionary scenarios are shown in Figure 6.16a)-c). The three scenarios, show a high degree of all the robustness measures and a similar trend towards more robust states. The development in the first and third scenario are somewhat stronger in the first generations, whereas the second scenario has the most constant development resulting in the highest robustness values of all three scenarios at the end. The former observation can be confirmed by the previous robustness analysis but the latter finding comes as a little surprise because so far the networks in the second scenario were believed to exhibit a lower degree of robustness. However, here the functional robustness towards rather small perturbations is observed. This type of robustness seems to profit particularly from modularity which is present at a higher degree in the networks of the second scenario due to the changing environment throughout the evolution of the simulations. For the structural robustness that was studied in the previous paragraphs, this appears to play a smaller role. Whether it also affects the functional robustness of larger perturbations will be investigated in the next paragraph concerning the minimum knockout set size distribution.

The third scenario exceeds the first scenario in the increase of robustness in the first generations and the second scenario reaches a higher level of robustness throughout evolution. It appears that changes either in the enzyme or metabolite set are beneficial for the evolution of robust systems. Although, the first scenario shows the most constant development without

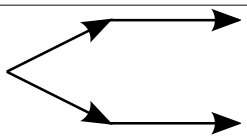
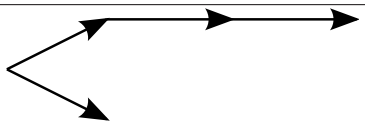
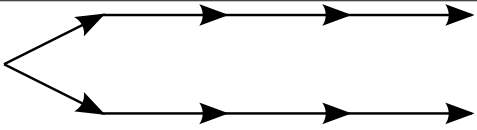
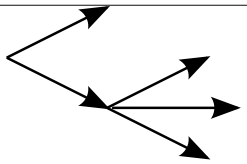
Networks	$\|r\|$	$\|m\|$	$\|EM\|$	R1	R2	R3(2)
	4	5	2	4/8	2/12	4/10
	4	5	2	4/8	3/12	4/10
	8	9	2	8/16	6/28	8/18
	5	6	4	13/20	15/40	13/24

Table 6.3: Elementary modes based robustness measures for four example networks. Reactions are represented as arrows in the graphs, metabolites are at the endpoints of these arrows. For the networks, the number of reactions ($|r|$), metabolites ($|m|$) and elementary modes is given as well as the robustness measures R1 for a single knockout, R2 for double knockouts and R3 for metabolite depletion. Adapted from [11]

drastic setbacks as in the other two scenarios. It should be noted, however, that this concerns mainly the robustness against small perturbations and it remains to be shown for overall robustness.

Another observation is that the average damage of a double knockout in the simulations is almost exactly twice that of a single knockout, which means that there are basically no synergetic effects of simultaneous knockouts of two enzymes. Furthermore, the damage due to a removal of one metabolite closely corresponds to that of a parallel removal of two reactions.

6.3.2 Minimal Knockout Sets

For an even more detailed analysis of the robustness of metabolic networks against multiple knockouts of enzymes, the minimal knockout set size distribution is a very useful measure. This concept of knockout sets is a related but opposing concept to the set of elementary modes, as described in Section 3.2.3. It is defined as a set of enzymes or reactions that are sufficient to block a certain target function (reaction) if knocked out simultaneously. It is now easy to see that the size of these knockout sets does provide some insights in the capabilities of the network to compensate for knockouts. If for instance, a vital function such as biomass production of a system has in the majority small knockout sets, let's say

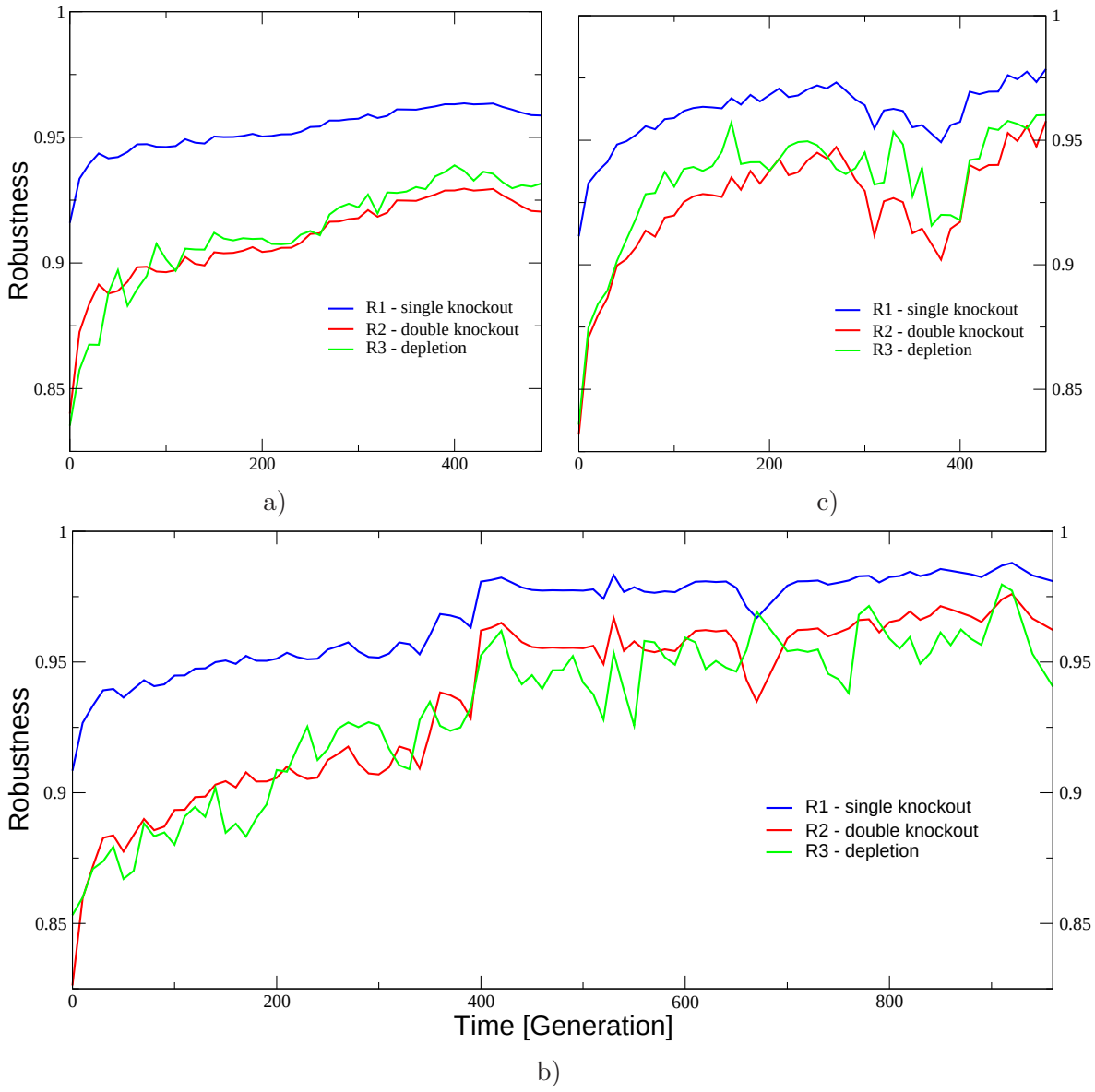


Figure 6.16: Elementary modes based robustness measures for the networks of the three scenarios. The averages of the robustness measures R1 for a single knockout (blue), R2 for double knockouts (red) and R3 for metabolite depletion (green) are depicted for the networks of a) the static scenario, b) the changing scenario and c) the HGT/MUT scenario, for all generations of the simulations.

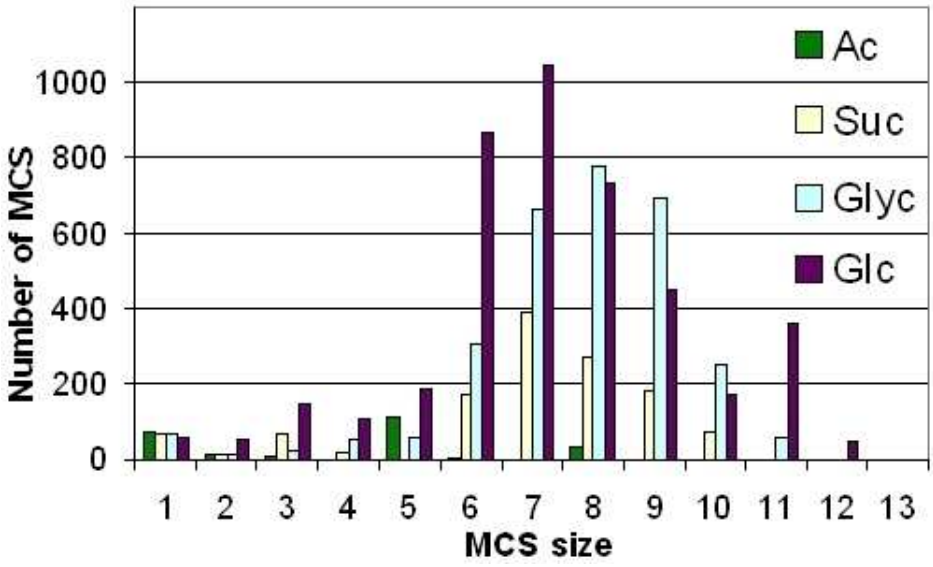


Figure 6.17: The cutset size distribution for four subnetworks of *E. coli*. The distributions are shown for the subnetworks of the *E. coli* metabolism under growth of acetate (green), succinate (yellow), glycerin (cyan) and glucose (purple). The subnetworks increase in complexity and robustness in this order. From [99].

of size one, then the knockout of only one of these single enzyme results in the blocking of the biomass production, having a possibly lethal consequence. Such a system would be considered rather fragile concerning knockouts. On the other hand, a minimum knockout set size distribution with mostly very large knockout sets, would be able to accept several simultaneous knockouts of enzymes. Only if all enzymes of such a large knockout set are knocked out through mutation or other perturbations, the target function is blocked, a rather unlikely event. Consequently, such a system can be said to be fairly robust against multiple knockouts.

The concept was introduced by [99] as so called cutsets, which they used for target and phenotype prediction as well as a measure for network fragility. They tested four different subnetworks of the *E .coli* metabolic network and compared the respective size distributions, see Figure 6.17. The example networks consist of certain parts of the central metabolism of *E .coli* that are related to growth on one of the four media: acetate, succinate, glycerin, glucose. The growth on glucose is clearly less fragile than acetate and also the other two media. These results are in agreement with a previous study based on the elementary modes of these subnetworks [151]. In a later work [86], cutsets were also used to prove epistasis between subnetworks of *E .coli* metabolism.

In the Figure 6.18, the minimal knockout set size distributions of all three scenarios are depicted for several times of the simulations. Interestingly, while the general network analysis of the first scenario did not show major changes in the structure of the networks from earlier to

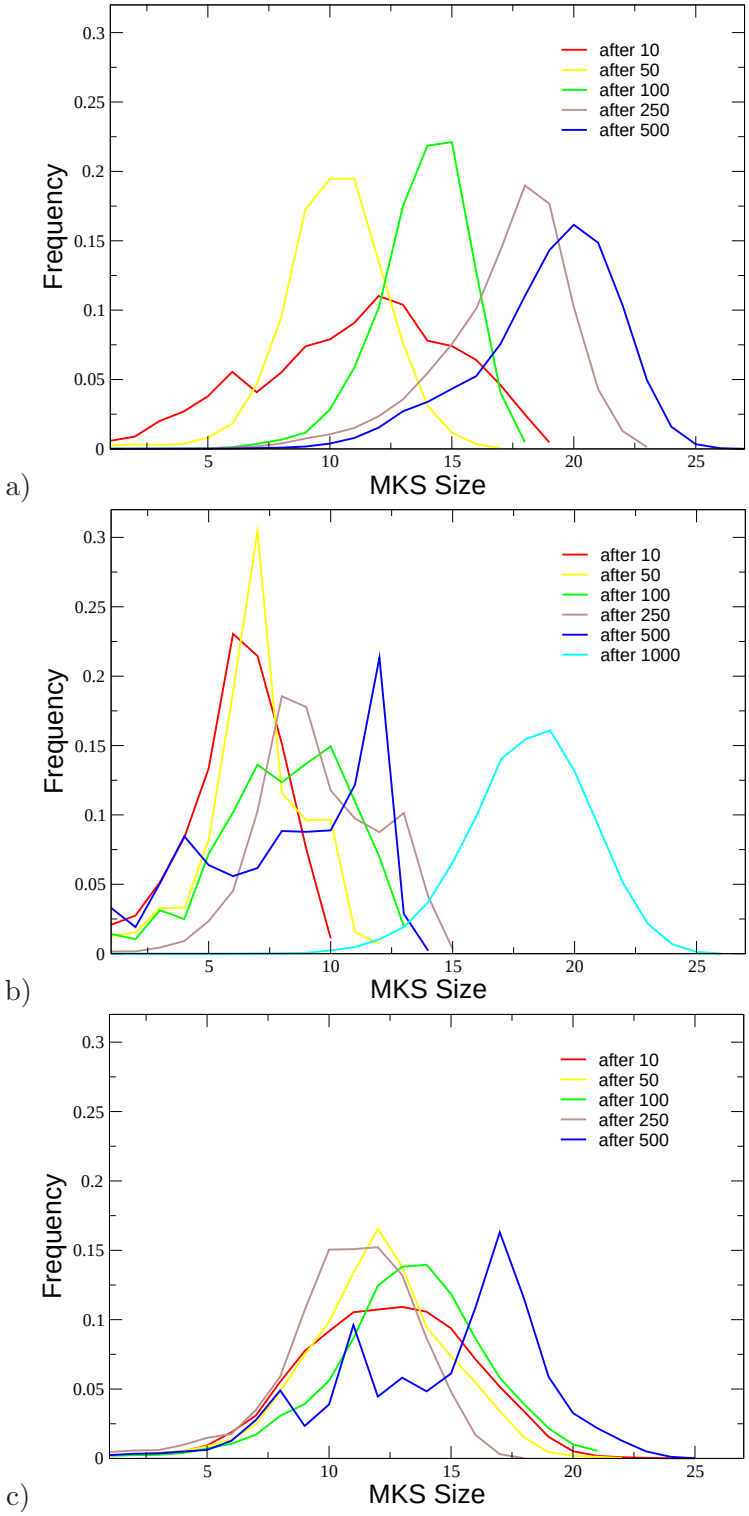


Figure 6.18: The minimal knockout set size distribution of the networks from the three scenarios, for several time steps. For a) the static scenario and c) the HGT/MUT scenario the distributions are shown for networks after generation 10 (red), 50 (yellow), 100 (green), 250 (brown) and 500 (blue). For b) the changing scenario also the distribution for networks from generation 1000 (cyan) is shown.

later generations, the constitution of the functional units is changing considerably throughout the evolution of this scenario. In the first generations, still a small fraction of knockout sets is of size one, revealing some extent of fragility in the system. In further generations not only the amount of one element knockout sets decreases towards zero, also the entire distribution moves toward an higher average and maximal set size, indicating a clear trend to a more and more robust system.

The development of the knockout set size distribution in the second scenario also sees a shift to larger sets. However, this process is weaker than that of the first scenario, especially in the first generations. Here, a not negligible amount of knockout sets are of size one. Even in later phases up to generation 500 such small knockout sets are present to similar extent. Only after 1000 generations, a distribution similar to that of the first scenario can be observed. The steady change of the environmental metabolites appears to slow down the evolution of the ability to accept multiple knockouts of enzymes. As explained in the beginning of this chapter, in the second phase (generation 500 - 1000) of this scenario the series of perturbations to the "food" set in the first phase are repeated again. This repetition might explain the faster adaptation observable in the later generations of the second scenario, because reactions, pathways and even modules that make use of the new (perturbed) situations already exist and do not have to be invented as in the first phase. So to say, the metabolic system has already been trained on these changes.

For the third scenario, again a much slower and weaker development in terms of the knockout set size distribution can be observed. Although, the distribution is comparable to that of the first scenario for the first generation, it does not follow a clear trend but rather undergoes setbacks in its development. This back and forth could be caused by disruptive genetic events, mutations or gene transfers that proved to be harmful only generations after they were introduced and therefore were not sorted out in the selection process.

It appears that perturbations of genetic as well as non-genetic nature have a slightly negative effect on the emergence and evolution of robustness against simultaneous knockout of multiple enzymes. Another similarity in the two scenarios with perturbations, is the shape of the minimal knockout set size distribution. While in the first scenario, the distribution corresponded to a smooth normal distribution, those of the second and third scenario exhibit several peaks. This may well be explained by the abundance of subsystems, modules, in those scenarios. Modules of different sizes and structure, while itself having normally distributed knockout set sizes, contribute to the overall distribution in a more random fashion.

6.3.3 Pathway Similarity

Pathway similarity measures are commonly used for the discovery of evolutionary relations among organisms, the evolution of metabolic pathways [56] and identifying conserved pathways [107]. In this paragraph, however, the similarity between a set of pathways from a single organism will be investigated. As mentioned above, it was shown by studies on the cutset

size distribution of *E .coli* that its subnetworks show some extent of epistasis. Consequently, there also has to be some overlap in the structure and functionality of the metabolic pathways, the smallest functional subnetworks. This means that the similarity between pathways has some relevance in the study of system properties such as modularity or flexibility. For instance, a cluster of pathways that are similar to each other but less to pathways outside the cluster indicate the presence of a module in the metabolic system. The nature of this module, whether it is functional or modular, depends on the way pathway similarity is defined (see Figure 6.19). Pathways that are similar in terms of their intermediary metabolites (Sim3 in Figure 6.19), represent a subset or module of pathways that are active under the same environmental conditions. Similarly, if pathway similarity is defined only as the similarity of their substrates and products (Sim2 in Figure 6.19), pathway clusters indicate modules fulfilling a common purpose, such as the production of a specific biomass metabolite. Another pathway similarity measure can be derived from the comparison of the enzymes and reactions within the metabolic pathways (Sim1 in Figure 6.19), which can be helpful in identifying structural modules. Furthermore, pathway clusters of this similarity measure would suggest a high degree of flexibility in this subnetwork. The pathways of such modules would be able to easily switch between each other because they differ only in one or few enzymes. If in addition, the enzymes which differ in the pathways are highly similar it is likely that they have a common origin in evolutionary history as well as the genome position. Finally, a combination of all three similarities can serve as a measure (Sim4 in Figure 6.19).

