Functional Evolution of Ribozyme-Catalyzed Metabolisms in a Graph-Based Toy-Universe

Alexander Ullrich and Christoph Flamm

University of Vienna, Institute for Theoretical Chemistry,
Waehringerstrasse 17, A-1090 Vienna, Austria

Abstract. The origin and evolution of metabolism is an interesting field of research with many unsolved questions. Simulation approaches, even though mostly very abstract and specific, have proven to be helpful in explaining properties and behavior observed in real world metabolic reaction networks, such as the occurrence of hub-metabolites. We propose here a more complex and intuitive graph-based model combined with an artificial chemistry. Instead of differential equations, enzymes are represented as graph rewriting rules and reaction rates are derived from energy calculations of the involved metabolite graphs. The generated networks were shown to possess the typical properties and further studied using our metabolic pathway analysis tool implemented for the observation of system properties such as robustness and modularity. The analysis of our simulations also leads to hypotheses about the evolution of catalytic molecules and its effect on the emergence of the properties mentioned above.

Keywords: metabolism, evolution, simulation, enzymes, robustness.

1 Introduction

Life, in the most basic sense, constitutes of interactions between chemical compounds building complex networks which in turn can be regulated and interacted with. 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 and flexibility. 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.

Metabolic networks are the best studied biochemical networks. We can reconstruct entire metabolisms because we have complete annotated genomes of model organisms at our disposal. Looking at pathways in these networks can in turn be interesting for functional genomics, e.g. gene expression data derived from DNA arrays may be better understood in terms of metabolic components, pathways or sub-networks. Insights about the metabolism of an organism can of course be useful for further biotechnological applications [1], e.g. the determination of pharmaceutical targets, metabolic engineering, changing direction and

yield of pathways. Before we can make use of all these applications, it is essential to gain an understanding of the network properties. It is often not enough to look at textbook-like metabolisms, since they do not resemble the actual behavior of these networks. Therefore, we need means to analyze the network's topology, structure, principle, plasticity and modules. Such means exist, e.g. metabolic flux analysis [2][3], and thus we are able to make important observations about the networks and systems of interest, such as the abundance of diversity among enzymes, hub-metabolites and small world networks. But it also has to be noted that there are still some limitations to the analysis and it remains unknown how these and other possible properties emerged and further evolved. The case is especially difficult if we regard properties that cannot be sufficiently explained by looking at a static network image [4], such as robustness or evolvability.

It can be assumed and it is believed by most biologists that all organisms which we can see today are evolved from one common ancestor [5]. This ancestor would be a single cell having properties similar to those observed in cells of modern organisms. Since we do not believe this cell to have been spontaneously and magically appeared on the earth's surface, it is fair to suggest that this cell in turn gradually evolved from simpler cells. Many theories on the origin of life and scenarios for the early evolution exist, but actually we cannot say anything with certainty until the point of the common ancestor, thus all the available modern molecular techniques will fail to provide a complete account of the emergence of some of the properties we are interested in.

Models for the simulation of the emergence of network properties exist and have provided explanations for some of the properties. For example, [6] showed that gene-duplication can account for the property of a network to be scale free. So far these models of biological systems use either differential equations [7], i.e. enzymes are not modeled as actual chemical entities but only as rates, or very abstract artificial chemistries. A more complex simulation integrating more functional constraints of the metabolism should provide further insights about the metabolism itself, properties of complex networks in general and also their emergence, so that these properties may be reproducible in other applications. For instance, artificial networks are desired to be robust and maybe even evolvable as well.

2   Model

In this section we will discuss the basic framework of the simulation -the model-, the basic ideas behind it and explain the individual components. All structures in the simulation are modeled as graphs and processes are performed through applications on the structure of graphs or the analysis of those. The choice to use graphs as the presentation for the structures in our chemical environment can be justified, firstly, by the fact that in chemistry molecules are for many years represented in graph form. Also it is the most intuitive way to regard chemical substances. Besides, networks are best understood by looking at its graph representation. And considering modern reaction classification systems[8,9,10,11],

even for reactions and thus enzymes graphs can be used as appropriate models. Furthermore, it is hoped and believed that using graphs for all parts of the model results in a more realistic behavior of the entire system. Also there exist versatile applications which can be performed on graphs to analyze and transform them, such as metabolic pathway analysis or graph-rewrite systems.

The graph-based model is supported by an artificial chemistry, ToyChem[12], completing the universe in which individuals and their metabolisms can evolve. The artificial chemistry uses a graph representation of the molecule for the energy calculation. ToyChem provides the look-and-feel of a real chemistry[12] and integrates a realistically chemical behavior into our simulation, which is sufficient for our purposes and more sophisticated than has so far been at the disposal of a comparable simulation approach[13].

