

# *In silico* Evolution of Early Metabolism

Alexander Ullrich<sup>1</sup>, Christoph Flamm<sup>2</sup>, Markus Rohrschneider<sup>3</sup> and Peter F. Stadler<sup>1,2,4,5,6</sup>

<sup>1</sup>Bioinformatics Group, Department of Computer Science, University of Leipzig, D-04107 Leipzig

<sup>2</sup>Institute for Theoretical Chemistry, University of Vienna, 1090 Wien

<sup>3</sup>Image and Signal Processing Group, Department of Computer Science, University of Leipzig, D-04109 Leipzig

<sup>4</sup>Max Planck Institute for Mathematics in the Sciences, D-04103 Leipzig

<sup>5</sup>Fraunhofer Institute for Cell Therapy and Immunology, D-04103 Leipzig

<sup>6</sup>Santa Fe Institute, Santa Fe, NM 87501

## Abstract

We developed a simulation tool for investigating the evolution of early metabolism, allowing us to speculate on the formation of metabolic pathways from catalyzed chemical reactions and development of characteristic properties. Our model consists of a protocellular entity with a simple RNA-based genetic system and an evolving metabolism of ribozyme-catalyzed enzymes that manipulate a rich underlying chemistry. Ensuring an almost open-ended and fairly realistic simulation is crucial for understanding the first steps in metabolic evolution. We show here, how our simulation tool can be helpful in arguing for or against hypotheses on the evolution of metabolic pathways. We demonstrate that seemingly mutually exclusive hypotheses may well be compatible when we take into account that different processes dominate different phases in the evolution of a metabolic system. Our results suggest that forward evolution shapes 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.

## Introduction

Understanding the evolutionary mechanisms of complex biological systems is an intriguing and important task of current research in biology as well as artificial life. 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  $\beta/\alpha$ -barrel fold architecture (Copley and Bork, 2000) shows that the evolution of modern metabolism is mainly driven by enzyme recruitment, as suggested by the patchwork model (Ycas, 1974; Jensen, 1976). Nevertheless, many aspects of the evolutionary machinery are still not well understood. In particular, the first steps in early metabolism evade observation by conventional approaches. Studies on hypotheses of pathway evolution (Caetano-Anollés et al., 2009; Morowitz, 1999) suggest that metabolism has evolved in different phases and only traces or “shadows” are still observable from the events in the very distant past. Thus, there is a need for realistic models of early metabolism that consider all its components and scales. Simulation approaches have shown to be useful in finding and challenging explanations for the

evolution of biological networks (Pfeiffer et al., 2005). We have recently proposed a computational framework for the evolution of metabolism (Flamm et al., 2010), modeling all its significant components in a realistic way. In this report we discuss first results from several simulation runs.

In the next section we recapitulate four scenarios of evolution that are of particular interest to understand the formation of metabolic pathways and assessing our own results. This will be followed by a brief introduction to our computational model that we use in this study. Then we will present some general results from a series of simulation runs and investigate some of the findings in more detail on two examples. We conclude with a short discussion on the comparison of our results with existing pathway evolution hypotheses.

## Scenarios of Evolution

In this section, we elucidate four relevant hypotheses on the evolution of metabolism in general and formation of metabolic pathways in specific. For more a more detailed discussion of the theories of pathway evolution we refer to the reviews by Caetano-Anollés et al. (2009) and Schmidt et al. (2003) discussing further theories of pathway evolution.

### Backward Evolution

Backward (or retrograde) evolution was one of the first theories for the evolution of metabolic pathways, proposed by Horowitz (1945). 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 going enzymes should be found (see Figure 1(a)).

