Modeling the complex dynamics of enzyme-pathway coevolution

Moritz Schütte
Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany and Bioinformatics Program, Boston University, Boston, MA 02215, USA

Alexander Skupin
Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany

Daniel Segrè
Bioinformatics Program, Department of Biology, Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

Oliver Ebenhöh
Institute of Medical Sciences, Institute of Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, U.K. and Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany

(Dated: December 26, 2010)

Metabolic pathways must have coevolved with the corresponding enzyme gene sequences. However, the evolutionary dynamics ensuing from the interplay between metabolic networks and genomes is still poorly understood. Here, we present a computational model that generates putative evolutionary walks on the metabolic network using a parallel evolution of metabolic reactions with their catalyzing enzymes. Starting from an initial set of compounds and enzymes, we expand the metabolic network iteratively by adding new enzymes with a probability that depends on their sequence-based similarity to already present enzymes. Thus, we obtain simulated time courses of chemical evolution in which we can monitor the appearance of new metabolites, enzyme sequences, or even entire organisms. We observe that new enzymes do not appear gradually but rather in clusters which correspond to enzyme classes. A comparison with Brownian motion dynamics indicates that our system displays biased random walks similar to diffusion on the metabolic network with long-range correlations. This suggests that a quantitative molecular principle may underlie the concept of punctuated equilibrium as enzymes occur in bursts rather than by phyletic gradualism. Moreover, the simulated time courses lead to a putative time-order of enzyme and organism appearance. Among the patterns we detect in these evolutionary trends is a significant correlation between the time of appearance and their enzyme repertoire size. Hence, our approach to metabolic evolution may help understand the rise in complexity at the biochemical and genomic levels.

Evolution is a dynamic process in which species become extinct and new species emerge all the time. It is a disputed question whether the emergence of new species proceeds with an approximately constant rate or whether new species rather evolve in short periods with a high speciation rate which are separated by long silent periods in which only few new species evolve. The latter scenario is referred to as 'punctuated equilibrium' and has recently experienced empirical evidence. Here, we present a model of metabolic evolution which suggests that punctuated equilibrium is also observed in the evolution of macromolecules. This supports the hypothesis that underlying molecular mechanisms are responsible for the phenomenon of punctuated equilibrium in the evolution of new species. Our model uses available amino acid sequences for thousands of enzymes present in several hundred different organisms. By comparing all these sequences, we estimate probabilities that sequences may have evolved from one another. This information allows for simulating putative scenarios how today's metabolism might have evolved. By time series analysis we demonstrate that the existing sequence information strongly supports a punctuated equilibrium behavior and we also demonstrate that this behavior is considerably less pronounced if sequence information is deliberately neglected.

Introduction

The evolution of the modern biochemical pathways from an early proto-metabolism must have been shaped by innovations concurrently involving enzymes and chemical compounds [1–3]. While it is generally assumed that

today's enzymes have evolved from a few ancestors that were able to catalyze the first reactions, the details of this evolutionary history are almost as uncertain as the details about the first self-replicating systems themselves [4–9]. Several scenarios have been proposed, the simplest suggesting a 'forward' evolution in which enzymes evolved that could make use of the end products of existing metabolic pathways [10]. In the reverse assumption of a retrograde evolution, a necessary precursor became depleted and enzymes have evolved that replenish this required resource from other, still abundant, substances [11]. While for both views supporting example pathways may be found, the more complex assumption of a patchwork evolution [12, 13] became more prominent when viewing metabolism as a whole. The method of network expansion [14, 15] provides a simple evolutionary model that extends the forward evolution scenario on the metabolic network comprising all biochemical reactions known to date. While this approach was useful to relate structural to functional properties [16] by tracing catalytic properties along the evolutionary tree [17], discovering hints for an early separation of DNA and RNA metabolism [18] and providing insight into the increase of complexity upon the rise of oxygen in the Earth's atmosphere [19], it is clearly too simple to reproduce realistic evolutionary paths. More recent models elaborating on these ideas include the toolbox model of metabolic evolution [20], which assumes that network evolution is driven by the need to explore new resources and can readily explain the apparent quadratic scaling of the numbers of transcription factors with the total number of genes. The view of metabolic evolution as a Markov process in which additions or removal of reactions depend on the numbers of neighboring reactions [21] allows to estimate parameters for the evolutionary dynamics and to assess possible evolutionary paths between two different network configurations.

The above mentioned examples all provide plausible arguments for a particular evolutionary path, but it is evident that, after the appearance of the first catalyzed reaction networks, the discovery of new chemical compounds is strongly linked to the evolution of new enzymes from existing ones. Hints that the evolution of the sequence space defining contemporary enzymes mirrors to some extent the gradual expansion of the chemical space, defined by the variety of metabolites, were found by correlating sequence similarities to a distance of the catalyzed reactions on the metabolic network [22].

It was argued [23] that such a coevolution promotes short term avalanches during which a large number of new enzymatic steps are invented, thus giving rise to a punctuated equilibrium behavior [24, 25]. In this paper, we present a model of metabolic evolution combining genome scale data, tools from bioinformatics, dynamic modeling and time series analysis with the goal of studying the apparent coevolution of small molecules and catalysts in further detail. As a basis for our exploration, we use the KEGG database [26, 27] which provides a comprehensive collection of biochemical reactions from several hundred organisms and information on amino acid sequences of the respective catalyzing enzymes. While existing models [28–30] investigate the evolution of artificial metabolic networks, our model explicitly considers available biological data and assumes that those enzymes are more likely to evolve for which a related enzyme has already been discovered. We systematically explore how the evolutionary dynamics depends on the coevolution of metabolites and enzymes by introducing a tunable parameter reflecting the importance of sequence similarity. Thus, we can separate the effects of a sequence-based evolution from one in which the discovery of new enzymes is only restricted by stoichiometric constraints. We find that simulations taking into account existing sequence data display a punctuated equilibrium behavior and thus support the view that evolution, also on the level of metabolic networks, occurs in bursts of rapid sequences of new inventions, rather than in a gradual fashion [24].

Model description

The enzymes found in contemporary organisms are highly efficient and usually very specific catalysts for chemical reactions. The amino acid sequences of present-day enzymes are the outcome of a long evolutionary history, in which they were subjected to random mutations and selective pressures favoring only particular sequences which may efficiently perform useful functions. A difficulty in modeling the evolutionary process of enzyme evolution is that neither sequences for early or extinct enzymes nor the precise criteria for the selective pressures are known.

Our proposed simple model for the evolution of metabolism takes these limitations into account. Instead of aiming at describing the evolution of networks of particular organisms, we focus on the network comprising reactions from several hundreds of species. We can thus focus on very general selective principles and ignore the specific pressures that were acting to support the evolution of highly specialized functions. Due to the lack of knowledge of early and now extinct protein sequences, our model is limited to all described biochemical reactions and sequence information available to date.