There are now also different ways of defining similarities between molecules (SimM) and enzymes or reactions (SimE). For the similarity of chemical molecules commonly the Tanimoto coefficient is determined [62]. For this, binary vectors for the two compared molecules are created, where each entry stands for a certain predefined structural feature, such as the widely used MACCS keys [61], where an entry is one if the feature is present in the respective molecule and zero if not. These binary vectors or fingerprints of both molecules (A, B) are then compared in the following way to get the Tanimoto similarity (τ)

$$\tau = \frac{|A \wedge B|}{|A| + |B| - |A \wedge B|} \tag{6.8}$$

where $|M|$ is the number of 1-entries in the binary vector and $A \wedge B$ is derived from the application of the and-operator on the vectors of A and B. For the use in this work due to the restricted chemistry, the MACCS keys are not the appropriate choice. Therefore, another similarity measure based on topological indices of the molecules graphs was applied. Here, the Wiener Number [177], the Platt Index [131], the Balaban Index, the Zagreb Index and the Connectivity Index were used (see Appendix B). First, for all five indices the similarity between the two molecules was computed and then all five measures multiplied together, see Table 6.4 for an example.

For the similarity between enzymes in studies of real-world metabolic networks the sequence information of the compared enzyme molecules is used. Thereby, a sequence alignment using

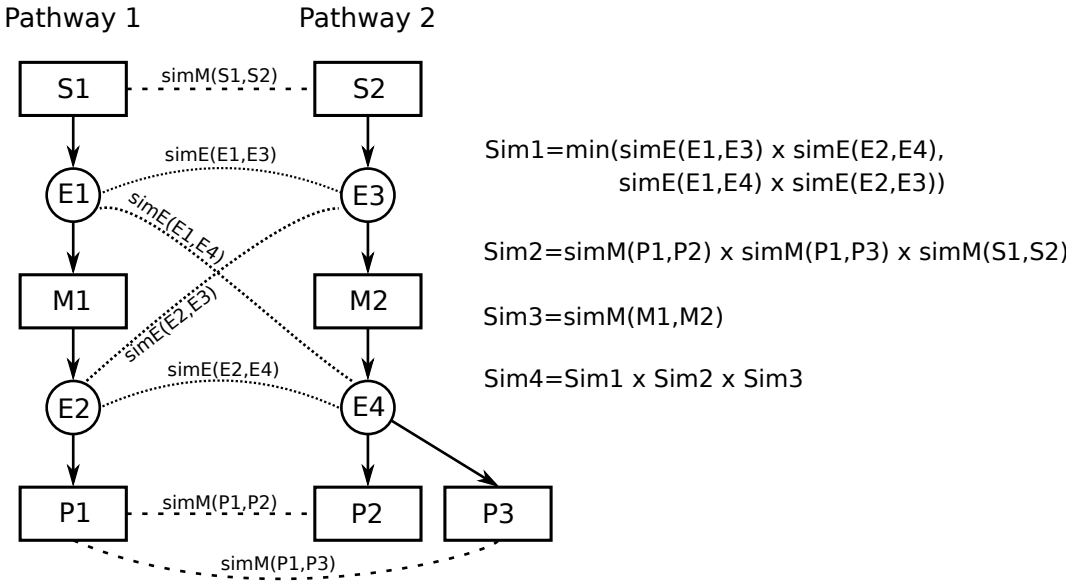


Figure 6.19: The pathway similarity measures shown on a simple pathway schema. The network consists of nodes for enzymes (E), inner metabolites (M), input molecules or substrates of the pathway (S) and products (P). Two simple pathways are compared, the dotted lines connect components of the pathways that are compared against each other. Three similarity measures of different components and one overall similarity is described in the equations on the right. R1 - enzyme similarity, R2 - input/output similarity, R3 - metabolite similarity and R4 - overall similarity. Adapted from [56].

substitution matrices such as BLOSUM [75] or PAM [34] are performed to compute a similarity score. In this work, a similarity measure based on the reaction mechanism will be used, as explained in Section 4.3. Here, also an alignment procedure is applied but on the transition states of the enzymes, minimizing the mismatches between both transition states. Possible substitutions that add to the score are atom substitutions, bond substitutions, change in the electron reordering and transition state size, see Figure 6.20. In case of specific reactions of enzymes are compared, then the similarities between the substrates and products are added, respectively.

Here three similarity measure for elementary modes will be investigated. The first measure gives the similarity between the enzyme sets of the elementary modes, its results are shown in Figure 6.21. The second measure compares the in- and output molecules of the metabolic pathways (Figure 6.22). The third measure corresponds to the similarity of the chemical reactions within the pathways in terms of their substrate and product molecules. All measures are recorded for the three evolutionary scenarios and several generations. Further, the pathways are clustered based on the similarity information, where pathways with a similarity above 90 percent belong to one cluster. The cluster information has been averaged over all generations since there was no significant change in the cluster distribution throughout the simulation history.

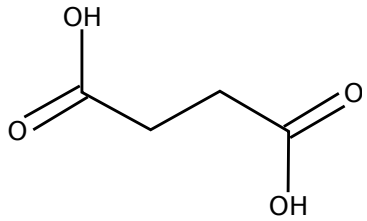
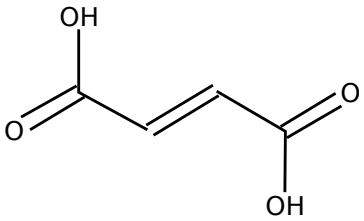
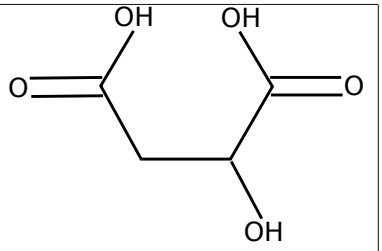
		
O=C(O)CCC(=O)O	O=C(O)C=CC(=O)O	O=C(O)CC(O)C(=O)O
WN= 74	74	96
PI= 16	16	20
BI= 22.3761	22.3761	28.4877
CI= 2.91421	2.61779	3.33493
ZI= 44	57	53
SIM(1,2)= 0.929391	SIM(2,3)= 0.812242	SIM(1,3)= 0.811256

Table 6.4: The molecule similarity measure for three example molecules. The molecules graphs, their SMILES notation and the topological indices WN=Wiener number, PI=Platt Index, BI=Balaban Index, CI=Connectivity Index and ZI=Zagreb index are given. The similarity measures below are computed from the similarity of the five topological indices of two molecules each.

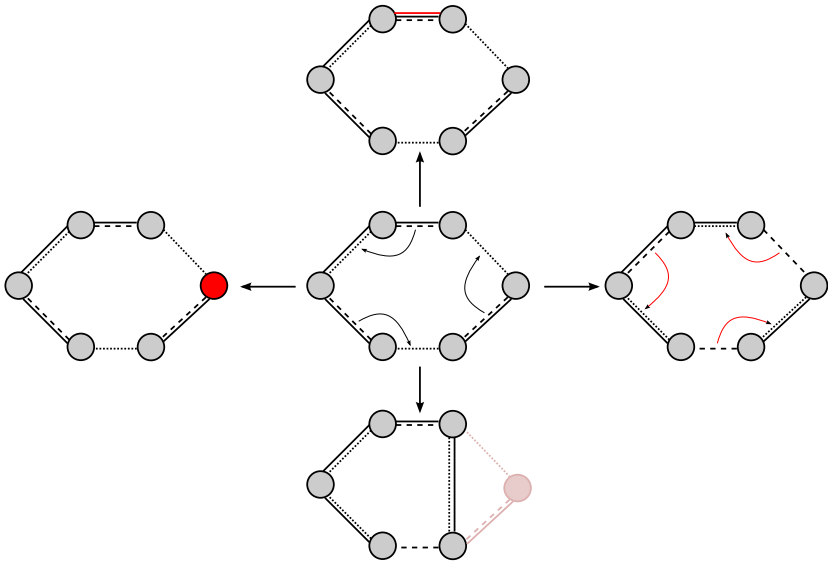


Figure 6.20: The enzyme similarity measure explained. In the center the original transition state of the enzyme is depicted. Around are the four possible changes to a transition state (marked in red), left: change in atom type, top:change in bond type, right: change in electron reordering (curved arrows inside transition state) and bottom: change in the transition state size (lost part is shown faded). The cost of the changes increase in this order.

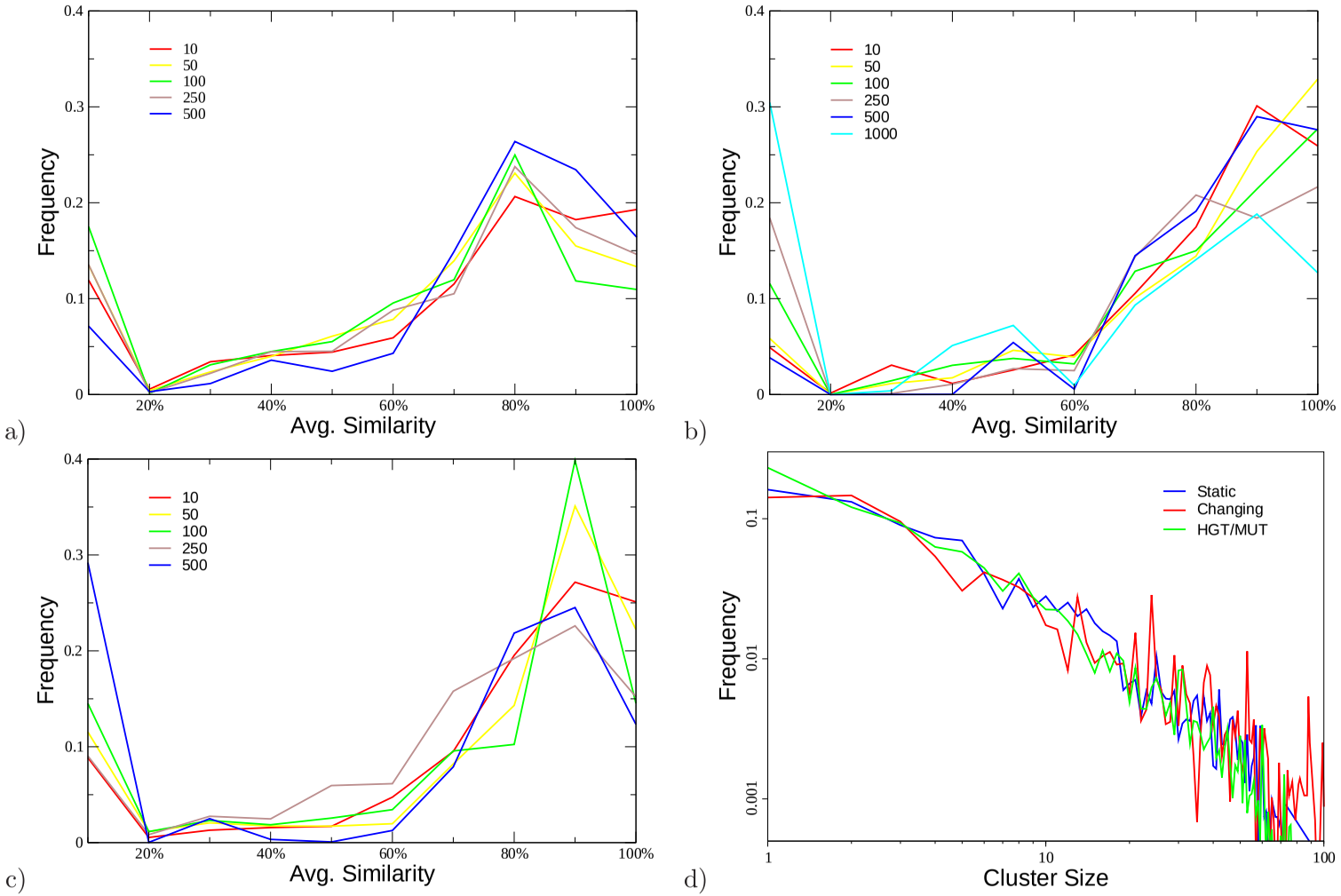


Figure 6.21: Enzyme similarity between the elementary modes of the simulated networks for several stages. For a) the static scenario and c) the HGT/MUT scenario the similarity measure is shown for networks after generation 10 (red), 50 (yellow), 100 (green), 250 (brown) and 500 (blue). For b) the changing scenario also the similarities for networks from generation 1000 (cyan) are shown. In d) the cluster size distribution for all three scenarios is depicted. Elementary modes with similarity $\geq 90\%$ belong to the same cluster.

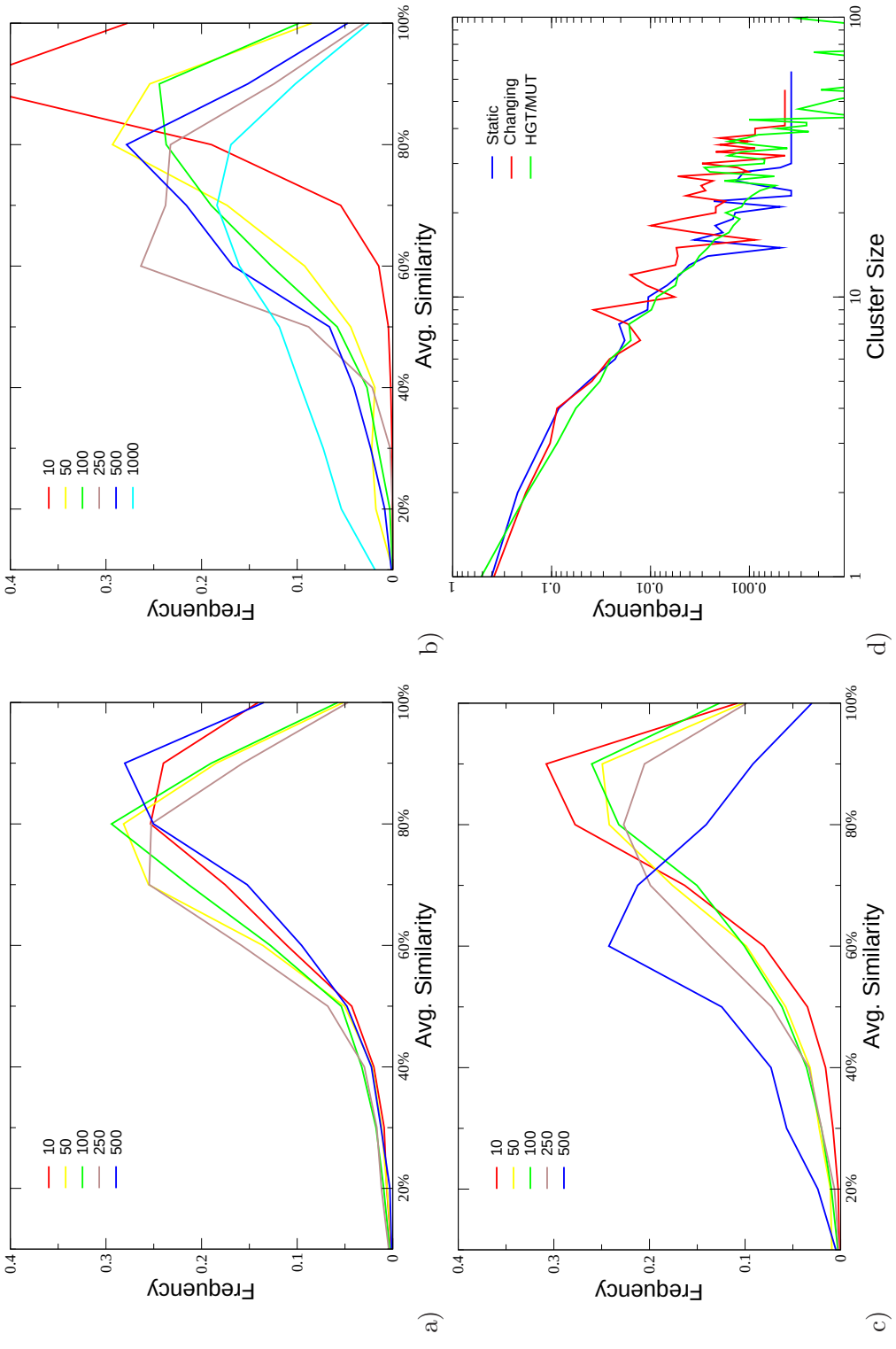


Figure 6.22: Input/Output similarity between the elementary modes of the simulated networks for several stages. For a) the static scenario and c) the HGT/MUT scenario the similarity measure is shown for networks after generation 10 (red), 50 (yellow), 100 (green), 250 (brown) and 500 (blue). For b) the changing scenario also the similarities for networks from generation 1000 (cyan) are shown. In d) the cluster size distribution for all three scenarios is depicted. Elementary modes with similarity $\geq 90\%$ belong to the same cluster.