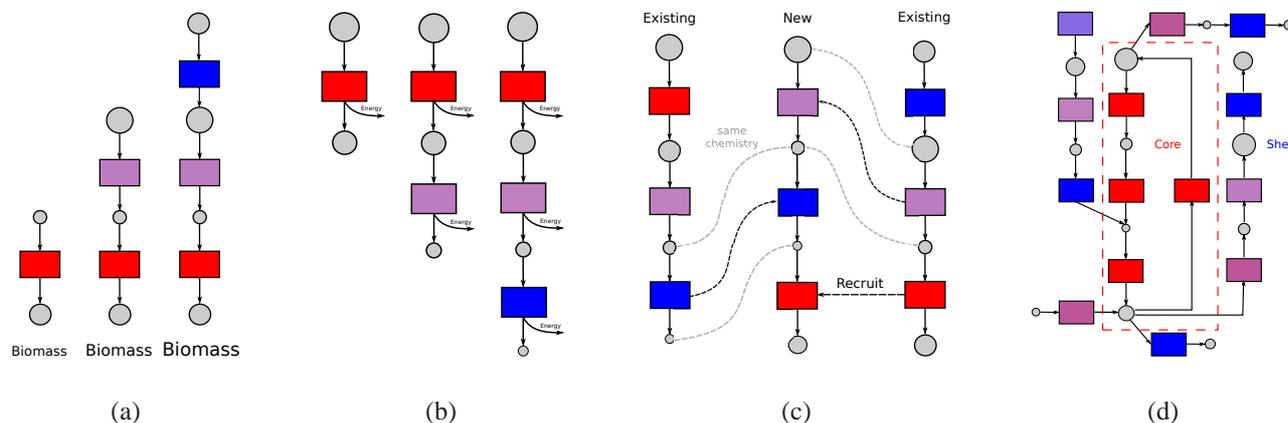


Figure 1: Hypotheses about the formation and evolution of metabolic pathways. (a) Backward evolution, (b) Forward evolution, (c) Patchwork model, (d) Shell hypothesis. Colored squares represent enzymes, gray circles are metabolites. Color encoding for enzymes stand for their age, red being older and blue being younger enzymes.

## Forward Evolution

Forward evolution could be seen as an extension or counterpart of the backward evolution hypothesis, reversing the direction of pathway evolution. Granick (1957), and later Cordon (1990), 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 1(b)).

## Patchwork Model

The patchwork model (Ycas, 1974; Jensen, 1976) 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 1(c)).

## Shell Hypothesis

The shell hypothesis was proposed by Morowitz (1999). 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 1(d)).

## Model

The computational model, summarized schematically in Figure 2, is composed of a genetic and a metabolic subsystem. The genetic subsystem is implemented as a cyclic RNA genome. A special sequence motif indicates the start of genes which are of constant length. The RNA sequence corresponding to the “coding sequence” of a gene is folded into the (secondary) structure using the Vienna RNA Package (Hofacker et al., 1994) (Step A in Figure 2).

During chemical reactions bond formation/breaking is confined to a small subset of atoms of the reacting molecules. A cyclic graph abstraction, called the imaginary transition state (ITS) (Fujita, 1986), can be used to capture the changes in the reactive center (Hendrickson, 1997). Furthermore, over 90% of all known organic reactions can be classified by their ITS (Hendrickson and Miller, 1990) and organized in a hierarchical structure (Herges, 1994). Sequence and structure features of the folded RNA gene products are mapped into the classification tree of organic reactions for functional assignment of the catalytic set (Step B in Figure 2). Thus we have implemented an evolvable sequence-to-function map (Ulrich and Flamm, 2009), allowing the metabolic organization to escape from the confines of the chemical space set by the initial conditions of the simulation.

The metabolic subsystem is built upon a graph-based artificial chemistry (Benkő et al., 2003) endowed with a built-in thermodynamics. To generate the metabolic reaction network, induced by the catalytic set on the set of metabolites, a rule-based stochastic simulation is performed (Faulon and Sault, 2001). Reaction rates are calculated from the chemical graphs of the reactants on the fly.

To identify the elementary flux modes, i.e., extreme pathways (Gagneur and Klamt, 2004), of the resulting reaction network, a metabolic flux analysis is performed. (Step D