We mimic the evolution of the network comprising the presently known metabolic reactions by a simple process in which the network grows in size by consecutive addition of single enzymes. The process is initiated by assuming that a certain combination of primitive metabolites are abundant in the environment. We assume that new enzymes may evolve from existing ones through a series of amino acid exchanges. Since the mechanism for such mutations is essentially a random process, we assume that the probability to discover a new functional enzyme from an existing one is higher, the more similar their corresponding sequences are. Thus, at any stage

of our simulated evolutionary process in principle every known enzyme may evolve. However, we assume that only those newly discovered enzymes will be positively selected which can perform a useful function. We therefore impose a selective pressure by accepting only those new enzymes which can catalyze a biochemical reaction from reactants that may in principle be produced from reactions already present in the network. A temporal scale is introduced by assuming that evolutionary events with a higher probability tend to occur faster.

The evolutionary simulation is implemented as a Gillespie algorithm [31] for the simulation of stochastic expansion processes and can be summarized in the following 7 steps:

1. A set of primitive compounds and first enzymes is selected. These comprise the initial network.

2. On the basis of the actual network structure, all enzymes that can catalyze a reaction utilizing only substrates present in the network are identified. For each enzyme i, a propensity p_i is calculated in dependence on the sequence similarity with already present enzymes (see below). The propensity describes the probability that the enzyme is discovered in one unit time.

3. Depending on the propensities, the time t_{next} of the next evolutionary event is determined by an exponentially distributed random variable with the mean given by $1/\sum p_i$.

4. Which particular enzyme is added at time t_{next} is determined by a uniformly distributed random number. The probability that enzyme j is selected is given by $p_j/\sum p_i$.

5. All reactions catalyzed by the selected enzyme as well as the corresponding products, are added to the network.

6. Due to the incorporation of new substances, new reactions catalyzed by enzymes already present in the network may be executable. These reactions and their products are added as well. The same holds true for any newly occurring spontaneous reactions.

7. The process is repeated with step 2 until no new enzymes can be added to the network.

Iterating this expansion process leads to a series of invented enzymes where their invention times depend on the underlying dynamics. Hence, we use the inter-enzyme intervals (IEI) which are defined by the sequence of t_{next} and correspond to waiting times to characterize the evolutionary process.

In contrast to the conceptually similar method of network expansion introduced in [15], in our model enzymes are considered to be the basic units of the networks rather than reactions. As a consequence, the discovery of a new enzyme leads to the addition of all reactions that it can catalyze. Moreover, whereas in the method of network expansion all reactions which can possibly occur are simultaneously added to the growing network in each step, we here only add a single enzyme in each expansion event. Defining probabilities for enzyme appearance introduces a stochastic component which is inherent to all evolutionary processes. Further, by assigning characteristic times for the single evolutionary events our model possesses an intrinsic definition of an evolutionary time coordinate.

Like in many applications of the method of network expansion (see e.g. [32, 33]), we also assume that common cofactors do not specifically have to be produced during the expansion process before they can be used (see Methods). The rationale for this is that their metabolic functions can in principle also be carried out by simpler pairs of molecules. For example, the transfer of phosphate groups by ATP/ADP is possible by pyrophosphate and phosphate, as demonstrated in the bacterial phosphotransferase system, the role of NADH/NAD^+ as electron carriers can in principle be performed by metal ions such as Fe^{2+}/Fe^{3+} with different oxidation states.

Sequence distances and propensities

One particular focus of our model is the investigation of the evolutionary dynamics for different assumptions on how strongly the evolvability of novel enzymes from existing ones depends on the respective sequence similarities.

For roughly half of all functionally different enzymes present in the KEGG database [27], sequence information is available. The amount of information for one particular enzyme commission (EC) number can vary between none and a couple of thousand different sequences. In total KEGG (release 53) provides around one million sequences from various organisms for about 3000 EC numbers. We reduced the space of possible sequences by construction of a consensus set using the clusters of orthologous groups of proteins (COG) database [34–37] as a benchmark. This enables to drop redundant sequences between functionally equivalent enzymes having the same EC number, see [22] and the Methods section. Herewith, we obtain a set of 11925 sequences that code for 3048 EC numbers for which we calculate all mutual distances using the '*score*' of the best BLAST alignment to assess the probability that sequence A evolves into B. The Blast '*score*' provides evolutionary knowledge about single amino acid substitutions, insertions, and deletions from which we define the distance between sequence A and B as

$$D_{AB} = 1 - \frac{2 \cdot score(A,B)}{score(A,A) + score(B,B)} \,. \qquad (1)$$

The pairwise distance ranges from 0 for identical sequences to 1 for sequences without any significant alignment. For EC numbers for which no sequence is avail-

able, we assign distances to all other enzymes randomly from the distribution of all calculated sequence distances. While it is certainly possible that this introduces a bias in our results, we consider this approach the best possibility under the circumstances of incomplete information.

It is plausible to assume that, during evolution, the probability to discover a new enzyme is higher, if a similar enzyme already exists. We denote by $d_i^{\min}$ the minimal distance for enzyme i to all enzymes that have already been found. To have a tunable parameter that weighs the strength of the influence of the protein sequences, we define the propensities for a new enzyme to be discovered by

$$p_i = \frac{1}{d_i^{\min\gamma}}. \quad (2)$$

This definition implies that we assume that the expected time to find a new enzyme depends only on the minimal distance to existing enzymes scaled by the exponent γ. The extreme assumption of $\gamma = 0$ leads to equal propensities for all possible new enzymes and thus reflects a hypothetical case in which sequence information has no influence on the selective process, and the evolution of the network is exclusively determined by chemical constraints. A value of $\gamma = 2$ corresponds to the assumption that the possible sequence space is explored in a process analogous to a random walk, for which the average distance covered is proportional to the square root of the elapsed time. The other extreme, $\gamma \to \infty$, reflects the hypothetical case of the path following least resistance in which always the enzyme with the closest distance to an existing enzyme will be discovered in the next step [23, 38].

Methods

Data for network structure and sequences

We use the KEGG database, release 53. The "genes.pep" in fasta format is downloaded for protein sequences and the ligand-file for reactions. In order to curate the data erroneous reactions, that are not balanced or contain unspecified parts like a rest group, are rejected. The irreversibility information is obtained by scanning the pathway maps [32, 39]. Such reactions that contain the cofactor pairs ATP/ADP, NAD/NADP, NADH/NADPH, or Co-A/Acetyl-CoA are added a second time pured of the cofactors, assuming that they are possible during expansion but no cofactor is produced.

Sequence distance and consensus set

In order to calculate the evolutionary distance between any two enzyme we use pairwise sequence alignment using BLAST (version 2.2.22, standalone blast, http://www.ncbi.nlm.nih.gov/blast/blast_overview.shtml): *bl2seq -i SequenceA -j SequenceB -p blastp -F F -o output* with the default substitution matrix BLOSUM62. Then we score the alignment by the best hit of the '*score*' from the output getting the distance D_{AB} given by Eq. (1). This is not a distance by mathematical definition as it does not fulfill the triangular inequality. It ranges from 0, identical, to 1, completely different [22]. KEGG release 53 contains around one million sequences containing an EC (enzyme commission) number in their description. We sort these sequences by EC numbers and choose representatives from each set by taking only into account those that have a mutual distance, defined similarly to (1), above 0.95 and drop all others. The cutoff has been chosen according to a benchmark using the clusters of orthologous groups of proteins (COG) database, [22, 34]. This reduces the sequence set to 11925 for 3048 EC numbers. For EC numbers that appear in the reaction set but for which we do not have a sequence we randomly pick a distance to any other sequence from the distribution of distances between all known ones.