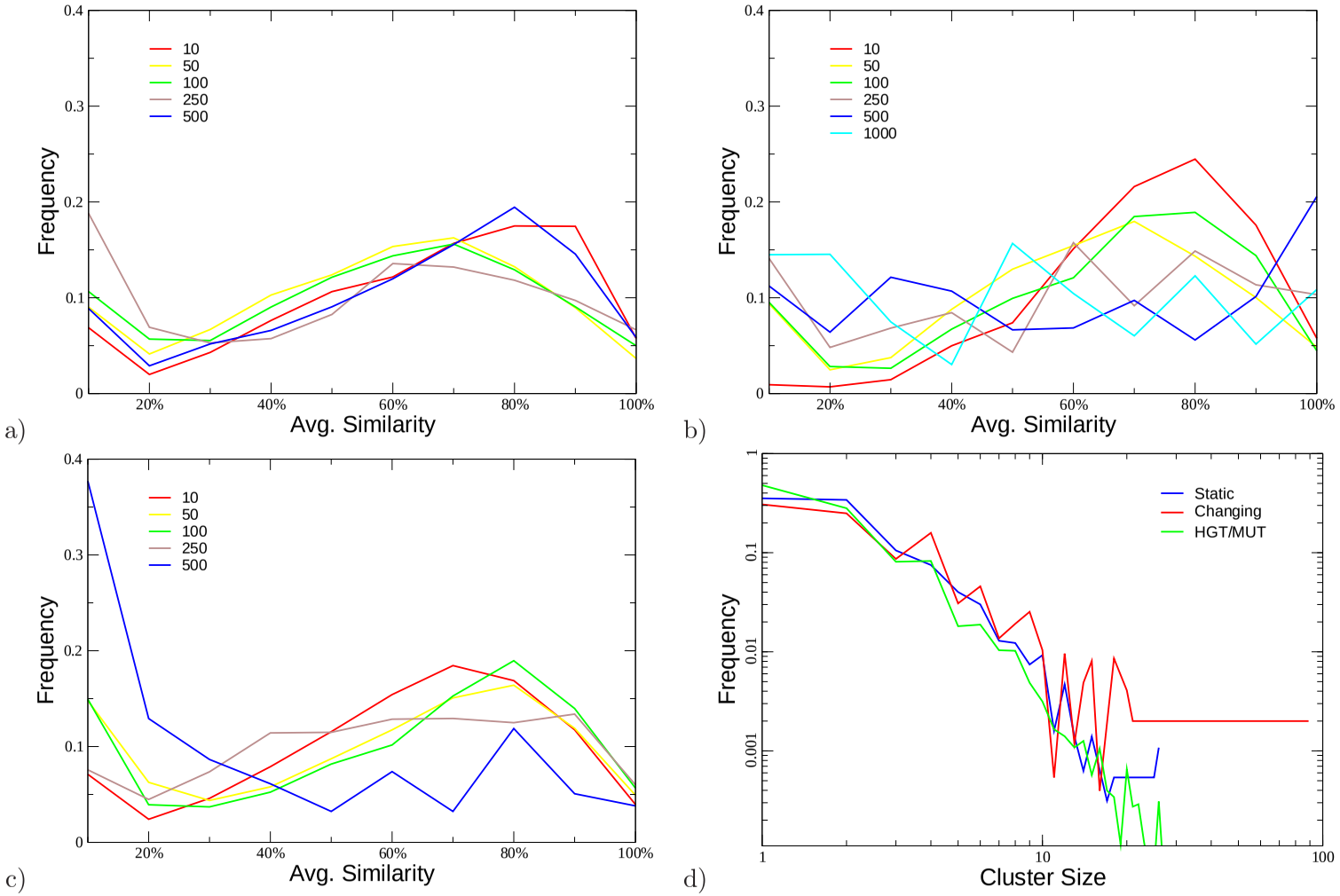


Figure 6.23: Metabolite similarity between the elementary modes of the simulated networks for several stages. For a) the static scenario and c) the HGT/MUT scenario the similarity measure is shown for networks after generation 10 (red), 50 (yellow), 100 (green), 250 (brown) and 500 (blue). For b) the changing scenario also the similarities for networks from generation 1000 (cyan) are shown. In d) the cluster size distribution for all three scenarios is depicted. Elementary modes with similarity $\geq 90\%$ belong to the same cluster.

Starting with the analysis of the enzyme set similarity among elementary modes, one can observe a high similarity in all scenarios. This observation indicates that enzyme recruitment has a leading role in the formation of metabolic pathways, as already suggested in the previous chapter. Furthermore, in this study it was found that in the second scenario enzyme recruitment was increased, due to its role in situations of environmental change enabling a backward evolution like pathway formation. Similarly, here a higher enzyme similarity can be observed for the second scenario also supporting the higher impact of enzyme recruitment on its evolution. The enzyme similarities in the first and third scenario are similar but the first scenario shows a slow and steady development towards more similar pathways, whereas in the third scenario the trend is more intense for the first generations but then undergoes a strong setback.

Turning to the similarity measure based on the comparison of input and output molecules of the pathways, a different picture can be observed. In the first scenario again a slow and weak trend. In fact the similarities stay constant or at least do not change significantly. Since we know that the set of "food" molecules in this scenario remains throughout the simulations, this suggests that almost the same molecules are produced at all stages. Thus, these simulations evolve toward optimization of a constant goal, the production of these constant product molecules. This can not be said about the other two evolutionary scenarios, which show heavier shifts in the similarity distributions. The second scenario more than the third scenario. This is explained by the changing environment in the second scenario which adds one more degree of freedom to the variability of the in- and output molecules of the simulations. In both scenarios, the development is heading to less similar pathways.

The differences in the pathway similarities between the first scenario and the other two scenarios becomes even more apparent when studying the third measure, the similarity of the specific chemical reactions comparing their substrate and product molecules. As for the distribution of the other similarity measures, the first scenario does not exhibit any significant changes, the pathways stay mostly very similar to each other. For the second and third scenario, there is a strong development to pathways which differ stronger and stronger in their used and produced metabolites. It can thus be concluded that in situations of genetic or non-genetic changes, the metabolite specificity is much less conserved than the reaction chemistry (the reaction mechanism of an enzyme) or even the products of a metabolic pathway.

Summarizing the results from the three similarity measures, it can be stated that the metabolic pathways of the second and third scenario are more similar to each other in terms of their enzymes than those of the first scenario. This is explained by the increased enzyme recruitment in these scenarios. On the other hand, when regarding the used and produced metabolites of the pathways as well as single chemical reactions, the pathways of the first scenario show a significantly higher similarity than those of the other two scenarios. The cause for this dissimilarity can be found in the modularity of the networks of the two scenarios. Another implication of these findings is that the modules in these networks differ particularly in the metabolites not in the enzymes. So even though enzymes are recruited and their reaction

chemistry is preserved throughout evolution, the metabolite specificity is not. In fact, there appears to be a high degree of enzyme promiscuity.

6.4 Chemical Organizations

Another steady-state based analysis of biological networks is possible through the study of so-called chemical organizations, self-maintaining and closed sets of metabolites and enzymes in the case of metabolic networks (see Section 3.2.4). Particularly interesting are the lattices that the different steady-states of a network form. Thereby, organizations are connected if one is a subset of the other, with the larger organization on a higher level. If an organization includes more than one other organization it will be placed one level above the largest organizations of all its subsets. Thus, in most cases on top of these organization hierarchies will be found an organization representing the full metabolic network, while the lowest organization is often an empty set. The organisms can then switch from one organization to another within the hierarchy, along the connections. In case of metabolite depletions or enzyme knockouts this will usually be a downward movement in the hierarchy. In contrast, if new metabolites enter the system, such as in the scenario of a changing "food" molecule set, the chemical evolution will have an upwards direction, heading to a "higher" organization with a possibly better metabolic yield. Consequently, the characteristics of the organization hierarchies can provide valuable information about the dynamics and properties of the studied system. For instance, if the hierarchy contains many organizations which are similar to the full metabolic network and thus are likely to have a close to optimal metabolic yield, the system becomes highly robust against knockouts and/or metabolite depletion due to these alternative states. On the other side, a large number of organizations close to the empty set or smaller organizations might be beneficial in situations of big changes such as drastic environmental shifts. Such a structure of the hierarchy increases the likelihood for the organism to find a smaller steady-state that at least produces some metabolic yield and from which other organizations more upwards can be reached through further chemical evolution. Another important factor is the similarity between organizations of adjacent levels, because most likely switching from one organization to the next will happen on adjacent levels. If two organizations are highly similar a few changes in the enzyme or metabolite set will cause such a switch. Hence, the average similarity between organizations of adjacent levels gives clues about the flexibility of the system to change steady-states to potentially better states.

At first, the average number of organization per hierarchy level will be investigated. In Figure 6.24 this distribution is pictured for all three scenarios and several generations. In the first generations of the first scenario, the hierarchies are comprised mainly of organizations close to the smallest organization. This changes constantly and dramatically throughout the evolution of the simulations, ending up in an almost contrary state where the majority of organization is closer to the full set or top organization in the hierarchy. These two opposing constitutions are illustrated on two example hierarchies from a simulation run of the first

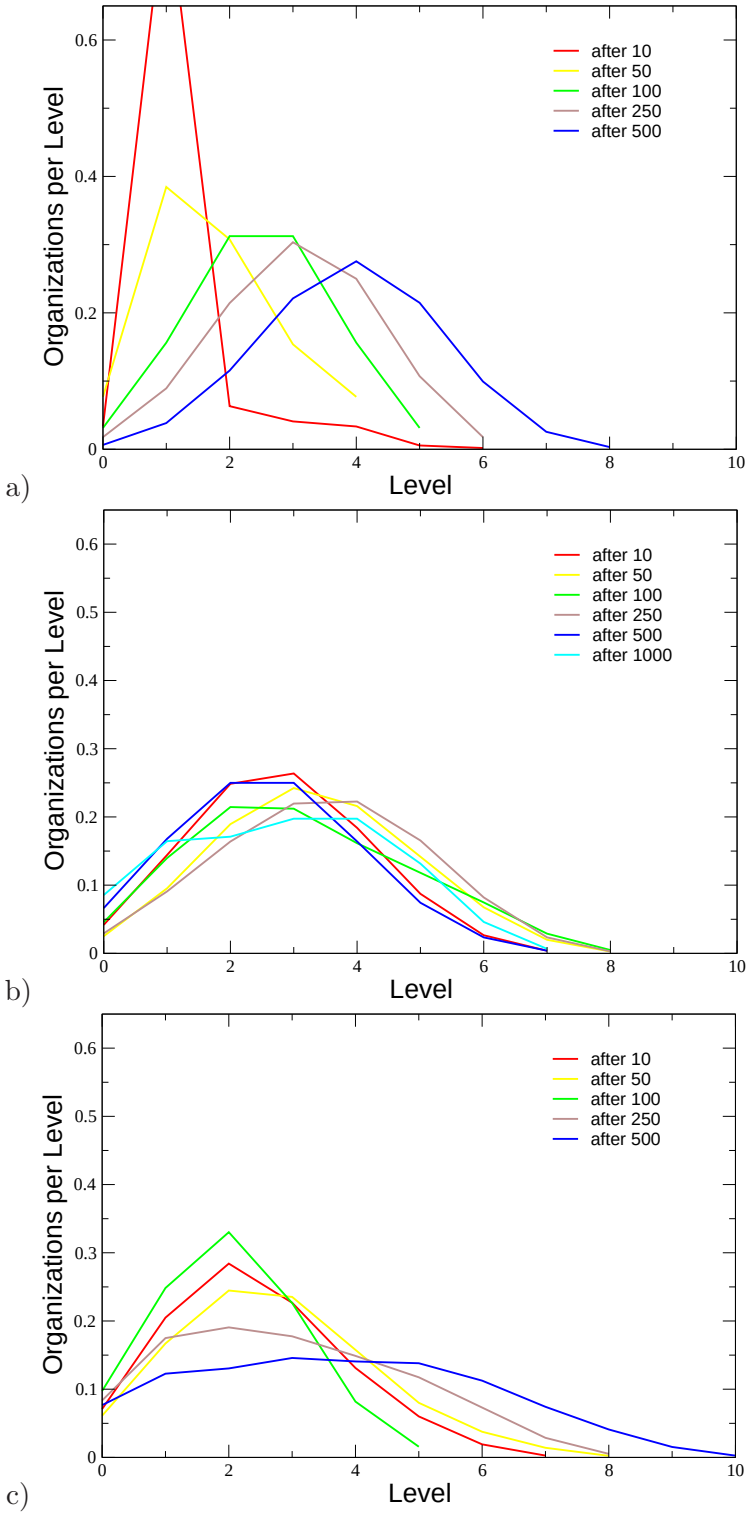


Figure 6.24: Frequency of organizations per hierarchy level, for the networks of all three scenarios. For a) the static scenario and c) the HGT/MUT scenario the relative frequency of organizations is shown for networks after generation 10 (red), 50 (yellow), 100 (green), 250 (brown) and 500 (blue). For b) the changing scenario also the distribution for networks from generation 1000 (cyan) is shown.

scenario in Figure 6.25. The first hierarchy (Figure 6.25a)) stems from the early phase and the second hierarchy (Figure 6.25b) from the later phase of the simulation. As mentioned for the average distribution, the hierarchy from the earlier phase has many organizations on the bottom, forming a pear-like shape, while the later hierarchy exhibits a more balloon-like shape with many organizations on the top. It has been already discussed in the beginning of this paragraph that such a pear-shape organization hierarchy can ensure at least a minimal metabolism for the system even if there are larger changes, while a balloon-shape hierarchy enables the maintenance of a metabolism close to the optimal in case of smaller perturbations. In the evolutionary context the transition from pear- to balloon-shape, therefore, makes sense because in the earlier phase the system might deal with drastic changes and can have a selective advantage if it can maintain at least some of its metabolic functionality. However, in later stages in which systems are more complex and adapted, perturbations might not have such a big effect on the overall structure of the metabolic network. Then, those systems that retain most of its metabolic functions would gain an advantage.

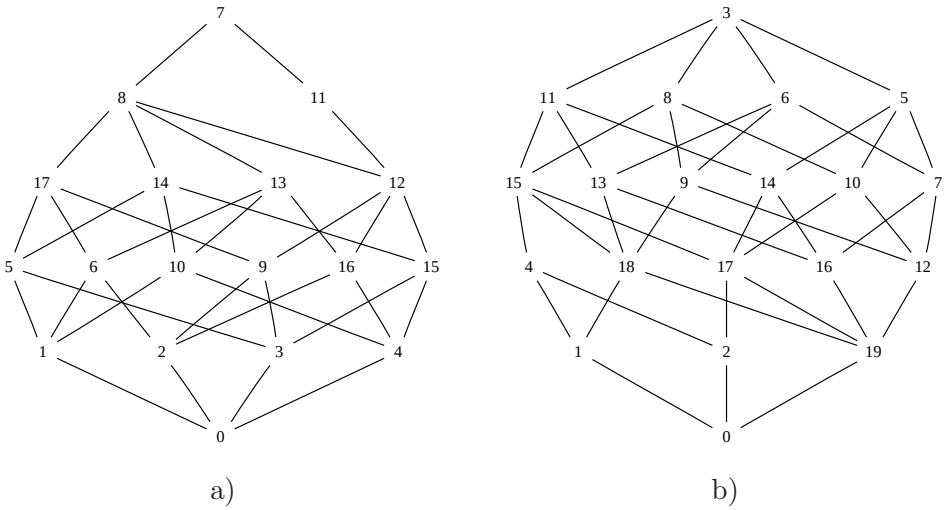


Figure 6.25: Organization hierarchies from two example networks of the static scenario. Nodes (numbers) are organizations of the network, a connection between two nodes represent an inclusion relation between two organizations, where the upper is the superset and the lower is the subset. Compared are networks from the a) early stage and b) later stage of the simulation.

The situation for the second scenario is somewhat different, there is no trend observable in the evolution of the organization hierarchies, maintaining the pear shape even in the latest generations. Also in the later stages of this scenario, rather disruptive events can occur through the steady changes in the constitution of the environment. Therefore, the ability to maintain a basic metabolism in such situations poses a selective advantage to a system, explaining the survival of the pear-shape hierarchy.

The development in the third scenario is similar to that of the first scenario to that effect

that the organization hierarchies comprise of more and more organizations from higher levels, but differs in that the hierarchies of the last generations are even flatter than those balloon-shaped hierarchies of the first scenario. The flatter shape might be a consequence of the higher number of levels which in turn could be caused by a an evolution towards system that can switch more easily from organizations of one level to those of the next higher level. A hierarchy with more levels suggests a higher similarity between organizations of adjacent levels. To investigate this idea, one has to observe and compare the features of organizations.

To this end, the average size of organizations on the same level are shown in Figure 6.26 for all three scenarios and several points in the simulation. Looking at the third scenario, the distribution of organization sizes per level corresponds to the expectation from the above discussion. The difference (in size) between organizations in adjacent levels is decreasing with time enabling an easier switching between those organizations. This adaptation might have resulted from the steady mutational noise that is present in this scenario. Thus, an increased mutation rate could cause a higher flexibility. For the first scenario with a more moderate mutation rate only a much weaker trend can be observed, supporting this hypothesis.

The development in the second scenario is contrary to that of the other two scenarios. Here, the organizations of the upper levels are considerably larger than their lower neighbors. On the one hand, this has the effect that switches between organization need bigger changes in the enzyme or metabolite set. On the other hand, this constitution is also an indicator of modularity. Modules in a system will have corresponding chemical organization and if they are combined with each other to build another organization (on a higher adjacent level), this organization will be considerably larger than its subsets because they are likely to have a rather small overlap. One example organization hierarchy from the second scenario is shown in Figure 6.27a). Noteworthy is here the fact that there are two organizations on top of the hierarchy, two favorable steady-states of the cell. Such parallel steady-states make the case for an adaptation to multiple conditions, in this case the constitution of the "food" set.