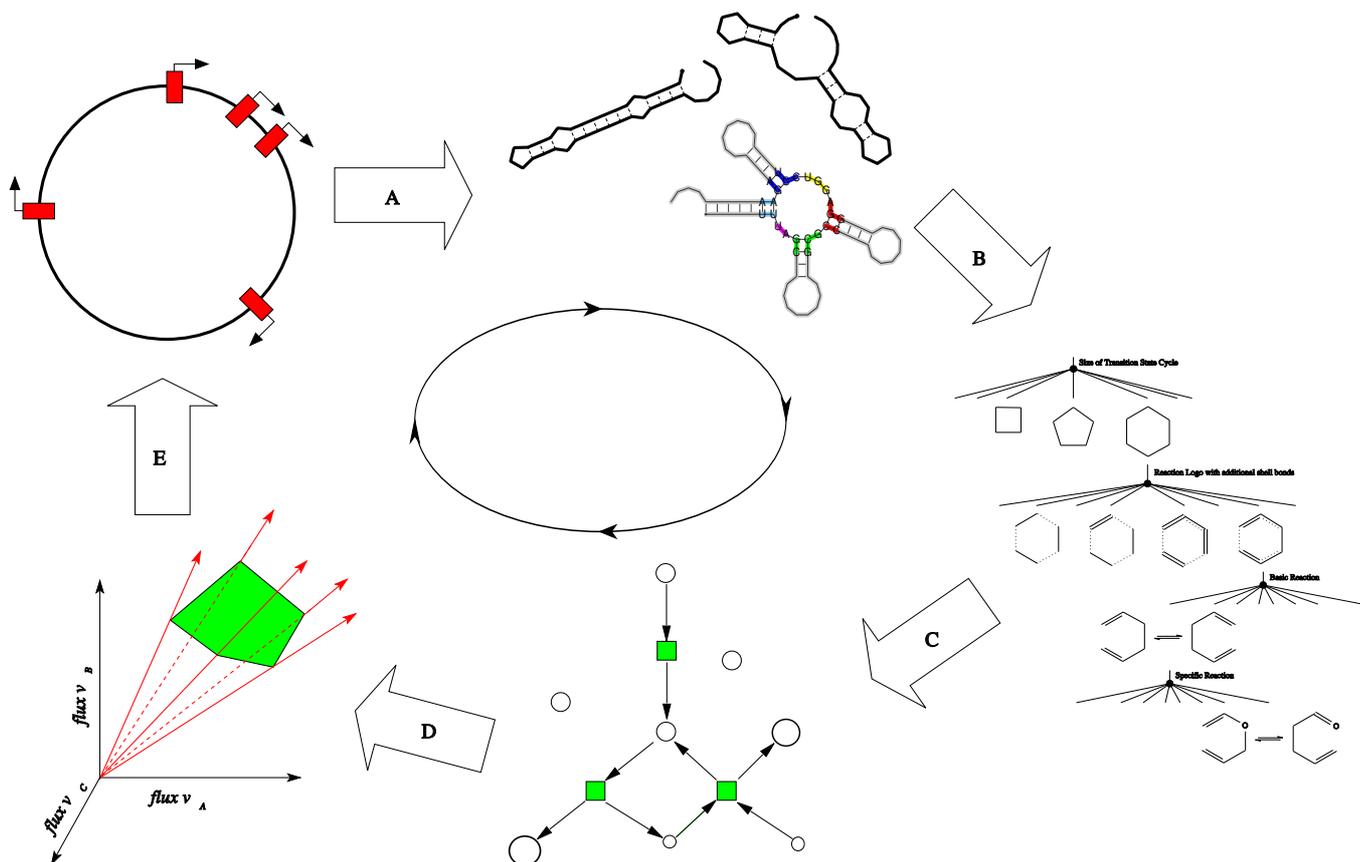


Figure 2: Scheme of the simulation system. (A) Decoding of genes to catalytic molecules; (B) Assignment of catalytic functions to “ribozymes”; (C) Construction and stochastic simulation of metabolic the network; (D) Metabolic Flux analysis and fitness evaluation; (E) Application of genetic variation operators.

in Figure 2). The fitness of an organism is computed as the maximum of the (linear) yield function over all extreme pathways. Finally, genetic variation operators are applied to the genome (Step E in Figure 2). For a detailed discussion of the various steps of the computational model we refer the reader to Flamm et al. (2010).

## Simulations and Results

In this section we use the computational model described above 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 paper 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).

## Quantitative Analysis

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

In a previous study (Ullrich and Flamm, 2008), we already showed that our metabolic networks evolved certain properties such as a scale-free node degree distribution and the existence of hub-metabolites. An investigation of the enzyme connectivity suggested that enzymes from early stages show a higher connectivity than those from later stages. Here, we confirm these findings with a much larger sample of 100 simulation runs. Figure 3(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 en-