2.1   Genome

Every individual in the simulation population contains a genome of a fixed length and a common TATA-box sequence. Furthermore, all genes have the same length. The genome is an RNA-sequence and the single genes represent RNA-enzymes and bear the function of a particular chemical reaction from the set of reactions defined as current chemistry, as will be explained in one of the next paragraphs. In each generation, new individuals are generated from the set of optimal (with respect to metabolic yield) individuals. Those new individuals contain a copy of the parent genome to which a point mutation was applied. The mutation can occur everywhere in the genome. There can be silent mutations, i.e. the mutation takes place in a non-coding region, or neutral mutations which change a nucleotide within a gene but not the function of the corresponding RNA-enzyme, i.e. it still performs the same chemical reaction. Accordingly, there can also be missense mutations which change the structure of the RNA-enzyme in such a way that it inhabits a different function than before, and there can be mutations which either destroy a TATA-box (nonsense mutation) or build a new TATA-box and, therefore, eliminate or add a new gene to the genome, respectively.

The genome is realized by a string containing the nucleotide sequence and a list of all genes which have to be transcribed to RNA-enzymes. The sequence is treated as circular. Consequently, there are as many genes as there are TATA-boxes and some of the genes may reach over the ends or overlap. Every gene is assigned an ID. The IDs for the currently expressed genes and that of the parental genes are stored and all the genes, currently expressed or not, are listed in the genome. With this information we are able to retrace the history of every single gene and determine whether it had a single or multiple origin and if it may have disappeared for some generations before reappearing. Furthermore, the entire history of mutations is kept in the genome and we can determine the exact time of change and analyze the means of these changes. This also allows us to compare sequential and structural changes with external events, such as changes in the environment or selection for certain properties.

2.2   Metabolites

Besides the genome, individuals also include a metabolite-pool. In the first generation this pool consists only of the metabolites of the environment. The user defines the content of this environment by providing an input-file with the SMILES[14] notations of the molecules that are to be included, otherwise a predefined set of molecules is used. In each generation the newly produced metabolites are added to the metabolite-pool. To avoid redundancy, every new metabolite graph has to be checked for isomorphism against the entire pool. Since the size of the metabolite-pool can increase quickly, graph-isomorphism checking could slow down the entire simulation, therefore, we keep the graphs in a hashmap with their unique SMILES notation[15] as key, reducing the problem to a string comparison. At the end of a simulation the metabolite-pool is printed to a file, using again SMILES since it is a concise and easily interpretable way of presenting chemical molecules.

The vertices of the metabolite graph are the atoms of the respective chemical molecule and edges exist between vertices whose atoms are connected, to represent the chemical bonds of the molecule. The labels for the vertices are the atom types: hydrogen, oxygen, nitrogen or carbon (H, O, N, C). For edges the labels are the chemical bond type: single bond, double bond or triple bond $(-, =, \#)$.

2.3   Enzymes

We now turn to the most important part of the metabolism and therefore also our model. Enzymes determine the metabolic fate of a cell and almost all processes with considerable contribution to the development of a cell need enzymes. So far, simulations modeled only the reaction rates, with differential equations, but the representation of the reaction itself was rather abstract and static. We use a more flexible and realistic approach. Flexible because the set of enzymes can be adjusted by the user initially. For different purposes one might want to choose different sets of chemical reactions, e.g. sometimes only reactions working on carbon-skeletons are of interest or in another experiment only reactions involving a small number of atoms is to be observed. Realistic because many different kinds of chemical reactions are available (around 15.000 in the current simulation), but only those that are chemically valid. Due to the sheer endless number of possible chemical reactions, the set of reactions which can be chosen in the simulation is restricted here to those containing hydrogen, oxygen, carbon or nitrogen atoms. We believe that these atoms suffice to build up the most important molecules necessary for a primitive metabolism as one would expect in the early evolution of metabolism and that the simulation still resembles a realistic account for the processes at such a phase or comparable situations and does not sustain a loss of expressiveness regarding robustness and other network properties. Furthermore, we consider only the class of pericyclic reactions, which is the most important one of the three organic reaction mechanisms[16] (the other two are ioinic and radical reactions). These reactions always have a cyclic transition structure. Pericyclic

reactions are used here because they are very clean, i.e. there are no unknown by-products, and they can be analyzed with frontier molecular orbital theory which is of use for the calculation of the reaction rates by the artificial chemistry. We further limit the set of reactions to those involving three to six atoms and not more than two metabolites. If we say that a certain number of atoms is involved in a reaction, then this does not mean that the metabolite which is to be worked on contains only this particular number of atoms, but rather that only the connections within a set of atoms of this particular size is changed by the reaction. Most of the already known chemical reactions lie in this range and it can be assumed that, accordingly, reactions crucial to simple metabolisms or the most basic pathways and networks underlying all metabolisms can be found there. Reactions involving more atoms or metabolites, account only for few interesting reactions and would simply add to the complexity of the computation.