In previous paragraphs it was suggested that also the third scenario evolves a certain degree of modularity, the observation of the organization hierarchies features did not deliver any evidence so far, though. Nevertheless, in Figure6.27b one example hierarchy of the this scenario does seem to show indications of modularity in that two sets (left and right half) of organizations are mostly connected among each other but have almost no connections to hierarchies of the other set. In such a case, the similarity between organizations of adjacent level is not able to provide evidence for modularity and therefore no further prediction about modularity for this scenario can be made other than the suggestion that the horizontal gene transfer could be a source of another type of modularity that environmental change. This modularity would be defined by few or no connections among the single modules, which was speculated before.

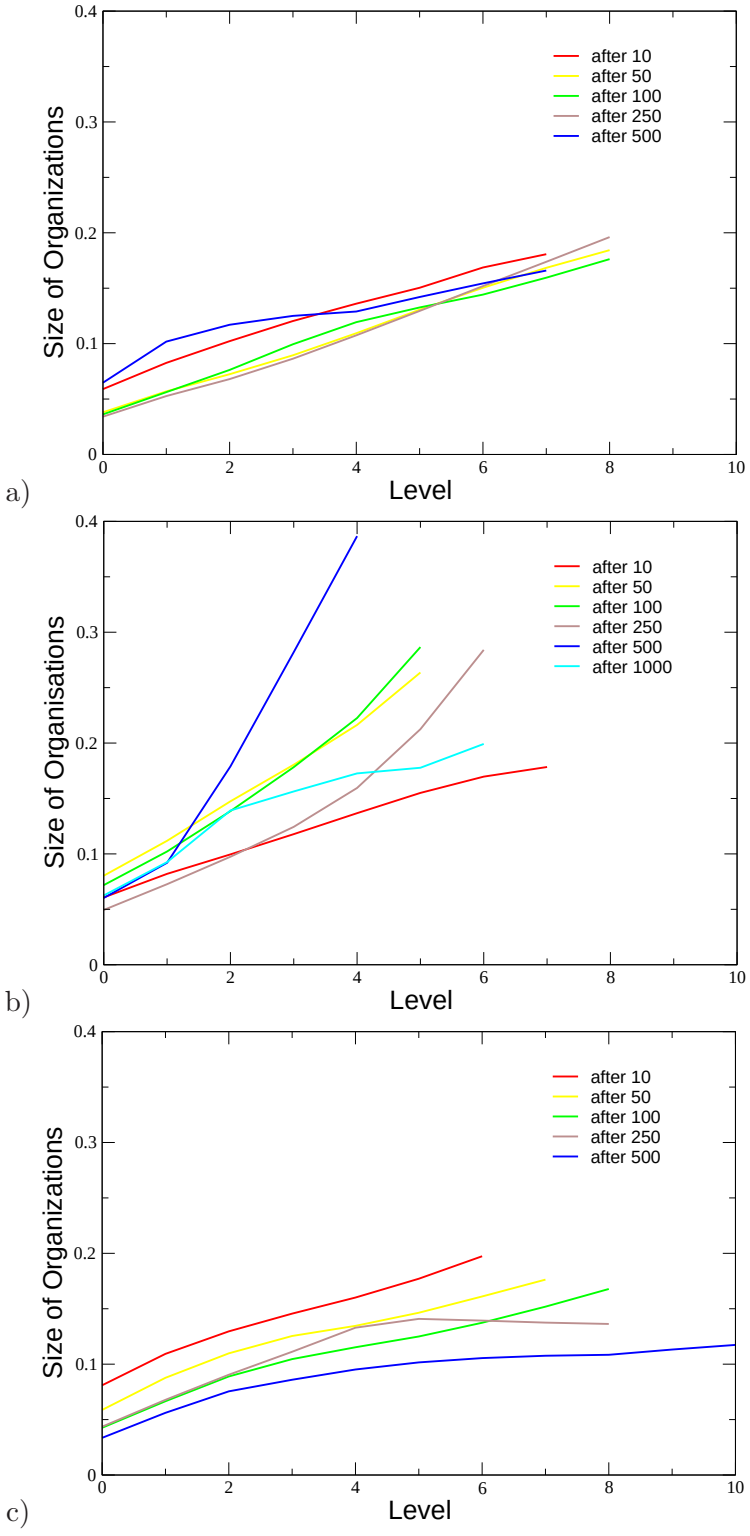
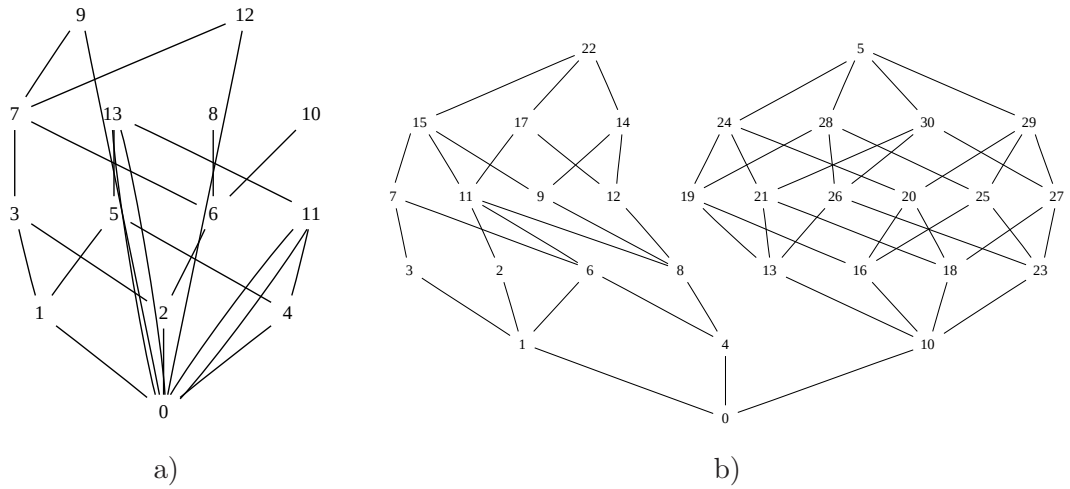


Figure 6.26: Relative size of organizations per hierarchy level, for the networks of all three scenarios. For a) the static scenario and c) the HGT/MUT scenario the relative organization sizes are shown for networks after generation 10 (red), 50 (yellow), 100 (green), 250 (brown) and 500 (blue). For b) the changing scenario also the size distribution for networks from generation 1000 (cyan) is shown.



a) b)

Figure 6.27: Organization hierarchies from two example networks. The networks stem from simulations of a) the changing scenario and b) the HGT/MUT scenario. Nodes (numbers) are organizations of the network, a connection between two nodes represent an inclusion relation between two organizations, where the upper is the superset and the lower is the subset.

6.5 Neutral Network Approach

For many years it is known that neutral mutations have a considerable influence on the evolution in molecular systems [98]. The RNA sequence-to-secondary map with its many-to-one property represents a system entailing considerable neutrality. It has been used successfully to shed light on the mechanisms of evolution [84, 85] and explain basic properties of biological systems, such as resolving the interplay between robustness and evolvability [169].

The concept of neutral networks has also been successfully applied in the field of regulatory networks [17] as well as metabolic networks [142] to investigate their structural properties and provide implications for the robustness of the studied systems. In the following, the structure-to-function map of the introduced model as well as the evolved metabolic networks from the simulations will be investigated on the neutral network level, to provide insights about their evolvability and robustness.

6.5.1 Genotype-to-Phenotype Mapping

In this paragraph the genotype-phenotype mapping (RNA-sequence to Ribozyme-function) which was introduced in the computational framework in Chapter 4 will be compared with other existing mapping approaches, in terms of their neutral network's constitution and the implications they have on their system properties. Mappings based on cellular automaton (CA) and random boolean network (RBN) have been proven to exhibit desirable properties [41] and, thus, will serve here as reference. In previous evolutionary models and for the

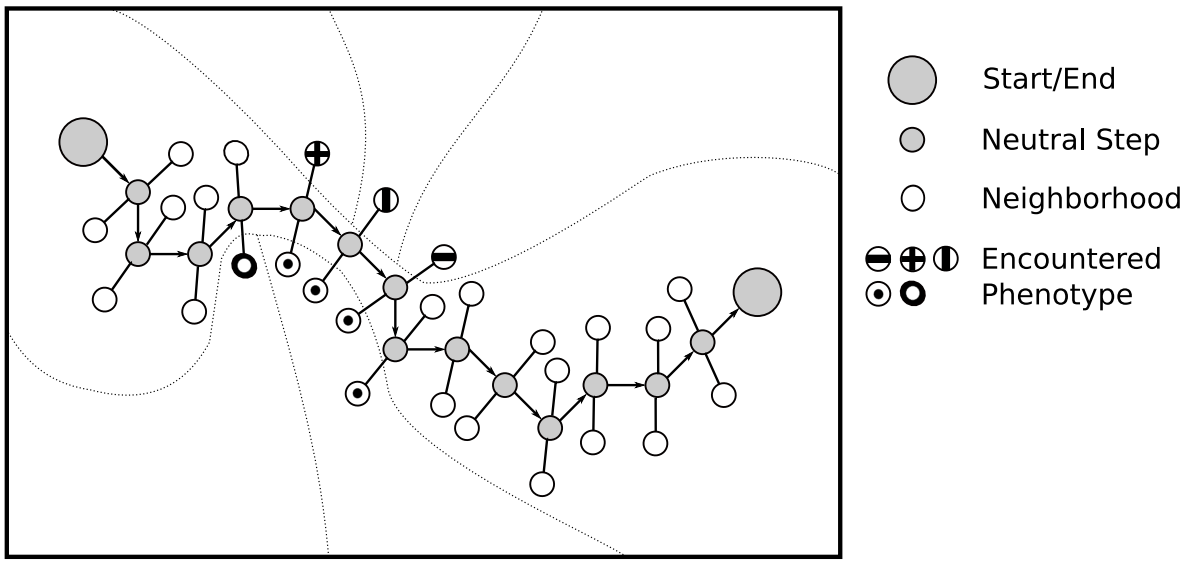


Figure 6.28: Random neutral walk from an arbitrary point in genotype space to random neutral neighbors until a certain length is reached. In each step, encountered phenotypes in the one-point mutation neighborhood are protocoled.

statistical analysis in this paper, other mappings based on the RNA sequence-to-structure map but varying structure-to-function maps have been developed. Here, only two of them will be mentioned briefly. For the first mapping, one target structure was assigned for each point in the phenotype space, the structure then maps to the phenotype of the closest target structure. The second mapping extracts structural features, similar to the presently used mapping, but for the entire structure.

Random Neutral Walk One way of gaining statistical results for the comparison of different genotype-phenotype mappings is through a random neutral walk on the genotype space. It starts from an arbitrary point of the mapping's genotype space. The first step is to choose randomly one neutral mutation out of the set of all possible one-point mutations. A mutation is neutral on a genotype if the phenotype of the original genotype is the same as the phenotype of the mutated genotype.

Applying the mutation, a new genotype is created that differs from its predecessor in only one position and maps to the same point in phenotype space. This procedure is repeated until the length of the walk reaches the pre-defined limit or no neutral mutation can be applied. The latter case occurs for example in one-to-one mappings where there exists only one genotype for each phenotype, therefore, contains no neutral elements. If neutral neighbors are found, the described procedure realizes the random walk on one particular neutral component of the genotype space.

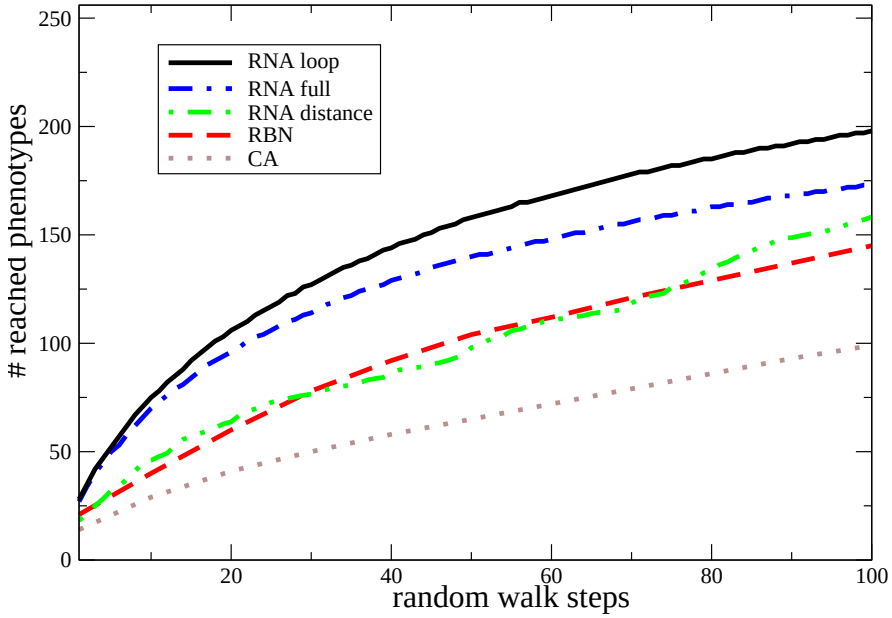
In each step, it is kept track of several statistics that give insight into the structure of the neutral network. For this aim, the one-point mutation neighborhoods of all genotypes of the

random neutral walk are observed. The fraction of neutral mutations within these neighborhoods will be referred to as the neutrality of the underlying neutral network. In terms of biological systems this can also be regarded as the robustness that is gained through the mapping, since neutrality protects from harmful mutations. Further, it is searched for new phenotypes that are encountered during the course of the walk, i.e. the number of different phenotypes that were found in the one-point neighborhoods. This measurement indicates the rate with which the mapping discovers new phenotypes and other neutral components. The faster a system can access different points in the phenotype space, the more evolvable is it. Thus, the discovery rate can be equated with the notion of evolvability. Other measures that are regarded as important here, are the extent and the connectivity of the neutral networks. Former, is given by the maximal distance between genotypes of the neutral walk. Latter, is defined by the fraction of possible phenotypes that can be reached from an arbitrary starting configuration.

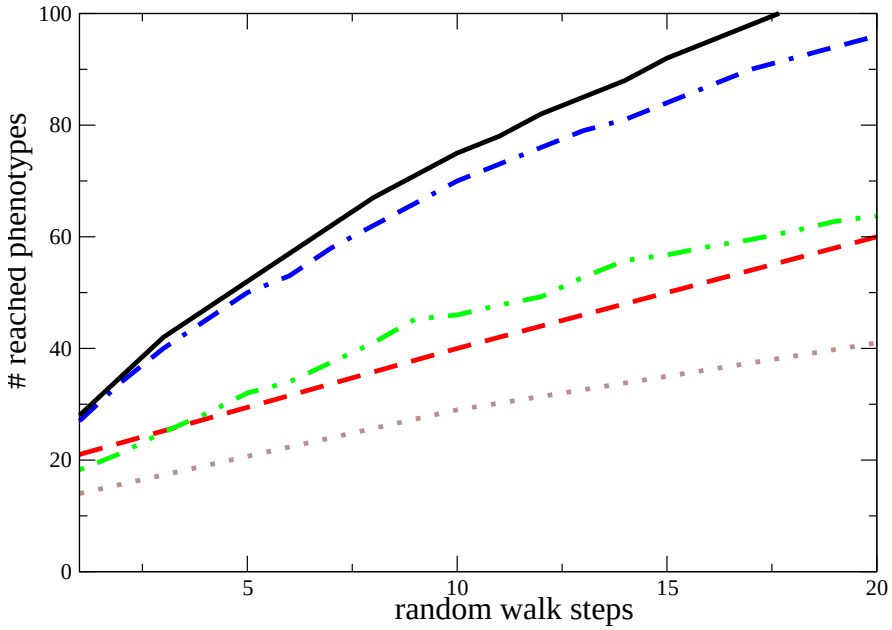
Comparison For the following study, 1000 random neutral walks of length 100 were performed for each mapping. The size of the genotype spaces vary for the different mappings. For the RNA-based mappings the size is 4^{100}, for the random boolean network mapping it is 2^{144} and for the cellular automaton mapping 2^{76}. The phenotype space is $2^8 = 256$ for all mappings.
Figure 6.29 shows the results of the performed random walks. It can be seen that the mapping introduced in this work (RNA loop) reaches the most phenotypes ($\approx$ 200 of 256), following the other RNA-based mappings ($\approx$ 175 and 150), the random boolean network (145) and the cellular automaton (100). It can also be seen that the difference in the first steps is even more drastically, indicating a faster discovery rate of this mapping compared to all others. This can be explained by the connectivity. Whereas CA and RBN have about 14 and 21 neighboring phenotypes, respectively, the RNA mappings have about 27 distinct phenotypes in their one-point neighborhood. Furthermore, the RNA loop mapping travels further in the genotype space, allowing a steady discovery of new phenotypes. However, in terms of neutrality the random boolean network mapping performs best, $\approx$ 58% of its mutations are neutral, for CA it is $\approx$44% and the RNA-based mappings are in between with around 50%.

6.5.2 Metabolic Network

The investigation now turns to the neutral networks and energy landscapes of the simulated metabolic networks. In particular the neutrality of the neutral network will be measured as well as the evolvability that it allows. At first, a series of 100 random walks of length 1000 is performed for the metabolic networks. To ensure comparability, only networks of similar size ($50 \leq |r| \leq 100$) are used for this study. Contrary to the above analysis, the random neutral walks do not start from a random situation. The starting point here is the empty network, which means that all reactions are inactive. A mutation in this context



a)



b)

Figure 6.29: Encountered phenotypes in a random neutral walk a) of length 100 and b) for the first 20 steps. RNA loop = mapping used in the model; RNA full = RNA-based mapping considering the entire structure; RNA distance = RNA-based mapping using target structures; RBN = random boolean network mapping; CA = cellular automaton mapping.

