

Evolution of metabolic networks

Eleni Giannoulatou

September 15, 2006

Abstract

Metabolism is one of the most complex cellular processes and a basal system for maintaining life of all organisms. Most of the components of a metabolic network can be identified and analysed due to large-scale sequencing projects and genome annotation efforts. However, understanding the organisational principles of the metabolism of living organisms is a major challenge in network biology. The process of evolution of these networks can be used in order to justify their principal design. Although models of network growth have been introduced, they do not apply directly to metabolic networks where the dynamics are determined by the gain or loss of enzymatic reactions. In this work, we introduce a simple stochastic model of the evolutionary process of metabolic networks. By calculating the likelihood of the phylogeny of metabolisms, we are able to infer important parameters, such as the rates of creation and deletion of interactions between the metabolites. Finally, we present a simple model with fitness-dependent link dynamics in order to provide a qualitative understanding of the evolutionary processes of metabolic networks.

Contents

1	Introduction	2
1.1	Properties of Metabolic Networks	2
1.2	Evolution of Metabolic Networks	4
1.3	Mathematical Models of Network Evolution	6
1.4	Phylogenies of Metabolic Networks	7
2	Methods	8
2.1	Metabolic Network representation	8
2.2	A stochastic model of evolution	8
2.2.1	Metabolism Enumeration	11
2.2.2	Definitions of Markov process theory	11
2.2.3	Simulation	12
2.2.4	Stationary Distribution	13
2.2.5	Time reversibility	13
2.2.6	Likelihood calculation	13
2.3	Model with Fitness-dependent link dynamics	15
3	Results	15
3.1	Stochastic model of evolution - Simulation results	16
3.2	Enumeration of all possible states	17
3.2.1	Rate matrix and Stationary distribution	18
3.3	Maximum Likelihood Estimates of a small phylogeny of networks	19
3.3.1	Estimation using parsimony	21
3.4	Model with Fitness-dependent link dynamics — Simulation results	22
4	Discussion and Further Work	24
A	Estimation of parameters	26

1 Introduction

Rapid advances in network biology aim to understand the structure and dynamics of the complex intracellular reactions that contribute to the function of a living cell (Barabasi and Oltvai, 2004). Comparative genomics aim to elucidate the design principles and the underlying evolutionary process that creates such a cellular organisation.

Metabolic networks have gained much attention since metabolism is one of the most complex cellular processes. It is comprised of a vast repertoire of enzymatic reactions and transport processes used to convert thousands of organic compounds into the various molecules necessary to support cellular life (Schilling et al., 2000). The collection of reactions and hence pathways that a metabolic network possesses determines the architecture and topology of the network.

This work presents a simplified stochastic model for the evolution of metabolic networks, in order to make estimates for the phylogeny of simple metabolisms based on likelihood approaches. It also introduces a simple model with fitness-dependent link dynamics in order to provide a qualitative understanding of the evolutionary processes of metabolic networks. This section describes the properties of metabolic networks based mainly on their structure and presents current approaches for modelling the evolution of networks as well as for their phylogenetic analysis. Section 2 describes the methods that were used, specifically the probabilistic model of evolution and the statistical methods for inferring phylogenetic relationships and parameters between simple metabolisms. Section 3 presents the simulation results as well as the likelihood estimates of “homologous” metabolic networks. Finally, section 4 concludes by giving an evaluation of the strengths and weaknesses of the methods used and suggesting some future work.

1.1 Properties of Metabolic Networks

Metabolic networks incorporate diverse data sets including genome annotations, biochemical characterisations and cell physiology experiments (Papin et al., 2003). Initially, high-throughput experimental technologies led to the precise characterisation of individual reactions (Figure 1a). Therefore, the stoichiometries of these reactions have been clearly defined. With the cataloguing of multiple reactions, shared metabolites have been grouped together and traditional pathways have been described (Figure 1b). Today, most of the components of a metabolic network can be identified and analysed due to large-scale sequencing projects and genome annotation efforts (Figure 1c). Once enzyme genes are identified in a genome based on sequence similarity and positional correlation of genes, organism-specific pathways can be constructed computationally by correlating genes in the genome with gene products (enzymes) in the reference pathways (Kanehisa and Goto, 2000). With this level of description, mathematically defined pathways enable the analysis of a complete metabolic system.

The organism specific metabolic networks are often very large and complex. Mathematical descriptions of cellular properties, such as cellular growth and reaction flux levels with given substrate input, have been used to describe the networks. These methods take into account properties of the enzyme chemistry by accounting for mass balance and thermodynamic constraints.

In the fields of bioinformatics, systems biology and statistical physics, efforts are also being made in order to analyse and understand the structure of these large-scale networks. These employ graph theory methods and address mainly the topological properties of these networks. Jeong et al. (2000) presented a systematic comparative mathematical analysis of the metabolic networks of 43 organisms representing all three domains of life. They used a graph theoretic representation of the biochemical reactions, where the substrates are the nodes that are connected to one another through links that are the actual reactions. It was shown that these networks have the same topological scaling properties among species; they have the feature of a scale-free network and a small-world character. Scale-free networks are characterised by a power-law degree distribution; the probability that a node has k links follows $P(k) \sim k^{-\gamma}$, where γ is the

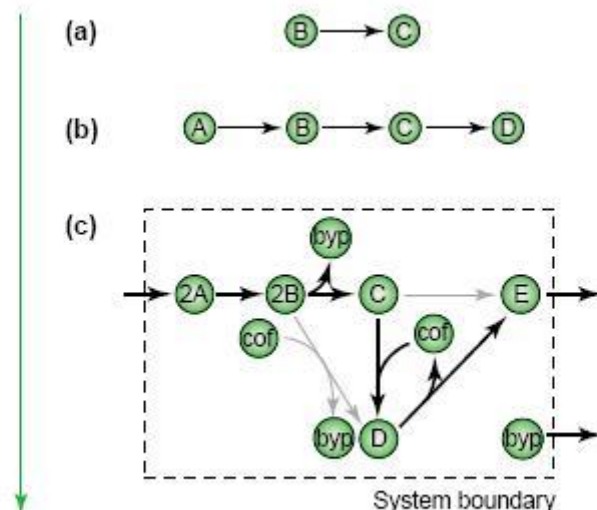


Figure 1: Development of the network-based pathway paradigm. (a) Characterisation of individual reactions. (b) Reactions grouped into traditional pathways. (c) Mathematically defined network-based pathways that account for cofactors and by-products. Taken from (Papin et al., 2003).

