JOURNAL OF SYSTEMS CHEMISTRY

RESEARCH ARTICLE Open Access

Evolution of metabolic networks: a computational frame-work

Christoph Flamm[1], Alexander Ullrich[2], Heinz Ekker[1,3], Martin Mann[4], Daniel Högerl[3], Markus Rohrschneider[5], Sebastian Sauer[1], Gerik Scheuermann[5], Konstantin Klemm[2], Ivo L Hofacker[1], Peter F Stadler[2,1,6,7,8*]

Abstract

Background: The metabolic architectures of extant organisms share many key pathways such as the citric acid cycle, glycolysis, or the biosynthesis of most amino acids. Several competing hypotheses for the evolutionary mechanisms that shape metabolic networks have been discussed in the literature, each of which finds support from comparative analysis of extant genomes. Alternatively, the principles of metabolic evolution can be studied by direct computer simulation. This requires, however, an explicit implementation of all pertinent components: a universe of chemical reactions upon which the metabolism is built, an explicit representation of the enzymes that implement the metabolism, a genetic system that encodes these enzymes, and a fitness function that can be selected for.

Results: We describe here a simulation environment that implements all these components in a simplified way so that large-scale evolutionary studies are feasible. We employ an artificial chemistry that views chemical reactions as graph rewriting operations and utilizes a toy-version of quantum chemistry to derive thermodynamic parameters. Minimalist organisms with simple string-encoded genomes produce model ribozymes whose catalytic activity is determined by an *ad hoc* mapping between their secondary structure and the transition state graphs that they stabilize. Fitness is computed utilizing the ideas of metabolic flux analysis. We present an implementation of the complete system and first simulation results.

Conclusions: The simulation system presented here allows coherent investigations into the evolutionary mechanisms of the first steps of metabolic evolution using a self-consistent toy universe.

Introduction

Computer models of the transition between an abiotic chemosphere and a primitive biosphere are plagued by the complexity of the systems and processes that need to be integrated into a coherent picture. Individual aspects and components, such as thermodynamic boundary conditions, the dynamics of self-replication, the effects of coding [1], the influence of spatial organization and compartmentalization, can be – and have been – tackled with their own specific minimal models. Much of the most successful modeling e orts have been invested in early systems of information propagation. The success of these approaches can at least in part be explained by the fact that generic behavioral regularities can be extracted independent of physical details. It is entirely sufficient to consider linear sequences that can be copied, mutated, ligated, and cleaved according to rules that do not have to recurse explicitly to underlying physics and chemistry [2-5].

We argue here that the situation becomes fundamentally different once we become interested in metabolic evolution. Then, chemistry (and in particular the complexities and subtleties of organic chemistry) is moved to center stage and any model must encapsulate the ground rules of chemical transformations: the conservation of mass and atomic types as well as conservation and dissipation of energy introduces constraints that critically determine the system's behavior. This does not mean, of course, that it is necessary to implement all of chemistry in all its natural beauty and with all its intricate details. Nevertheless, it calls for a frame-work that

* Correspondence: studla@bioinf.uni-leipzig.de
[2]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
Full list of author information is available at the end of the article

goes much beyond most implementations of artificial chemistries or the string models of polymer systems.

In principle, so we argue, we eventually will need to understand the transition to life, and the first steps in the evolution of primitive life-like systems, in terms of their chemical organization. Nevertheless, it appears prohibitively inefficient to even attempt an atom-level simulation, and even if it were feasible, it is not clear what insights were to gain from it. Instead, we would like to understand, and implement, information molecules and other hetero-polymer components by their sequence, at least in part because we already understand their behavior at that level. Practical simulations in this field, therefore, will necessarily have to have components at different scales and implement them using different modeling paradigms, leaving us with the question how to bridge the gaps between these different layers.

In this contribution we describe a particular framework in which questions about the most primitive "life-forms" and their evolution can be addressed. As we shall see, it combines a grossly simplified Chemical Universe (see Figure 1) with a very minimal, RNA-World style, genetics, and simple fitness function linked to metabolic efficiency.

The Chemical Model Universe

Artificial Chemistries

Many models of artificial chemistries have been explored in recent years. The spectrum ranges from chemically accurate quantum mechanical simulations to abstract computational models. Walter Fontana's `AlChemy`[6,7], for example represents molecules as λ-calculus expressions and reactions are defined by the operations of "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 [8-14], see also the reviews [15,16]. The abstract computational models are very useful for understanding algebraic properties of reaction systems; the notion of a *self-maintaining set* and the development of a theory of Chemical Organizations [17] emphasizes the success of such approaches.

Reaction energies pose severe constraints on chemical reactions networks, by selecting one or a few preferred reaction pathways from a plethora of logically possible reaction channels. An energy function, that behaves as a state variable and allows the modeling of transition states, is therefore indispensable for any model that is

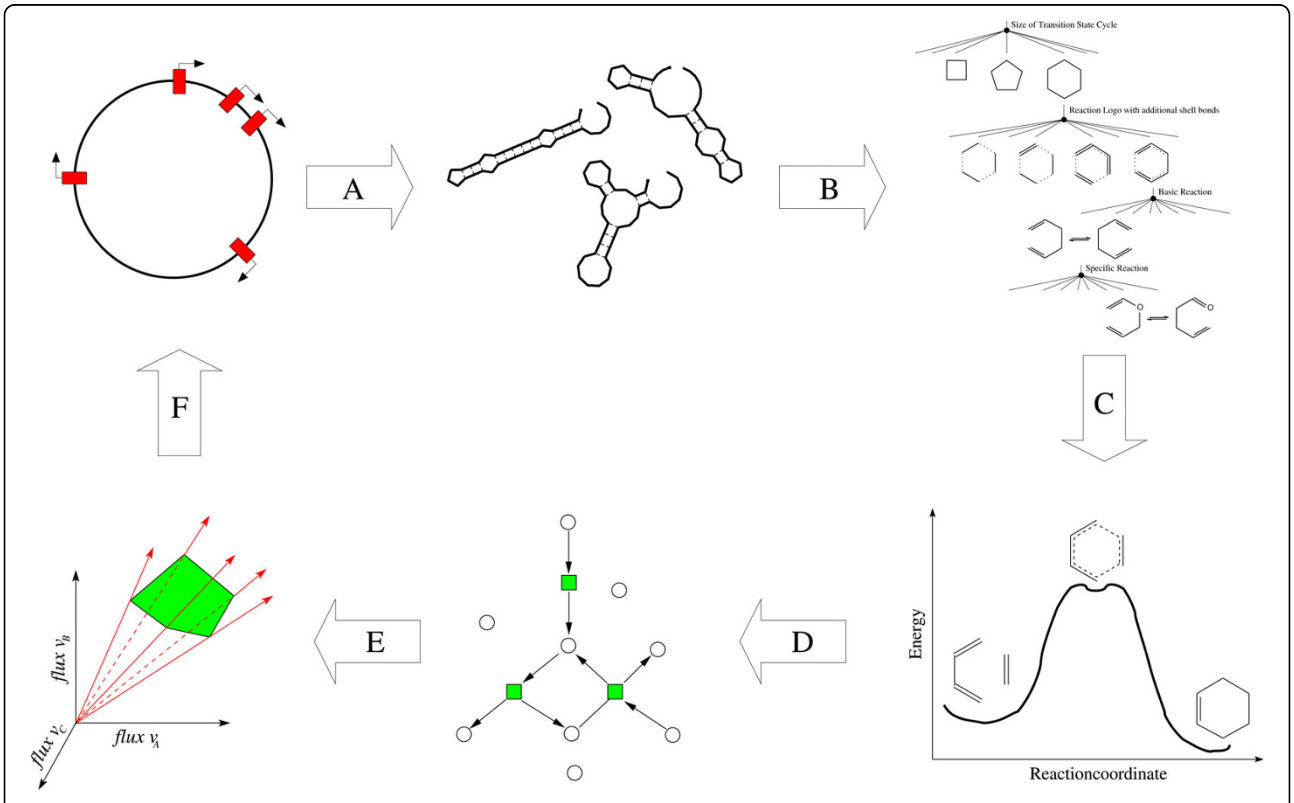

Figure 1 Overview of the simulation system. (A) Transcription of genes into catalytic ribozymes; (B) Assignment of catalytic functions to each ribozyme; (C) Estimation of reaction rates; (D) Construction and stochastic simulation of the metabolic network; (E) Metabolic Flux analysis and fitness evaluation; (F) Application of genetic variation operators.

realistic enough to allow us to consider, say, the differences between a bacterial metabolic network and atmospheric chemistry of the planet Mars. 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 at 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 [18] provided a view biased by the history of chemical research. Knowledge-based approaches hence appear less attractive for our endeavor.