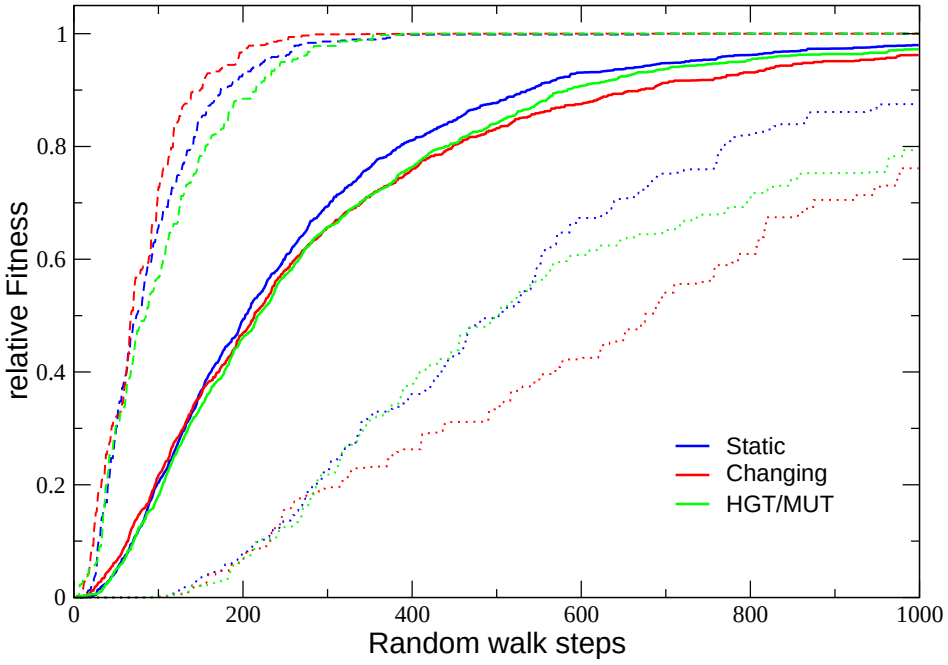


Figure 6.30: Random neutral walk statistics for metabolic networks from the simulations. The average, maximal and minimal fitness values for the networks of all scenarios are shown for each random walk step. A relative fitness of one is the optimal metabolic yield of the full metabolic network, a value of zero indicates that there is no biomass production at all.

represents an activation or inactivation of a reaction. The phenotype of a metabolic network is the fitness value as it is derived during the simulation, thus, the metabolic yield computed through metabolic flux analysis. During the random walk, mutations are accepted if the new metabolic network has the same or an even higher fitness than the present network. With every step the relative fitness, the ratio between the fitness of the perturbed network and the fitness of the full network, is protocoled. For every random walk, the number of steps until the optimum is reached, i.e. when the relative fitness is one, is recorded. These measurements are averaged for all 100 random walks of a metabolic network. For the relative fitness, also a maximal and minimal value for each step is determined. These averages are then again averaged over the studied metabolic networks.

In Figure 6.30 the average, maximal and minimal relative fitness per random walk step is depicted for the three evolutionary scenarios. The average fitness increases in a rather similar way for all scenarios, but after 100 steps the first scenario evolves a bit faster and reaches higher average fitness values than the other two scenarios. Somewhat bigger differences between the three scenarios exist for the minimal and maximal fitness values, here the second scenario holds the extremes on both ends, indicating a great variety of networks, a side effect of the strong and steady perturbations during the evolution of this scenario. Consequently, the statically evolved first scenario has the narrowest distribution.

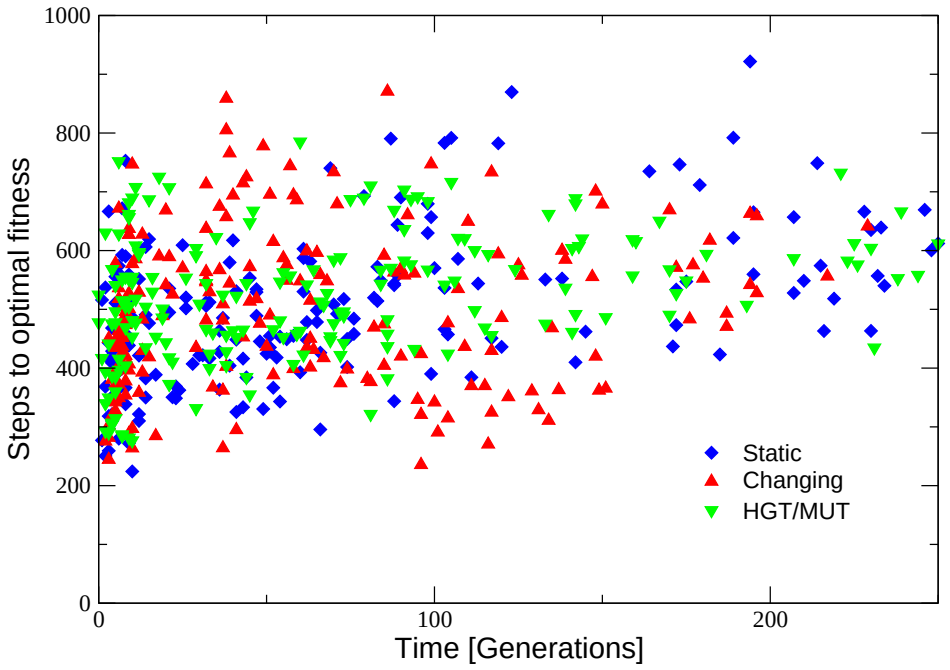


Figure 6.31: Random neutral walk statistics for the history of the simulations. Shown are the average number of steps towards the optimal fitness value of the full metabolic network.

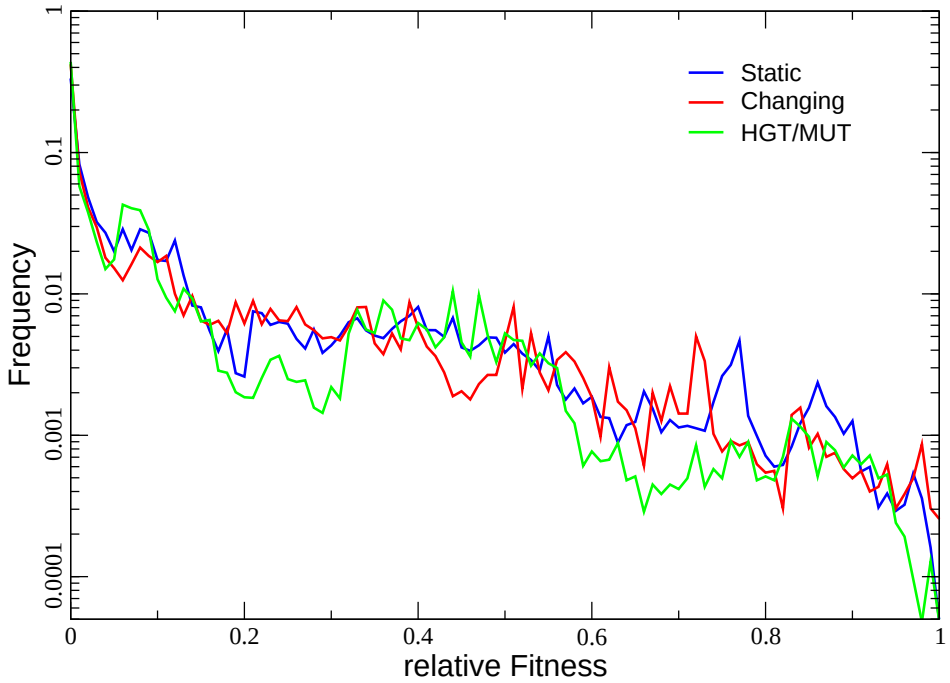


Figure 6.32: Random sampling statistics for metabolic networks from the simulations. For all three scenarios the frequency of reaction combinations with a certain relative fitness value are shown. A relative fitness of one is the optimal metabolic yield of the full metabolic network.

To see the changes of the fitness distributions over time, the average number of random walk steps to the optimal fitness value is depicted for the history of the simulations from all three scenarios in Figure 6.31. Again, the distributions of all scenarios are similar to each other and the second scenario shows the greatest variety. One noteworthy observation in the evolution of the underlying neutral networks of all scenarios is that the time to adapt stays almost constant over the simulation time. One might expect that system would evolve towards faster adaptation but the increasing optimal fitness and increasing complexity of the metabolic networks counters this development.

To get an impression of the constitution of the fitness landscapes underlying the metabolisms of the different scenarios, 1000 samples of reaction combinations for each metabolic network were chosen and their fitness value determined. The results of this analysis are shown in Figure 6.32. Similarly to the results of the random neutral walk, the fitness landscapes of all three scenarios are highly similar an do not change over time (not shown here). The majority of the combinations of reactions possess only little metabolic functionality, while only a tiny percentage of combinations come close to the optimized metabolic yield of the original metabolic network.

6.6 Summary of Results

Several measures for the robustness and modularity of networks, stemming from graph theory, pathway analysis or neutral network analysis, were applied on metabolic networks from simulations of three different evolutionary scenarios. Investigated was the evolution of metabolic networks for a static development, development under steady environmental change and under increased mutation rates and horizontal gene transfer.

The results did confirm some of the expectations from biological theory. For instance, a considerable degree of modularity was only found in networks that evolved under environmental change or horizontal gene transfer, both believed to be sources of modularity. It was further discovered that both factors seem to evolve different kind of modular structures. While environmental change lead to strongly interconnected modules, horizontal gene transfer in turn produced more isolated modules. However, a more detailed proof of these observations has still to be provided.

The findings concerning the robustness of the networks were also mostly as expected. One hypothesis was that networks that evolve under steady non-genetic noise would also adapt better to handle such metabolite fluctuations, with the interpretation of the analysis made for this chapter this suggestion can be supported. These networks are on the one hand more robust to single depletion events and on the other hand exhibit the possibility to maintain a basic metabolism in case of larger disruptions in terms of the metabolite set. One expected result was, that these networks also showed a slightly higher robustness towards small genetic errors, here the networks from the static evolution and the evolution scenario with higher

mutation rates were believed to develop a higher resistance. Nevertheless, the networks of these scenarios showed a significantly higher robustness towards larger genetic perturbations.

Chapter 7

Conclusions

The evolution of early metabolism is a key part in the origin of life research. It is still an open questions how metabolic pathways formed from chemical reactions and how the metabolic networks took the form that we can observe in today's organisms. There do exist several theories that approach these topics and have evidence from experimental studies. However, it seems clear that the possibilities of experimental biology do not suffice to uncover the mystery of the origin of life and metabolism itself. To this end, computational studies simulating the prebiotic events that lead to their emergence have been introduced. These approaches have shown considerable success in offering new explanations of many aspects in this field of research, but also general network evolution.

The modeling of metabolism as well as chemical reaction systems in general is a challenging task and faces the modeler with many difficult choices of representations and modeling techniques. Many scales of such a system can be modeled in great detail and realism, but sometimes a simple and abstract approach can yield better results. One always has to consider the consequences of such a modeling decision. Ignoring important details of the modeled system may question the overall predictability of a model. On the other hand, it is not possible and probably also not desirable to reproduce the original one-to-one. All parts on all scales must be well concerted among each other and importantly also with the purpose of the overall model. Last but not least, the model has to be computationally feasible and produce interpretable results.

In this thesis, a computational framework was introduced, which was developed with the above mentioned criteria in mind. Diverse new and existing modeling techniques from computer science as well as biological and chemical modeling were incorporated and combined. Thereby, the computational applicability and efficiency were as important in the modeling procedure as the biological and chemical soundness. If some parts had to be defined rather abstractly because there does not exist enough knowledge about it to describe it more pre-

cisely (structure-to-function map of ribozymes) or a realistic account would simply be to expensive, then emphasis was put on an at least metaphorical connection to the real system. This allows to incorporate some of the facts and assumptions of reality into the model and helps to interpret its results in terms of the original problem.

The simulations based on this model yield metabolic networks that resemble real world metabolic networks to a high degree in basically all investigated properties and behaviors, giving the proposed interpretations and underlying results more credibility. The simulation system was successfully applied to approach different questions of the evolution of early metabolism, its properties and components. However, some problems were left unsolved. The scope of this model did not allow to ask for the origin of key components such as ribozymes, the genetic system and the protocell itself. The evolution of all three can be viewed in detail, evolutionary history of enzymes and metabolites were studied on the network level as well as for single instances, showing a great variety of evolutionary mechanisms at work. Nevertheless, the question about their emergence is off-topic per se, since they were assumed from the beginning. In the future, the system could be modeled to produce by itself the macromolecules that serve as catalysts, information carriers or membranes of the compartment. So instead of evolving these macromolecules in order to maximize a fitness function, they might adapt by developing exactly these components. One could also start off with a metabolism-only scenario, where catalytic elements would not be contained in a protocell but rather concentrate on a surface and instead of being transcribed from a genome simply replicate itself and affect the replication of other macromolecules. Such a system would allow a look even further back in prebiotic history and deliver potential answers to the origin and evolution of catalysts.

The investigation on the formation of metabolic pathways led to interesting results that make sense of the different and seemingly conflicting hypotheses that are around for many decades. Thereby, the metabolic pathways of the simulated networks were analyzed in terms of their evolutionary direction, i.e. how and where new chemical reactions were added to the metabolic pathway. This was then observed for the entire history of the simulation. It showed that the different evolutionary theories can be compatible when considering that different processes dominate different phases in the evolution of a metabolic system. For instance, it was found that forward evolution shapes metabolic network in the very early steps of evolution. In later and more complex stages where a certain reaction repertoire is reached, enzyme recruitment supersedes forward evolution, however, keeping a core set of pathways from the early phase. Backward evolution,on the other hand, could only be observed under conditions of steady environmental change. Several more hypotheses on metabolic evolution exist. A future task should be to address these alternative explanations. For instance, the thread followed by Pfeiffer stating that metabolism started with a few multi-functional but not very efficient enzymes evolving to a series of specialized and faster processing enzymes. Although the present model allows to differentiate degrees of specificity of enzymes, the extension could be increased. For

this, one would have to define certain starting conditions representing the situation in the hypothesis. The tools to analyze such changes in specificity already exist and have been used here.

A great strength of the introduced model is its flexibility, that enable testing of evolutionary scenarios with different conditions. In the study of the emergence of complex properties, three such scenarios were tested. One scenario describes a comparably static evolution with a fixed environment and low mutation rates. The second scenario resembles a system under steady environmental change, while the third scenario differs in mutation rate and horizontal gene transfer to the first. Already with this set of scenarios, interesting results could be gained from the simulations. Concerning network robustness, the two static evolutionary scenarios performed best for almost all measures. But it should be mentioned that the basis of these measure may be more favorable to the conditions of these two scenarios, somewhat neglecting the different strategy of the networks evolving with steady environmental change. Thus, even more diverse measures of robustness should be explored in future. Further, a higher degree of modularity in the scenarios underlying environmental change and horizontal gene transfer events was observed. However, the modularity of the two scenarios seemed to be quite different in the interconnectivity between modules. Networks that evolved with regular gene transfers showed the formation of isolated modules, whereas those networks under steady environmental change had strongly interconnected or overlapping modules. A detailed analysis of single simulation runs to observe the actual network structure and properties is necessary to verify the statistical observations. A further study of chemical organization hierarchies is planned, as well as in another project, the investigation of the energy landscapes with the help of barrier tree has already been started with the aim of predicting the metabolic cost and, thus, the flexibility of switching between the different steady-states of a system.

Appendix A

Formats

A.1 SMILES

For the input and output of metabolites, a notation for chemical compounds is needed. SMILES, Simplified Molecular Input Line Entry System, provides a notation that is short and readable for chemists [174]. Another advantage of SMILES is that it is a unique notation, an important feature for the graph rewriting process where every newly generated metabolite is checked against the entire metabolite pool for graph isomorphism. An exhaustive procedure of graph isomorphism checking with a large number of graphs would lead to an explosion in computational cost. For each graph now only its SMILES notation has to be generated once and the graph isomorphism comes down to sequence comparison, which is significantly cheaper.

To generate a unique SMILES notation, the graph nodes have to be assigned with canonical labels first. This is accomplished through the CANON algorithm [175]. First, some graph invariants for each node are obtained, included are the number of connections of an atom, the number of non-hydrogen bonds, the atomic number and the number of hydrogen bonds. Based on these invariants, the atoms are sorted and ranked. If no two atoms share a rank, the labeling is done. Otherwise, the ranks are replaced by the product of primes corresponding to the neighbors' ranks and the atoms are sorted and ranked again. In some very symmetrical molecule graphs, it can happen that some atoms are always ranked the same, then this tie has to be broken and one atom is chosen to be in a lower rank. Uniqueness is still ensured. After the canonical labeling is derived, the final unique SMILES string is generated through a depth-first search through the metabolite graph [175], starting with the node having the minimal canonic label. The depth-first search branches always to the node with the lower label, except in a ring it is branched toward a multiple bond if existing, this is basically done for the clarity of the notation. In Figure A.1 two examples are shown, the upper one illustrating the branching problem and the lower depicting the generation of a unique SMILES in case of a cyclic and completely symmetric molecule on the example of cubane.