degree component (Barabasi and Oltvai, 2004). The probability that a node is highly connected is statistically more significant than in a random graph. In the case of metabolic networks, most metabolic substrates participate in only one or two reactions, but a few, such as pyruvate or coenzyme A, participate in dozens and function as metabolic hubs. It was also shown that any two metabolites in the system can be connected by relatively short paths along existing links. This suggested that metabolism in every organism behaves like a small-world network, which has also been shown by analysis of networks of specific organisms, such as *E. coli* (Wagner, 2001). Finally, it was found that the measured diameters of the metabolic networks are almost the same for different organisms, regardless of the number of metabolites found in the given species.

Ma and Zeng (2003a) repeated the analysis of Jeong et al. (2000) after removing small molecules and cofactors — in order to make the path length analysis physiologically more meaningful — and accounting for irreversible reactions. Their results suggested that the degree of connectivity follows a scale-free distribution, but differences exist in the network structure of the three domains of organisms. Eukaryotes and archaea have a longer average path length than bacteria. The results of both of these network studies provide useful insights into the properties and evolution of metabolism. They also demonstrate the importance of considering the specific properties of the enzyme chemistry and the need for the consistent definition of “what is a pathway” when cell-wide metabolic networks are analysed (Hatzimanikatis et al., 2004).

Other studies have suggested additional characteristics for the structure of metabolic networks. These include the use of functional modules in the networks, and the classification of nodes into universal roles according to their pattern of intra- and inter-module connections (Guimera and Nunes Amaral, 2005). Specifically, it was found that nodes with different roles are affected by different evolutionary constraints and pressures. Remarkably, metabolites that participate in only a few reactions but that connect different modules are more conserved than hubs whose links are mostly within a single module. Another study has showed that the metabolic networks of various organisms are not fully connected (Ma and Zeng, 2003b). A “bow-tie” similar connectivity structure that was previously found for the web network connection structure

also exists in metabolic networks. A giant strong component, represents the most complex part and the core of a metabolic network and it exhibits characteristics of a small world network.

The characteristics that have been suggested for the structure of metabolic networks are under debate. An example is the structure of the metabolism of *E. Coli*, which has been recently suggested not to follow the properties of a small-world network (Arita, 2004). Biological networks can have differences in their functional states and the topological properties suggested are derived from network structure, but not from their functional, or phenotypic, states (Mahadevan and Palsson, 2005). For example, for the metabolic networks it is observed that, functionally speaking, the essentiality of reactions in a node is not correlated with node connectivity as structural analyses can suggest. Therefore, we would expect to have differences in the findings of studies that use different representations of the networks.

1.2 Evolution of Metabolic Networks

A variety of theories have been developed to explain the evolution of metabolic networks from early life. First, pathways might have evolved spontaneously without adopting existing enzymes (Schmidt et al., 2003) (Figure 2a). For example, different tRNA synthetases seem to have evolved independently and later catalyse steps in different pathways such as in protein translation. Second, the retrograde model of evolution, proposed by Horowitz in 1945, suggests that pathways evolve “backwards” from a key metabolite (Rison and Thornton, 2002) (Figure 2b). The model assumes a chemical environment in which both key metabolites and potential intermediates are available. When a key metabolite A is exhausted, an organism capable of synthesising molecule A from an environmental precursor B will have a selective advantage. Therefore, any mutant evolving an enzyme that catalyses this synthesis will survive and spread through the environment quickly. This model of evolution has been proposed for the glycolytic and the mandelate pathway. Third, multifunctional enzymes might be responsible for pathway evolution. A multifunctional enzyme that catalyses consecutive steps in a pathway can get duplicated and diversified to more specific enzymes that catalyse only one step each in a pathway (Figure 2c). An example of an existing multifunctional enzyme is carbamoyl phosphate synthase which is used in diverse pathways, that can be precursors to new ones. Moreover, whole pathways can get duplicated and diverted instead of simple enzymes (Figure 2d). Tryptophan and histidine biosynthesis are two pathways with similar reaction chemistry that are catalysed by homologous enzymes; these might be the results of pathway duplication. Finally, a “patchwork” model of evolution suggests that metabolic pathways evolve from the serial recruitment of enzymes that are relatively inefficient in their existing pathway and have broad specificity in order to react with a wide range of chemically related substrates (Lazcano, 1999). It has been found, specifically in *E. coli* metabolism, that one type of enzyme fold or superfamily that catalyse similar reactions can occur in different pathways due to widespread recruitment.

A number of different theories have been proposed depending on the specific pathways under study. It is not necessary for all the metabolic pathways to arise in the same manner. Thus, it has been suggested that some of the earliest pathways may have arisen from the retrograde model, while recent studies suggest that enzyme recruitment seems to be the main driving force for the evolution of new pathways, sometimes followed by specific pathway duplications and then by other more rarely observed mechanisms.