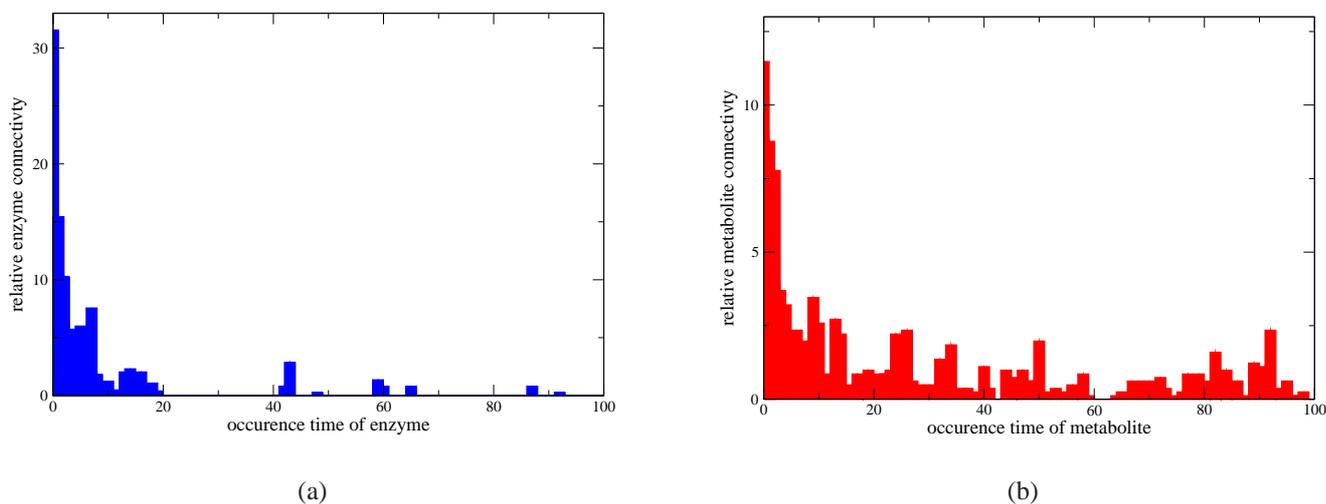


Figure 3: Average relative connectivity of (a) enzymes and (b) metabolites introduced in the same generation, for 100 generations. The height of the bars shows the fraction of the overall connections that are accounted for by enzymes/metabolites from a particular generation. All values are averages over 100 simulation runs. Input molecules are not considered in the statistic, they account for nearly 50 percent of metabolite connectivity.

zymes 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 3(b)), we still find the highly connected nodes in the early steps, especially if we consider environment metabolites that are always abundant which account for about 50 percent of connectivity. However, there is constant production of metabolites potentially becoming highly connected.

In order to find arguments for some of the evolution hypotheses, we study the occurrence time (age) of reactions and metabolites along pathways. 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 new metabolites. We will use the term forward (backward) link if, in a pair of reactions in a pathway, the successor is evolutionary older (younger). 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, we define forward (backward) pathways as pathways in which there is at least one forward (backward) link and no backward (forward) link. Given these definitions, we compute the set of extreme pathways for every generation and all cells. For each pathway we then determine the percentage of forward and backward links and pathways, for both reactions and metabolites.

For this study, we performed 100 runs with the following settings: a population size of 100 cells running for

100 generations and performing 100 network expansion steps per generation, the input molecules were cyclobutadiene, ethenol, phthalic anhydride, methylbutadiene, and cyclohexa-1,3-diene. In Figure 4 we see the change from generation to generation in the constitution of the metabolic networks regarding our measures of forward/backward links and pathways. Considering the reactions of the networks, one can see that in the first generations, the networks consist mainly of links and pathways conforming to the forward evolution scenario. However, in later generations we observe a much more mixed mosaic like picture 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 in the last stages, which could mean that they form a core of pathways that are not subject of evolutionary change. This supports the shell hypothesis. So far, our simulation do not provide any support for the backward evolution scenario. However, so far we have not simulated an environment with temporary depletion of “food” metabolites, which is one of the major assumption of this theory. A future study considering this impact of variations in resource abundances might bring new insights on this matter.

### Example

In the following we illustrate some of our findings from the previous study in more detail for an example simulation. We use data from a simple simulation run, including only few enzymes and two input molecules, for the visualization of an evolutionary time series (see Figure 5) an