Seeds

What are the first metabolites and enzymes? We use the following primordial seed of compounds: H_2O, CO_2, H_2SO_4,H_2PO_3, NH_3 and H^+, [1, 40]. In order to identify the putative first enzymes, we use work by Y. Sobolevsky [41–43] who identified common conserved protein fragments in 131 proteomes. One particularly long fragment *LSGGQQQRVAIARAL* was found in bmn:BMA10247_1739 and tpe:Tpen_0904 and we added the remaining two of the same function 3.6.3.21, eco:b2306 and hpa:HPAG1_0922, to the seed of enzymes.

Results

The expansion: process and enzyme sequences

The expansion process starts from a given set of metabolites and enzymes, called the seed [14–17]. This set represents a putative prebiotic chemical environment. A necessary requirement for the evolution of a substantial reaction network is the presence of all essential chemical elements in the seed. Here, we consider only the atoms H, C, O, N, P and S, because 80% of all metabolites in the KEGG database are composed of these elements. As seed, we choose H_2O, CO_2, H_2SO_4, H_2PO_3, NH_3 and H^+ [1, 40]. The choice of a first enzyme sequence is made by using conserved sequence fragments [41–43] to be enzymes of the function 3.6.3.21, see Methods.

To study the effect of sequence information, controlled by the parameter γ introduced in Eq. (2), we perform simulations with five values $\gamma = 0, 2, 10, 20, 100$. We account for the stochasticity of the simulated evolutionary

walks by performing 200 simulation runs for each selected value of γ, which correspond to scenarios in which sequence information is completely ignored ($\gamma = 0$) to the case in which a strict order of enzyme appearance is imposed by the sequence relatedness ($\gamma = 100$). As a direct consequence of the definition of the propensities of Eq. (2), simulations with different γ proceed on very different time scales, with the total time required to explore the entire network (inlet in Fig. 1A) being roughly 1000 times longer for the random scenario when compared to the scenario with high γ. In order to compare the velocities of the evolutionary processes between scenarios with different γ's, we normalize for each γ the time by the average final time of the respective 200 simulations and term the resulting temporal measure the *normalized time*, as opposed to the non-normalized *absolute time*. Fig. 1A provides a comparison of the expansion processes on both time scales.

The expansion process with maximum sequential order, $\gamma = 100$, leads to the quickest exploration of the network also on the normalized time (Fig. 1A). The behavior in terms of the number of metabolites as a function of time looks qualitatively similar [44], and obeys particularly the same ranking in dependence on γ.

Which novel sequences may actually perform a useful function by catalyzing a biochemical reaction depends on the specific structure of the metabolic network at any given time during the evolutionary process. The number of these potential new enzyme sequences can be understood as a measure of the evolvability of the network [45]. To compare networks of identical sizes for scenarios with different γ, we introduce a third time measure, the *enzyme time*, defined by the current network size determined by the number of contained enzymes. In Fig. 1B the evolvability is shown in dependence of the enzyme time for different values of γ. For all values of γ, the temporal change of the evolvability can be divided into three phases. Until enzyme time 1500–2000, it increases rapidly before it reaches a plateau which is more pronounced for higher values of γ. In the final phase after enzyme time 5000 the limitation of the enzyme pool results in a rather constant decrease. The off-set on the y-axis for enzyme time 0 in Fig. 1B results from a peculiarity of reaction R00086, ATP + $H_2O \Longleftrightarrow$ ADP + P_i. This reaction can be catalyzed by enzymes of 66 EC numbers and is associated with 355 different sequences. Considering that ATP is treated as a cofactor, for which we do not explicitly require that it can be produced by the present network, this reaction can be added even to the initial seed network at enzyme time 0. Addition of this reaction does not expand the chemical functions of the network, but increases the variety of sequences from which potentially new sequences may evolve. Measuring the evolvability in numbers of new executable reactions results in qualitatively similar curves, with the main differences that the offset is not observed and that the curves exhibit a negative skewness instead the positive one, see [44].

For both measures, it is remarkable that the evolvability is systematically larger for scenarios with lower γ in which new sequences are added more randomly. This observation suggests that in this case consecutively added enzymes are rather unrelated in their chemical function, leading to a high metabolic diversification. In contrast, in the scenarios in which sequence information is important, the preferential discovery of enzymes similar to existing ones leads to an evolutionary exploration of local neighborhoods and thus to denser and more functional networks.

To support this hypothesis, we investigate the appearance of metabolites which only occur in a single reaction of the KEGG database and which can thus be seen as the border of the currently known metabolism. Overall, in the KEGG subnetwork that is reached by our simulated evolutionary processes, 29.5% of the metabolites (661 of 2237, see [44]) belong to this class. In Fig. 1C, the appearance of these border metabolites is depicted over enzyme time. Evidently, the influence of sequence information leads to a quicker exploration of the border. This supports the notion that, as a tendency, for large γ pathways are completed in a consecutive order, whereas for small γ pathways tend to be explored in parallel.

In Fig. 1D we depict how the actual minimal sequence distance for the selected enzymes changes with enzyme time. For large values of γ, in which a strong preference for sequences with a low minimal distance to existing enzymes prevails, the curve exhibits a characteristic U-shape. The initial drop is explained by the low number of enzymes within the evolving network and the restricted choice of new functional enzymes; the increase late in the process results from the fact that only those sequences remain unattached which have no noticeable sequence similarity to any other sequence. For smaller γ enzymes are picked at random and the pure increase in network size results in a lower average minimal distance.

Dynamic bursting in evolution

The invention of new classes of enzymes often goes along with a completely new sequence structure and may open a new branch in the evolutionary process. Such a novel enzyme can have deep impact on the evolutionary dynamics, since once a new reaction is found, similar reactions may evolve in close temporal neighborhood. Hence, a strong sequence dependency is expected to lead to a bursting like behavior of enzyme attachment. This would reflect the principle of punctuated equilibrium on a molecular level [23–25]. In the framework of punctuated equilibrium, new species do not appear gradually at equally spaced time points but in rapid successions followed by silent intervals. We exploit the capability of the current model to investigate the evolutionary dynamics to test if the sequence dependency of the network expansion may substantiate this hypothesis.