Pfeiffer et al. (2005) indicated that metabolic networks with a large number of highly specialised enzymes may evolve from a few multifunctional enzymes. This was shown by simulating a model that assumes the patchwork model of evolution for novel pathways and that a core metabolic network with specialised enzymes is already present. Specifically, the key evolutionary mechanism in this scenario is the duplication of enzymes followed by specialisation of the copies for different biochemical reactions. This also suggests that biochemical reactions and intermediate metabolites are lost during the evolution of the metabolic networks. Another important suggestion of this study is related with the presence of highly connected metabolites in the networks, i.e. the metabolic hubs. These can emerge as a consequence of selection for

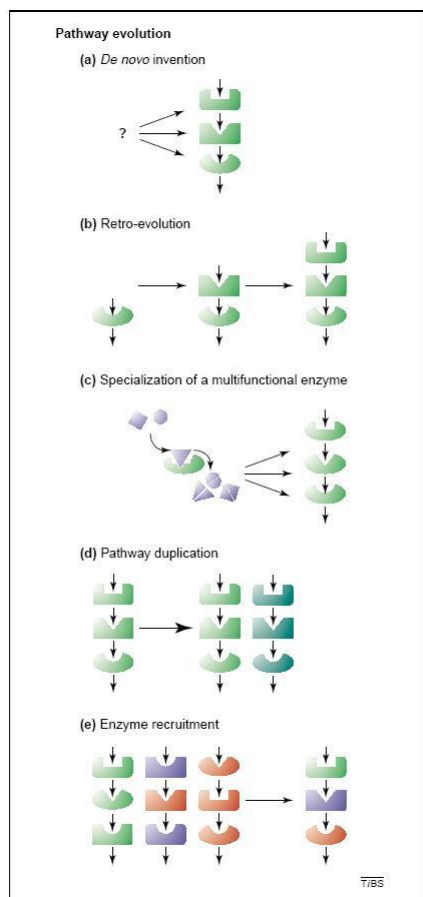


Figure 2: Theories for the evolution of metabolic pathways. (a) De novo, all reactions evolved independently, (b) Retrograde model, (c) Evolution by specialisation of a multifunctional enzyme, (d) Duplication of a complete pathway and (e) Patchwork model. Taken from (Schmidt et al., 2003).

growth rate, since they are the result of group transfer reactions. This supports the view that hubs are not necessarily the result of selection for robustness as it has been suggested previously (Jeong et al., 2000).

A recent study on the causes of evolution of yeast metabolism suggested that the apparent dispensability of many enzymes is not due to network robustness, but to the fact that many enzymes are only required under specific environmental conditions (Papp and Pál, 2004). The authors found that many of these dispensable genes are compensated for by isozymes, and a smaller fraction is compensated for by alternative metabolic pathways. It was also showed that a better explanation for the possession of multiple isozymes is the need for high flux rates through specific reactions, rather than the provision of redundancy for essential genes. This again supports the selection for growth rate.

In the case of bacteria it has been shown that most changes to the metabolic networks are due to horizontal gene transfer, with little contribution from gene duplication (Pal et al., 2005). Specifically, the networks evolve in response to changing environments, not only by changes in enzyme kinetics through point mutations, but also by the uptake of peripheral genes and operons through horizontal gene transfers. Additionally, a more recent study on the evolution of metabolic networks of endosymbiotic bacteria suggests that differences between minimal networks based on lifestyle are predictable and therefore the gene content of an organism can be predicted by using knowledge of its distant ancestors and its current lifestyle (Pal et al., 2006).

Finally, we should mention that most of these findings are based on the analysis of one or few organisms, and there are limitations related to the pathway definitions. By using a

network approach where many pathways are branched into each other and considering many fully sequenced genomes, it was found that functional blocks of similar chemistry have evolved within metabolic networks (Alves et al., 2002). Specifically, homologous pairs of enzymes (which therefore catalyse similar type of reactions) are roughly twice as likely to have evolved from enzymes that are less than three steps away from each other in the reaction network than pairs of non-homologous enzymes.

1.3 Mathematical Models of Network Evolution

Since networks are complex and large, it is difficult to analyse them, find properties of their structure and their evolution. Typically, the analysis of networks is restricted to the use of summary statistics such as degree distribution and clustering coefficient. However, since there are many examples of real networks in which the structural changes are ruled by the dynamical evolution of the system, a class of models has been developed whose primary goal is to reproduce the growth process taking place in real networks (Boccaletti et al., 2006).

The *Barabasi-Albert*(BA) model is a model of network growth based on two basic ingredients: growth and preferential attachment. It is inspired by the formation of the World Wide Web where popular vertices with high number of links are more attractive for new connections than vertices with few links – *popularity is attractive* (Dorogovtsev and Mendes, 2002). Therefore, this model explains the structure of scale-free networks, where hubs are present. In the case of biological networks its main application is on protein interaction networks rather than metabolic ones: the growing procedure can be understood as the appearance of duplicated genes. The new protein obtained by such duplication inherits the connections of the old one, while some of these connections are later eliminated. The more connections a node acquires, the stronger is the selective pressure to make it more connectable. This can be understood as a tendency to increase the number of protein attachment domains, such proline-rich, or to improve the existing domains such that they bind to more target proteins (Eisenberg and Levanon, 2003). Hence, due to selection the proteins that are highly connected tend to keep the new connections. Therefore, this process can be modelled as preferential attachment to the existing proteins that are more connected. Starting with m_0 isolated nodes, at each time step $t = 1, 2, 3, \dots, N - m_0$ a new node j with $m \leq m_0$ links is added to the network. The probability that a link will connect j to an existing node i is linearly proportional to the degree of i (Boccaletti et al., 2006):

$$P_{j \rightarrow i} = \frac{k_i}{\sum_l k_l}$$

Because every new node has m links, the network at time will have $N = m_0 + t$ nodes and $K = mt$ links, corresponding to an average degree $\langle k \rangle = 2m$. The BA model has been solved by means of rate equation and master equation approaches giving a degree distribution $P(k) \sim k^{-\gamma}$, (with an exponent $\gamma = 3$) in the limit $t \rightarrow \infty$.

Similar to the BA model, the *Dorogovtsev-Mendes-Samukhin*(DMS) model considers a linear preferential attachment of the form:

$$P_{j \rightarrow i} = \frac{k_i + k_0}{\sum_l (k_l + k_0)}$$

with $-m < k_0 < \infty$. It is more general than the BA model, as the presence of k_0 plays the role of initial node attractiveness. The power law degree distribution has exponent $\gamma = 3 + k_0/m$.

The *Albert and Barabasi*(AB) model suggests a generalisation of the initial model where connections can be lost and added. At each time step, with probability p , one selects an existing node and adds m new links from it to m nodes chosen preferentially. Then, with probability q , one rewires m arbitrarily selected edges. Finally, with probability $1 - p - q$, one adds a new node to the network and connects it with m existing nodes preferentially. The model shows a regime in which $P(k)$ is a power law with an exponent whose value is determined by p and q (Boccaletti et al., 2006). Many other models, similar to the above, have been proposed.