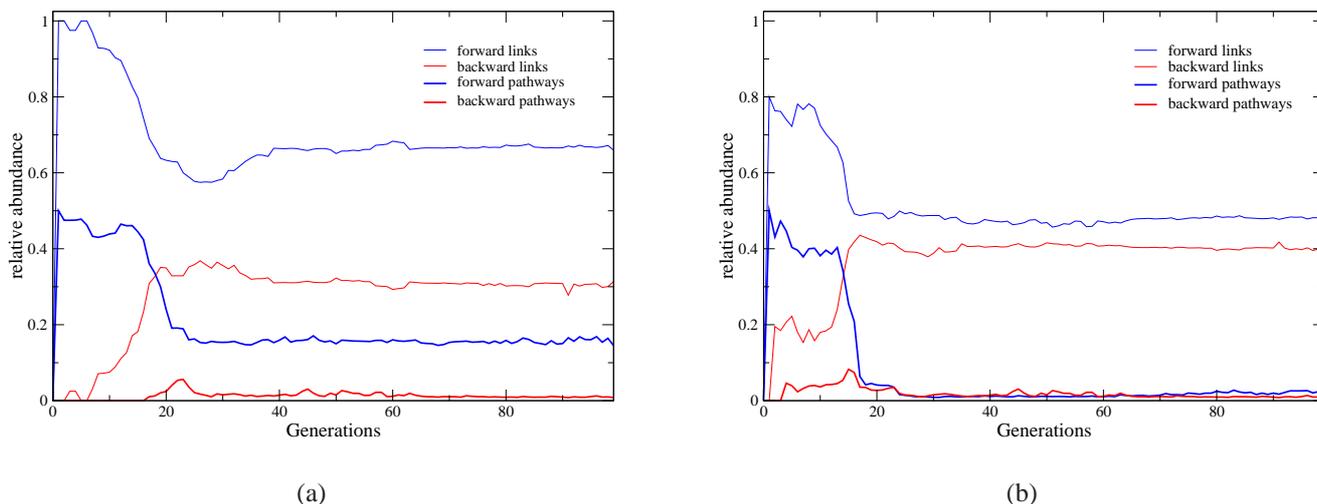


Figure 4: Evolutionary history of simulated metabolic networks. For the first 100 generations, we show the number of links and pathways that conform to the forward and backward evolution scenarios, respectively. Links are pairs of (a) consecutive reactions or (b) consecutive metabolites along a pathway. 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.

imation of the network evolution (see Additional Files) and the reaction- and metabolite-lifetime overviews (see Figure 6). The genome, and hence the set of enzymes, is chosen at random in the beginning. The two input molecules of this simulation are cyclic and sequential glucose. The simulation run is kept to 100 generations. We focus 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 showing the metabolic network in different stages are aligned to a union graph over all generations (Rohrschneider et al., tted). Thus, we can see that in the first steps 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. This observation is further supported by the interval graph for all chemical reactions in Figure 6. The reactions are here ordered according to their position in the graph. There is a clear trend of older reactions being on the top (upstream) and younger ones following more downstream. The colored bar next to the interval graph shows the pattern of the relation between age and position of reactions and metabolites for our example simulation run. The other three bars show the patterns for backward, forward evolution and the patchwork model, respectively. The forward evolution pattern comes closest 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 we do not see 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 we investigate the evolutionary history of the involved genes/enzymes, depicted in the catalytic function genealogy for all generations (Figure 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. Our simulation framework allows us to study the of divergence and convergence of catalytic functions (Almonacid et al., 2010) since we can record the genealogy of each gene (reaction catalyst) throughout a simulation run, and we 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 happens when catalysts from genealogically unrelated genes independently accumulate mutations resulting in the catalysis of the same reaction (or class of reactions). In Figure 7 convergence events are marked by circles. A small selection of divergence events, which are very frequent in our simulations, are marked by broken cir-