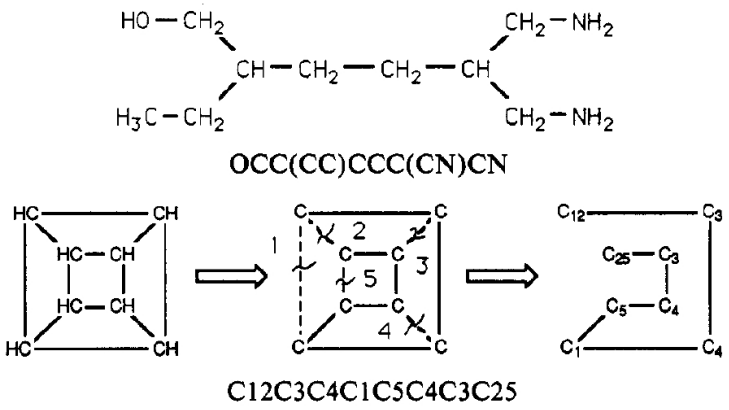


Figure A.1: Examples for the generation of unique SMILES, from [175]

A.2 GML

For the input of the set of chemical reaction graphs, constituting the chemistry of the system, an appropriate format is needed. The `GML`, Graph Modelling Language [78], is a flexible format with a simple syntax, meeting all requirements on a graph input format. `GML` is in `ASCII` representation, thus, ensuring portability regarding the platform and simplicity concerning the use of parsers. Furthermore, the structure of `GML` can be regarded simple. It consists basically of hierarchical key-value lists, where keys are alphanumeric characters, such as `graph` or `node`. Values can be integers, float numbers, strings, or another key-value list and can contain attributes. Further attributes and keys can be specified since `GML` intends to be implemented for many different data-structures.

For the modeling of graphs, there exist three different keys. The top level key -`graph`- containing the other two, `node` and `edge`. For the purposes of the simulation model, the syntax of `GML` was extended. Instead of `graph` the new key `rule` is used to define a reaction graph. The node `rule` always contains the keys `context`, `left` and `right`, all of which are newly added. In `GML`, graphs can contain keys `node` and `edge`. The topological structure is modeled through the `node id`'s used in the `edge`'s `source` and `target`. The reaction graphs are defined in the same way, but `node` keys are exclusively listed in `context`. Both, `left` and `right` contain `edge` keys. All `node`'s have the attribute `label` representing the atom type, whereas, `edge`'s are labeled with the bond type of the respective edge in the reaction graph. Below, the Diels-Alder reaction is shown in the changed `GML` format. The `context` specifies the atoms of the pericyclic reaction; `left` defines the edges of the graph resembling the substrate molecule; `right`, accordingly, defines the product molecule graph.

```
# ID 414141404140
rule [
     context [
             node [ id 0 label "C" ]
             node [ id 1 label "C" ]
             node [ id 2 label "C" ]
             node [ id 3 label "C" ]
             node [ id 4 label "C" ]
             node [ id 5 label "C" ]
     ]
     left [
          edge [ source 0 target 1 label "=" ]
          edge [ source 1 target 2 label "-" ]
          edge [ source 2 target 3 label "=" ]
          edge [ source 4 target 5 label "=" ]
     ]
     right [
           edge [ source 0 target 1 label "-" ]
           edge [ source 1 target 2 label "=" ]
           edge [ source 2 target 3 label "-" ]
           edge [ source 3 target 4 label "-" ]
           edge [ source 4 target 5 label "-" ]
           edge [ source 5 target 0 label "-" ]
     ]
]
```

Appendix B

Topological Indices

A topological index is a number characterizing the constitution of a graph [160]. The value of the index does not depend on the labeling of the graph or the way it is presented, thus, it can also be seen as a graph invariant. There are several groups of indices. One includes indices that are based on the vertex or edge connectivity such as the Zagreb and Connectivity Index. Another group contains indices that rely on distance information, such as the Wiener number, the Platt number and the Balaban index. All indices are supposed to resemble a different chemical or graph-theoretical property.

Topological indices are interesting for Quantitative structure-activity relationship (QSAR), Quantitative structure-property relationship (QSPR) and other applications in the pharmaceutical industry [103]. QSAR, QSPR are methods to correlate the structure of a chemical molecule with its biological activity or property, respectively. Therefore, graph indices are the appropriate tools in these areas since they are simple descriptors that do not need empirically derived measurements and that are rapidly computed. This proves to be important for drug design which perform QSAR studies of billions of structures before the actual synthesis of the designed drug targets. Here, they are used as similarity measures for molecules and in the case of the Wiener number also for a heuristic energy calculation.

B.1 Zagreb Index

The Zagreb index is a topological index based on the connectivity [160]. It was defined such that it correlates with the π-electron energy. There are actually two Zagreb indices: the original one, M_1, can be seen in equation B.1, and the one used in the metabolite evaluation of the simulation, M_2, in Equation B.2, with $D(i)$ being the valency of the vertex i.

$$M_1 = \sum_i D^2(i) \tag{B.1}$$

$$M_2 = \sum_{\{i,j\}} D(i)\, D(j) \tag{B.2}$$

B.2 Connectivity Index

The Connectivity index [133] is calculated in a similar fashion as the Zagreb index. It also uses the valency of all vertices (Equation B.4). Despite the similarity between the two indices, the Connectivity index was supposed to be an indicator of molecular branching of a chemical structure, as opposed to the π-electron energy as is the case for the Zagreb index. However, it is also correlated with the π-electron energy of its metabolite. There are two more Connectivity indices that are commonly used. The original Connectivity index can be seen as of first-order and sums over all bonds in the molecule graph, the other two are the Connectivity index of zero order over all vertices(Equation B.3) and second-order over all paths of length two (Equation B.5).

$${}^0\chi = \sum_i [D(i)]^{-1/2} \tag{B.3}$$

$${}^1\chi = \sum_{\{i,j\}} [D(i)\, D(j)]^{-1/2} \tag{B.4}$$

$${}^2\chi = \sum_{\{i,j,k\}} [D(i)\, D(j)\, D(k)]^{-1/2} \tag{B.5}$$

B.3 Wiener Number

The Wiener number was one of the first indices based on distances and was first introduced as path number [177], which is the number of bonds in a molecule. The Wiener number is calculated using the distance matrix D, as in Equation B.6. It is an indicator for the compactness of a molecule graph, as is illustrated in Figure B.1. In combination with other graph invariants, such as the polarity number p (equation B.7 with p_3 being the number of paths with length 3), it can also indicate other physical properties of molecules, e.g. boiling point.

$$W = (1/2) \sum_k \sum_l (D)_{kl} \tag{B.6}$$

$$p = (1/2) \sum_i (p_3)_i \tag{B.7}$$

B.4 Platt Number

The Platt Number [131], although, being a distance based index, is rather similar to the two indices described first. It is defined by the sum of edge-degrees in the molecule graph, as in

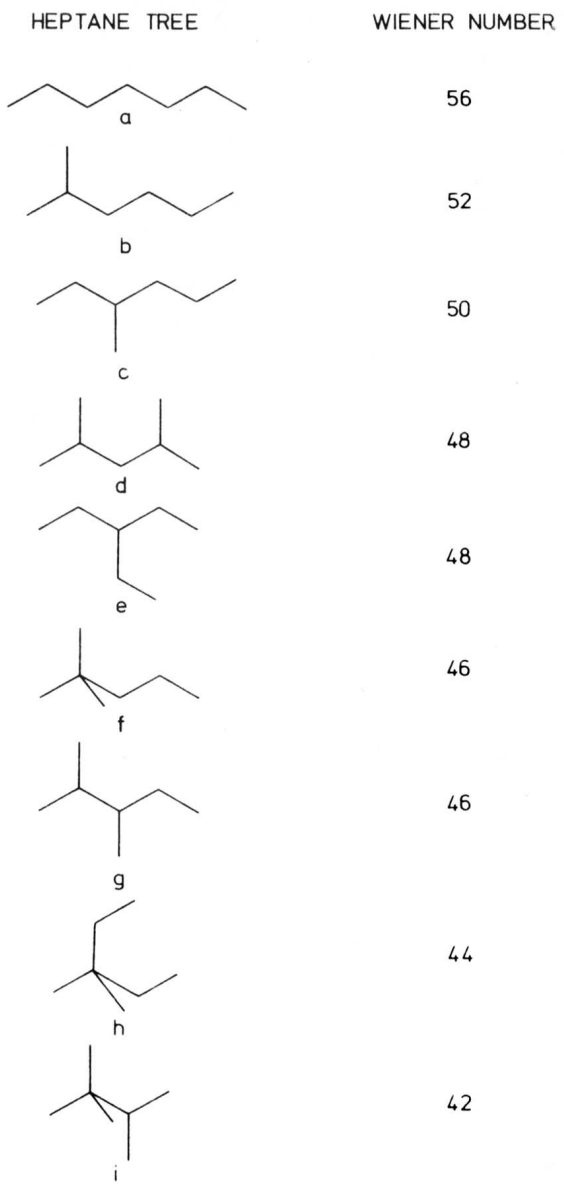


Figure B.1: Ordering of heptane trees based on their Wiener number [160]

equation B.8 with $D(e)$ being the edge-degree of edge e. Since $D(e)$ is the number of adjacent edges of e, it can be easily obtained because the graph interface used in the simulation provides a function. Thus, it is not necessary to calculate it as in Equation B.9, where $D(i)$ is the valency of vertex i, which shows the similarity to Zagreb and Connectivity index. The Platt number was intended as a measure for molar volume and some other physical properties as the heat of formation or vaporization.

$$F = \sum_{i} D(e_i) \tag{B.8}$$

$$F = \sum_{\{i,j\}} [D(i) + D(j) - 2] \tag{B.9}$$

B.5 Balaban Index

The Balaban index is defined as the average distance sum connectivity [4]. It is calculated by Equation B.10, where d_i is the distance sum for vector i to all other vertices, M is the number of edges and μ is the cyclomatic number, thus, $\mu = M - N + 1$ with N being the number of vertices in the graph. For the sample graph in Figure B.2, the Balaban index can be computed from $M = 9$, $N = 8$, $\mu = 2$ and the distance sums $d_1 = 14$, $d_2 = 16$, $d_3 = 16$, $d_4 = 14$, $d_5 = 12$, $d_6 = 16$, $d_7 = 16$, $d_8 = 12$, resulting in $J = 1.9215$. The distance sums itself can be used as index being a measure for compactness of the respective area of the vertex. The Balaban index, however, is an indicator for the ramification of the molecule graph.

$$J = \frac{M}{\mu + 1} \sum_{\{i,j\}} (d_i d_j)^{-1/2} \tag{B.10}$$

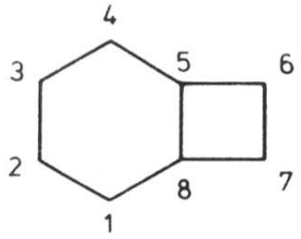


Figure B.2: Example graph for the calculation of the Balaban index [160]

Appendix C

Computer Programs

C.1 Manual - SimCell

Usage

SimCell [OPTIONS]

Description

SimCell performs an agent-based simulation of an *in silico* evolution of reaction networks, starting from user-defined sets of graph-rewrite rules (`GML`) and graphs (`SMILES`). Agents are so-called protocells with a randomly initiated genome (String of flexible Size), a set of genes (String of fixed length), a subset of the user-defined graph-rewrite rules and a reaction network. Reaction networks are generated in a stochastic simulation procedure, alternating for several rounds between the application of the graph-rewrite rules of the protocell in every possible way to the given graphs and a reduction step discarding graphs with concentrations below a certain multiplicity. This multiplicity or concentration is calculated through randomly choosing one of the produced reactions with a probability based on its reaction rate. The resulting reaction networks are evaluated through a flux balance analysis, defining the fitness of the protocells. A certain number of protocells is selected based on this fitness value. The selected protocells create copies of themselves, applying genetic operations to the copy's genome. The entire procedure is performed for several generations.

Options

```
--rulefile            File with GML style graph rewriting rules
                      Default=rules.gml
```

```
--metabolitefile      File with molecules in SMILES format
                      Default=metabolites.dat
--maxcells            Maximal number of cells
                      Default=2
--startcells          Number of initial cells
                      Default=1
--runs                Number of simulations
                      Default=1
--generations         Number of generations
                      Default=100
--iterations          Number of iterations per generation
                      Default=10
--mcsteps             Number of monte carlo steps per iteration
                      Default=10
--network_generation  Create network graphs
                      Default=true
--print_network       Write network files
                      Default=false
--flux_computation    Compute flux distribution of the network
                      Default=true
--print_fluxes        Write elementary modes in files
                      Default=false
--analyse_network     Analyse basic network properties
                      Default=false
--print_net_analysis  Write network properties in files
                      Default=false
--analyse_robustness  Compute robustness of network
                      Default=false
--print_rob_analysis  Write robustness measure in files
                      Default=false
--analyse_flux_evol   Analyse direction of pathway evolution
                      Default=false
--print_flux_evol     Write pathway evolution results in files
                      Default=false
--analyse_flux_sim    Compute pathway similarities
                      Default=false
--print_flux_sim      Write pathway similarity results in files
                      Default=false
--print_gene_hist     Write gene phylogeny in a file
                      Default=false
--background          Apply all reactions as background
```

```
                    Default=false
--all_reactions     Apply all reactions as catalyzed
                    Default=false
--test              0 - simulation, 1 - network generation
                    Default=0
--threshold         Similarity threshold for flux-clustering
                    Default=0.5
--mutation          Mutations per generation
                    Default=1.0
--duplication       Duplications per Mutation
                    Default=0.1
--gene_transfer     Horizontal gene transfer per generation
                    Default=0.0
--seed              Random number generator seed, 0-random
                    Default=0
```

Requirements

For the use of this software the following packages and libraries have to be installed on the computer.

```
ViennaRNA   RNA Secondary Structure Prediction Package
            http://www.tbi.univie.ac.at/~ivo/RNA/
Boost       C++ source library
            http://www.boost.org/
lp_solve    Mixed Integer Linear Programming solver
            http://lpsolve.sourceforge.net/5.0/
GGL         Graph Grammar Library
            http://www.tbi.univie.ac.at/~xtof/software/GGL/
```

Files

The following files can be specified by the user to configure the simulation, subsequent analysis and output.

Inputfile The inputfile is for the comfortable input of all the options stated above, avoiding the effort of adding options through the command line. Each options has to be written on a separate line, starting with the option name followed by the value for this option separated with a space or tabulator.

```
--option  value
--option2 value2
```

Rulefile The rulefile specifies the universe of graph-rewriting rules (chemical reactions) which can be applied on the graphs (molecules) in the protocells during the simulation. A graph- rewrite rule is a graph in `GML` format with a context, a left side and a right side. The rule's id precedes the rule description.

```
# ID 404040-
rule [
     context [
             node [ id 0 label "C" ]
             node [ id 1 label "C" ]
             node [ id 2 label "C" ]
     ]
     left [
          edge [ source 1 target 2 label "-" ]
     ]
     right [
           edge [ source 0 target 1 label "-" ]
           edge [ source 2 target 0 label "-" ]
     ]
]
# ID 204040
rule [
     context [
             node [ id 0 label "O" ]
             node [ id 1 label "C" ]
             node [ id 2 label "C" ]
     ]
     left [
          edge [ source 0 target 1 label "-" ]
          edge [ source 2 target 0 label "-" ]
     ]
     right [
           edge [ source 1 target 2 label "-" ]
     ]
]
```

Metabolitefile The metabolite file lists all the graphs (molecules) that are available at all times for all cells in `SMILES` notation. Each `SMILES` is written on a separate line. For simu-

lations with change in the environment, several sets can be specified in one file. For this, the generation in which the respective set is supposed to be introduced has to be stated.

```
#50
C=CO
O=C1C=CC(=O)O1
C1=CCCC=C1
#100
C1=CC=C1
C=CC(=C)C
```

Output

Simcell produces several files and outputs, described in the following paragraphs.

Networkfiles The network file contains the graph representation of one metabolic network in the GraphML format. If the option –print_network is enabled, networkfiles for the metabolic networks of all protocells are generated in every generation. Furthermore, list files for each protocell keeping track of all network files corresponding to the respective protocell and an overall list file showing the networks with the highest fitness value for each generation.

```
<graphml>
  <key id="key0" for="node" attr.name="caption" attr.type="string"/>
  <key id="key1" for="node" attr.name="concentration" attr.type="string"/>
  <key id="key2" for="edge" attr.name="label" attr.type="string"/>
  <key id="key3" for="edge" attr.name="ulabel" attr.type="string"/>
  <graph id="G" edgedefault="directed">
    <node id="n0">
      <data key="key0">C</data>
      <data key="key1">6</data>
    </node>
      .
      .
      .
    <edge id="e73" source="n71" target="n4">
      <data key="key2">25</data>
      <data key="key3">O=C(O)C(C(=O)O)C(=O)O4020400=C(O)C(O)C(=O)OC=O</data>
    </edge>
  </graph>
</graphml>
```

Fluxfile The fluxfile comprises of all elementary modes of a given metabolic network, where each row corresponds to an elementary mode and each column to a reaction. This is followed

by specifying which reactions are reversible (rev 0) or irreversible (rev 1) and whether they are input (exc 1), output (exc -1) or internal reactions (exc 0). If the option –print_fluxes is enabled, fluxfiles for the metabolic networks of all protocells are generated in every generation.

```
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 -1 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 -1 -1 0 0 0 0 0 0
0 0 0 0 0 0 -1 1 -1 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 -1 0 0 0 0 0 0 0 0
0 0 0 0 1 -1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
0 0 0 1 -1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
0 0 1 -1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 -1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 -1
0 0 0 0 0 0 1 -1 1 0 0 0 1 -1 0 0 0 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 0 0 0 0 -1 0 -1 0 0 0 0 0 0 0 0 0
0 -1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 -1 0 0 0 0 -1 0 0 0 0 0 0 -1 1 0 0 0 0
0 0 0 0 0 0 0 0 0 0 1 -1 0 0 0 -1 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0 -1 1 0 -1 0 1 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 -1 1 -1 0 0 0 0 0 0 0 0 0 -1 0 0 0
1 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 1 -1 1 0 0 0
rev 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1
exc 1 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 -1 0 -1 -1
```

C.2 Manual - ElPathway

Usage

ElPathway [OPTIONS] file

Description

ElPathway computes the set of elementary modes from a stoichiometric matrix of a metabolic network. It performs the binary null-space approach and some reduction steps. Optionally, ElPathway can calculate the minimal knockout sets from the set of elementary modes from the previous computation (–ko) or from user input (–onlyko), using an improved depth-first Berge algorithm.