For biological networks, such as protein networks, a model of network of evolution has been proposed which is based on the observed rates of link and duplication dynamics (Berg et al., 2004). According to this model, the link dynamics is the dominant evolutionary force shaping the statistical structure of the network, while the slower gene duplication dynamics mainly affects its size. At the observed rates, the model predicts the important structural features of the networks are shaped solely by the link dynamics. Finally, in the case of social networks, the changes can be modelled as stochastic continuous time Markov processes with specific rates of deletion and insertion of edges (Snijders and Van Duijn, 1997).

The research focused on modelling the evolution of metabolic networks is limited. Typically, models used for many other real networks are applied in the case of any biological networks. However, according to a recent study, a “rich-travel-more” mechanism can generate the observed scale-free organisation of metabolic networks, rather than the “rich-get-richer” mechanism proposed by the growth with preferential attachment models (Ueda and Hogenesch, 2005). This analysis showed that the evolutionary process of metabolic networks follows the same and surprisingly simple principles in Archaea, Bacteria and Eukaryotes: highly linked metabolites change their chemical links more dynamically than less linked metabolites.

1.4 Phylogenies of Metabolic Networks

Phylogenetic trees represent the evolutionary histories of organisms. However, phylogenetic trees based on single genes represent the history of that gene, not necessarily that of the whole organism. Specifically, they do not provide a complete and accurate picture of evolution as they do not account for evolutionary leaps due to gene transfer, duplication, deletion and functional replacement. Thus, a tree based on metabolic pathways can represent both the evolutionary time scale (changes in genetic content) and the evolutionary process (changes in metabolism) (Hong et al., 2004). However, one of the drawbacks of classical phenotypic analysis is lack of a method of quantification. Since the evolution of the metabolic pathways has been under debate, a phylogenetic analysis that also incorporates the sequence information will expand the understanding of the evolutionary processes molding their form and structure.

Recently, studies have tried to infer phylogenies of metabolic networks without modelling the actual evolutionary process. First, a method based on a combination of sequence information with graph topology of the underlying pathway was applied to pathways related to electron transfer and to the Krebs citric acid cycle. (Forst and Schulten, 1999). A distance between two pathways was calculated using the distance between sequences of the same functional role in the pathway. Typical functional roles are enzymes that process substrates in a specific reaction or substrates that are processed by specific enzymes. A functional role also describes how a gene product functions in a protein complex. The analysis revealed a close relationship between pathways of organisms within the same genus.

Instead of comparing the metabolic networks according to the organisms’ enzyme contents, the structures of the networks can be used (Forst et al., 2006). Specifically, a framework borrowed from set algebra can provide natural definitions of unions, intersections, and differences that can be used to compare the metabolic networks of different organisms. Again a distance measure is used on the set of reaction networks to infer their phylogenetic relationship.

Graph matching algorithms have also been used for alignment of metabolic pathways. Pinter et al. (2005) presented a pathway alignment tool that, given a query pathway and a collection of pathways, finds and reports all approximate occurrences of the query in the collection, ranked by similarity and statistical significance. Other simpler methods calculate the Hamming distances between networks, i.e. the number of discrepancies scored in the corresponding elements of the two networks (Tun et al., 2006). This is used as a dissimilarity measure for creating a phylogenetic tree.

The above methods use distance-based measures to infer the phylogenetic relationships of the metabolic networks. Using probabilistic models of evolution of the networks, standard statistical methods, such as maximum likelihood methods, can be used to make estimates of the phylogeny.

The parameters to be inferred are the phylogenetic tree, deletion or insertion rates or ancestor networks. So far, only a recent study of (Wiuf et al., 2006) allows to calculate the likelihood of a full network under a given mathematical model. In this work, the likelihood of a graph is calculated conditional on a initial graph isomorphism. The basic evolutionary model assumed is a typical duplication with random attachment model, applied in protein interaction networks. The networks are so complex that the computations can be too demanding to be practical for even moderate networks. A recursion, combined with Importance Sampling technique was used for the calculation of the likelihood.

2 Methods

This section describes the methods and the theoretical considerations used for introducing the mathematical models of evolution as well as the statistical methods used to make estimates for the phylogeny of simple metabolisms.

2.1 Metabolic Network representation

A metabolic network can be mapped in a graph that contains the metabolites and their reactions, catalysed by specific enzymes. Two complementary ways of representation have been commonly used for mapping a metabolic network in a graph using either enzymes or metabolites as vertices (Zhang et al., 2006; Tun et al., 2006). In this work, nodes correspond to metabolites and edges to possible reactions between them. Specifically, an edge between two metabolites corresponds to the presence of one or more enzymes catalysing a specific reaction that transforms one of the metabolites into the other.

The graph topology of a metabolic network is represented as an adjacency matrix N , an $n \times n$ matrix for n metabolites. Each element of the matrix takes values 0 or 1 depending on the presence or absence of a reaction between two metabolites. Since the graph is simple and undirected the adjacency matrix is symmetric with zero entries in the diagonal. This representation framework was very useful for the network analysis, allowing for straightforward manipulation of the network by matrix algebra. For example, an interesting property is that the (i, j) entry of the matrix product of k copies of N , i.e. the matrix N^k , gives the number of paths of length k between node i and node j .

A metabolic pathway is a special case of a metabolic network with distinct start and end point, initial and terminal vertices and a unique path between them (Forst and Schulten, 1999). It is part of the metabolic network of an organism and therefore it is connected to other parts of it, however it can be isolated when we are interested in its specific function. In this work the graph and adjacency matrix representation is also used for the analysis of specific pathways or smaller metabolisms.

As an example, figure 3 shows the D-alanine metabolism, important for the cell-wall preparations obtained from a variety of microorganisms. The green-coloured pathway is organism specific; in this case the organism is the bacterium *Pseudomonas fluorescens* (strain Pf0-1). Figure 4 presents the graph that corresponds to the metabolic pathway of D-alanine metabolism of this bacterium together with its adjacency matrix.