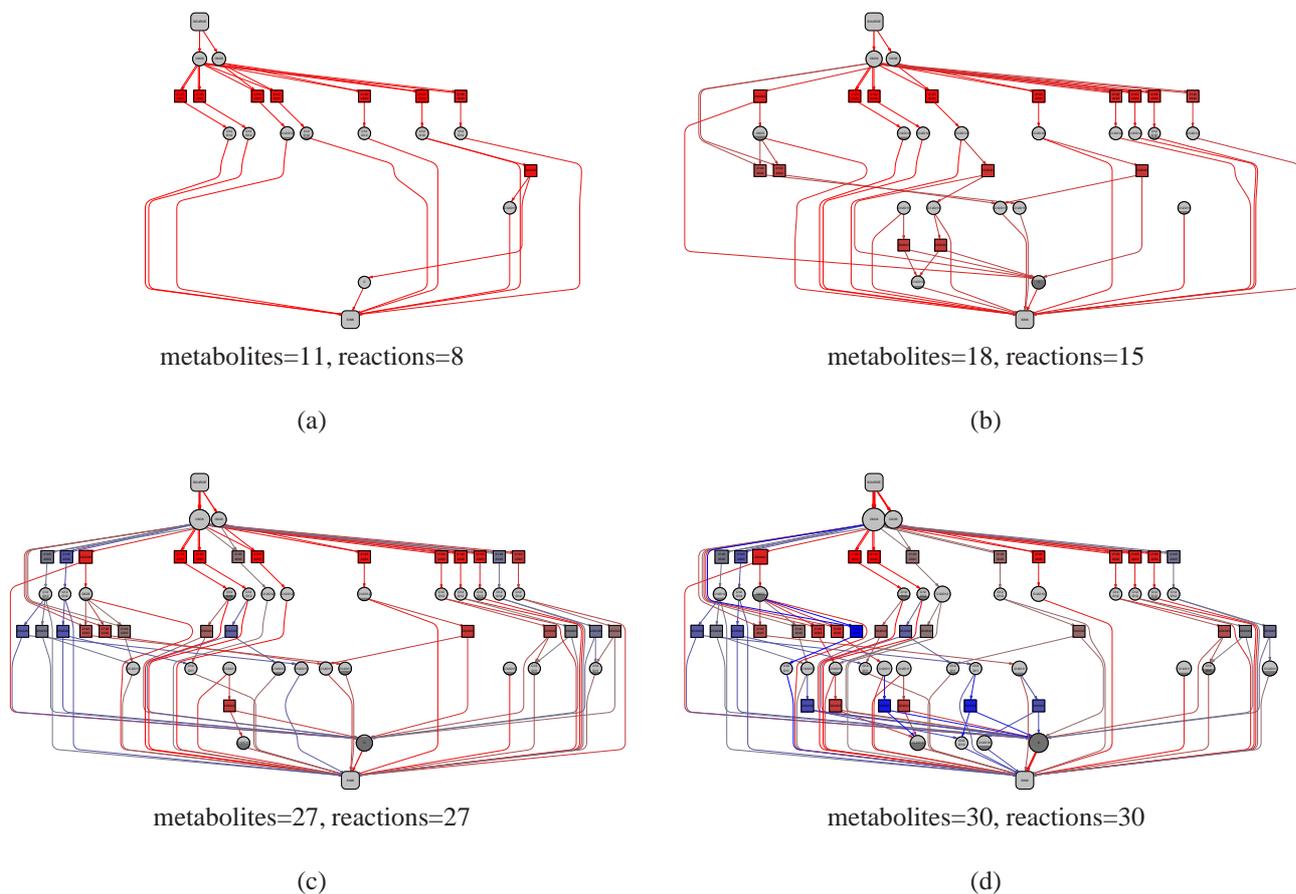


Figure 5: A series of simulated metabolic networks after (a) 10, (b) 30, (c) 66, and (d) 100 generations. Colored squares represent chemical reactions, gray circles represent metabolites. Metabolites involved in a reaction are connected to it in the network graph. The size of the nodes and the width 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 and blue for newer reactions.

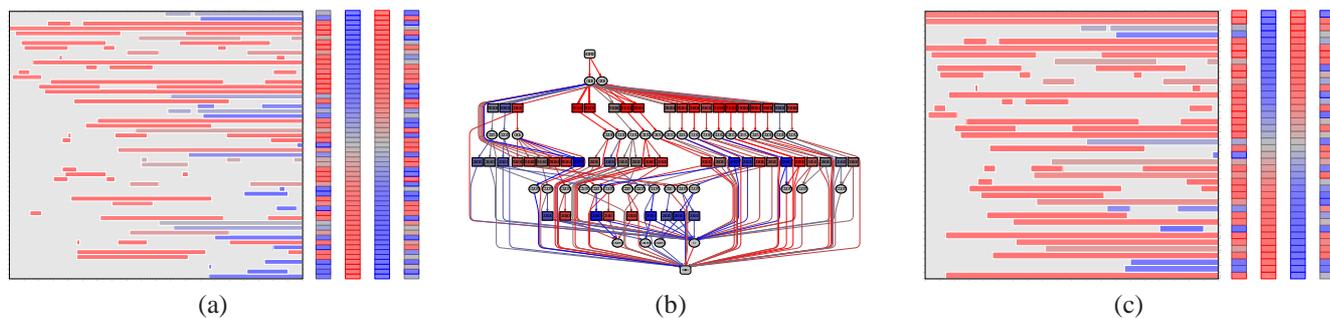


Figure 6: Life-time diagram for reactions and metabolites. (a) Life-time of reactions, (b) union network graph over all 100 generations, (c) life-time of metabolites. The reactions and metabolites (rows) in the life-time diagrams are positioned corresponding to their position in the union network graph, i.e. from top to bottom in the life time overview the reactions/metabolites are enumerated in a breadth first manner on the network graph. 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.