First, we determine the appearance times for a new

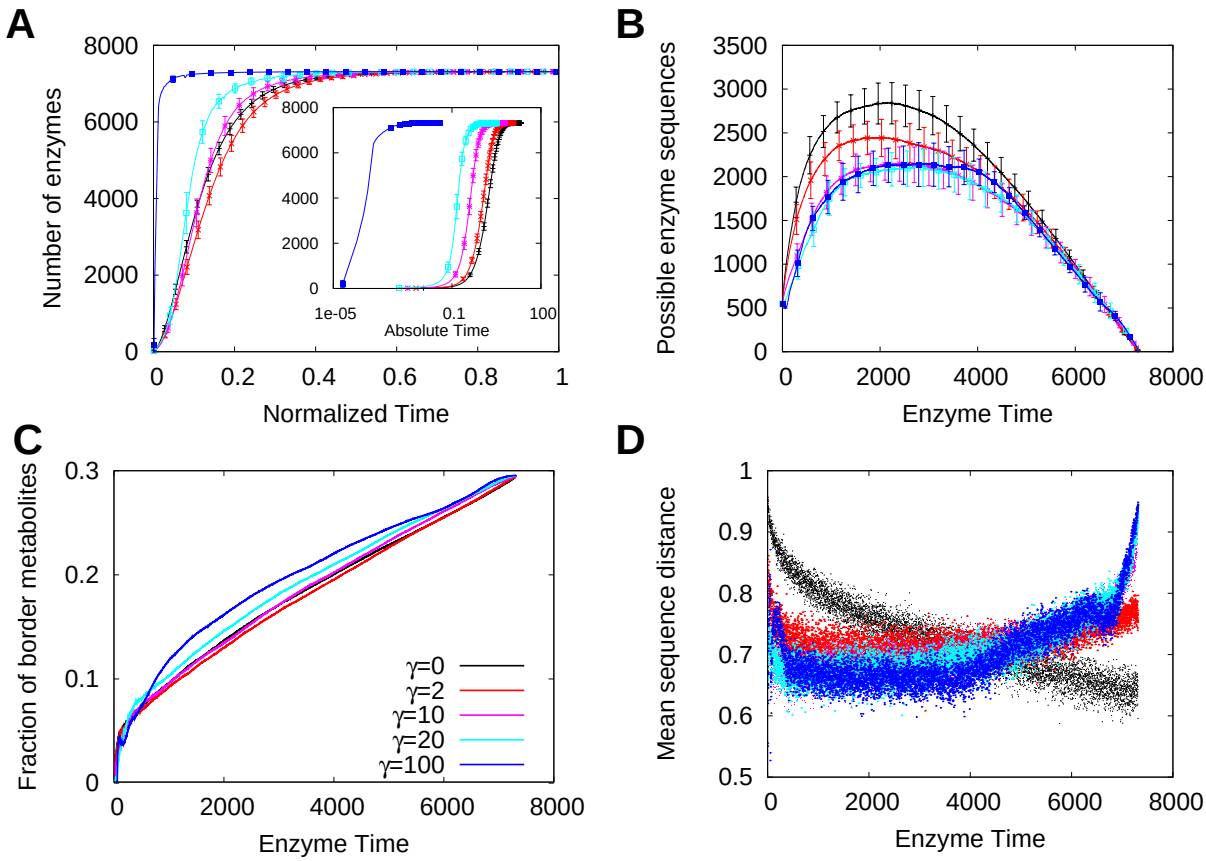

FIG. 1: Comparison of the expansion process for different strengths of the sequence-information parameter γ. Means of 200 simulations are shown and for all panels the color code given in panel C holds. **A**: The network size measured by the number of enzymes attached to the network is shown over time. The expansion velocity increases with γ in both normalized time and in absolute time (inlet) obtained from the Gillespie algorithm. **B**: The number of attachable enzymes at every step in the expansion process can be understood as the evolvability of the network. Using sequential information leads to a less evolvable but thus denser network. **C**: How quickly do we expand to the border of existing knowledge. At every step in enzyme time we plot the number of detected metabolites which only participate in one reaction in KEGG. Higher γ approach the border faster supporting the assumption of a smarter expansion. **D**: Mean sequence distances between every new enzyme and its duplication partner. The $\gamma = 0$ curve decreases since by chance for any new enzyme on average a similar sequence can be found if more enzymes are present in the current network. For higher γ isolated sequences without any similarity to all others are preferentially found at the end resulting in an increase.

enzyme in dependence on γ as shown in Fig. 2A. Here the invention of 500 enzymes (enzyme time 1500–2000) is plotted for 3 different values of γ by a vertical black line. Again it is obvious that the sequence dependence leads to an acceleration of evolution as can be seen by comparing the normalized time of each panel. A closer view reveals a more homogeneous structure for smaller γ. For $\gamma = 0$, the 500 events are rather homogeneously distributed over time with only few gaps. For $\gamma = 20$, the number and size of visible silent intervals increases because the 500 enzymes are invented more clustered. In case of very strong sequence dependence with $\gamma = 100$, the dynamics exhibit an even stronger clustering of events. The bursting dynamics leads to relatively large inter-cluster distances and subsequently to short intervals within an enzyme-class cluster because the number of enzymes is constant for all three scenarios, see [46].

These observations are subsumed in Fig. 2B where the logarithm of the frequency of inter-enzyme intervals (IEI) in normalized time are plotted. First of all, larger γ lead to shorter IEI in normalized time corresponding to faster evolutionary dynamics. The bursting like behavior leads to multiple peaks in the distribution for larger γ and a flat plateau for $\gamma = 100$ which has similarly been observed in earlier models [47, 48]. Summarizing, the clustered appearance of new sequences hints at a molecular principle of punctuated equilibrium. Interestingly, these distributions deviate from previous studies about self-organized criticality [49, 50]. The differences are probably caused by the different generating processes. While in the former investigations the number of possible events was unlimited, our model has a finite number of events since it is restricted to existing enzymes. This may be seen as a disadvantage of the model but at the same time it may