2.2 A stochastic model of evolution

A metabolic network is assumed to evolve down the branches of a phylogenetic tree according to a continuous time Markov process with finite states. Markov models have been widely used for modelling the evolution of sequences in order to estimate parameters such as the tree topology, substitution rates or inferring ancestral sequences. In the case of networks, this simplistic framework is applied not as a suggested method for predicting future structure, but as a way of mathematically comprehending how changes in a metabolic network could arise.

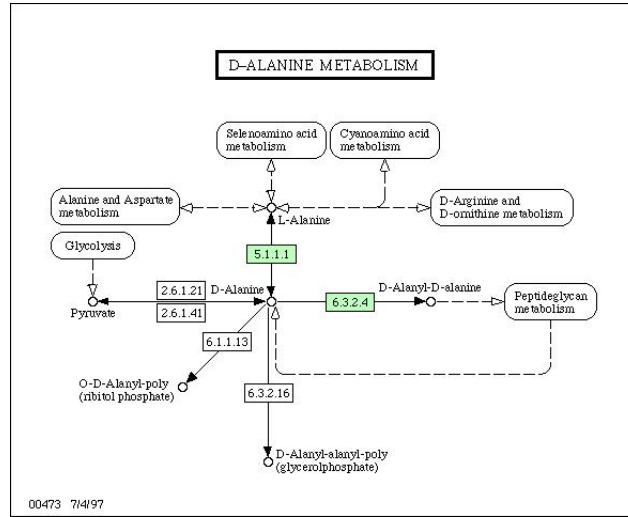


Figure 3: D-alanine metabolism. The green-coloured pathway is specific for the organism *Pseudomonas fluorescens* (strain Pf0-1). Taken from the KEGG database (Kanehisa and Goto, 2000).

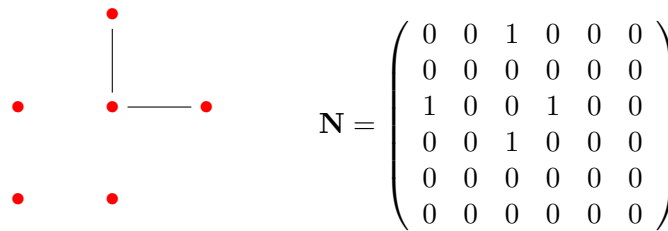


Figure 4: The graph and the adjacency matrix that corresponds to the metabolism of D-alanine in *Pseudomonas fluorescens* (strain Pf0-1).

The dynamics that are defined on the network and that give rise to the different states of our Markov chain, consist of creation or deletion of interactions between the metabolites. Given a set of fixed metabolites, and a set of present reactions between them, one of the edges of the graph can be either created with rate μ if it was absent or deleted with rate λ . The next change will take place after an exponentially distributed length of time. According to the Markov property the current state of the network determines the future dynamics. The average waiting time of the network before changing to another state is a function of the entire network; it depends on the number of edges that can be inserted or deleted and their rates. All of the current reactions (or their absence) of a metabolic network influence its evolution.

Another property that is applied here is that, at any given time, the pairs of metabolites (basically the reactions between them) are conditionally independent of each other given the current state of the process. Therefore, the possibility of simultaneous changes between two pairs of metabolites (i.e. the change in two edges) has probability zero. As the metabolism is gradually adapting to the environment, it is reasonable to assume that changes, such as the deletion of two enzymes, happen independently of each other.

The latter property can be modified if we supplement the above dynamics with two extra features, in order for the network to maintain its biological functionality. According to the first requirement, a metabolic network contains a core of reactions, which cannot be changed under evolution. This is because some reactions are essential to the organisms, such as reactions involved in some amino acid metabolisms. These cannot be changed through time as the organisms could not survive.

The second feature involves the connectedness requirement of the network over time. An edge

2.2.1 Metabolism Enumeration

One of the major problems that this model faces is that the number of states increases exponentially with the number of metabolites. The maximum number of states for networks with n metabolites is given by $\alpha_n = 2^{\frac{n(n-1)}{2}}$, as this specifies the possible changes that can happen to all edges. It is a maximum that is never reached because of the restrictions for the allowed operations specified in the model. However, the computation can get impractical for large networks, as it is not possible to efficiently enumerate all the states, let alone to store the corresponding rate matrix.

Specifically, we can calculate the number of connected graphs for a given number of vertices, where isolated nodes are allowed. The number of possible graphs for n vertices is $\alpha_n = 2^{\frac{n(n-1)}{2}}$. Let c_n be the number of connected graphs of size n with no isolated nodes. Then α_n can be written as:

$$\alpha_n = \sum_{k=1}^n \binom{n-1}{k-1} c_k \alpha_{n-k}$$

since the graph of size k has no isolated nodes while the rest can be any possible graph. Therefore α_n can be calculated by summing over all possible k . We can calculate c_n recursively since we have:

$$\begin{aligned} \alpha_n &= \binom{n-1}{n-1} c_n \alpha_{n-n} + \sum_{k=1}^{n-1} \binom{n-1}{k-1} c_k \alpha_{n-k} \\ c_n &= \alpha_n - \sum_{k=1}^{n-1} \binom{n-1}{k-1} c_k \alpha_{n-k} \end{aligned}$$

Let t_n be the number of possible connected graphs where the isolated nodes are allowed (isolated graph components are not allowed). We can calculate t_n using c_n by summing over the number of all possible isolated nodes:

$$t_n = \sum_{j=1}^n \binom{n}{j} c_j \quad (1)$$

The above equation is used to calculate the number of possible states for a metabolism of a given size.