The particular Chemical Universe that underlies our simulation is motivated by the way how chemical reactions are explained in introductory Organic Chemistry classes: in terms of structural formulae (labeled graphs) and reactions mechanisms (rules for modifying graphs).

Molecules

Historically, the description of molecular structures was one of the roots of graph theory [19,20]. 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 we avoid the most time-consuming computation step: embedding the atoms in 3 D by means of finding the minima on a potential energy surface [21]. 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 `C = C` double bond. (2) there is no notion of asymmetric atoms and chirality.

Energy

As argued in the introduction, a consistent energy function is indispensable in a meaningful model of chemistry since all chemical transformations are associated with well-defined energy differences that determine e.g. the direction in which a chemical reaction will proceed. The `ToyChem` model [22] 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 [23]. Application of a simplified version of the Extended Hückel. Theory (EHT) [24] 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 we refer to Ref [22].

The `ToyChem` model was used to study the generic graph-theoretic properties [25] 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 [26].

The Klopman-Salem equation [27-29] connects the wave function to the reactivities of molecules, paving the way to study the kinetic properties of chemical reaction networks. However, it turned out, that the reaction rate estimates calculated with the Extended Hückel method implemented in the ToyChem model are too inaccurate especially to study time-scale separation in the time evolution of pre-biotic reaction networks. The reason for this problem is that the accuracy of the rate constant depends exponentially on the quality of the energy predictions. More realistic estimates of reaction rates therefore 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. For our purposes, semi-empirical methods like PM3 (implemented for example in `Mopac` and `GAUSSIAN`) might be better suited, although the results are not very reliable. 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 [30,31].

Three tasks are required to automate reaction rate calculations when using any one of the quantum chemistry packages: (i) a fast and high-quality 3 D embedding of the molecular graphs, (ii) the correct pre-orientation of the educt structures in bi-molecular reactions as well as a good guess of the transition state geometry, and (iii) a fast and reliable reaction mapping assigning atoms from the educts to the respective atoms of the products. The knowledge-based program `CORINA`[32] is used to generate high-quality 3 D structures from molecular graphs. The *Imaginary Transition State* (ITS) of the reaction [33,34], see below and Figure 2, guides the construction of a transition structure analogon, which is then 3D-embedded by CORINA. Splitting the embedded transition state analogon into the educts results in properly pre-orientated reactants.

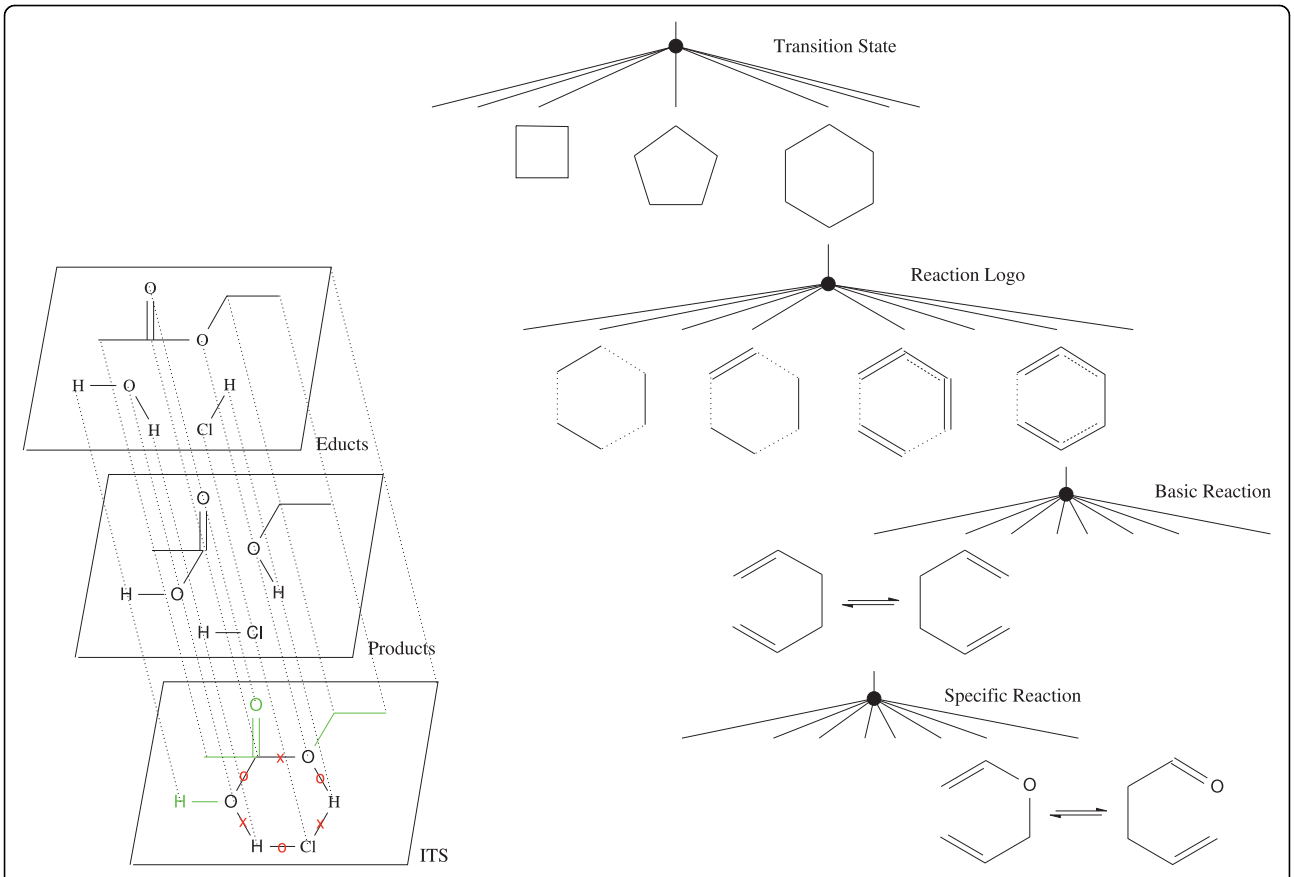

Figure 2 Imaginary Transition State (ITS) and their hierarchical organization. Superimposition of educt and product molecular graphs and subsequent removal of all atoms and bonds which do not directly participate in the chemical reaction (marked in green) yields a cyclic ITS for a chemical reaction (e.g. acidic hydrolysis of ethylacetate). Bonds which are broken/formed during the reaction are marked with a red x/o. Left: 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. Mono-cyclic ITS account for over 90% of all known chemical reactions [34].

For situations in which even 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, we use the QSPR and the approach for activation energy computation from Faulon, delivering still realistic results [35], for the calculation of the rate constants. We gain the final reaction rate, 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).

Chemical Reactions: Graph Rewriting

Once we represent the molecules as (labeled) graphs it becomes natural to view reactions as graph trans-formations. 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. The part of chemistry that does not deal with energy, therefore, can be modeled and understood as a graph grammar. The applicability of rewriting-based approaches to metabolic network data was demonstrated recently in an analysis of KEGG data [36].