Options

```
--ko       Compute elementary modes + minimal knockout sets
```

```
--onlyko   Compute only minimal knockout sets
--target   Target reaction for MKS computation
           Default=last reaction
```

Input

By default, for the computation of the set of elementary modes, the inputfile should contain a stoichiometric matrix, where the stoichiometric coefficients for one metabolite are in the rows, that of reactions are in columns. The stoichiometric matrix should be followed by a line indicating which reactions are reversible or irreversible and one line with information whether reactions are input, output or inner reactions. If this information is missing, all reactions are considered irreversible and the stoichiometric matrix is analyzed to determine potential input or output reactions. If no such reactions are found, the first reaction is defined as input and the last reaction as output reaction.

```
1 -1 -1 0 0 0 0
0 1 0 -1 -1 -1 0
0 0 1 0 1 -1 0
0 0 0 0 0 1 -1
rev 1 1 1 1 0 1 1
exc 1 0 0 0 0 0 -1
```

If knockout sets are computed from the user input, the inputfile has to be a set of elementary modes (see Output) and the information about reversibility and transport reactions.

Output

The standard output is the set of elementary modes, where each elementary mode is written on one line with entries for all reactions.

```
1 1 0 1 0 0 0
2 2 0 0 1 1 1
2 0 2 0 -1 1 1
2 1 1 0 0 1 1
1 0 1 1 -1 0 0
```

If one of the options –ko or –onlyko is selected, than the output contains the minimal knockout sets of the network. Each line represents one minimal knockout set and shows the column number of the included reactions in the stoichiometric matrix.

```
1
6
7
2 3
2 5
3 5
```

List of Figures

List of Tables

Bibliography

[1] Albert, R., Jeong, H., and Barabási, A. Error and attack tolerance of complex networks. *Nature 406*, 6794 (2000), 378–382.

[2] Almonacid, D. E., Yera, E. R., Mitchell, J. B., and Babbitt, P. C. Quantitative comparison of catalytic mechanisms and overall reactions in convergently evolved enzymes: implications for classification of enzyme function. *PLoS Computational Biology 6*, 3 (2010).

[3] Bagley, R. J., and Farmer, J. D. Spontaneous emergence of a metabolism. In *Artificial Life II* (Redwood City, CA, 1992), C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen, Eds., Santa Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, pp. 93–141.

[4] Balaban, A. T. Highly discriminating distance-based topological index. *Chemical Physicls Letters 89* (1982), 399–404.

[5] Banerjee, A., and Jost, J. Spectral plot properties: Towards a qualitative classification of networks. In *In European Conference on Complex Systems* (2007).

[6] Banzhaf, W., Dittrich, P., and Eller, B. Self-organization in a system of binary strings with spatial interactions. *Physica D 125* (1999), 85–104.

[7] Barabási, A., and Albert, R. Emergence of scaling in random networks. *Science 286*, 5439 (1999), 509.

[8] Barabasi, A.-L., Albert, R., and Jeong, H. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications 281*, 1-4 (June 2000), 69–77.

[9] Barabási, A. L., and Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics 5*, 2 (Feb. 2004), 101–113.

[10] Bauer, F. *Eigenvalues of directed and undirected graphs and their applications.* PhD thesis, Universität Leipzig, 2011.

[11] Behre, J., Wilhelm, T., von Kamp, A., Ruppin, E., and Schuster, S. Structural robustness of metabolic networks with respect to multiple knockouts. *Journal of Theoretical Biology 252* (2008), 433–41.

[12] Benkö, G., Flamm, C., and Stadler, P. F. Generic properties of chemical networks: Artificial chemistry based on graph rewriting. In *Advances in Artificial Life* (Heidelberg, Germany, 2003), W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, and J. Ziegler, Eds., vol. 2801 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 10–20. Proceedings of the 7th European Conference of Artifical Life, ECAL 2003, Dortmund, Germany, September 14-17, 2003.

[13] Benkö, G., Flamm, C., and Stadler, P. F. A graph-based toy model of chemistry. *Journal of Chemical Information and Computer Science 43* (2003), 1085–93.

[14] Benkö, G., Flamm, C., and Stadler, P. F. Multi-phase artificial chemistry. In *The Logic of Artificial Life: Abstracting and Synthesizing the Principles of Living Systems* (Berlin, 2004), H. Schaub, F. Detje, and U. Brüggemann, Eds., IOS Press, pp. 16–22. Proceedings of GWAL, Bamberg 14-16 April 2004.

[15] Bernal, J. *The Origin of Life*. W. Clowes and Sons, London, 1967.

[16] Blank, L. M., Kuepfer, L., and Sauer, U. Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome biology 6*, 6 (Jan. 2005), 49.

[17] Boldhaus, G., and Klemm, K. Regulatory networks and connected components of the neutral space. *The European Physical Journal B 77*, 2 (June 2010), 233–237.

[18] Bornscheuer, U. T., and Kazlauskas, R. J. Catalytic promiscuity in biocatalysis: using old enzymes to form new bonds and follow new pathways. *Angewandte Chemie (Int. Ed. Engl.) 43*, 45 (2004), 6032–40.

[19] Brittain, D. R. B., Lin, C. Y., Gilbert, A. T. B., Izgorodina, E. I., Gill, P. M. W., and Coote, M. L. The role of exchange in systematic DFT errors for some organic reactions. *Physical Chemistry Chemical Physics 11* (2009), 1138–1142.

[20] Caetano-Anollés, G., Yafremava, L. S., Gee, H., Caetano-Anollés, D., Kim, H. S., and Mittenthal, J. E. The origin and evolution of modern metabolism. *International Journal Biochemistry & Cell Biology 41* (2009), 285–297.

[21] Cayley, A. On the mathematical theory of isomers. *Philosophical Magazine 47* (1874), 444–446.

[22] Centler, F., Kaleta, C., di Fenizio, P. S., and Dittrich, P. Computing chemical organizations in biological networks. *Bioinformatics (Oxford, England) 24*, 14 (July 2008), 1611–8.

[23] Chen, I. A., Roberts, R. W., and Szostak, J. W. The emergence of competition between model protocells. *Science 305*, 5689 (2004), 1474–1476.

[24] Chen, X., Li, N., and Ellington, A. D. Ribozyme catalysis of metabolism in the RNA world. *Chemistry & Biodiversity 4* (2007), 633–655.

[25] Ciliberti, S., Martin, O. C., and Wagner, A. Innovation and robustness in complex regulatory gene networks. *Proceedings of the National Academy of Sciences USA 104* (2007), 13591–13596.

[26] Copley, R. R., and Bork, P. Homology among $(\beta\alpha)_8$-barrels: implications for the evolution of metabolic pathways. *Journal of Molecular Biology 303* (2000), 627–641.

[27] Copley, S. D. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Current Opinions in Chemical Biology 7*, 2 (Apr 2003), 265–72.

[28] Cordella, L. P., Foggia, P., Sansone, C., and Vento, M. Performance evaluation of the VF graph matching algorithm. In *Image Analysis and Processing (ICIAP)* (1999), pp. 1172–1177.

[29] Cordella, L. P., Foggia, P., Sansone, C., and Vento, M. An improved algorithm for matching large graphs. In *3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition* (2001), pp. 149–159.

[30] Cordon, F. *Tratado evolucionista de biologí*. Aguilar Ediciones, Madrid, Spain, 1990.

[31] Costa, L., Rodrigues, F., Travieso, G., and Boas, P. Characterization of complex networks: A survey of measurements. *Advances in Physics 56*, 1 (2007), 167–242.

[32] da Silva, M. Centrality, Network Capacity, and Modularity as Parameters to Analyze the Core-Periphery Structure in Metabolic Networks. *Proceedings of the IEEE 96*, 8 (Aug. 2008), 1411–1420.

[33] Darwin, C. *The origin of species*. Random House, 1993.

[34] Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed. 1978, pp. 345–352.

[35] de Figueiredo, L. F., Podhorski, A., Rubio, A., Kaleta, C., Beasley, J. E., Schuster, S., and Planes, F. J. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics (Oxford, England) 25*, 23 (Dec. 2009), 3158–65.

[36] di Fenizio, P. S. A less abstract articial chemistry. In *Artificial Life VII* (Cambridge, MA, 2000), M. Bedau, J. McCaskill, N. Packard, and S. Rasmussen, Eds., MIT Press, pp. 49–53.

[37] di Fenizio, P. S. *Chemical organisation theory.* PhD thesis, University of Jena, Computer Science Department, 2007.

[38] Dittrich, P., and di Fenizio, P. S. Chemical organisation theory. *Bulletin of Mathematical Biology 69* (2007), 1199–1231.

[39] Dittrich, P., Ziegler, J., and Banzhaf, W. Artificial chemistries–a review. *Artifical Life 7* (2001), 225–275.

[40] Eakin, R. An approach to the evolution of metabolism. *Proceedings of the National Academy of Sciences of the United States of America 49*, 3 (1963), 360–6.

[41] Ebner, M., Shackleton, M., and Shipman, R. How neutral networks influence evolvability. *Complexity 7* (2001), 19.

[42] Eigen, M., and Schuster, P. *The Hypercycle, a Principle of Natural Self-Organization.* Springer-Verlag, 1979.

[43] Ellington, A. D., and Szostak, J. W. In vitro selection of rna molecules that bind specific ligands. *Nature 346*, 6287 (1990), 818–822.

[44] Erdos, P., and Renyi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci 5* (1960), 17–61.

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

[46] Faulon, J.-L., and Sault, A. G. Stochastic generator of chemical structure. 3. Reaction network generation. *Journal of Chemical Information and Computer Science 41* (2001), 894–908.

[47] Félix, L., Rosselló, F., and Valiente, G. Efficient reconstruction of metabolic pathways by bidirectional chemical search. *Bulletin of Mathematical Biology 71* (2009), 750–769.

[48] Flamm, C., Endler, L., Müller, S., Widder, S., and Schuster, P. A minimal and self-consistent *in silico* cell model based on macromolecular interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences 362* (2007), 1831–1839.

[49] Fontana, W. Algorithmic chemistry. In *Artificial Life II* (Redwood City, CA, 1992), C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen, Eds., Addison-Wesley, pp. 159–210.

[50] Fontana, W., and Buss, L. W. What would be conserved if 'the tape were played twice'? *Proceedings of the National Academy of Sciences USA 91* (1994), 757–761.

[51] Fontana, W., Konings, D. A., Stadler, P. F., and Schuster, P. Statistics of RNA secondary structures. *Biopolymers 33* (Sep 1993), 1389–404.

[52] Fontana, W., Schnabl, W., and Schuster, P. Physical aspects of evolutionary optimization and adaption. *Physical Review A 40* (1989), 3301–3321.

[53] Fontana, W., and Schuster, P. Continuity in evolution: On the nature of transitions. *Science 280* (1998), 1451–1455.

[54] Fontana, W., and Schuster, P. Shaping space: the possible and the attainable in rna genotype-phenotype mapping. *Journal of theoretical biology. 194* (1998), 491.

[55] Fontana, W., Stadler, P. F., Tarazona, P., Weinberger, E. D., and Schuster, P. RNA folding and combinatory landscapes. *Physical Review E 47* (1993), 2083–2099.

[56] Forst, C. V., and Schulten, K. Evolution of metabolisms: A new method for the comparison of metabolic pathways using genomics information. *Journal of Computational Biology 6*, 3/4 (1999).

[57] Fothergill-Gilmore, L. A., and Michels, P. A. Evolution of glycolysis. *Progress in Biophysics and Molecular Biology 59*, 2 (1993), 105–235.

[58] Freilich, S., Kreimer, A., Borenstein, E., Gophna, U., Sharan, R., and Ruppin, E. Decoupling Environment-Dependent and Independent Genetic Robustness across Bacterial Species. *PLoS computational biology 6*, 2 (Feb. 2010).

[59] Fujita, S. Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. *Journal of Chemical Information and Computer Science 26* (1986), 205–212.

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

[61] Gasteiger, J., Ed. *Handbook of Chemoinformatics: From Data to Knowledge*, vol. 1. Wiley-VCH, August 2003.

[62] Gasteiger, J., and Engel, T. *Chemoinformatics: a textbook*. Wiley-VCH, 2003.

[63] Gerlt, J. A., and Babbitt, P. C. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annual Reviews in Biochemistry 70* (2001), 209–246.

[64] Gesteland, R. F., Cech, T. R., and Atkins, J. F. *The RNA World*, 3rd ed. Cold Spring Harbor Laboratories Press, Woodbury, NY, 2006.

[65] Gibson, M. A., and Bruck, J. Efficient exact stochastic simulation of chemical systems with many species and many channels. *Journal of Physical Chemistry A 104* (2000), 1876–1889.

[66] Gilbert, W. Origin of life: The RNA world. *Nature 319* (1986), 618.

[67] Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry 81*, 25 (1977), 2340–2361.

[68] Gillespie, R. J., and Nyholm, R. S. Inorganic Stereochemistry. *Quarterly Reviews of the Chemical Society 11* (1957), 339–380.

[69] Granick, S. Speculations on the origins and evolution of photosynthesis. *Annals of the New York Academy of Sciences 69* (1957), 292–308.

[70] Grzybowski, B. A., Bishop, K. J. M., Kowalczyk, B., and Wilmer, C. E. The 'wired' universe of organic chemistry. *Nature Chemistry 1* (2009), 31–36.

[71] Handorf, T., Ebenhöh, O., and Heinrich, R. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *Journal of molecular evolution 61*, 4 (Oct. 2005), 498–512.

[72] Heidrich, D., Kliesch, W., and Quapp, W. *Properties of Chemically Interesting Potential Energy Surfaces*, vol. 56 of *Lecture Notes in Chemistry*. Springer-Verlag, Berlin, 1991.

[73] Hendrickson, J. B. Comprehensive system for classification and nomenclature of organic reactions. *Journal of Chemical Information and Computer Science 37* (1997), 852–860.

[74] Hendrickson, J. B., and Miller, T. M. Reaction indexing for reaction databases. *Journal of Chemical Information and Computer Science 30* (1990), 403–408.

[75] Henikoff, S., and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences 89*, 22 (Nov. 1992), 10915–10919.

[76] Herges, R. Coarctate transition states: The discovery of a reaction principle. *Journal of Chemical Information and Computer Science 34* (1994), 91–102.

[77] Herges, R. Organizing principle of complex reactions and theory of coarctate transition states. *Angewandte Chemie (Int. Ed. Engl.) 33* (1994), 255–276.

[78] Himsolt, M. *GML: A portable Graph File Format*. Universität Passau.

[79] Hofacker, I. L., Fontana, W., F, S. P., Bonhoeffer, S., Tacker, M., and Schuster, P. Fast folding and comparison of RNA secondary structures. *Chemical Monthly 125* (1994), 167–188.

[80] Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *Journal of Chemical Physics 39* (1963), 1397–1412.

[81] Horowitz, N. H. On the evolution of biochemical syntheses. *Proceedings of the National Academy of Sciences USA 31* (1945), 153–157.

[82] Hrmova, M., De Gori, R., Smith, B. J., Fairweather, J. K., Driguez, H., Varghese, J. N., and Fincher, G. B. Structural basis for broad substrate specificity in higher plant beta-D-glucan glucohydrolases. *Plant Cell 14* (2002), 1033–1052.

[83] Hult, K., and Berglund, P. Enzyme promiscuity: Mechanism and applications. *Trends in Biotechnology 25*, 5 (May 2007), 231–8.

[84] Huynen, M. A. Exploring phenotype space through neutral evolution. *Journal of Molecular Evolution 43* (1996), 165.

[85] Huynen, M. A., Stadler, P. F., and Fontana, W. Smoothness within ruggedness: the role of neutrality in adaptation. *Proceedings of the National Academy of Sciences USA 93* (1996), 397–401.

[86] Imielinski, M., and Belta, C. Exploiting the pathway structure of metabolism to reveal high-order epistasis. *BMC systems biology 2* (Jan. 2008), 40.

[87] Jain, S., and Krishna, S. Autocatalytic sets and the growth of complexity in an evolutionary model. *Physical Review Letters 81*, 25 (1998), 5684–5687.

[88] Jain, S., and Krishna, S. Crashes, recoveries, and core shifts in a model of evolving networks. *Physical Review E 65*, 2 (Jan. 2002), 20–23.

[89] Jensen, R. A. Enzyme recruitment in evolution of new function. *Annual Reviews in Microbiology 30* (1976), 409–425.

[90] Kanehisa, M., and Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research 28*, 1 (Jan. 2000), 27–30.

[91] Kashtan, N., and Alon, U. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences USA 102*, 39 (2005), 13773–13778.

[92] Kauffman, K. J., Prakash, P., and Edwards, J. S. Advances in flux balance analysis. *Current Opinion in Biotechnology 14* (2003), 491–496.