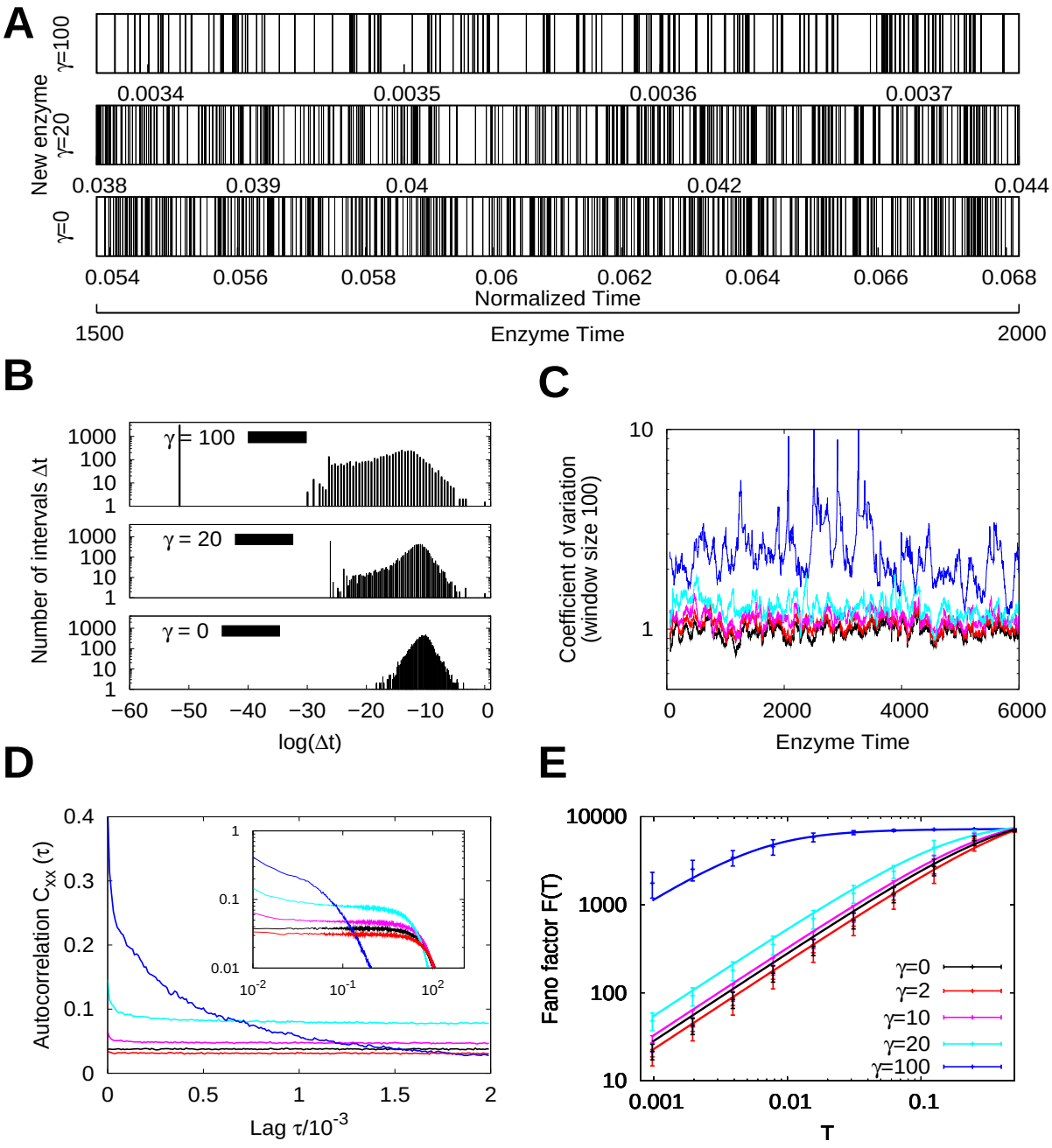

FIG. 2: The acquisition of new enzymes happens in bursts of increasing strength for larger sequence sensitivity. Here we show one example run in **A**–**C** and means of 200 runs in **D** and **E**. **A**: Spike train with one bar at every incident of a new enzyme. The panel shows a window of 500 new enzymes for each γ on its particular normalized time. While for $\gamma = 0$ the enzymes appear almost equidistantly, larger γ leads to enzyme bursts. **B**: Distribution of time intervals between any two new enzymes (IEI). For higher γ the distributions are shifted to smaller distances and exhibit multiple peaks. **C**: The coefficient of variation, $C_v = \sigma/\mu$, measured in sliding frames of 100 enzymes indicates multiple characteristic time scales. The peaks point to times of evolutionary explosions. **D**: The autocorrelation $C_{xx}(\tau)$ of IEIs supports the bursting behavior further. For large γ IEI are strongly correlated on a short time scale whereas small γ lead to no significant correlation. **E**: The fit of the data to the Fano factor of biased Brownian motion enables to estimate the correlation time τ_{corr}. (For all color panels the legend of panel E holds.)

γ	C_v	$D/10^5$	τ_{corr}	$D \cdot \tau_{\mathrm{corr}}/10^3$
0	1.14 ± 0.03	0.57 ± 0.05	0.20 ± 0.03	11.4 ± 2.0
2	1.2 ± 0.04	0.46 ± 0.04	0.29 ± 0.06	13.3 ± 3.0
10	1.4 ± 0.07	0.66 ± 0.06	0.16 ± 0.03	10.6 ± 2.2
20	1.9 ± 0.1	1.1 ± 0.08	0.081 ± 0.009	8.9 ± 1.2
100	4.8 ± 0.8	25.6 ± 2.3	0.0028 ± 0.0003	7.2 ± 1.0

TABLE I: Coefficients of variation and parameters of Fano factor fits averaged over 200 runs. The coefficient of variation is measured in the domain of the first 6000 enzymes. The data is fitted to the Fano factor Eq. (4) via parameters diffusion coefficient D and correlation time τ_{corr}.

reflect biological constraints such as a limited number of functional protein sequences.

For a further analysis of the evolutionary dynamics, we characterize the process in terms of the IEI by the Coefficient of variation C_v [51]. We use the resulting spike trains shown in Fig. 2A to determine the average IEI μ and the corresponding standard deviation σ and calculate the Coefficient of variation $C_v = \sigma/\mu$. For an unbiased evolution ($\gamma = 0$), we expect the characteristics of a Poisson process as a generating process since the time step determined by the Gillespie algorithm is independent of the history and purely random. A Poisson process leads to an exponential distribution of the waiting times [52] implying $C_v = 1$ [53].

Since we are interested in the temporal characteristics of the expansion, we use a sliding window of 100 enzymes to calculate C_v in dependence on the evolutionary steps. Indeed, the C_v for $\gamma = 0$ (black line) fluctuates around 1 as shown in Fig. 2C. Increasing the influence of sequence information by increasing γ, leads to systematically increased $C_v > 1$. This is a strong indicator for multiple characteristic time scales [51, 53]. These are given here on the one hand by the typical time to explore a new 'class' of enzymes, a slow process in which a novel sequence, unrelated to existing ones, evolves, and on the other hand by the characteristic time to invent an enzyme with a sequence similar to an already present one. For the shown window size of 100, the C_v for $\gamma = 100$ (blue) exhibits several peaks and reaches values up to 10 indicating strong bursting. The peaks may hint at important points of evolutionary explosion.

This analysis is further confirmed by the comparison of C_vs determined with different sliding window sizes (compare Fig. 2C and [54]). The comparison clearly demonstrates that the peaks of C_v are not an effect of the limited window size since even for larger window sizes the C_v reaches comparable values [54] and exhibits peaks. In Table I the systematic increase of the asymptotic C_v with increasing γ is given for all IEIs up to an Enzyme Time of 6000. This demonstrates the different characteristic time scales of the evolutionary process.

To substantiate this analysis we also calculated the autocorrelation function $C_{xx}(\tau)$ of the normalized IEIs for each γ as shown in Fig. 2D. For $\gamma = 100$ we observe strong correlations for small time lags τ indicating bursting. For unbiased evolution ($\gamma = 0$), no significant correlation on any time scale τ is observed which is in accordance with our assumed reason for bursting, the sequence information. In the inset of Fig. 2D, $C_{xx}(\tau)$ is plotted on a log-log scale. In this representation, the autocorrelation decreases linearly at the beginning as it is observed in other models of self-organized criticality [50]. But due to the limited enzyme pool size there is a strong cross over to the pure random behavior for large τ. Thus, our enzyme based model can quantitatively support the bursting dynamics of punctuated equilibrium.

While the Coefficient of variation allows for the analysis of dynamical variations on the scale of the average IEI μ, the Fano factor [55] characterizes variability in IEI on all accessible time scales T [56]. Therefore, the normalized time is divided in M non-overlapping windows and in each window the number of invention events N is determined. The Fano factor is defined as

$$F(T) = \frac{\langle N^2 \rangle - \langle N \rangle^2}{\langle N \rangle} , \tag{3}$$

where the time scale $T = T_{\mathrm{tot}}/M$ is given by the ratio between total time T_{tot} and the number of windows M.

For $\lim M \to \infty$, i.e. $T \to 0$, F equals 1. The dependence of $F(T)$ is shown in Fig. 2E and exhibits an increasing and saturating behavior. The increase is an indicator of long-range correlations. Because an increase is observed for all values of γ, these correlations are most likely a result of biochemical constraints given by the underlying metabolic network structure.