A graph rewriting rule is specified as a triple consisting of left *graph, context, and right graph*, see Figure 3. 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 a affected by the reaction itself. Note that in proper chemical reactions all vertices (atoms) involved in the reaction are part of the context of the rewrite rule because they neither disappear nor are newly created. The ITS of the reaction is intimately connected to the left and right graphs. It is obtained

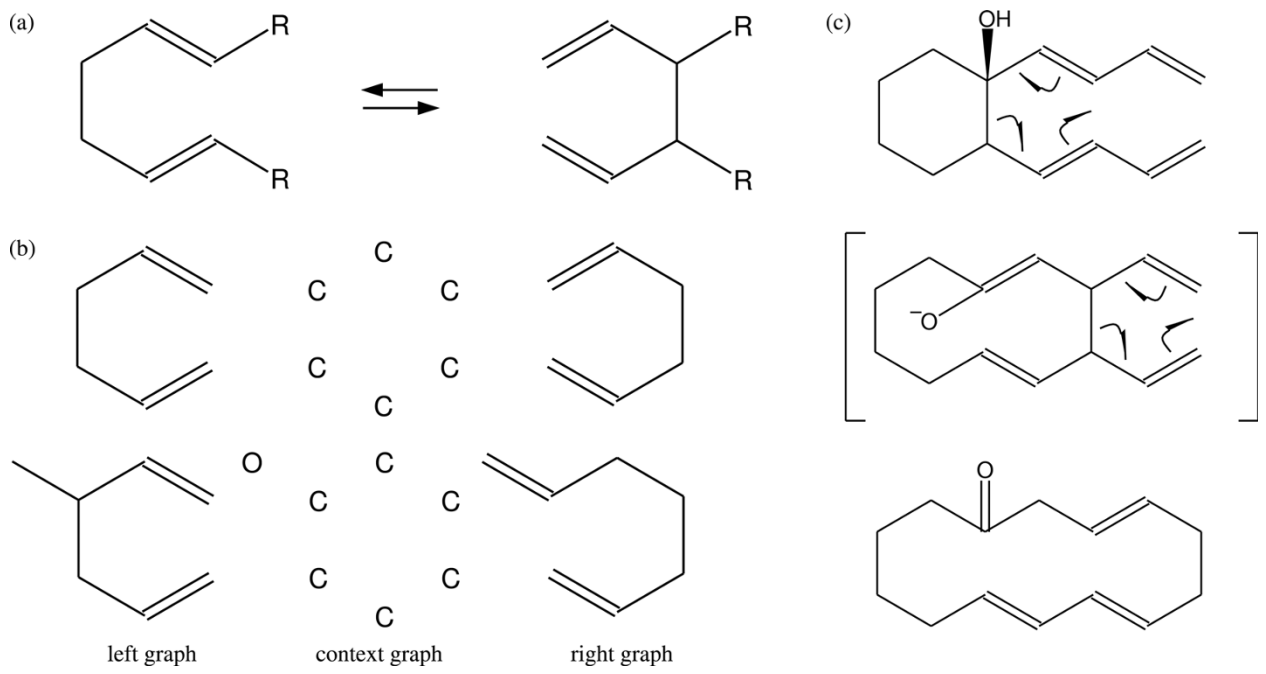

Figure 3 Reactions as Graph Grammars. Chemical transformations very naturally translate into graph transformation rules. As an example the Cope rearrangement, a concerted pericyclic [3,3]-sigmatrope rearrangement, is shown (a). A graph rewrite rule consists of 3 graphs: (i) the left graph which is composed 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 (b). 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 (c).

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. The ITS thus can be derived from the rewriting rule provided the mapping of the vertices (atoms) between the left graph and the right graph is known. A variant of the cut successive largest (CSL) algorithm [37] is used to predict the atom mapping from the educt and product molecular graphs automatically. The performance of the original CSL algorithm was drastically improved by replacing the expensive complete subgraph isomorphism check with an efficient subgraph isomorphism check [38] augmented by a chemically motivated heuristic scoring schema for bond breaking energies. The correctness of the automatic atom mappings were validated against the KEGG RPAIR database [39].

A graph rewrite system [40] interprets the graph rewrite rule and performs the graph rewriting step if the graphical pre-condition is matched in a host graph. We utilize here a generic graph rewrite engine. 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 we apply the dedicated state-of-the-art VF-algorithm, freely available in the C++ VFlib-2.0 library [41,42]. 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 [43] to test for duplicates. The initial molecule(s) and the resulting ones are utilized to generate the ITS (Figure 2) needed to calculate the transition rate for the applied reaction. Our fully generic object oriented C++ implementation is freely available as the Graph Grammar Library (GGL) [44].

Reaction Networks

Once molecules and reactions have been implemented, it is conceptually trivial to construct the complete chemical reaction network 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 [35,45], Figure 4.

Suppose we are given a list of reaction mechanisms and an initial list $\mathcal{L}_0$. Performing all unimolecular

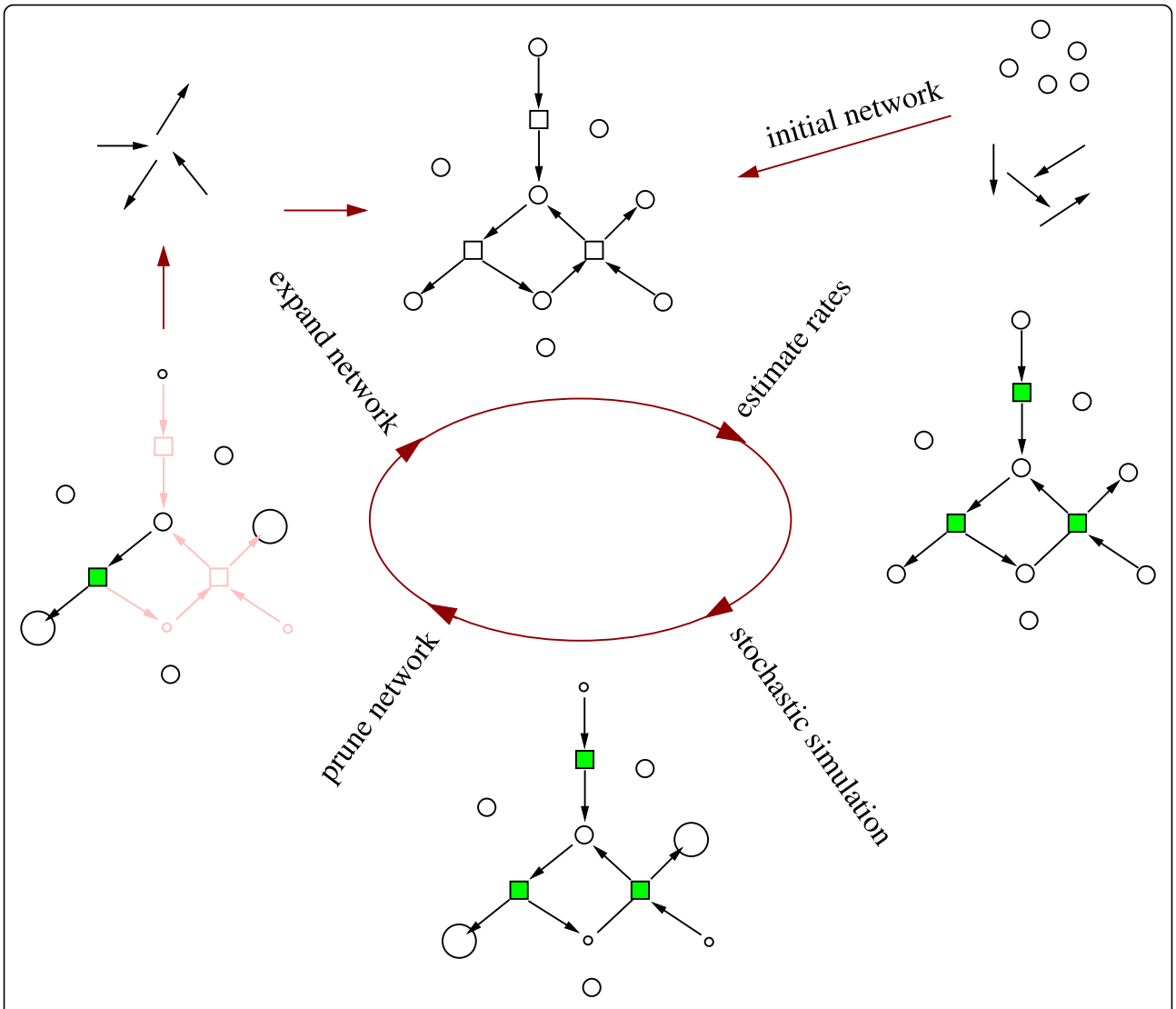

Figure 4 Generating Reaction Network. To avoid combinatorial explosion during reaction network generation a filtering step, which prunes unproductive parts from the reaction network, is needed after each application of the reaction set (arrows) to the (current) set of molecules (circles). The network usually quickly converges in size if the filtering is performed based on reaction kinetics. 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.

reactions on each molecule $M \in \mathcal{L}_0$ and all bimolecular reactions with each pair of molecules $(M_1, M_2) \in \mathcal{L}_0 \times \mathcal{L}_0$ we obtain a new list $\mathcal{L}_1'$ and a list of new molecules $\mathcal{L}_1 = \mathcal{L}_1' \setminus \mathcal{L}_0$. The recursion then proceeds in the obvious way:

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

and $\mathcal{L}_{k+1} = \mathcal{L}'_{k+1} \setminus \cup \mathcal{L}_k$. This type of strategy [35] 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 we also exclude all molecules with more than 30 atoms in order to keep the e orts 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 [46] 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 [47], which then are compared as strings.