[93] Kauffman, S. *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press, May 1993.

[94] Kaufmann, M., and Wagner, D., Eds. *Drawing Graphs, Methods and Models (the book grow out of a Dagstuhl Seminar, April 1999)* (2001), vol. 2025 of *Lecture Notes in Computer Science*, Springer.

[95] Khersonsky, O., Roodveldt, C., and Tawfik, D. S. Enzyme promiscuity: evolutionary and mechanistic aspects. *Current Opinions in Chemical Biology 10*, 5 (Oct 2006), 498–508.

[96] Khersonsky, O., and Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annual Reviews in Biochemistry 79* (July 2010), 471–505.

[97] Kim, J., Kershner, J. P., Novikov, Y., Shoemaker, R. K., and Copley, S. D. Three serendipitous pathways in e. coli can bypass a block in pyridoxal-5'-phosphate synthesis. *Molecular Systems Biology 6* (Nov 2010), 436.

[98] Kimura, M. *Nature 217* (1968), 624.

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

[100] Klipp, E. *Systems biology in practice: concepts, implementation and application.* Wiley-VCH, 2005.

[101] Klukas, C., and Schreiber, F. Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics 23*, 3 (Feb. 2007), 344–350.

[102] Kreimer, A., Borenstein, E., Gophna, U., and Ruppin, E. The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America 105*, 19 (May 2008), 6976–81.

[103] Krier, L., and Hall, L. Molecular connectivity in chemistry and drug research. *Academic Press* (1976).

[104] Kun, A., Papp, B., and Szathmary, E. Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. *Genome Biology 9*, 3 (2008), 51.

[105] Lawrence, J. G., and Roth, J. R. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics 143*, 4 (Aug. 1996), 1843–1860.

[106] Lenski, R. E., Ofria, C., Pennock, R. T., and Adami, C. The evolutionary origin of complex features. *Nature 423*, 6936 (May 2003), 139–44.

[107] Li, Y., de Ridder, D., de Groot, M. J. L., and Reinders, M. J. T. Metabolic pathway alignment between species using a comprehensive and flexible similarity measure. *BMC systems biology 2* (Jan. 2008), 111.

[108] Lipson, H., Pollack, J. B., and Suh, N. P. On the origin of modular variation. *Evolution 56*, 8 (2002), 1549–1556.

[109] Margulis, L., and Sagan, D. *What is life?* University of California Press, 2000.

[110] Martin, W., Baross, J., Kelley, D., and Russell, M. J. Hydrothermal vents and the origin of life. *Nature Reviews Microbiology* (Sept. 2008).

[111] Mathews, D. H., Sabina, J., Zuker, M., and Turner, H. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *Journal of Molecular Biology 288* (1999), 911–940.

[112] McCaskill, J. S., and Niemann, U. Graph replacement chemistry for DNA processing. In *DNA Computing*, A. Condon and G. Rozenberg, Eds., vol. 2054 of *Lecture Notes in Computer Science*. Springer, Berlin, D, 2000, pp. 103–116.

[113] Mendel, G. Versuche über pflanzenhybriden. *Verhandlungen des Naturforschenden Vereines in Brünn IV* (1866), 3–47.

[114] Miller, S., and Orgel, L. *The origins of life on the earth*. Prentice-Hall Biological Science Series. Prentice-Hall, 1974.

[115] Miller, S. L. A production of amino acids under possible primitive earth conditions. *Science 117*, 3046 (1953), 528–529.

[116] Miller, S. L., and Urey, H. C. Organic compound synthes on the primitive eart. *Science 130*, 3370 (1959), 245–251.

[117] Morowitz, H. J. A theory of biochemical organization, metabolic pathways, and evolution. *Complexity 4* (1999), 39–53.

[118] Müller, U. F. Re-creating an RNA world. *Cellular and Molecular Life Sciences 63* (2006), 1278–1293.

[119] Nacher, J., Ueda, N., Yamada, T., Kanehisa, M., and Akutsu, T. Study on the Clustering Coefficients in Metabolic Network Using a Hierarchical Framework. In *International workshop on bioinformatics and systems biology* (2004), pp. 34–35.

[120] Nagl, M. *Graph-Grammatiken, Theorie, Implementierung, Anwendung*. Vieweg, Braunschweig, 1979.

[121] Nussinov, M. Formation of the early earth regolith. *Nature 275* (1978), 19–21.

[122] O'Brien, P. J., and Herschlag, D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chemical Biology 6*, 4 (Apr 1999), 91–105.

[123] of Helicobacter pylori 26695 Christophe H. Schilling, G.-S. M. M., Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., and Palsson, B. O. Genome-scale metabolic model of helicobacter pylori 26695. *J. Bacteriol 184* (2002), 4582–4593.

[124] Oro, J. A novel synthesis of polypeptides. *Nature 186* (1960), 156 – 157.

[125] Orr, H. A. The evolutionary genetics of adaptation: a simulation study. *Genetics Research 74* (1999), 207–214.

[126] Ourisson, G., and Nakatani, Y. The terpenoid theory of the origin of cellular life: the evolution of terpenoids to cholesterol. *Chemical Biology 1*, 1 (1994), 11–23.

[127] Palsson, B. O. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, New York, NY, USA, 2006.

[128] Parter, M., Kashtan, N., and Alon, U. Environmental variability and modularity of bacterial metabolic networks. *BMC evolutionary biology 7* (Jan. 2007), 169.

[129] Petsko, G. A., Kenyon, G. L., Gerlt, J. A., Ringe, D., and Kozarich, J. W. On the origin of enzymatic species. *Trends in Biochemical Sciences 18*, 10 (1993), 372–6.

[130] Pfeiffer, T., Soyer, O. S., and Bonhoeffer, S. The evolution of connectivity in metabolic networks. *PLoS Biology 3* (2005), 228.

[131] Platt, J. Influence of neighbor bonds on additive bond properties in paraffins. *Journal of Chemical Physics 15* (1947), 419–420.

[132] Powner, M. W., Gerland, B., and Sutherland, J. D. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature 459*, 7244 (May 2009), 239–42.

[133] Randic, M. Characterization of molecular branching. *Journal of the Americal Chemical Society 97*, 23 (1975), 6609–6615.

[134] Ravasz, E., Somera, a. L., Mongru, D. a., Oltvai, Z. N., and Barabási, a. L. Hierarchical organization of modularity in metabolic networks. *Science (New York, N.Y.) 297*, 5586 (Aug. 2002), 1551–5.

[135] Read, R. C. Every one a winner. *Annals of Discrete Mathematics 2* (1978), 107–120.

[136] Regina, S., Christine, R., and Beate, S. Selex–a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomolecular Engineering 24*, 4 (2007), 381 – 403.

[137] Roberts, J. C. Exploratory visualization with multiple linked views. In *Exploring Geovisualization*, A. MacEachren, M.-J. Kraak, and J. Dykes, Eds. Amsterdam: Elseviers, December 2004.

[138] Robinson, R. Jump-starting a cellular world: Investigating the origin of life, from soup to networks. *PLoS Biol 3*, 11 (11 2005), 396.

[139] Rohrschneider, M., Heine, C., Reichenbach, A., Kerren, A., and Scheuermann, G. A novel grid-based visualization approach for metabolic networks with advanced focus and context view. In *Graph Drawing (GD 2009)* (Berlin, 2010), E. Gansner and D. Eppstein, Eds., vol. 5849 of *Lecture Notes in Computer Science*, Springer, pp. 268–279.

[140] Rohrschneider, M., Ullrich, A., Kerren, A., Stadler, P. F., and Scheuermann, G. Visual network analysis of dynamic metabolic pathways. In *Advances in Visual Computing (ISVC 2010)* (Berlin, 2010), G. Bebis, R. Boyle, B. Parvin, D. Koracin, R. Chung, R. Hammoud, M. Hussain, T. Kar-Han, R. Crawfis, D. Thalmann, D. Kao, and L. Avila, Eds., vol. 6453 of *Lecture Notes in Computer Science*, Springer, pp. 316–327.

[141] Rossellá, F., and Valiente, G. Chemical graphs, chemical reaction graphs, and chemical graph transformation. *Electronic Notes in Theoretical Computer Science 127* (2005), 157–166.

[142] Samal, A., Rodrigues, J. F. M., Jost, J., Martin, O. C., and Wagner, A. Genotype networks in metabolic reaction spaces. *BMC Systems Biology 4* (2010), 30.

[143] Schilling, C. H., Letscher, D., and Palsson, B. O. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of theoretical biology 203*, 3 (Apr. 2000), 229–48.

[144] Schmidt, S., Sunyaev, S., Bork, P., and Dandekar, T. Metabolites: a helping hand for pathway evolution? *Trends in Biochemical Sciences 28* (2003), 336–341.

[145] Schultes, E. A., and Bartel, D. P. *Science 289* (2000), 448.

[146] Schuster, P., Fontana, W., Stadler, P. F., and Hofacker, I. L. From sequences to shapes and back: A case study in RNA secondary structures. *Proceedings of the Royal Society B: Biological Sciences 255* (1994), 279–284.

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

[148] Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages (VL '96)* (1996), pp. 336–343.

[149] Stadler, B. M. R., Stadler, P. F., Wagner, G., and Fontana, W. The topology of the possible: Formal spaces underlying patterns of evolutionary change. *Journal of Theoretical Biology 213* (2001), 241–274.

[150] Stadler, P. F. Fitness landscapes arising from the sequence-structure maps of biopolymers. *Journal of Molecular Structure 463* (1999), 7–19.

[151] Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J., and Doyle, J. Robustness of cellular functions. *Cell 118*, 6 (Sept. 2004), 675–85.

[152] Stephan-Otto Attolini, C., Stadler, P. F., and Flamm, C. `CelloS`: a multi-level approach to evolutionary dynamics. In *Advances in Artificial Life: 8th European Conference, ECAL 2005* (2005), M. S. Capcarrere, A. A. Freitas, P. J. Bentley, C. G. Johnson, and J. Timmis, Eds., vol. 3630 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 500–509. Canterbury, UK, September 5-9, 2005.

[153] Suzuki, H., and Dittrich, P. Artificial chemistry. *Artificial Life 15* (2009), 1–3.

[154] Sylvester, J. J. On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics, with three appendices. *American Journal of Mathematics 1* (1878), 64–128.

[155] Talini, G., Gallori, E., and Maurel, M.-C. Natural and unnatural ribozymes: Back to the primordial RNA world. *Research in Microbiology 160* (2009), 457–465.

[156] Teichmann, S., Rison, S., Thornton, J., Riley, M., Gough, J., and Chothia, C. Small-molecule metabolism: an enzyme mosaic. *TRENDS in Biotechnology 19*, 12 (2001), 482–486.

[157] The MANET project. Pyrimidine Metabolism (map00240). `http://manet.illinois.edu/pathways.php`.

[158] Thürk, M. *Ein Modell zur Selbstorganisation von Automatenalgorithmen zum Studium molekularer Evolution*. PhD thesis, Universität Jena, Germany, 1993.

[159] Tollis, I. G., Di Battista, G., Eades, P., and Tamassia, R. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, July 1998.

[160] Trinajstic, N. *Chemical Graph Theory, Second Edition (New Directions in Civil Engineering)*. CRC, February 1992.

[161] Ugi, I., Stein, N., Knauer, M., Gruber, B., Bley, K., and Weidinger, R. New elements in the representation of the logical structure of chemistry by qualitative mathematical models and corresponding data structures. *Topics in Current Chemistry 166* (1993), 199–233.

[162] Ullman, J. An algorithm for subgraph isomorphism. *Journal of The ACM* (1976).

[163] Ullrich, A. Evolution of metabolism in a graph-based toy-universe. Diploma Thesis, Universität Leipzig, 2008.

[164] von Neumann, J. The general and logical theory of automata. In *Cerebral Mechanisms in Behaviour*, L. A. Jeffress, Ed. Wiley, 1951.

[165] Wächtershäuser, G. Before enzyme and templates: theory of surface metabolism. *Microbiological Reviews 52* (1988), 452–484.

[166] Wächtershäuser, G. Evolution of the first metabolic cycles. *Proc Natl Acad Sci USA 87* (1990), 200–204.

[167] Wächtershäuser, G. Origin of life: Life as we don't know it. *Science 289* (2000), 1307.

[168] Wagner, A. *Robustness and evolvability in living systems.* Princeton studies in complexity. Princeton University Press, 2007.

[169] Wagner, A. Robustness and evolvability: a paradox resolved. *Proceedings. Biological Sciences / The Royal Society 2758, 91–100* (2008).

[170] Wagner, G. P., and Altenberg, L. Complex Adaptations and the Evolution of Evolvability. *Evolution* (1996).

[171] Watson, J. D., and Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature 171*, 4356 (1953), 737–738.

[172] Watts, D. J., and Strogatz, S. H. Collective dynamics of small-world networks. *Nature 393*, June (1998), 440–442.

[173] Weberndorfer, G., Hofacker, I. L., and Stadler, P. F. On the evolution of primitive genetic codes. *Origins Life and Evolution of the Biosphere 33* (2003), 491–514.

[174] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Science 28* (1988), 31–36.

[175] Weininger, D., Weininger, A., and Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Science 29* (1989), 97–101.

[176] Whitacre, J. M. Degeneracy: a link between evolvability, robustness and complexity in biological systems. *Theoretical biology & medical modelling 7* (Jan. 2010), 6.

[177] Wiener, H. J. Structural determination of paraffin boiling points. *Journal of the Americal Chemical Society 69* (1947), 17–20.

[178] Wikipedia. Millerurey experiment — Wikipedia, the free encyclopedia, 2011. `http://en.wikipedia.org/wiki/Miller-Urey_experiment` [Online; accessed 07-July-2011].

[179] Wikipedia. Reaction coordinate — Wikipedia, the free encyclopedia, 2011. http://en.wikipedia.org/wiki/Reaction_coordinate [Online; accessed 07-July-2011].

[180] Wilke, C. O., and Adami, C. Evolution of mutational robustness. *Mutation Research Frontiers 522*, 1-2 (Jan. 2003), 3–11.

[181] Wodrich, M. D., Corminboeuf, C., Schreiner, P. R., Fokin, A. a., and von Ragué Schleyer, P. How accurate are DFT treatments of organic energies? *Organic Letters 9* (2007), 1851–1854.

[182] Ycas, M. On earlier states of the biochemical system. *Journal of Theoretical Biology 44* (1974), 145–160.

[183] Zhao, J., Ding, G.-H., Tao, L., Yu, H., Yu, Z.-H., Luo, J.-H., Cao, Z.-W., and Li, Y.-X. Modular co-evolution of metabolic networks. *BMC Bioinformatics 8* (2007), 311.

Curriculum vitae

Education:

Since 04/2008	PhD student at the University of Leipzig, Germany, Department of Computer Science, Bioinformatics Group
02/2008 - 11/2008	Scientific assistant at the University of Vienna, Austria, Institute for Theoretical Chemistry
10/2001 - 03/2008	Study of Computer Science and Bioinformatics at the University of Leipzig, Germany, Degree: Diploma
06/2006 - 09/2006	Scientific assistant at the State University of New York in Binghamton, USA, Computer Science Department
08/2005 - 05/2006	Study of Computer Science and Bioengineering at the State University of New York in Binghamton, USA
07/2000	Abitur in Leipzig, Germany

Scientific Publications

Ullrich A., Rohrschneider M., Scheuermann G., Stadler P.F. and Flamm C. (2011). In silico evolution of early metabolism. *Artificial Life*, Vol. 17:2 - Spring 2011

Rohrschneider M., **Ullrich A.**, Kerren A., Stadler P.F. and Scheuermann G. (2010). Visual Network Analysis of Dynamic Metabolic Pathways. Advances in Visual Computing, *Lecture Notes in Computer Science*, Vol. 6453

Ullrich A., Flamm C., Rohrschneider M. and Stadler P.F. (2010). In Silico Evolution of Early Metabolism. *Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems*, pp.57-64, MIT Press, 2010.

Flamm C., **Ullrich A.**, Ekker H., Mann M., Hgerl D., Rohrschneider M., Sauer S., Scheuermann G., Klemm K., Hofacker I.L. and Stadler P.F. (2010). Evolution of Metabolic Networks: A Computational Framework. *Journal of Systems Chemistry*, Vol. 1/1/4, 2010

Ullrich A. and Forst C.V. (2009). k-PathA: k-shortest Path Algorithm. *In Proceedings of High Performance Computational Systems Biology (HiBi 2009)*, pp. 23-30, 2009.

Ullrich A. and Flamm C. (2009). A Sequence-to-Function Map for Ribozyme-catalyzed Metabolisms. *ECAL, Lecture Notes in Computer Science (LNAI)*, Vol. 5778

Ullrich A. and Flamm C. (2008). Functional Evolution of Ribozyme-Catalyzed Metabolisms in a Graph-Based Toy-Universe. *CMSB, Lecture Notes in Computer Science (LNBI)*, Vol. 5307

Selected Scientific Presentations

In Silico Evolution of early Metabolism.
ALIFE XII, Odense (DK) 2010

k-PathA: k-shortest Path Algorithm.
HiBi, Trento (IT), 2009

A Sequence-to-Function Map for Ribozyme-catalyzed Metabolisms.
ECAL, Budapest (HU), 2009

Functional Evolution of Ribozyme-Catalyzed Metabolisms in a Graph-Based Toy-Universe.
CMSB, Warnemuende (GER), 2008

Using the RNA sequence-to-structure map for functional evolution of ribozyme catalyzed artificial metabolisms.
ALIFE XI, Winchester (UK), 2008

Eigenständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder Dienstleistungen als solche gekennzeichnet.

Leipzig, 11. Juli 2011

Alexander Ullrich