The atoms and bonds of the reaction center of the chemical reaction, corresponding to the enzyme, constitute the vertices and edges of the enzyme graph. Each vertex in an enzyme graph is connected to two other vertices in such a way that the atoms build a cycle. The vertex label is equal to that of the metabolite graph, but the edge-labeling differs somewhat. In the enzyme graph, every edge has two labels for bond-types: one for the substrate molecule and the other for the product molecule. Also the bond-types for enzymes are extended by the empty symbol, indicating that two atoms are not connected. Below, the Diels-Alder reaction is shown in the GML[1][17] format which is used as the input format.

```
# 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 "-" ]
```

[1] www.infosun.fim.uni-passau.de/Graphlet/GML/

```
    edge [ source 4 target 5 label "-" ]
    edge [ source 5 target 0 label "-" ]
  ]
]
```

The function of the enzyme is performed through a graph-rewriting mechanism. The graph that is transformed is that of a metabolite. The graph-rewrite rules are pairs of graphs that are gained from the enzyme graph. As explained above, an edge in the enzyme graph has two labels, one is for the substrate graph (the left side of the rule) and the other is for the product graph (the right side of the rule). First, we search for subgraphs in the metabolite that match the substrate graph and then replace it with the product graph. Note, that the atoms and the number of connections stays the same but the connections are reordered. Following, the energy of the substrate and the product are calculated as well as the reaction rate. The enzymatic reaction is only applied if the product molecule is energetically more favorable. Besides the energy calculation, metabolite graphs can be measured and assessed in terms of topological graph indices[18]. So far we use these indices (e.g. Connectivity Index[19], Platt Number[20] and Balaban Index[21]) to produce networks differing in the set of selected enzymes although starting in the same environment and with the same chemistry. We can use this to check whether the common network properties depend on lower level properties and also analyze if the selected enzymes spezialise directed or rather randomly. Furthermore, the additional use of topological indices could allow us to select for metabolites and enzymes with certain characteristics, such as very stable or very reactive metabolites and enzymes building long chains of the same molecule.

2.4   Mapping

As mentioned before, the genes encoding the enzymes are RNA-sequences and the enzymes, consequently, are modeled to act as ribozymes. To ensure a realistic behavior and evolution of our enzymes, we developed a novel genotype-phenotype mapping for the transition from the gene (genotype) to the catalytic function of the enzyme (phenotype). We use the RNA sequence-to-structure map as the basis for the mapping and consider two observations from the study of evolution and enzymes.

Firstly, it is known that neutral mutations, leading to a redundant genotype-phenotype mapping, have a considerable influence on the evolution in molecular systems. The folding of RNA-sequences to secondary structures with its many-to-one property represents such a mapping entailing considerable redundancy. Various extensive studies concerning RNA-folding in the context of neutral theory yielded to insights about properties of the structure space and the mapping itself. Thus, we will get a better understanding of some of these properties and especially of the evolution of RNA-molecules as well as their effect on the evolution of the entire molecular system.

The second observation we use is that enzymes typically have an active site where only few amino acids or bases determine its catalytic function and the

remaining structure has mostly stabilization purposes. Accordingly, we extract structural and sequence information only from a restricted part of the fold. We decided to focus on the longest loop of the folded RNA since most RNA-aptamers are known to contain a loop region as their catalytic center. The idea for mapping the extracted information directly to a specific chemical reaction was inspired by the fact that many enzymes catalyze a reaction by stabilizing its transition state and the work on reaction classification systems, in particular Fujita's imaginary transition structures (ITS) approach [10,22,23].

The mapping from the structure and sequence information to the pericyclic reaction that resembles the function of the enzyme is generated as following. The length of the longest loop in the secondary structure of the RNA-enzyme determines the number of atoms that are involved in the chemical reaction to which the gene will be mapped. A statistical analysis was performed to ensure that the different reaction types occur in appropriate proportions. Further, the loop is divided in as many parts as the number of atoms involved in the reaction. The mapping to the atom types of the reaction is derived simply from the sequence information in the different parts of the loop, each corresponding to one atom. The exact mapping from sequence information to atom type here is not important since not biologically meaningful. It suffices to notice that all atom types are chosen with the same rate. The bond type of the reaction logo is derived from the structural information of the different parts of the loop, in particular, the stems contained in these parts. The number of stems in a loop region, the length of these stems, and the sequence of the first two stem pairs accounts for the decision to which bond type will be mapped. Again the exact procedure of the mapping will not be discussed because it is a rather technical detail. An example for the mapping is explained in figures 1,2 and table 1 for better understanding.

Finally, we performed several statistical tests commonly used in neutral theory. We compared it with results of approaches using cellular automatons, random boolean networks and other RNA-fold-based mappings. It exceeds all non-RNA mappings in extent and connectivity of the underlying neutral network. Further, it has a significantly higher evolvability and innovation rate than the rest. Especially interesting is the highly innovative starting phase in RNA-based mappings. This shows that the use of such a genotype-phenotype mapping contributes greatly to a more realistic modeling of evolution.

2.5   Artificial Chemistry

We will use ToyChem, an artificial chemistry with the look-and-feel of a real chemistry, for the calculation of the energy of metabolites and reaction rates. In ToyChem molecules are represented in another type of graph -the orbital graph-. From the orbital graph of a molecule all the necessary properties needed for the energy calculation can be derived. The ToyChem package also provides functionality for the computation of solvation energies [24] and reactions rates[25].