Artificial Molecular Biology

Minimalist Genomics and Genetics

The "players" in our Simulation Universe are modeled as complex agents that are characterized by individual *genomes*. This genome is necessary and sufficient to encode the individual's metabolic, i.e., catalytic capabilities.

We are interested here primarily in the earliest stages of metabolic evolution, which arguably took place in the setting of the Early RNA World [48]. In this setting, RNA has the double role of genetic material and serves as catalysts. Both the analysis of naturally occurring ribozymes and a wide variety of artificial selection experiments have shown that RNA molecules of about 100 nt are capable of catalyzing most important types of chemical transformations that occur in a modern organism, see [49-51] 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, we use a very simple genomic organization: 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.

Our organisms are though to be haploid. As genetic operators we currently use only point mutations as well as gene duplication. More sophisticated modes of genome evolution such as rearrangements, recombination, or lateral gene transfer are excluded at present but could easily be incorporated into the computational framework.

We refrain from modeling in detail any form of gene regulation to reduce the computational e orts. Again, such refinements could be included e.g. along the lines of [52,53]. Our minimal organisms 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.

Artificial Biochemistry

Ribozyme Catalysis

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:

$$\text{sequence} \mapsto \text{structure} \mapsto \text{function}$$

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 [54], is well approximated by the usual RNA folding algorithms. RNA molecules form secondary structure by folding back onto themselves to form double helical regions interspersed with unpaired regions termed "loops". The resulting secondary structure can be rep-resented by an outer planar graph with nucleotides as vertices and base pairs as edges. A well established energy model [55], 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 we use the folding routines as implemented in the `Vienna RNA package`[56-58].

For the structure-to-function mapping, unfortunately, we do not have a well-understood physically realistic model. Instead, we employ a simple purely computational model based on structural features motivated by early models of RNA evolution [59]. Catalytic structures typically depend on the molecular details of an active center, which we abstract here to a local motif contained in a secondary structure. We use here the longest "loop" (cycle) of the secondary structure as a computationally easily accessible feature of this type.

Without any claim of physical realism, we interpret this cycle 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 reaction by stabilizing its transition state and the work on reaction classification systems, in particular Fujita's imaginary transition structures (ITS) approach [33], in which cycles also play a central role. All common homo- and ambivalent reactions, which account for over 90% of all known reactions [60], can be described by a mono-*cyclic* ITS [34]. The rest of the reactions are usually composites of successive mono-cyclic reactions in sequence (rarely more than two [61]) with unstable intermediates like carbene or nitrene.

In order to construct and evaluate the structure-to-function map we utilize a hierarchical classification of imaginary transition states [62]. 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 5. The "reaction logo" specifies in addition the bond types in the transition state. We use the length and the type of the

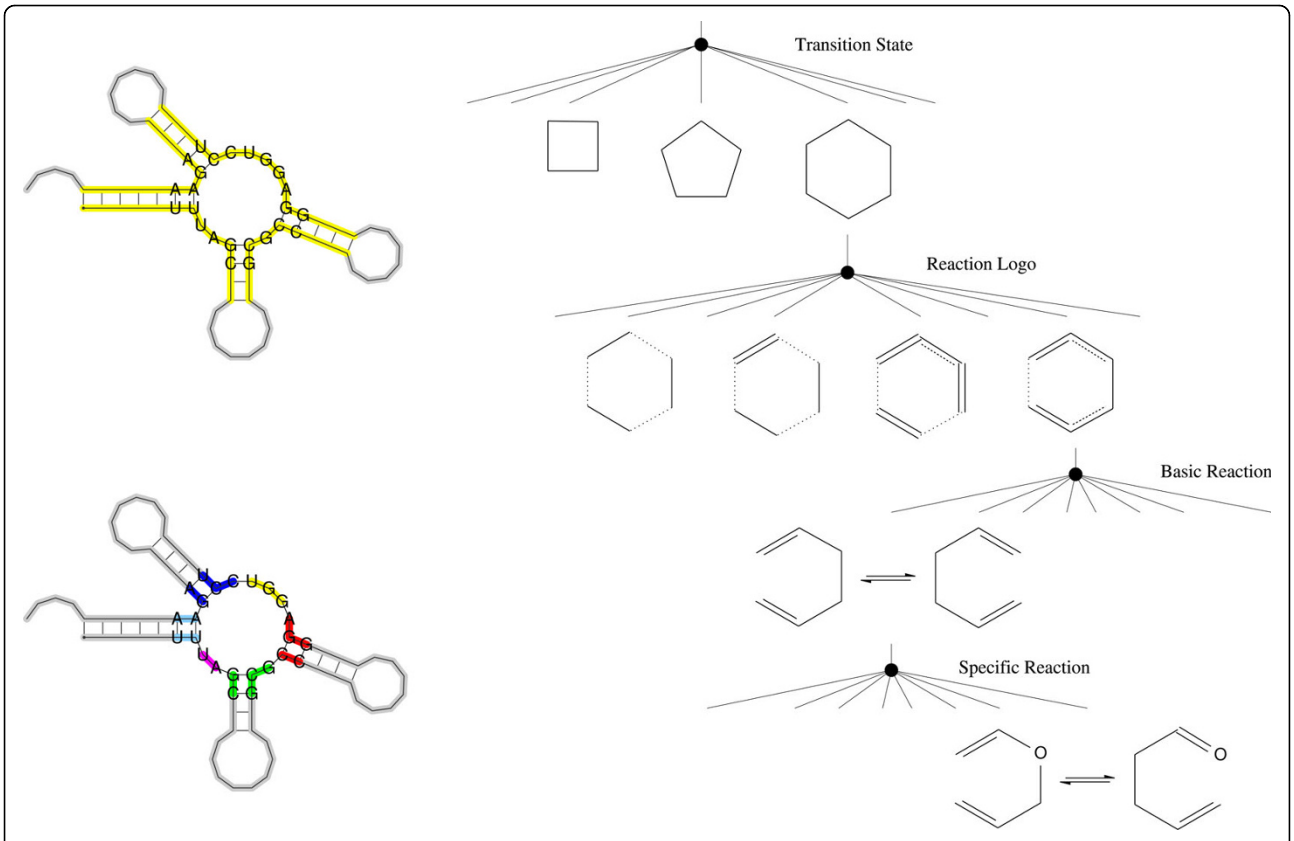

Figure 5 The structure-to-function map. (left) The colored regions of the ribozyme fold determine the catalytic function i.e. which leaf in the ITS-tree is picked; (right) Along the levels of the ITS-tree the amount of chemical detail increases.

enclosing base pairs of the adjacent stems to determine the bond types. 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 regions stands for one part in the transition state, here the atoms. The details of the mapping are specified in [63].

Since the structure-to-function map is not based on an approximation of physico-chemical principles but on an *ad hoc* model, we need to investigate its statistical properties. To this end, we consider in particular its autocorrelation function of the sequence-to-function map and compare it to the autocorrelation function of the sequence-to-structure map of RNA folding [64]. To this end, we need distance measures on the spaces of RNA structures and transition states, 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 [54]. 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 [65]. Edit operations include substitution of atoms, rearrangement of electron re-ordering, 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.