2.2.2 Definitions of Markov process theory

For a continuous time Markov chain we denote by $P_{ij}^t(s)$ the transition probability from state i to state j and from time t to time $t+s$. If the chain is time-homogeneous we can denote $P_{ij}(s)$ the transition probability from i to j over s time period. Due to Markov property the time being spent in any state is memoryless; the distribution of the holding time depends only on the state and not on the time already spent on the state. Therefore the time is exponentially distributed. As $P_{ij}(t)$ correspond to transition probability we must have:

$$\sum_j P_{ij}(t) = 1$$

Also we have:

$$P_{ij}(t+s) = \sum_k P_{ik}(t) P_{kj}(s) = \sum_k P_{ik}(s) P_{kj}(t)$$

as in order to enter state j after $t+s$ time period, the Markov chain has to occupy certain states k that can be any state in the state space. Since the time is continuous, we can calculate P'_{ij} ,

which is the instantaneous rate of change. It can be calculated from the above equation for a small time interval that tends to zero:

$$P'_{ij}(t) = \sum_{k \neq j} P_{ik}(t)P'_{kj}(0) + P_{ij}(t)P'_{jj}(0) = \sum_{k \neq i} P'_{ik}(0)P_{kj}(t) + P'_{ii}(0)P_{ij}(t)$$

These equations are the forward and backward Chapman-Kolmogorov equations. Furthermore, we indicate:

$$q_{ij} = P'_{ij}(0), \quad v_i = -q_{ii}, \quad Q = (q_{ij})$$

Therefore the rate that the chain enters states j from state i is q_{ij} when $i \neq j$ and the rate that the chain leaves state i is given by v_i , which is equal to $\sum_{j \neq i} q_{ij}$. Using the matrix notation we have:

$$P'(t) = P(t)Q, \quad P'(t) = QP(t)$$

where P is the matrix with the transition probabilities. When the state space is finite, as in our case, we can always solve $P'(t) = P(t)Q$, $P(0) = I$ (I is the identity matrix) as:

$$P(t) = \exp(tQ) = \sum_{m=0}^{\infty} \frac{(tQ)^m}{m!} \quad (2)$$

If we are interested in the transition probability the chain enters state j when it leaves state i , then we can describe the process (the jump process) by its embedded discrete Markov chain. In this case we are not interested in when the chain moves to another state, but in where the chain moves to when the transition happens. This transition probability can be given by:

$$P_{ij} = \frac{q_{ij}}{v_i}$$

This means that when a chain leaves state i , it enters other states proportionally to the transition rates into those rates. The staying time of the chain in a specific state is exponentially distributed with parameter v_i .

2.2.3 Simulation

Using the above specifications, computer simulation of the stochastic process of the metabolic network evolution was used that was based on the Gillespie algorithm for chemical reactions (Gillespie, 1977). The algorithm for the simulation is the following:

- Begin with an initial state: a given network with a specified adjacency matrix N
- Set $t = 0$ and complete the following steps until $t < \text{time of evolution}$
 1. Enumerate all possible insertions
 2. Enumerate all possible deletions
 3. Calculate the insertion and deletion rates (q_{ij}) for all possible operations, and the total rate of all operations v_i
 4. Randomly draw the waiting time before the next operation $\Delta t \sim \exp(v_i)$, by generating a random number $u \sim U(0, 1)$ and taking $\Delta t = -\log(u)/v_i$
 5. Randomly choose an operation, by generating a random number $r \sim U(0, 1)$ and choosing j such that $\sum_{k=1}^{j-1} \frac{q_{ik}}{v_i} < r < \sum_{k=1}^j \frac{q_{ik}}{v_i}$
 6. Carry out operation (either insertion or deletion according to 5) and update N
 7. Update $t = t + \Delta t$

2.2.4 Stationary Distribution

The Markov process that we have described is by definition aperiodic, since the time between state changes is exponentially distributed, making it impossible to restrict state changes to occur only at regularly spaced intervals. Moreover, it was found to be positive recurrent. Therefore as t goes to infinity, the probability that a network is in a specific state is non-zero and independent from the starting state. We have:

$$\lim_{t \rightarrow \infty} P_{ij}(t) = \pi_j$$

Therefore:

$$P'(t) = 0 \Rightarrow \pi Q = 0 \quad (3)$$

A second equality that must be true for the stationary distribution is:

$$\sum_i \pi_i = 1 \quad (4)$$

The stationary distribution, also called equilibrium distribution or equilibrium frequencies, is very useful for inferring parameters such as the phylogeny of a tree given this model of evolution.

2.2.5 Time reversibility

A Markov chain is time reversible if the probability of sampling state i from the stationary distribution and going to state j is the same as the probability of sampling j from the stationary distribution and going to state i :

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t) \quad (5)$$

Alternatively, another way to test whether the model is time-reversible is to test if the following equality is true:

$$P_{ij}(t)P_{jk}(t)P_{ki}(t) = P_{ik}(t)P_{kj}(t)P_{ji}(t) \quad (6)$$

It is clear that the time reversibility is very important for a model of evolution on a phylogeny. The probability of starting with network i at one end of a phylogenetic tree, and ending with j , is the same as the probability of starting with j and evolving to i . This is not so much a realistic biological assumption, but a mathematical convenience in order to infer unrooted trees without having to place the root (the common ancestor) of the phylogenetic tree.

2.2.6 Likelihood calculation

Felsenstein (1981) introduced the likelihood calculation in molecular phylogeny. Maximum likelihood techniques have been widely used for the estimation of evolutionary trees from DNA sequences. The same methods can also be applied for the estimation of a phylogenetic tree that captures the evolution between metabolic networks of different species.

The aim is to compute the probability of a particular set of networks on a given tree and maximizing this probability over all evolutionary trees. The probability of obtaining a given set of networks at the tips of a given tree can be computed using the above model that specifies the probability that network N_1 changes to network N_2 during evolution along a segment of the tree of length (in time) t (using equation 2).