Since the analysis of the C_v has already shown the stochastic character of the expansion process, we hypothesize that the evolutionary process basically represents a diffusion process on the network. The evidence for long-range correlation suggests an Ornstein-Uhlenbeck process [52, 57] as approximative dynamics. For such a process the Fano factor can be expressed as [56]

$$F(T) = D \cdot \tau_{\mathrm{corr}} \left(1 - \frac{\tau_{\mathrm{corr}}}{T} \left[1 - \exp\left(-\frac{T}{\tau_{\mathrm{corr}}} \right) \right] \right) , \tag{4}$$

where D denotes a scaled diffusion coefficient and τ_{corr} is the correlation time. In order to characterize the dynamics on the network, we fit Eq. (4) to the Fano factor determined by Eq. (3) from simulations. We find a very good agreement for all γ values as shown in Fig. 2E. From the fitting procedure, we can estimate the diffusion coefficients D and correlation times τ_{corr} for each γ. The acceleration due to the sequence information leads to an increase of D accompanied by larger C_vs. The correlation time decreases in the units of relative time. This is caused by the faster expansion for larger γ. In this case, the invention of a novel sequence, representing a new class of enzymes, triggers the discovery of related sequences in short evolutionary time and thus the correlation time is shorter. For smaller γ, new classes are invented before all enzymes with a similar sequence structure are included

and thus correlation ranges over enzyme classes leading to larger $\tau_{\rm corr}$. From Eq. (4) we expect that the product $D\tau_{\rm corr}$ should stay rather constant what is verified in Table I.

Appearance: order of enzymes, compounds and organisms

From the observation of enzyme bursts and of the correlation time for large γ we expect that similar organisms appear at similar times, what would be a further confirmation of punctuated equilibrium in organismic evolution. Following our previous results indicating bursting evolutionary behavior in the time series of new enzymes, we will now conclude on the biological and biochemical outcomes of the appearance and order of metabolic compounds, enzymes, and even entire organisms.

Not all evolutionary paths are possible. Rather, the order of enzyme appearance is constrained by two factors. First, the selection criteria that only useful reactions are positively selected implies a chemical constraint. Some enzymes require other enzymes to be present since otherwise their required substrates could not be provided. The second constraint results from sequence similarity. It is conceivable that the sequence structure favors a certain order of enzyme evolution to avoid large jumps in sequence space.

Our model allows to distinguish the biochemical and evolutionary constraints which have shaped the metabolic map. To achieve this, we determine for $\gamma = 10$ and $\gamma = 0$ all pairs of enzymes which appear in the same temporal order in all 200 runs, excluding the seed enzymes which appear by definition before all others. Ordered pairs found for $\gamma = 0$ can only result from biochemical constraints since in this case sequence information is ignored. To identify those ordered pairs which result as a consequence of sequence similarities, we remove the pairs found for $\gamma = 0$ from the pairs determined for $\gamma = 10$. The remaining ordered pairs define a tree with 7117 nodes and 1348709 edges. Since visualization of such a large tree is impractical, we concentrate on all paths of length three or higher from root to leaf node (see Fig. 3A and [58] for a larger fraction of the tree).

Most of the enzymes on the first hierarchy level belong to essential pathways of central carbon metabolism. Enzymes in lower levels tend to belong to biosynthesis pathways of more specialized compounds. The tree gives insight into an enzyme's role in an evolutionary context. For example, enzymes 2.1.1.128 (a methyltransferase) or 1.2.1.38 (an oxidoreductase) appear only after a considerable number of precursors (5 and 32 respectively) have evolved. Apparently their sequences could have only evolved after many sequences for enzymes of central carbon metabolism had arisen. Interestingly, the opposite observation can be made for another methyltransferase, enzyme 2.1.1.116. The discovery of five enzymes directly dependent on the evolution of this particular sequence makes it plausible that this enzyme has presented an evolutionary bottleneck.

Highly important for the origin of life is the synthesis of amino acids as building blocks of proteins, and nucleotides for DNA and RNA. The amino-acids appear in good correlation (rank correlation 0.7) with previous results investigating the robustness of *E. coli's* network against reaction removal [33] (see Fig. 3B). This is not surprising and can be explained by stoichiometric effects. If more metabolic paths allow for the synthesis of a particular amino acid, it is likely to be discovered earlier. At the same time, one would expect that its production will be more robust against removal of reactions. The order of appearance of amino acids also reflects the commonly known biochemical synthesis pathways. Glutamate as a precursor of proline and arginine is synthesized first. In bacteria aspartate is the common precursor for lysine, threonine, and methionine. For all used γ values except $\gamma = 100$ this order is reproduced in the evolutionary scenarios. However, for $\gamma = 100$ threonine appears slightly before aspartate. The pyruvate family of leucine, isoleucine, and valine is detected in close proximity. Furthermore, the aromatic amino acids, phenylalanine, tryptophan, and tyrosine, labeled by asterisks appear rather late (position 16, 17, and 19 for $\gamma = 100$) as a result of their more complex chemical structure.

Additionally, we investigate the relation between the simplicity of synthesis and the actual usage of amino acids. For this, we compare the time of appearance to the frequency of the amino acids in the enzyme sequences of our consensus set and find a significant correlation for $\gamma = 100$ (Spearman 0.51, p-value 0.02, see [59]). Assuming that metabolites detected earlier in evolution are cheaper to synthesize, supports the hypothesis that cost minimization is an important factor for amino acid usage in protein synthesis.

Organisms own a specific metabolism depending on resources and living environment. Studying when the metabolic networks of various species have evolved allows speculations about the evolutionary tree of life. Clearly, the discovery of a complete set of an organism's reactions does not necessarily reflect its appearance during evolution. However, it presents a prerequisite for the respective metabolism to have evolved. Fig. 3C presents the discovery of the metabolic enzymes of 1097 organism specific networks retrieved from the KEGG database. The size of the respective networks are plotted versus the average enzyme time when 80% of an organism's enzymes were found ($\gamma = 10$). For higher organisms the enzyme time of appearance correlates well with the network size (Pearson correlation: animals=0.88, plants=0.95, fungi=0.75, protists=0.77, archaea=0.055, bacteria=0.25). Also, similar organisms tend to appear closely together, see [60]. For example, eight species of Drosophila occur from enzyme time 3571 to 3639, six Plasmodium species from enzyme time 3179 to 3317, or seven Mycoplasma from enzyme time 2956 to 3171.