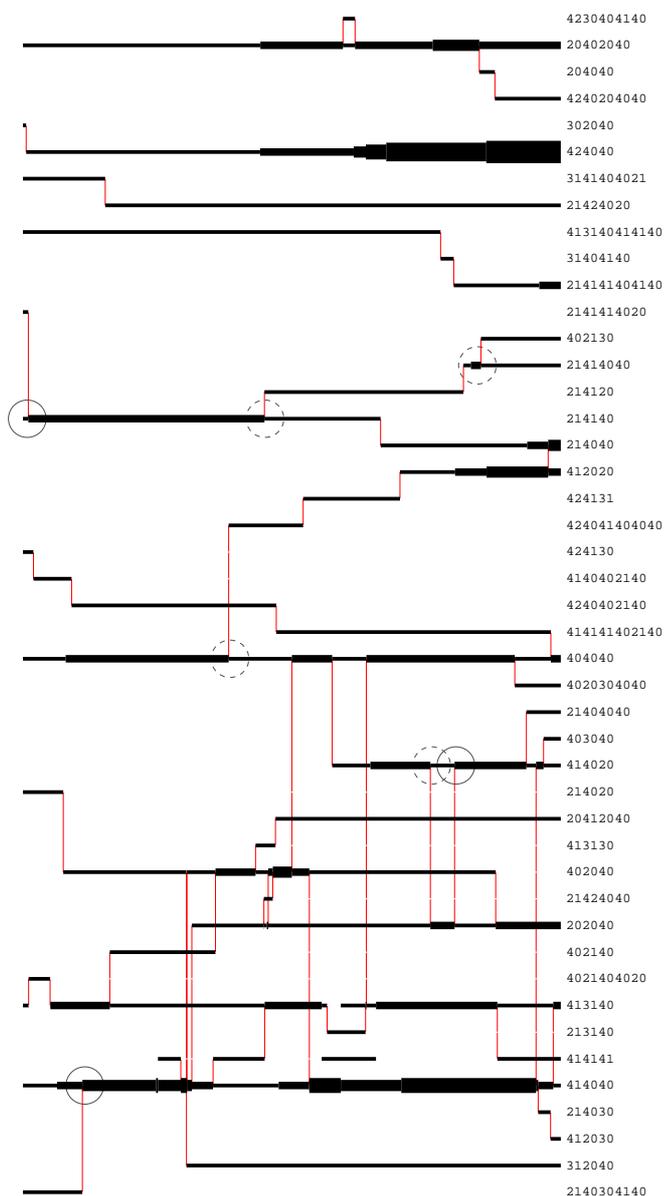


Figure 7: Genealogy of catalytic functions and gene dosage over 2000 generations. Each row represents an observed catalytic function. Black horizontal lines indicate time intervals in which genes coding for that catalytic function were present in the genome (0-200: from left to right). The thickness 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.

cles. 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 of the context of the graph rewrite rule, i.e. the necessary precondition for the applicability of the graph transformation rule.

## Conclusions

We have introduced a simulation tool that models the early evolution of metabolism in a quite realistic setting and provides many tools for the detailed investigation of metabolic evolution. Using both simple example and a series of more complex simulation runs, the evolution of the components on the small scale (metabolites, enzymes) as well as on systems (pathways, networks) was investigated. The simulations allow to discriminate between different scenarios for the evolution of metabolic pathways. Based on the observations from this study, we argue that the different evolutionary hypotheses can be reconciled, in that they act in different phases of evolution, i.e. in different scenarios we might observe another strategy at work. Here, we 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.

To further test these hypotheses, we intend to simulate a number of different scenarios with changing parameters (mutation rate, duplication rate, depletion), define other goals for the organisms (production of one specific metabolite, biomass or energy) and increase the complexity of the simulation runs (length and number of input molecules).

Albeit our simulation environment is still a drastic simplification of chemistry, it is realistic enough to investigate the evolution of early metabolism. Computer simulations like this one are likely to provide new insights about the general evolutionary mechanisms governing biological systems in particular in regimes that are not readily observable. Our approach of a realistic, yet computationally feasible, model appears to be a promising step in this direction.