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

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

where $\langle D^2\rangle$ denotes the expected distance between the images $\varphi(x)$ and $\varphi(y)$ of two independent elements $x, y \in X$ [54]. Figure 6 shows that the composite sequence-to-function map behaves much like the underlying sequence-to-structure map. This is not surprising:

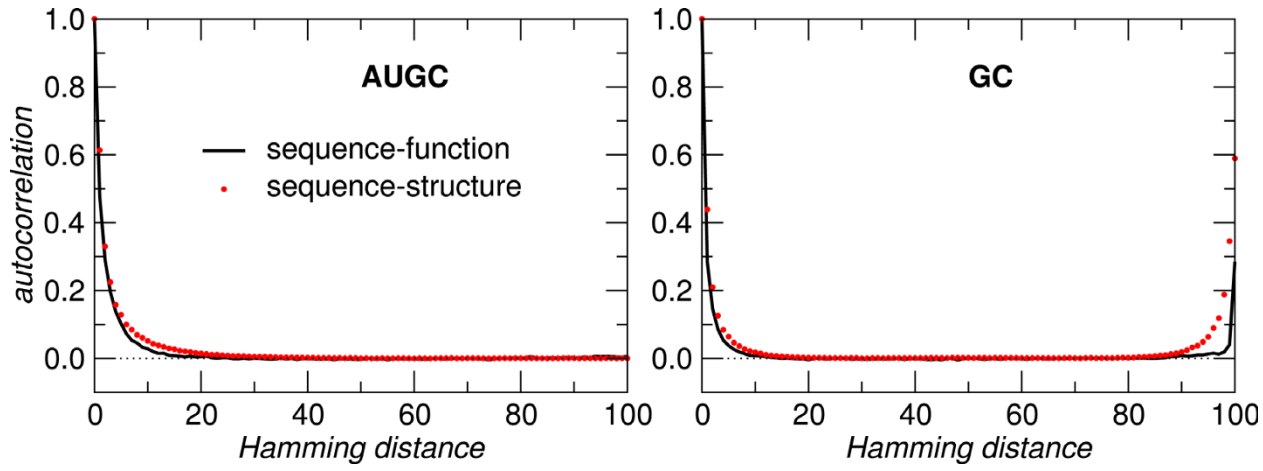

Figure 6 Autocorrelation functions for sequences of length $n = 100$ for secondary structure landscape and transition-state landscape, with alphabets AUGC (left) and GC (right). For each of 1000 randomly generated reference sequences we produced 1000 mutants for each of the 100 Hamming distance classes.

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 [66]. This observation in particular justifies the use of an *ad hoc* artificial structure-to-function map in our simulation setting.

In other work [67] we showed that the composite map, of RNA sequence-to-structure map and our novel structure-to-function map, 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. Thus, making it the preferable choice for our model and possibly other similar optimization tasks.

Fitness and Selection

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

The fitness of our minimal organisms 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. We determine the pathway distribution of the metabolic network under the steady-state assumption using metabolic pathway analysis (MPA) [68]. 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, our simulations represent metabolic networks as bipartite graphs composed of metabolite and reaction nodes.) The steady state assumption implies that we are interested in non-negative flux vectors $\vec{v}$ in the null-space of **S**, i.e., $S\vec{v} = \vec{o}$. We assume that catalyzed reactions have a non-zero flux only in one direction. Our implementation of MPA 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 [69]. The fitness is therefore computed as the maximum of the (linear) yield function over all extreme pathways.

This fitness function depends on our 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 we offer several different topological indices such as Balaban-Index [70] or Wiener-Number [71]. 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".

In principle, selection could be handled in an agent-based framework [72], using e.g. tournaments of pairs of individuals taken from a population of competing model organism. Due to the computational e orts necessary to construct and evaluate each metabolic network, however, we are currently limited to small populations. In many cases we work with a single individual, resorting to the simplest possible model of *adaptive walks*, which applies in the limit of strong selection, weak mutation, negligible interactions between individuals, and constant

environment [73]. 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 [74] that explain the emergence of hub metabolites.

Visualization

The analysis of complex simulations is impossible without efficient visualization tools, in particular in the current exploratory phase of research, where it is not clear at all which evolutionary patterns we will encounter and which aggregate statistics can be used to summarize the simulation results.

A suitable representation for a metabolic network is a directed bipartite graph with two types of nodes, i.e. reactions and molecules. The adjacent nodes of a reaction are the substances consumed – incoming edges - or produced - outgoing edges. For intuitive visual distinction, we represent reactions as yellow squares and molecule nodes as white circles. The number of potential fluxes through a graph element is represented by the node size, or edge width, respectively. We used here the orthogonal grid based layout [75]. Nodes lie on even integer grid positions. Edges run along grid segments and may bend at any grid position. The layout algorithm allows multi-edges but no loops. In the first step of the algorithm, the nodes are placed at crossing sections of a regular grid minimizing the global stress. The edge routing step places edges on a sequence of grid lines minimizing a global edge cost function taking into account the number of bends, crossings, edge length and segment densities. The last step displaces edges running along the same segments to avoid overlapping edge routes. This approach does not take edge directions into account. When visualizing flux dynamics of the network, this may not be the most intuitive method. Nevertheless, the produced drawings are fairly compact and appear to be more aesthetic even for larger graphs.

Evolution of Metabolisms

Simulation-based studies of the evolution of metabolisms so far pre-suppose the presence of an elaborate complement of metabolic enzymes and focused on the structural changes of network of catalyzed reactions under the action of mutations that change enzyme specificities, see e.g. [74]. On the other hand, there is mounting evidence that biochemical aspects, such as similarities among catalyzed reactions and their coupling in pathways influences the evolutionary patterns of the many gene families that encode metabolic enzymes [76,77].

The simulation system described in the previous sections attempts to address this point head on. Instead of artificial high-level proxies of the underlying chemical entities, we strive to simulate the entire chemical universe as completely as possible. As a consequence, we can in particular address the origin of metabolism itself. We may start from a primitive proto-cell that is just about to invent its first enzyme, and ask how the internalization of chemical reactions into metabolic pathways proceeds in the most early steps of molecular evolution.

Several different scenarios have been discussed in the literature, reviewed recently in [78]:

1. The *retrograde hypothesis* [79] postulates that internalization starts with the last reaction in the path-way and stepwisely generates, via gene duplication, enzymes the extend the pathway to more an more distant starting points. Examples include histidine biosynthesis and nitrogen fixation.
2. The *forward hypothesis* [80], in contrast, suggest that evolution proceeds from simple to complex biochemical compounds, so that the oldest enzymes are those at the beginning of pathways. A good example is the isoprene lipid pathway.
3. The *patchwork hypothesis* [81,82] suggests that metabolic pathways may have been assembled through the recruitment of primitive enzymes that could react with a wide range of chemically related substrates.
4. The *semi-enzymatic hypothesis* [83] posits the emergence of a limited number of "starters", novel protein classes that later diversified into large paralog groups. These starters would have taken over originally uncatalyzed reactions of stable abundant chemical species in the environment or of (by-)products of already established pathways.

Although all these mechanisms appear to have left traces in the extant metabolic networks (see [78] and the references therein), their relative importance in early evolution remains unclear. First, replacement of enzymes by non-homologous functional analogs may have been a frequent process in the early phases of evolution (which may have been dominated by horizontal gene transfer [84]). This would superimpose a patchwork pattern on the metabolic network that eventually eradicates genomic traces of more ancient states. If LUCA was predated by a ribo-organism with an elaborate ribozyme-catalyzed metabolism, the ancestral catalysts have been completely replaced by peptide-based innovations. This would naturally produce a pattern consistent with the predictions of Lazcano and Miller, with novel protein families rapidly replacing functionally analogous ribozymes throughout the system.

We argue, therefore, that insights into the earliest evolutionary history of metabolism cannot be safely based on comparative studies of the extant enzyme repertoires,

since the latter may have emerged much later than the metabolic pathways themselves. As an alternative we propose here to explore systematically the selection pressures and advantages associated with the genetic internalization of reaction pathways that arise from the underlying chemistry itself. Although we cannot - yet - report on a coherent scenario of metabolic evolution, our first simulations show that a simulation approach utilizing a full- edged artificial chemistry and complex model of biopolymers is feasible, Figure 7.