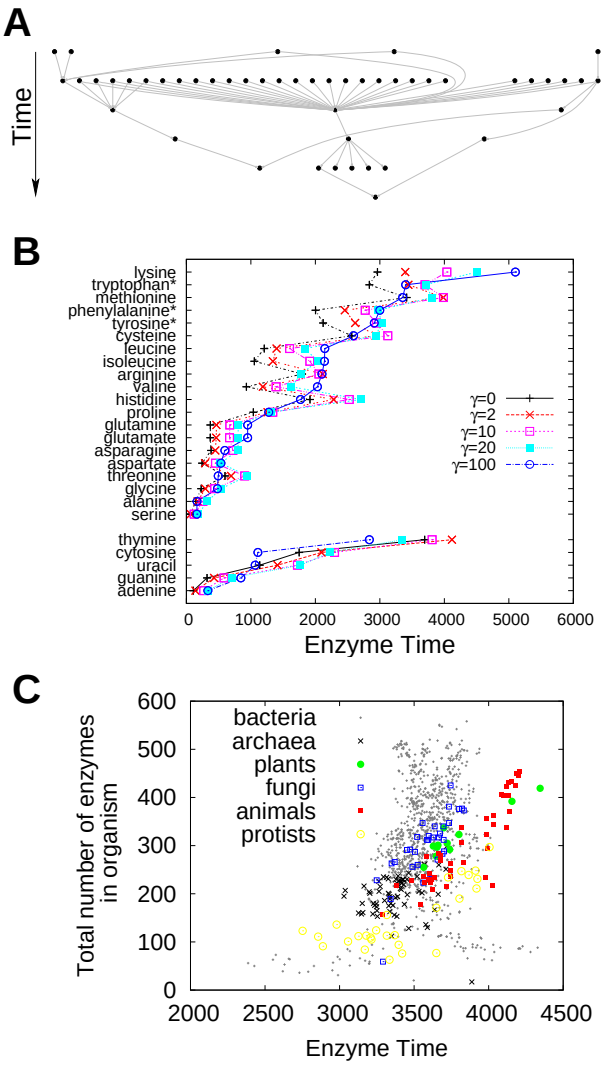

FIG. 3: Time order of appearance of enzymes, amino acids and nucleotides, and entire organisms. **A**: Time-ordered ranking of enzyme appearance for $\gamma = 10$. From the graph of all time-ordered pairs of enzymes with $\gamma = 10$ pairs also appearing in the $\gamma = 0$-case are removed and only the paths of length 3 of higher are shown (order-precision 100%). Time runs from top to bottom; the seed enzymes as root nodes are omitted for simplicity. Every enzyme is linked to its entry in KEGG. **B**: Appearance of amino acids (top part) and nucleotides (bottom part) sorted by the $\gamma = 100$ appearance and averaged over 200 runs. The order is very similar (rank correlation 0.7) to the order of robustness observed in the *E. coli* network [33]. Further, aromatic amino acids (labeled by *) are synthesized late. The γ-curves look similar indicating that the order strongly originates from stoichiometry rather than from sequence relations. **C**: Every enzyme defined by its EC number is mapped to its genes and thus to the corresponding organisms. An organism is assumed to have evolved if 80% of its annotated enzymes are discovered. The x-axis depicts the mean enzyme time of birth of a new organism while the y-axis shows the size of the organisms given by the enzyme repertoire. For higher organisms, the appearance time correlates well with the size of the organisms but this is not the case for bacteria and archaea. See [60] for a list of all organisms and the appearance time.

Conclusion

We developed a model that allows for an analysis of metabolic evolution by using a Systems Biology approach that combines experimental data, bioinformatic tools, modeling techniques, and time series analysis. Starting from an initial seed of prebiotic metabolites and simple enzyme sequences exhibiting a large amount of conserved proteome fragments, the metabolic network is extended by iterative invention of novel enzymes that catalyze putative new reactions and thus increase the number of present metabolites. Thereby, we focused on the role of sequence information as a source of evolutionary memory. The assumption that new enzymes with a similar sequence to already explored enzymes have a higher probability to appear was implemented by the use of the inverse BLAST-based enzyme distance, Eq. (1), determining the corresponding propensities of the Gillespie algorithm. For a quantitative analysis, the propensities were scaled by the power of the weighting factor γ, where $\gamma = 0$ corresponds to a pure random invention process and large γ to a highly sequence similarity dependent process, respectively. Previous models have explicitly investigated the effect of gene duplications and identified these events as important components of evolutionary innovation [29, 30]. Including gene duplications in our model is difficult because we mimic the evolution of the entirety of all enzymes, not those of a single species. We assume that the probability to evolve one gene from another is determined by sequence similarity. This dependence holds regardless whether one assumes gene duplications as an underlying mechanism or not. In this respect, different assumptions about the frequency of gene duplication events could be reflected in our model by applying different functional dependencies of the propensities on the sequence similarities.

The model generates temporal network dynamics, where the sequence-similarity driven expansion process leads to an acceleration of evolution. The implemented process of mutation and selection may be seen as a concrete realization of the network-based reconciliation [61]. In this framework, neutral evolution with mutations which do not lead to new phenotypes are combined with positive selection. From this conceptual model, we expect that evolutionary changes often occur in cycles of neutral diversity expansion and selective diversity contraction leading to a boom and bust behavior which was shown in simulations of RNA evolution [62] and experimentally by the analysis of the evolution of the human influenza virus antigen *hemagglutinin* [63, 64]. A phylogenetic analysis of *hemagglutinin* has revealed multiple short evolutionary branches corresponding to accumulation of neutral diversity.

This kind of behavior can be observed in our model for metabolic evolution as well. Here, neutral mutations occur whenever a new sequence is added that codes for an enzyme which is already present in the network. Sequence information leads to a bursting like behavior,

where enzymes of one class are invented within short intervals whereas discovery of a new enzyme class needs more and larger mutations and thus occurs only rarely, confirming a boom and bust behavior. Therefore our model gives a first molecular description of punctuated equilibrium in metabolic evolution. This is quantified by the coefficient of variation C_v which increases with increasing sequence information dependency of the expansion process. Using a sliding window for the calculation of C_v indicates events of evolutionary explosion. A high autocorrelation function of the IEIs for small time lags in the case of large γ provides further evidence for the bursting behavior. Moreover, the good agreement of the Fano factor with the analytical result of biased Brownian motion points to the diffusive character of network evolution and allows for an estimation of typical correlation times of the evolutionary process. High sequence dependence leads to shorter correlation times, since once a functional sequence is found, all directly related enzymes are invented as well.

From our simulations, we could extract typical time orders of enzyme appearances which start with carbon metabolism that is needed for all subsequent processes. Although the model is rather elementary and neglects putative transient enzyme sequences or metabolites, the obtained order of amino acid appearance fits nicely with their robustness. This illustrates that many evolutionary paths lead to the development of simple but essential building blocks, whereas complex structures which depend on the previous discovery of simpler ones occur later. Interestingly, mapping enzymes to organisms by the EC number leads to results that match the intuition, despite the simplifying assumptions made. Bacteria appear as first species and plants and animals rather late in the artificial evolution. Moreover, occurrence time and complexity are strongly correlated for animals and plants.

Since the aim of our model is not to explain the early origins of metabolism, we assume that catalysts have already evolved and neglect mechanisms for the production of the enzymes themselves. As a consequence, we do not expect the model to produce realistic evolutionary paths for these early stages. Only after a core metabolism was assembled and the protein synthesis machinery has evolved, a closely intertwined coevolution of metabolites and enzymes can be seen as plausible. Moreover, the agreement of our mathematical model with phenomenological observations transforms the qualitative description of evolution into a quantitative level. This might be used in future work for deeper insights into enzymatic functions and their role in interactions of organisms.

Acknowledgments

We acknowledge financial support, in particular for a five-months research visit at Boston University, from the International Research Training Group *Genomics and Systems Biology of Molecular Networks* IRTG 1360 (MS), the German Federal Ministry of Education and Research, Systems Biology Research Initiative *Go*FORSYS (AS), the Scottish Universities Life Science Alliance SULSA (OE), the NASA Astrobiology Institute, and the US Department of Energy (DS). We would like to thank Nils Christian, Zoran Nikoloski, and Thomas Handorf for useful discussions.