Energy Calculation. The most accurate way to calculate arbitrary properties or reaction rates for molecules is to derive a wave-function from the 3D-space
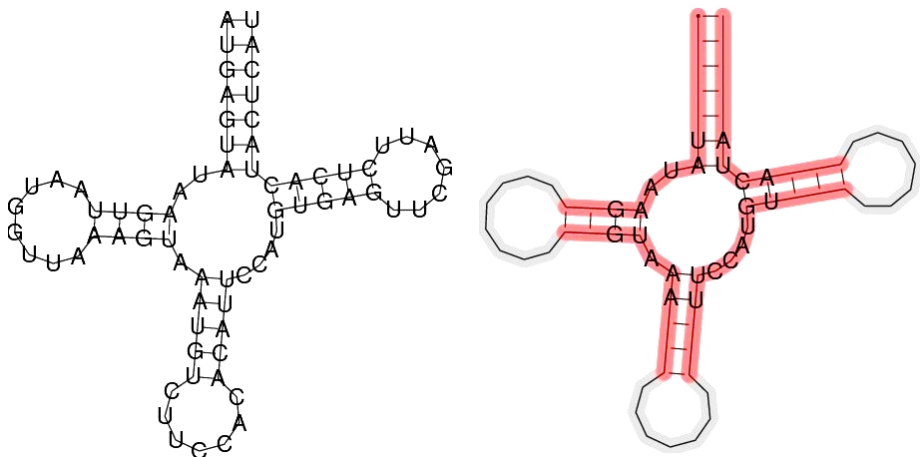
Fig. 1. Example: Reaction mapping. Left: The folded RNA. Right: Longest loop of the folded RNA and the relevant sequence and structure information (marked red).

Table 1. Example: Information derived from the longest Loop of the folded RNA-sequence. The mapped reaction contains four carbon atoms, one oxygen and one nitrogen and is unimolecular. The substrate contains one triple bond, one double bond and three single bonds, whereas the product molecule contains three double bonds and two single bonds, both are not closed.

Section	Loop	C-G pair	Neighbor > 5 bp	**Bond**	Valence	Seq. (loop)	**Sequence**
1 (red)	yes	0	yes (+1)	**1** ” –”	3	4	**4 = C**
2 (blue)	yes	1	yes (+1)	**2** ” =”	4	1	**4 = C**
3 (gray)	no	-	no	**0** ” ”	3	4	**4 = C**
4 (yellow)	yes	0	no	**0** ” ”	1	4	**4 = C**
5 (pink)	no	-	yes (+1)	**1** ” –”	2	2	**2 = O**
6 (green)	yes	1	no	**1** ” –”	3	3	**3 = N**

embedding of the molecular structures with a subsequent application of quantum mechanical methods. This approach is however rather demanding in terms of computational resources. We therefore resign to a computationally tractable approach called ToyChem [12]. This method constructs an analog of the wavefunction from the adjacency relations of a graph followed by a simplified quantum mechanical treatment called extended Hückel Theory (EHT)[26]. In the EHT the Hamilton matrix is parametrized in terms of the atomic ionization potentials and the overlap integrals between any two orbitals. The overlaps between orbitals are gained unambiguously from the molecular graph by applying the valence shell electron pair repulsion theory (VSEPR) [27]. The resulting information can conveniently be stored in the orbital graph who's vertices represent atom orbitals (labeled by atom type and hybridization state of the atom) and the edges denote overlaps of interacting orbitals. Within the ToyChem
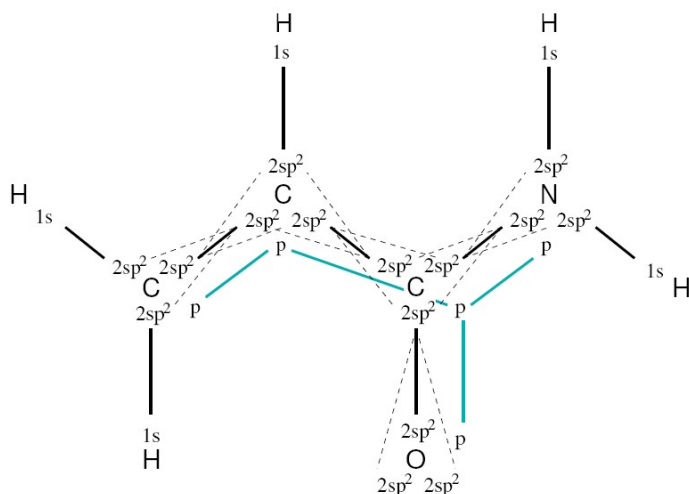
Fig. 2. Example: Extracted information maps to this particular pericyclic reaction, represented by its reaction logo. This is an example of a sigmatrophic rearrangement. The coloring indicates which information was used for the respective part of the reaction logo, the colors correspond to the labeling in the table above.

framework the atomic ionisation potentials and the overlap integrals are tabulated as functions of the atomic type and the type of the hybrid orbitals for a subset of atom which frequently occure in organic molecules. This information allows a fast construction of the Schrödinger equation from the orbital graph. Solving of the Schrödinger equation yields the eigenvectors and eigenvalues from which any physical properties of a ToyChem molecule can be calculated.