For this contribution, we performed two simulation runs for a length of 100 generations. Both runs were

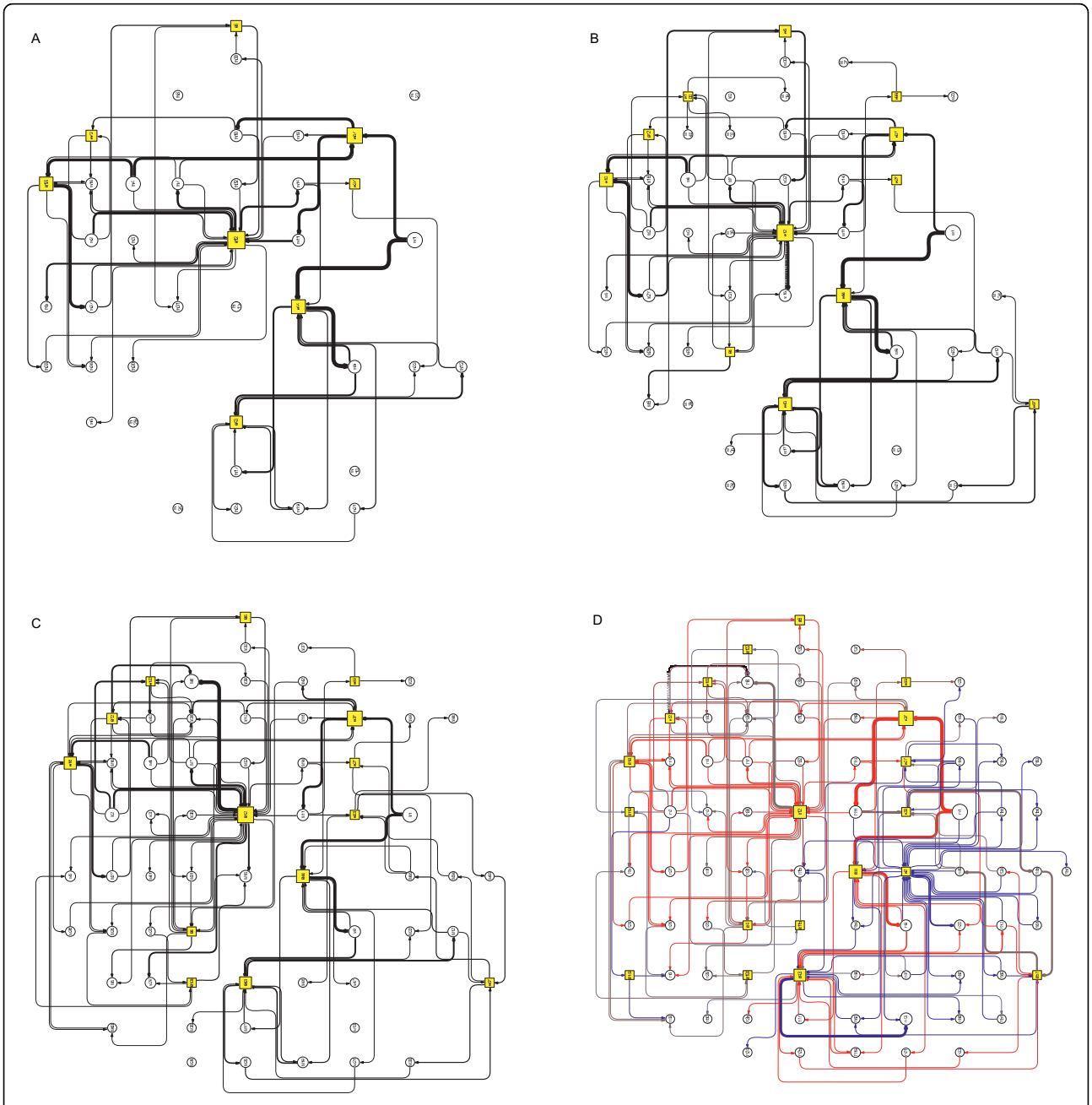

Figure 7 A series of simulated metabolic networks after (a) 10, (b) 40, (c) 50, and (d) 100 generations. Yellow squares represent enzymes, gray circles represent metabolites. An edge leading from a metabolite to an enzyme indicates that the respective metabolite is an educt in the reaction. An edge from an enzyme to a metabolite marks it as a product. The size of the nodes and the width of the edges encode for the number of minimal pathways in which the respective object is involved. The colors in panel (d) encode the age of the reactions, where red stands for old and blue for newer reactions.

initialized with the full set of chemical reactions to choose from, the same configurations for genome length (5000 bases), TATA-box constitution ("UAUA") and gene length (100 bases) but with different starting conditions and different selection criteria.

The first simulation run (Figure 7) starts with a population of ten cells in an environment constituting of a set of five chemical molecules, namely, cyclobutadiene, ethenol, phthalic anhydride, methylbutadiene, and cyclohexa-1,3-diene. In each generation cells are selected based on their production of molecules with maximal Balaban-Index (this leads to molecules with a high degree of ramification). The selected cells produce one copy of themselves, which might include a mutation or duplication event.

The second simulation run (see animation in [Additional file 1]) with a starting population of just one cell, increasing in size up to 100 cells. The environment consists of glucose only. Fitness values correspond here to production of molecules with a maximal Wiener number. Until the population reached 100 cells, no selection is performed. Afterwards the same procedure of selection and multiplication as above is applied.

In these two computer experiments we observe that enzymes and metabolites introduced in earlier stages of the simulation are involved in more reactions. In a previous study [63] we showed for a series of long simulations (1000 generations) that the resulting networks have structural properties similar to those of real-world metabolic networks. In particular, the node degree distribution follows a power law and consequently there are hub metabolites. This effect could either constitute a generic feature of network evolution that can be explained by preferential attachment [85], or it could be governed by chemical properties that single out highly connected metabolites. Figure 7(d) shows that most of the more recent reactions (blue) have a smaller flux (thin line), while older reactions (red) have a bigger flux (thick line). This effect might support the patchwork hypothesis provided the flux increase is associated with enzyme recruitment. In order to clarify the interrelations of chemical properties, evolutionary age, strength of flux, and to quantify a possible preference for attachment to high connectivity and/or high flux nodes, more extensive simulations with different initial compositions will be necessary. Currently, we are investigating the similarity to metabolic networks from pathway databases, in terms of more dynamical properties based on the set of extreme pathways [86] and minimal knockout sets [87].

The simulation system described here demonstrates that computational investigations into major organizational transitions – here the transition from a chemical reaction system to a genetically controlled metabolism – are feasible in practice. Our work also exemplifies that progress in this direction requires the construction of multi-scale models in which different components (chemical reactions, catalytic biopolymers, and genetic machineries) are represented as different levels of abstraction. A major difficulty with such models lies in the construction of the interfaces between the levels of description, highlighted here by the sub-system modeling the catalytic function of our "ribozymes".

Additional material

Additional file 1: Animation of metabolic network evolution. An animation of the evolution of a metabolic network, using data from a sample simulation run over 100 generations.

Acknowledgements

This work was supported in part by the Vienna Science and Technology Fund (WWTF) proj. no. MA07-030, the Austrian GEN-AU project "bioinformatics integration network III", the Volkswagen Stiftung proj. no. I/82719, and the COST-Action CM0703 "Systems Chemistry".

Author details

[1]Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Wien, Austria. [2]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany. [3]FH Campus Wien, Diplom-Studiengang Bioengineering (Diploma Degree Course), Muthgasse 18, 1190 Wien, Austria. [4]Bioinformatics Group, Institute for Computer Science, Albert-Ludwigs-University of Freiburg, Georges-Köhler-Alle 106, 79110 Freiburg, Germany. [5]Image and Signal Processing Group, Department of Computer Science, University of Leipzig, Johannisgasse 26, D-04109 Leipzig, Germany. [6]Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany. [7]Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany. [8]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA.