We assume that after speciation two lineages evolve independently, and that the same stochastic process of network evolution applies in all lineages. Felsenstein has proved a general expression for the likelihood of a tree, but initially it is more useful to present the expression for a simple particular case: The task is to calculate the likelihood of homology that is, the

likelihood that two networks have evolved from a common ancestor. The likelihood of a pair of networks, N_1 and N_2 , separated from a common ancestor N by divergence time t is:

$$P_t(N_1, N_2) = \sum_N P_\infty(N) P_t(N_1|N) P_t(N_2|N)$$

The probability of the ancestor N , $P_\infty(N)$, may be taken to be the probability that, at a random point on an evolving lineage, network N would be seen. We assume that evolution has been proceeding for a very long time according to the particular model, so it is reasonable to take $P_\infty(N)$ to be the equilibrium probability of network N , $\pi(N)$, under that model. Because of the time-reversibility of the model (equations 5 and 6 are true) the joint probability of N_1 and N_2 is equivalent to the probability of one network evolving to the other, in twice the time that separates the ancestor from the two descendants:

$$P_t(N_1, N_2) = \sum_N P_\infty(N_1) P_t(N|N_1) P_t(N_2|N) = P_\infty(N_1) \sum_N P_t(N|N_1) P_t(N_2|N) \Rightarrow$$

$$P_t(N_1, N_2) = P_\infty(N_1) P_{2t}(N_2|N_1) \quad (7)$$

In the case of 3 or more networks, the calculation of likelihood involves the sum, over all possible networks that may have existed at the interior nodes of the tree, of the probabilities of each scenario of events. A recursion can be used (Felsenstein, 2004). It computes the likelihood $L_k(s)$ at each node on the tree from the same quantities in the immediate descendant nodes. Let node k have immediate descendants l and m , which are at the top ends of branches of length t_l and t_m . The likelihood can be given by:

$$L_k(s) = \sum_x P_{t_l}(x|s) L_l(x) \sum_y P_{t_m}(y|s) L_m(y) \quad (8)$$

According to the above equation the probability of everything at or above node k , given that node k has state s , is the product of the events taking place on both descendant lineages. In each lineage, it sums over all the states to which s could have changed, and for each of those computes the probability of changing to that state multiplied by the probability of everything at or above that node (e.g. l), given that state has changed to state x . This algorithm starts at the node that has all of its immediate descendants as tips. The result is L_0 for the common ancestor in the tree. The likelihood for this network is given by calculating a weighted average of these over all possible networks, weighted by their prior probabilities:

$$L = \sum_x \pi_x L_0(x) \quad (9)$$

As it has been already discussed, the number of states that a network can change to, increases exponentially with the number of its nodes. Therefore the computation of the exact likelihood gets impractical for large networks. In order to estimate rates for larger networks we used parsimony reasoning; it is assumed that evolution between two metabolisms has taken the shortest possible path, instead of the most likely one. Given two or more metabolisms, we enumerate possible changes between them accounting only for the most parsimonious ones. Hence, only shortest paths between metabolisms are considered, and the likelihood calculation is done over those.

A solution to the full likelihood calculation problem could be given, if the number of states are not enumerated, but the corresponding transition probabilities are calculated using a recursive formula similar to the one used by Wiuf et al. (2006). In that case, the likelihood of a network is calculated conditional on an initial graph isomorphism. As the model proposed there is a general duplication-attachment model, this recursion is doable. There was no need to specify a distribution on the initial graph, because any two paths that result in the same network initiate

with isomorphic graphs. However, in our case, since deletions of edges are allowed, and the model is not based on duplication of nodes, a recursion starting from the final network will not end in the same initial network.

Alternatively, if we are interested in the estimation of parameters by maximising the likelihood of a phylogeny, then we can use heuristics or sampling Markov Chain Monte Carlo methods, as described in Appendix A. In this case, we need to sum over all possible paths that a network N_1 can follow in order to evolve to network N_2 , without enumerating all the states and creating the rate matrix Q .

2.3 Model with Fitness-dependent link dynamics

The model described above presents a simplistic framework for network evolution. As the environment plays a key role for adaptation and evolution, we present here a model that accounts for the fitness of the organism changing its metabolism.

By quantifying the concentration of species that have a specific structure of metabolism, we are interested in its rate of change. The following model is inspired by the quasispecies model used to describe the process of the Darwinian evolution of self-replicating entities first proposed by Eigen et al. (1988). According to this theory, new sequences enter the system as a result of a copy process either correct or erroneous, of other sequences that are already present, while existing sequences can decay. Mutations occur through errors made in the copying process and selection arises as different sequences tend to replicate more at different rates. Due to the ongoing production of mutant sequences, selection does not act on single sequences, but on mutational “clouds” of closely related sequences, referred to as quasispecies.

In the case of metabolic networks, we assume that a species mutates to other species by insertions or deletions of edges in their metabolism, causing the concentration of the rest of the species to change. Here by species, we mean sets of them with the same metabolism. Therefore we can define: $m(s, s')$ the instantaneous mutation rate from species with network s to species with network s' that depends on the structure of s and s' and the rates of insertion and deletion of edges, and $f(s)$ the fitness of species with network s that depends on the structure of s and the environment. We have:

$$\frac{ds}{dt} = f(s)s(1 - \sum_{s'} m(s, s')) + \sum_{s'} m(s', s)f(s')s' \quad (10)$$

The first term on right of the equation is gain in concentration by the growth of species s ignoring mutations and also loss of concentration from mutations of species with network s to species with network s' . The second term is gain in concentration from species mutating to ones with network s . The fitness function can be chosen to be the concentrations of some key metabolites that can change due to environmental conditions. However, the fitness also depends on the structure of the network s . Therefore, if we are interested in a specific pathway, we can only use the concentrations of the metabolites that affect the output of the pathway according to the structure of the graph.

Using this model we can see how the environment can affect the dynamics by changing the concentration of key metabolites. Moreover, the importance of the network structure can be confirmed, as different networks have different fitness functions. Finally, questions concerning the effects of changing environment versus stable environment can be answered.

3 Results

In this section, we present the simulation results of both models described above, as well as maximum likelihood estimates of evolutionary model parameters, i.e. rates of gain and loss of enzymatic reactions.

3.1 Stochastic model of evolution - Simulation results

As we have already described, the dynamics that are defined on a network consist of creation or deletion of interactions between the metabolites. In other words, the enzymes that catalyse the reactions get turned on and off during evolution. We simulated the generation and evolution of a continuous-time Markov chain of adjacency matrices for a time period (t_m, t_{m+1}) , where t_m can be set to 0, starting from an initial state N_0 , which is the observed configuration of the network at t_m . As described in section 2, for the simulation, the Markov process is defined in terms of its jump chain, the consecutive states visited by the Markov chain, and the waiting times between consecutive changes.