Oribtal Graph. In the orbital graph of a molecule, nodes are the atom orbitals and edges indicate overlapping orbitals. From the four atom orbitals $2p_x$, $2p_y$, $2p_z$ and $2s$, three hybrid orbitals with different geometry can be formed. The hybrid orbitals sp (linear geometry), sp^2 (trigonal geometry) and sp^3 (tetrahedral geometry) combined with the respective atom type constitute the node labels of the orbital graph. The edge labels depend on the orientation of the two interacting orbitals relative to each other. In ToyChem, three types are regarded. Therefore, there are three different edge labels, direct σ-overlap, semi-direct σ-overlap and π-overlap.

2.6   Metabolic Reaction Network

The central subject of the simulation is the metabolism, thus, we need a representation that we can easily observe and also use for analysis tools. In particular, the metabolic flux analysis but also other forms of network, graph or even grammar

Fig. 3. Orbital graph of propenamide. Direct, semi-direct σ-overlaps, and π-overlaps are represented by solid black, dashed, and solid gray lines.

analysis as well. The most intuitive solution seems to be to use a network graph. In case of a metabolic network this could be a bidirectional and bipartite labeled graph. Bipartite because enzymes are only connected to metabolites and not to other enzymes and, vice versa, metabolites are only connected to enzymes. Bidirectional because one metabolite may at one time be the product of an enzyme and another time be the substrate of the same enzyme, therefore, the direction of a connection is important. The nodes are labeled with IDs for metabolites and enzymes. The edge labels contain information about the specific reactions in which an enzyme-metabolite pair was involved. This is necessary because we can identify the exact parts of a reaction which can be up to four metabolites and one enzyme. Further, a metabolite can be the substrate or product of an enzyme in more than one reaction. From this graph the stoichiometry matrix can easily be derived and it has the advantage that no information is lost and can be extracted in a straight forward way for almost all objectives. For example, the single reactions can be listed with substrates and products. Consequently, it is possible to analyze from which different sources a metabolite can be gained. Looking at the enzyme graph may even enable us to specify the exact regions which were joint, split or changed. The interpretability and expressiveness of this network graph, therefore, allows for a very detailed manually analysis as well as the typical computational approaches.

3   Results

We performed several simulation runs and analyzed their results to gain information about the properties of the evolved metabolic networks. Discussed will

be ten simulations, differing in the topological index that is used as selection criteria for reactions and metabolites. We use five different indices and for each of them, two simulations were performed, where one is aiming to reduce the respective index and the other tries to maximize it. We believe the addition of these indices makes it easier to observe different behaviors among the networks. All simulations are initialized with a population of six individuals. The genome for the individuals is chosen randomly, but for all simulations the random number generator (RNG [28]) used for building a random genome and generate random mutations is set with the same seed number. The set of metabolites constituting the environment is the same in all simulations as well. Thus, the simulations start with equal preconditions. Furthermore, in every generation, half the population is selected and from each of the selected individuals a new individual is produced and a mutation in its genome is performed.

Metabolic networks are small world networks. Therefore, the metabolite connectivity distribution follows the power law. In other words, in a realistic metabolic network, a few highly connected metabolites, called hubs, should be observed, whereas the majority of metabolites is involved in only one or two reactions. In order to prove whether this property can be found in the networks produced by the simulation tool, the distribution of the metabolite connectivity was derived. Since the different simulations do not result in networks of equal size and it is known that in small networks, around 50 metabolites, the connectivity distribution does not follow exactly the power law and contains fewer hubs than could expected in a scale free network, we consequently group the networks into sets of networks with similar size. The values of the connectivity distribution are listed in table 2 and illustrated in figure 4.

In all networks, the majority of metabolites is involved in one or two reactions, but only larger networks ($m > 150$) contain enough highly connected metabolites to satisfy the small world property. A similar observation as in [7] can be made. In small networks, the number of hubs in the range between eight and twelve is higher than for scale-free networks, but too few hubs of higher connectivity exist. Most real world metabolic networks contain more than fifty metabolites, thus are large networks. The conclusion about the small-world property of metabolic networks, therefore, was drawn with the assumption that the network of investigation is large ($m \geq 100$). The connectivity distribution of smaller real world metabolic networks, actually, exhibit the same deviations from the power law as the networks gained from the example simulations. Accordingly, we can not state that all produced networks are scale-free, but we can assume them to resemble realistic metabolic networks.

For further analysis, one of the simulations is studied in more detail. We will discuss exemplarily the simulation that uses the minimal Balaban index [21] as selection criteria. In figure 5, 6 and 7 network graphs for generation one, two and 87 are depicted. Enzymes are drawn in light blue circles and metabolites in light gray boxes. The enzyme and metabolite indices are defined in the protocol.

From the network graph in figure 7 it can be derived, that some enzymes catalyze many reactions (e2, e12, e44, e45) and others participate in very few

Table 2. Connectivity of metabolites in networks of different sizes. Frequency in %.