Authors' contributions

CF, KK, ILH, and PFS designed the study, AU, HE, MM, DH, MR, SS implemented various components of the software and performed the simulations, MR and GS contributed the visualization. CF and PFS drafted the manuscript. All authors contributed to and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 28 January 2010 Accepted: 18 August 2010
Published: 18 August 2010

References

1. Bentley P, Kumar S: **Three Ways to Grow Designs: A Comparison of Evolved Embryogenies for a Design Problem.** *Genetic and Evolutionary Computation Conference*. Massachusetts: Morgan Kaufmann 1999, 35-43.
2. Banzhaf W: **On the Dynamics of an Artificial Regulatory Network.** *Advances in Artificial Life, of LNCS*. Heidelberg, Germany: Springer-VerlagBanzhaf W, Christaller T, Dittrich P, Kim JT, Ziegler J 2003, **2801**:217-227.
3. Eggenberg P: **Evolving morphologies of simulated 3 D organisms based on differential gene expression.** *Proc. ECAL97*. The MIT Press/Bradford Books. Husbands P, Harvey I 1997, 205-213.
4. Geard N, Wiles J: **Structure and dynamics of a gene network model.** *Proc. CEC2003*. IEEE Press. Sarker R, Reynolds R, Abbass H, Tan KC, McKay B, Essam D, Gedeon T 2003, 199-206.

5. Reil T: **Dynamics of gene expression in an artificial genome - inplications for biological and artificial ontogeny.** *Proc. ECAL99, of Lecture Notes in Computer Science.* Berlin: Springer-Verlag. Floreano D, Nicoud JD, Mondada F 1999, **1674**:457-466.
6. Fontana W: **Algorithmic Chemistry.** *Artificial Life II.* Redwood City, CA: Addison-Wesley. Langton CG, Taylor C, Farmer JD, Rasmussen S 1992, 159-210.
7. Fontana W, Buss LW: **What would be conserved if `the tape were played twice'?** *Proc Natl Acad Sci USA* 1994, **91**:757-761.
8. Bagley RJ, Farmer JD: **Spontaneous emergence of a metabolism.** *Artificial Life II, Santa Fe Institute Studies in the Sciences of Complexity.* Redwood City, CA: Addison-Wesley. Langton CG, Taylor C, Farmer JD, Rasmussen S 1992, 93-141.
9. Banzhaf W, Dittrich P, Eller B: **Self-organization in a system of binary strings with spatial interactions.** *Physica D* 1999, **125**:85-104.
10. Speroni di Fenizio P: **A less abstract artificial chemistry.** *Artificial Life VII.* Cambridge, MA: MIT Press. Bedau M, McCaskill J, Packard N, Rasmussen S 2000, 49-53.
11. Ugi I, Stein N, Knauer M, Gruber B, Bley K, Weidinger R: **New Elements in the Representation of the Logical Structure of Chemistry by Qualitative Mathematical Models and Corresponding Data Structures.** *Top Curr Chem* 1993, **166**:199-233.
12. Thürk M: **Ein Modell zur Selbstorganisation von Automatenalgorithmen zum Studium molekularer Evolution.** *PhD thesis.* Universität Jena, Germany 1993.
13. McCaskill JS, Niemann U: **Graph Replacement Chemistry for DNA Processing.** *DNA Computing, of Lecture Notes in Computer Science.* Berlin, D: Springer. Condon A, Rozenberg G 2000, **2054**:103-116.
14. Rossellá F, Valiente G: **Chemical graphs, chemical reaction graphs, and chemical graph transformation.** *Electron Notes Theor Comput Sci* 2005, **127**:157-166.
15. Dittrich P, Ziegler J, Banzhaf W: **Artificial chemistries-a review.** *Artif Life* 2001, **7**:225-75.
16. Suzuki H, Dittrich P: **Artificial chemistry.** *Artif Life* 2009, **15**:1-3.
17. Centler F, Kaleta C, di Fenizio PS, Dittrich P: **Computing chemical organizations in biological networks.** *Bioinformatics* 2008, **24**:1611-1618.
18. Grzybowski BA, Bishop KJM, Kowalczyk B, Wilmer CE: **The wired universe of organic chemistry.** *Nature Chemistry* 2009, **1**:31-36.
19. Cayley A: **On the Mathematical Theory of Isomers.** *Philos Mag* 1874, **47**:444-446.
20. Sylvester JJ: **On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics, with three appendices.** *Amer J Math* 1878, **1**:64-128.
21. Heidrich D, Kliesch W, Quapp W: *Properties of Chemically Interesting Potential Energy Surfaces, of Lecture Notes in Chemistry.* Berlin: Springer-Verlag 1991, **56.**
22. Benkö G, Flamm C, Stadler PF: **A graph-based toy model of chemistry.** *J Chem Inf Comp Sci* 2003, **43**:1085-93.
23. Gillespie RJ, Nyholm RS: **Inorganic Stereochemistry.** *Quart Rev Chem Soc* 1957, **11**:339-380.
24. Hoffmann R: **An Extended Hückel Theory. I. Hydrocarbons.** *J Chem Phys* 1963, **39**:1397-1412.
25. Benkö G, Flamm C, Stadler PF: **Generic Properties of Chemical Networks: Artificial Chemistry Based on Graph Rewriting.** *Advances in Artificial Life, of Lecture Notes in Computer Science.* Heidelberg, Germany: Springer-Verlag. Banzhaf W, Christaller T, Dittrich P, Kim JT, Ziegler J 2003, **2801**:10-20.
26. Benkö G, Flamm C, Stadler PF: **Multi-Phase Artificial Chemistry.** *The Logic of Artificial Life: Abstracting and Synthesizing the Principles of Living Systems.* Berlin: IOS Press, Akademische Verlagsgesellschaft. Schaub H, Detje F, Brüggemann U 2004, 16-22.
27. Klopman G: **Chemical reactivity and the concept of charge- and frontier-controlled reactions.** *J Am Chem Soc* 1968, **90**:223-243.
28. Salem L: **Intermolecular Orbital Theory of the Interaction between Conjugated Systems. I. General Theory.** *J Am Chem Soc* 1968, **90**:543-552.
29. Salem L: **Intermolecular Orbital Theory of the Interaction between Conjugated Systems. II. Thermal and Photochemical Calculations.** *J Am Chem Soc* 1968, **90**:553-566.
30. Wodrich MD, Corminboeuf C, Schreiner PR, Fokin AA, von Ragué Schleyer P: **How accurate are DFT treatments of organic energies?** *Org Lett* 2007, **9**:1851-1854.
31. Brittain DRB, Lin CY, Gilbert ATB, Izgorodina EI, Gill PMW, Coote ML: **The role of exchange in systematic DFT errors for some organic reactions.** *Physical chemistry chemical physics: PCCP* 2009, **11**:1138-1142.
32. Gasteiger J, Rudolph C, Sadowski J: **Automatic Generation of 3 D Atomic Coordinates for Organic Molecules.** *Tetrahedron Comp Method* 1990, **3**:537-547.
33. Fujita S: **Description of Organic Reactions Based on Imaginary Transition Structures. 1. Introduction of new concepts.** *J Chem Inf Comput Sci* 1986, **26**:205-212.
34. Hendrickson JB: **Comprehensive System for Classification and Nomenclature of Organic Reactions.** *J Chem Inf Comput Sci* 1997, **37**:852-860.
35. Faulon JL, Sault AG: **Stochastic generator of chemical structure. 3. Reaction network generation.** *J Chem Inf Comput Sci* 2001, **41(4)**:894-908.
36. Félix L, Rosselló F, Valiente G: **Efficient Reconstruction of Metabolic Pathways by Bidirectional Chemical Search.** *Bull Math Biol* 2009, **71**:750-769.
37. Crabtree JD, Mehta DP: **Automated Reaction Mapping.** *J Exp Algor* 2009, **13**:1-29.
38. Ullmann JR: **An Algorithm for Subgragraph Isomorphism.** *J ACM* 1976, **23**:31-42.
39. Kotera M, Hattori M, Oh MA, Yamamoto R, Komeno T, Yabuzaki J, Tonomura K, Goto S, Kanehisa M: **RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions.** *Genome Inform* 2004, **15**:P062.
40. Nagl M: *Graph-Grammatiken, Theorie, Implementierung, Anwendung.* Braunschweig: Vieweg 1979.
41. Cordella LP, Foggia P, Sansone C, Vento M: **Performance Evaluation of the VF Graph Matching Algorithm.** *ICIAP* 1999, 1172-1177.
42. Cordella LP, Foggia P, Sansone C, Vento M: **An Improved Algorithm for Matching Large Graphs.** *3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition* 2001, 149-159.
43. Weininger D: **SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules.** *J Chem Inf Comp Sci* 1988, **28**:31-36.
44. Mann M, Flamm C: **Graph Grammar Library (GGL).** 2010 [http://www.tbi.univie.ac.at/TBI/software.html].
45. Read RC: **Every one a winner.** *Ann Discr Math* 1978, **2**:107-120.
46. Kun A, Papp B, Szathmary E: **Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks.** *Genome Biol* 2008, **9**:R51.
47. Weininger D, Weininger A, Weininger JL: **SMILES. 2. Algorithm for generation of unique SMILES notation.** *J Chem Inf Comput Sci* 1989, **29**:97-101.
48. Gesteland RF, Cech TR, Atkins JF: *The RNA World.* Woodbury, NY: Cold Spring Harbor Laboratories Press, 3 2006.
49. Müller UF: **Re-creating an RNA world.** *Cell Mol Life Sci* 2006, **63**:1278-1293.
50. Chen X, Li N, Ellington AD: **Ribozyme Catalysis of Metabolism in the RNA World.** *Chemistry & Biodiv* 2007, **4**:633-655.
51. Talini G, Gallori E, Maurel MC: **Natural and unnatural ribozymes: Back to the primordial RNA world.** *Res Microbiol* 2009, **160**:457-465.
52. Stephan-Otto Attolini C, Stadler PF, Flamm C: **CelloS: a Multi-level Approach to Evolutionary Dynamics.** *Advances in Artificial Life: 8th European Conference, ECAL 2005, of Lect. Notes Comp. Sci.* Berlin: Springer Verlag. Capcarrere MS, Freitas AA, Bentley PJ, Johnson CG, Timmis J 2005, **3630**:500-509.
53. Flamm C, Endler L, Müller S, Widder S, Schuster P: **A minimal and self-consistent *in silico* cell model based on macromolecular interactions.** *Philos Trans R Soc Lond B Biol Sci* 2007, **362**:1831-1839.
54. Fontana W, Konings DA, Stadler PF, Schuster P: **Statistics of RNA secondary structures.** *Biopolymers* 1993, **33(9)**:1389-404.
55. Mathews DH, Sabina J, Zuker M, Turner H: **Expanded Sequence Dependence of Thermodynamic Parameters Provides Robust Prediction of RNA Secondary Structure.** *J Mol Biol* 1999, **288**:911-940.
56. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures (The Vienna RNA Package).** *Monatsh Chem* 1994, **125**:167-188.
57. Hofacker IL: **The Vienna RNA Secondary Structure Server.** *Nucl Acids Res* 2003, **31**:3429-3431.
58. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL: **The Vienna RNA Websuite.** *Nucl Acids Res* 2008, **36**:W70-W74.