## Acknowledgments

This work has been funded by the Volkswagen Stiftung under grant I/82 719 and by the Vienna Science and Technology Fund (WWTF) MA07-30, and the COST-Action CM0703 “Systems Chemistry”.

## Additional Files

An animated movie of an example network evolution simulation, can be found here <http://www.bioinf.uni-leipzig.de/~alexander/animation.avi>.

## References

Almonacid, D. E., Yera, E. R., Mitchell, J. B., and Babbitt, P. C. (2010). Quantitative comparison of catalytic mechanisms and

- overall reactions in convergently evolved enzymes: implications for classification of enzyme function. *PLoS Comput Biol*, 6(3):e1000700.
- Benkő, G., Flamm, C., and Stadler, P. F. (2003). A graph-based toy model of chemistry. *J Chem Inf Comput Sci*, 43:1085–93.
- Caetano-Anollés, G., Yafremava, L. S., Gee, H., Caetano-Anollés, D., Kim, H. S., and Mittenthal, J. E. (2009). The origin and evolution of modern metabolism. *Inter J Biochem & Cell Biol*, 41:285–297.
- Copley, R. R. and Bork, P. (2000). Homology among ( $\beta\alpha$ )<sub>8</sub>-barrels: implications for the evolution of metabolic pathways. *J Mol Biol*, 303:627–641.
- Cordon, F. (1990). *Tratado evolucionista de biología*. Aguilar Ediciones, Madrid, Spain.
- Faulon, J.-L. and Sault, A. G. (2001). Stochastic generator of chemical structure. 3. Reaction network generation. *J Chem Inf Comput Sci*, 41:894–908.
- Flamm, C., Ullrich, A., Ekker, H., Mann, M., Högerl, D., Rohrschneider, M., Sauer, S., Scheuermann, G., Klemm, K., Hofacker, I. L., and Stadler, P. F. (2010). Evolution of metabolic networks: A computational framework. *J. Syst. Chem.* in press.
- Fujita, S. (1986). Description of organic reactions based on imaginary transition structures. 1. introduction of new concepts. *J. Chem. Inf. Comput. Sci.*, 26:205–212.
- Gagneur, J. and Klamt, S. (2004). Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, 5.
- Granick, S. (1957). Speculations on the origins and evolution of photosynthesis. *Ann NY Acad Sci*, 69:292–308.
- Hendrickson, J. B. (1997). Comprehensive system for classification and nomenclature of organic reactions. *J Chem Inf Comput Sci*, 37:852–860.
- Hendrickson, J. B. and Miller, T. M. (1990). Reaction indexing for reaction databases. *J. Chem. Inf. Comput. Sci.*, 30:403–408.
- Herges, R. (1994). Coarctate transition states: The discovery of a reaction principle. *J Chem Inf Comput Sci*, 34:91–102.
- Hofacker, I. L., Fontana, W., F, S. P., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Mh. Chem.*, 125:167–188.
- Horowitz, N. H. (1945). On the evolution of biochemical syntheses. *Proc Natl Acad Sci USA*, 31:153–157.
- Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*, 30:409–425.
- Morowitz, H. J. (1999). A theory of biochemical organization, metabolic pathways, and evolution. *Complexity*, 4:39–53.
- Pfeiffer, T., Soyer, O. S., and Bonhoeffer, S. (2005). The evolution of connectivity in metabolic networks. *PLoS Biol*, 3:e228.
- Rohrschneider, M., Ullrich, A., Kerren, A., Stadler, P. F., and Scheuermann, G. (2010 (submitted)). Visual network analysis of dynamic metabolic pathways. manuscript submitted for publication.
- Schmidt, S., Sunyaev, S., Bork, P., and Dandekar, T. (2003). Metabolites: a helping hand for pathway evolution? *Trends Biochem. Sci.*, 28:336–341.
- Ullrich, A. and Flamm, C. (2008). Functional evolution of ribozyme-catalyzed metabolisms in a graph-based toy-universe. In Istrail, S., editor, *Proceedings of the 6th International Conference on Computational Methods in Systems Biology (CSMB)*, volume 5307 of *Lect. Notes Bioinf.*, pages 28–43.
- Ullrich, A. and Flamm, C. (2009). A sequence-to-function map for ribozyme-catalyzed metabolisms. In *ECAL*, volume 5777/5778 of *Lect. Notes Comp. Sci.*
- Ycas, M. (1974). On earlier states of the biochemical system. *J Theor Biol*, 44:145–160.