$\frac{Connectivity}{|Metabolites|}$	1	2	3	4	5	6	7	8 - 12	>12
avg(m)=47.5	43.91	12.61	13.03	8.61	8.82	4.83	2.31	4.41	1.47
avg(m)=104.5	50.89	17.1	6.02	5.2	5.61	5.2	3.28	5.06	1.64
avg(m)=156.5	56.21	16.36	5.85	3.56	2.92	3.29	2.1	4.02	5.67
avg(m)=249.5	58.59	12.89	8.13	6.01	2.69	2	1.37	3.67	4.64
avg(m)=570	62.8	12.1	9.12	5.49	2.89	1.2	1.46	2.34	2.63
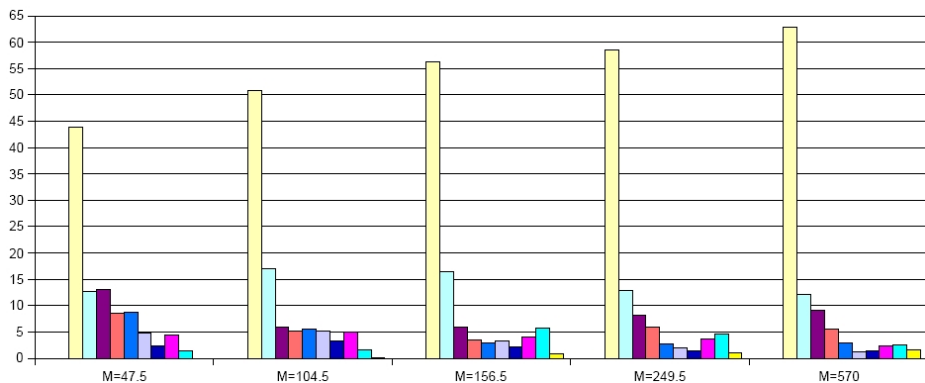
Fig. 4. Connectivity of metabolites in networks of different sizes. Frequency in %. Connectivity of 1, 2, 3, 4, 5, 6, 7, 8-12, >12, >30 for all 5 network classes.

Table 3. Specificity of enzymes in the example network

Enzyme	e37	e45	e12	e44	e18	e27	e6	e82	e4	e62	e124	e130	e2
Generation	1	1	1	2	3	3	7	10	14	42	44	51	53
Connectivity	4	5	12	9	4	2	2	2	2	2	3	1	6

reactions. In the different theories about the evolution of enzymes, it was stated that enzymes of low specificity evolve to highly specific enzymes. This is expected for the simulations as well. To study the evolution of enzymes within the simulation run, we looked at the generation in which the respective enzyme participated for the first time. This information is listed in table 3 for all enzymes in the example network in generation 87. A tendency for early enzymes to be less specific can be observed, in fact, three of the four enzymes with relatively low specificity are from generation one or two. With the exception of $e2$, all enzymes in later generations are more specific. The same observation is made for the other simulations.

The observation stated above is explained with the following scenario. In the beginning, every reaction producing valid metabolites is beneficial for the yield of the network. Since more metabolites are generated, more of the already existing enzymes find reactants. At some point, metabolites will protrude from the
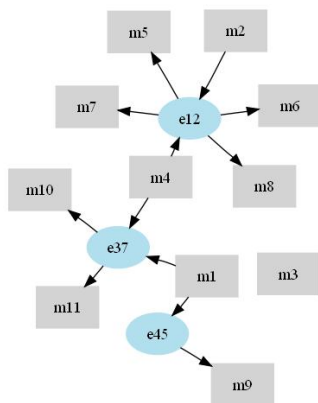
Fig. 5. Example: Network graph from simulation Balaban in generation 1. Light blue circle = enzyme, light gray box = metabolite.
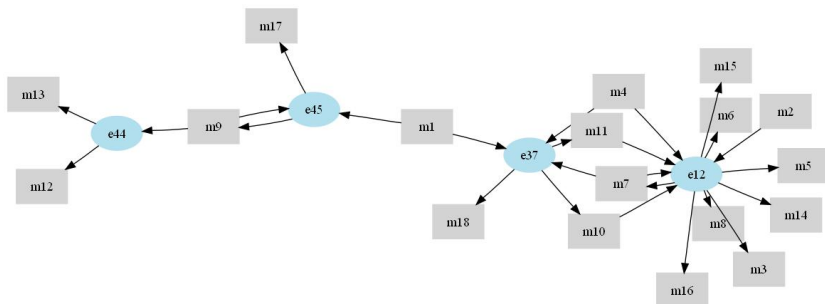

Fig. 6. Example: Network graph from simulation Balaban in generation 2. Light blue circle = enzyme, light gray box = metabolite.