59. Fontana W, Schuster P: **Continuity in evolution: on the nature of transitions.** *Science* 1998, **280**:1451-5.
60. Herges R: **Organizing Principle of Complex Reactions and Theory of Coarctate Transition States.** *Angew Chem Int Ed* 1994, **33**:255-276.
61. Hendrickson JB, Miller TM: **Reaction indexing for reaction databases.** *J Chem Inf Comput Sci* 1990, **30**:403-408.
62. Herges R: **Coarctate Transition States: The Discovery of a Reaction Principle.** *J Chem Inf Comput Sci* 1994, **34**:91-102.
63. Ullrich A, Flamm C: **Functional Evolution of Ribozyme-Catalyzed Metabolisms in a Graph-Based Toy-Universe.** *Proceedings of the 6th International Conference on Computational Methodes in Systems Biology (CSMB), of Lect. Notes Bioinf.* Berlin: Springer. Istrail S 2008, **5307**:28-43.
64. Fontana W, Stadler PF, Tarazona P, Weinberger ED, Schuster P: **RNA folding and combinatory landscapes.** *Physical Review E* 1993, **47**:2083-2099.
65. Ullrich A: **Evolution of Metabolism in a graph-based Toy-Universe.** *PhD thesis.* Universität Leipzig, Germany 2008.
66. Stadler PF: **Fitness Landscapes Arising from the Sequence-Structure Maps of Biopolymers.** *J Mol Struct* 1999, **463(1-2)**:7-19.
67. Ullrich A, Flamm C: **A Sequence-to-Function Map for Ribozyme-catalyzed Metabolisms.** *ECAL, Lect Notes Comp Sci* 2009.
68. Palsson BO: *Systems Biology: Properties of Reconstructed Networks.* New York, NY, USA: Cambridge University Press 2006.
69. Gagneur J, Klamt S: **Computation of elementary modes: a unifying framework and the new binary approach.** *BMC Bioinformatics* 2004, **5**:175.
70. Balaban AT: **Highly discriminating distance-based topological index.** *Chem Phys Lett* 1982, **89**:399-404.
71. Wiener H: **Structural Determination of Para n Boiling Points.** *J Am Chem Soc* 1947, **69**:17-20.
72. Axelrod R: *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration.* Princeton, NJ: Prince-ton University Press 1997.
73. Orr HA: **The evolutionary genetics of adaptation: a simulation study.** *Genet Res Camb* 1999, **74**:207-214.
74. Pfeiffer T, Soyer OS, Bonhoeffer S: **The Evolution of Connectivity in Metabolic Networks.** *PLoS Biol* 2005, **3**:e228.
75. Rohrschneider M, Heine C, Reichenbach A, Kerren A, Scheuermann G: **A Novel Grid-based Visualization Approach for Metabolic Networks with Advanced Focus and Context View.** *17th International Symposium on Graph Drawing (GD09), Lect. Notes Comp. Sci.* Springer. Emden Gansner DE 2009.
76. Díaz-Mej a JJ, Pérez-Rueda E, Segovia L: **A network perspective on the evolution of metabolism by gene duplication.** *Genome Biol* 2007, **8**:R26.
77. Papp B, Teusink B, Notebaart RA: **A critical view of metabolic network adaptations.** *HFSP J* 2009, **3**:24-35.
78. Fani R, Fondi M: **Origin and evolution of metabolic pathways.** *Phys Life Rev* 2009, **6**:23-52.
79. Horowitz NH: **On the evolution of biochemical syntheses.** *Proc Natl Acad Sci USA* 1945, **31**:153-157.
80. Granick S: **Speculations on the origins and evolution of photosynthesis.** *Ann NY Acad Sci* 1957, **69**:292-308.
81. Ycas M: **On earlier states of the biochemical system.** *J Theor Biol* 1974, **44**:145-160.
82. Jensen RA: **Enzyme recruitment in evolution of new function.** *Annu Rev Microbiol* 1976, **30**:409-425.
83. Lazcano A, Miller SL: **The origin and early evolution of life: Prebiotic chemistry, the Pre-RNA world, and time.** *Cell* 1996, **85**:793-798.
84. Doolittle RF: **Evolutionary aspects of whole-genome biology.** *Curr Opin Struct Biol* 2005, **15**:248-253.
85. Newman MEJ: **Power laws, Pareto distributions and Zipf's law.** *Contemporary Physics* 2005, **46**:323-351.
86. Behre J, Wilhelm T, von Kamp A, Ruppin E, Schuster S: **Structural robustness of metabolic networks with respect to multiple knockouts.** *J Theor Biol* 2008, **252**:433-41.
87. Haus UU, Klamt S, Stephen T: **Computing knock out strategies in metabolic networks.** *J Comp Biol* 2008, **15**:259-68.

doi:10.1186/1759-2208-1-4
Cite this article as: Flamm *et al.*: Evolution of metabolic networks: a computational frame-work. *Journal of Systems Chemistry* 2010 **1**:4.