[1] S. L. Miller, *Science* **117**, 528 (1953).
[2] W. Martin, M. J. Russell, *Philos Trans R Soc Lond B Biol Sci* **362**, 1887 (2007).
[3] S. D. Copley, E. Smith, H. J. Morowitz, *Bioorg Chem* **35**, 430 (2007).
[4] P. A. Bachmann, P. L. Luisi, J. Lang, *Nature* **357**, 57 (1992).
[5] P. Walde, R. Wick, M. Fresta, A. Mangone, P. L. Luisi, *Journal of the American Chemical Society* **116**, 11649 (1994).
[6] D. Segrè, D. Ben-Eli, D. W. Deamer, D. Lancet, *Orig Life Evol Biosph* **31**, 119 (2001).
[7] S. A. Kauffman, *The origins of order: self-organization and selection in evolution* (Oxford University Press, 1993).
[8] H. J. Morowitz, *Beginnings of Cellular Life* (Yale University Press, 2004).
[9] F. Dyson, *Origins of Life* (Cambridge University Press, 1999).
[10] S. Granick, *Ann N Y Acad Sci* **69**, 292 (1957).
[11] N. H. Horowitz, *Proc Natl Acad Sci U S A* **31**, 153 (1945).
[12] M. Ycas, *J Theor Biol* **44**, 145 (1974).
[13] R. A. Jensen, *Annu Rev Microbiol* **30**, 409 (1976).
[14] O. Ebenhöh, T. Handorf, R. Heinrich, *Genome Inform* **15**, 35 (2004).
[15] T. Handorf, O. Ebenhöh, R. Heinrich, *J Mol Evol* **61**, 498 (2005).
[16] O. Ebenhöh, T. Handorf, R. Heinrich, *Genome Inform* **16**, 203 (2005).
[17] O. Ebenhöh, T. Handorf, D. Kahn, *Syst Biol (Stevenage)* **153**, 354 (2006).
[18] F. Matthäus, C. Salazar, O. Ebenhöh, *PLoS Comput Biol* **4**, e1000049 (2008).
[19] J. Raymond, D. Segrè, *Science* **311**, 1764 (2006).
[20] S. Maslov, S. Krishna, T. Y. Pang, K. Sneppen, *Proc Natl Acad Sci U S A* **106**, 9743 (2009).
[21] A. Mithani, G. M. Preston, J. Hein, *Bioinformatics* **25**, 1528 (2009).
[22] M. Schütte, N. Klitgord, D. Segrè, O. Ebenhöh, *Genome Inform* **22**, 156 (2010).
[23] P. Bak, K. Sneppen, *Phys Rev Lett* **71**, 4083 (1993).
[24] N. Eldredge, J. G. Gould, *Models in Paleobiology* (T.Schopf, 1972), chap. Punctuated equilibria: an alternative to phyletic gradualism, pp. 82–115.
[25] S. F. Elena, V. S. Cooper, R. E. Lenski, *Science* **272**, 1802 (1996).

[26] M. Kanehisa, S. Goto, *Nucleic Acids Res* **28**, 27 (2000).
[27] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa, *Nucleic Acids Res* **38**, D355 (2010).
[28] H. Kacser, R. Beeby, *J Mol Evol* **20**, 38 (1984).
[29] T. Pfeiffer, O. S. Soyer, S. Bonhoeffer, *PLoS Biol* **3**, e228 (2005).
[30] A. Hintze, C. Adami, *PLoS Comput Biol* **4**, e23 (2008).
[31] D. Gillepsie, *J. Phys. Chem.* **8**, 2340 (1977).
[32] T. Handorf, N. Christian, O. Ebenhöh, D. Kahn, *J Theor Biol* **252**, 530 (2008).
[33] N. Christian, P. May, S. Kempa, T. Handorf, O. Ebenhöh, *Mol Biosyst* **5**, 1889 (2009).
[34] R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (1997).
[35] R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, *Nucleic Acids Res* **28**, 33 (2000).
[36] R. L. Tatusov, *et al.*, *Nucleic Acids Res* **29**, 22 (2001).
[37] R. L. Tatusov, *et al.*, *BMC Bioinformatics* **4**, 41 (2003).
[38] E. Wilkinson, J. Willemsen, *J Phys A* **16**, 3365 (1983).
[39] T. Handorf, O. Ebenhöh, *Nucleic Acids Res* **35**, W613 (2007).
[40] Q. W. Chen, C. L. Chen, *Current Organic Chemistry* **9**, 989 (2005).
[41] Y. Sobolevsky, E. N. Trifonov, *J Mol Evol* **61**, 591 (2005).
[42] Y. Sobolevsky, E. N. Trifonov, *J Mol Evol* **63**, 622 (2006).
[43] Y. Sobolevsky, Z. M. Frenkel, E. N. Trifonov, *J Mol Evol* **65**, 640 (2007).
[44] Supplementary, *See supplementary material at [URL will be inserted by AIP] for additional figures illustrating the expansion process* .
[45] A. Wagner, *Proc Biol Sci* **275**, 91 (2008).
[46] Supplementary, *See supplementary material at [URL will be inserted by AIP] for heatmaps that show the appearance of enzyme classes* .
[47] H. F. Chau, *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **49**, 4691 (1994).
[48] S. Paczuski, M.and S. Maslov, P. Bak, *Phys Rev E* **53**, 414 (1995).
[49] C. Adami, *Phys Let A* **203**, 29 (1995).
[50] M. Usher, M. Stemmler, Z. Olami, *Phys. Rev. Lett.* **74**, 326 (1995).
[51] D. Cox, L. Lewis, *The Statistical Analysis of Series and Events* (Wiley, New York, 1966).
[52] C. Gardiner, *Handbook of stochastic methods* (Springer, Berlin, 1985).
[53] D. Cox, V. Isham, *Point Processes* (Chapman and Hall, London, 1980).
[54] Supplementary, *See supplementary material at [URL will be inserted by AIP] for C_v using sliding windows of different sizes* .
[55] Fano, *Phys. Rev.* **72**, 26 (1947).
[56] J. W. Middleton, M. J. Chacron, B. Lindner, A. Longtin, *Phys Rev E Stat Nonlin Soft Matter Phys* **68**, 021920 (2003).
[57] N. van Kampen, *Stochastic processes in physics and chemistry* (North-Holland, Amsterdam, 2001).
[58] Supplementary, *See supplementary material at [URL will be inserted by AIP] for a larger fraction of the enzyme-ranking network* .
[59] Supplementary, *See supplementary material at [URL will be inserted by AIP] for a table of the amino acid abundance in the consensus set* .
[60] Supplementary, *See supplementary material at [URL will be inserted by AIP] for a table of all organisms listed with names as they appear in enzyme time* .
[61] A. Wagner, *Nat Rev Genet* **9**, 965 (2008).
[62] W. Fontana, P. Schuster, *Science* **280**, 1451 (1998).
[63] K. Koelle, S. Cobey, B. Grenfell, M. Pascual, *Science* **314**, 1898 (2006).
[64] D. J. Smith, *et al.*, *Science* **305**, 371 (2004).