metabolite pool, that is, some metabolites become more beneficial than others. This in turn means that not all of the enzymes increase the network yield. If an enzyme with low specificity overlaps in functionality with another enzyme which is specialized on the few common reactions, then the impact of the former enzyme on these reactions is very low. Since the rate of a reaction depends on how many reactions the involved enzyme performs, the remaining reactions of the lowly specific enzyme are performed on a relatively low rate. A highly specific enzyme which can catalyze the remaining non-overlapping reactions, can do so at much higher rate. Overall, it can be stated that enzymes which have a unique function have an advantage in natural selection. In later generations, most beneficial reactions are already realized by existing enzymes and only few are left. It follows that only specific enzymes can find their niche. However, sometimes lowly specific enzymes enter the network at later stages because they express a completely new functionality, e.g. different atoms, bond type or reaction type. This scenario does not comply exactly with retrograde evolution[29], since it
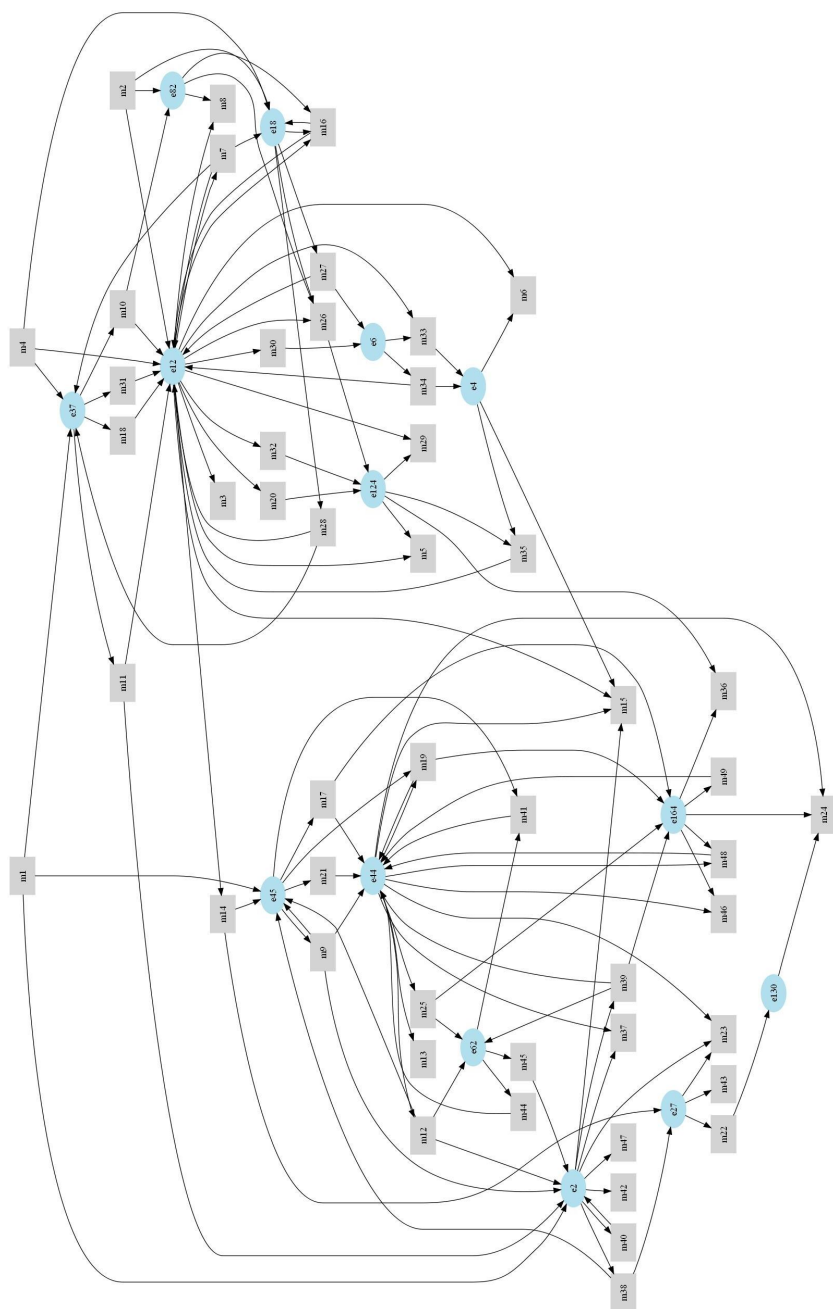
Fig. 7. Example: Network graph from simulation Balaban in generation 87. Light blue circle = enzyme, light gray box = metabolite.

does not need a metabolite depletion, but it integrates to a certain extent the idea of patchwork evolution[30] and the theory of [5].

4   Conclusions and Outlook

The presented simulation model can be a tool in the study of the evolution of metabolism and enzymes, as well as research on properties of complex networks. The underlying graph concept in combination with a sophisticated artificial chemistry and redundant genotype-phenotype map ensures a realistic behavior of the evolution. The resulting metabolic networks exhibit the characteristic properties of real world metabolisms. Various options, from the constitution of the environment and chemistry to selection properties, such as the number of descendants or the use of topological indices as additional criteria, can be adjusted. An extensive amount of information about the simulation can be gained from its protocol. The data about metabolic networks and their evolution over generations, is expressive and meaningful, so that it can be used to formulate new hypotheses or test existing theories.

For the future we are working on the integration of regulatory elements to the model which would add to the complexity of the simulated individuals, leading to a more realistic characteristic of the metabolism properties. It is also planned to interconnect the individuals of a population. Individuals would have to compete for metabolites or could cooperate and build higher-level systems. This would require a change in the modeling of the metabolite pool, so far we do not consider a depletion or shortage of metabolites. Besides the changes of the model, a lot of the future work will consist of developing ways to analyze the simulations and study the network properties in more detail. The focus will be on the emergence of robustness and flexibility.

Acknowledgements

References

1. Liao, J.: Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. Biotechnol. Bioeng. 52, 129–140 (1996)
2. Schuster, S.: Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends Biotechnol. 17, 53–60 (1999)
3. Pfeiffer, T., Sanchez-Valdenabro, I., Nuno, J., Montero, F., Schuster, S.: Metatool: for studying metabolic networks. Bioinformatics 15.3, 251–257 (1999)
4. Papin, J., Price, N., Wiback, S., Fell, D., Palsson, B.O.: Metabolic pathways in the post-genome era. Trends Biochem. Sci. 28(5), 250–258 (2003)
5. Kacser, H., Beeby, R.: Evolution of catalytic proteins. J. Mol. Evol. 20, 38–51 (1984)
6. Diaz-Mejia, J.J., Perez-Rueda, E., Segovia, L.: A network perspective on the evolution of metabolism by gene duplication. Genome. Biol. 8(2) (2007)

7. Pfeiffer, T., Soyer, O.S., Bonhoeffer, S.: The evolution of connectivity in metabolic networks. PLoS Biology 3/7, 228 (2005)
8. Arens, J.: Rec. Trav. Chim. Pays-Bas. 98, 155–161 (1979)
9. DeTar, F.: Modern approaches to chemical reaction searching. Comput. Chem. 11, 227 (1986)
10. Fujita, S.: Description of organic reactions based on imaginary transition structures. 1. introduction of new concepts. J. Chem. Inf. Comput. Sci. 26(4), 205–212 (1986)
11. Hendrickson, J.: Comprehensive system for classification and nomenclature of organic reactions. J. Chem. Inf. Comput. Sci. 37(5), 852–860 (1997)
12. Benkö, G., Flamm, C.: A graph-based toy model of chemistry. J. Chem. Inf. Comput. Sci. 43(4), 1085–1095 (2003)
13. Benkö, G.: A toy model of chemical reaction networks. Master's thesis, Universität Wien (2002)
14. Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28(1), 31–36 (1988)
15. Weininger, D.: Smiles. 2. algorithm for generation of unique smiles notation. J. Chem. Inf. Comput. Sci. 29(2), 97–101 (1989)
16. Houk, K.N., Gonzalez, J.: Pericyclic reaction transition states: Passions and punctilios, 1935-1995. Accounts of Chemical Research 28, 81–90 (1995)
17. Himsolt, M.: GML: A portable Graph File Format (Universität Passau)
18. Trinajstic, N.: Chemical Graph Theory (New Directions in Civil Engineering), 2nd edn. CRC, Boca Raton (1992)
19. Randic, M.: Characterization of molecular branching. J. Am. Chem. Soc. 97(23), 6609–6615 (1975)
20. Platt, J.: Influence of neighbor bonds on additive bond properties in paraffins. J. Chem. Phys. 15, 419–420 (1947)
21. Balaban, A.: Highly discriminating distance-based topological index. Chem. Phys. Lett. 89, 399–404 (1982)
22. Fujita, S.: Description of organic reactions based on imaginary transition structures. 2. classification of one-string reactions having an even-membered cyclic reaction graph. J. Chem. Inf. Comput. Sci. 26(4), 212–223 (1986)
23. Fujita, S.: Description of organic reactions based on imaginary transition structures. 3. classification of one-string reactns having an odd-membered cyclic reaction graph. J. Chem. Inf. Comput. Sci. 26(4), 224–230 (1986)
24. Benkö, G., Flamm, C.: Multi-phase artificial chemistry. In: The Logic of Artificial Life: Abstracting and Synthesizing the Principles of Living Systems (2004)
25. Benkö, G., Flamm, C.: Explicit collision simulation of chemical reactions in a graph based artificial chemistry. In: Capcarrère, M.S., Freitas, A.A., Bentley, P.J., Johnson, C.G., Timmis, J. (eds.) ECAL 2005. LNCS (LNAI), vol. 3630, pp. 725–733. Springer, Heidelberg (2005)
26. Hoffmann, R.: An Extended Hückel Theory. I. Hydrocarbons. J. Chem. Phys. 39(6), 1397–1412 (1963)
27. Gillespie, R.J., Nyholm, R.S.: Inorganic Stereochemistry. Quart. Rev. Chem. Soc. 11, 339–380 (1957)
28. Matsumoto, M., Nishimura, T.: Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model Comput. Simul. 8(1), 3–30 (1998)
29. Horowitz, N.: On the evolution of biochemical syntheses. Proc. Nat. Acad. Sci. 31, 153–157 (1945)
30. Jensen, R.: Enzyme recruitment in evolution of new functions. Ann. Rev. Microbiol. 30, 409–425 (1976)