In figure 6, an example is given of a simple theoretical metabolic network, of which the connectivity changes over time. The example presents the network observed at two time points $t_m = 0$ and $t_{m+1} = 10$ (million years). For the insertion and deletion rates, we used $\lambda = 3 * 10^{-3}$ and $\mu = 5 * 10^{-3}$ per million years. We should note that these are rates of insertion or deletion of a single edge, therefore the total rate of change of the network in the initial state N_0 is much higher ($v_i = \sum_{j \neq i} q_{ij} = 0.042$). The network goes through two states during its evolution for the specific time interval between the two observations. This presents how the enzymatic reactions can get lost or gained depending on the diversification of the species.

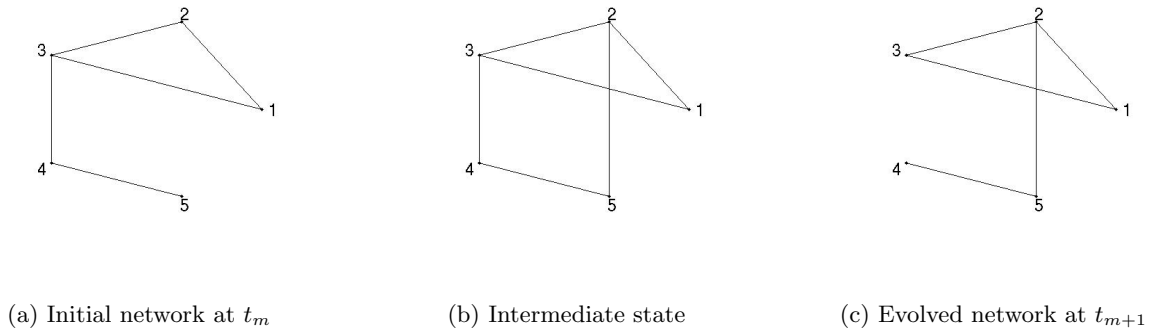


Figure 6: Example of the evolution of connectivity in a simple metabolic network. $\lambda = 3 * 10^{-3}$, $\mu = 5 * 10^{-3}$, $t_m = 0$ and $t_{m+1} = 10$.

Figure 7 shows the evolution of the simple theoretical metabolic network for $t_{m+1} = 100$. In this case, the network goes through more states during its evolution, as the time interval between the two observations is longer. During this time the network reaches a state which is its core metabolism as specified in the beginning of the simulation. This is the set of interactions that cannot be removed, thus the next state can only be the result of an insertion of an edge.

As we can see in figure 8, the number of highly connected networks increases when the rate of insertion is greater than the rate of deletion of an edge. Respectively, the number of less connected networks increases when the rate of deletion is greater than the rate of insertion. When the two rates are equal, the network changes states stochastically, having though almost the same average total degree distribution. The network, in this example, has 4 metabolites and it starts with no interactions. The total number of states is 61. The chain is run for 500 hundred iterations and for $t_{m+1} = 100$.

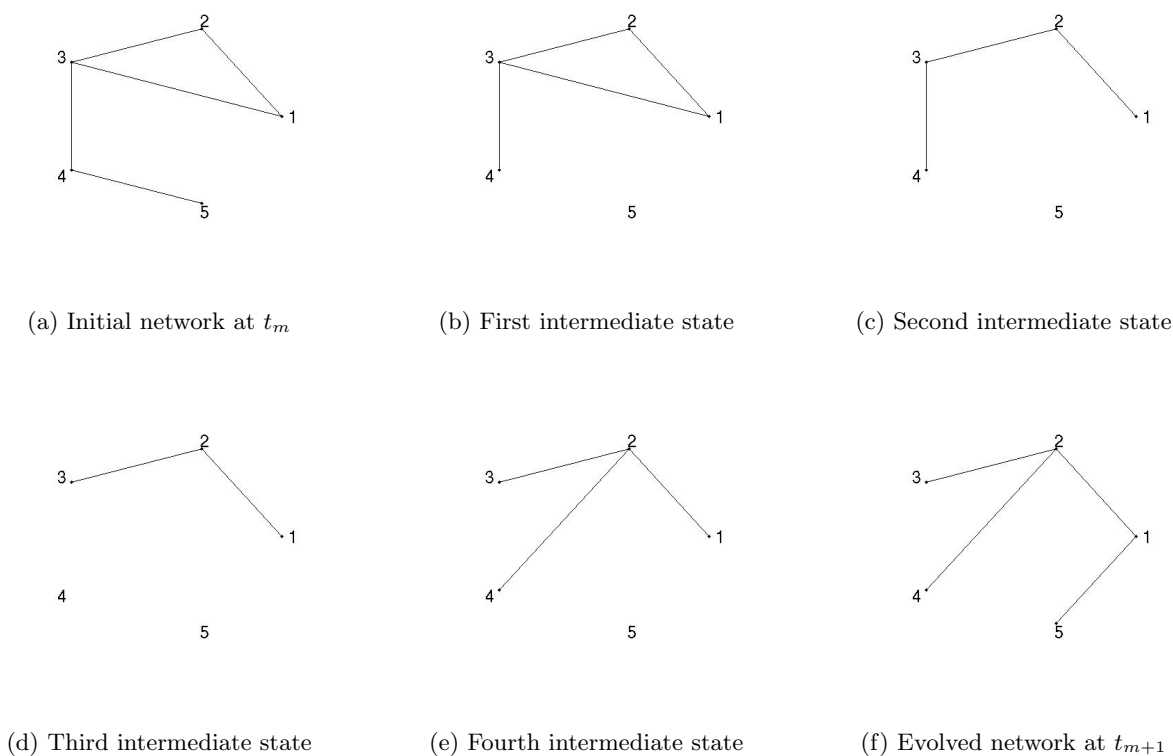


Figure 7: Example of the evolution of connectivity in a simple metabolic network. $\lambda = 3 * 10^{-3}$, $\mu = 5 * 10^{-3}$, $t_m = 0$ and $t_{m+1} = 100$.

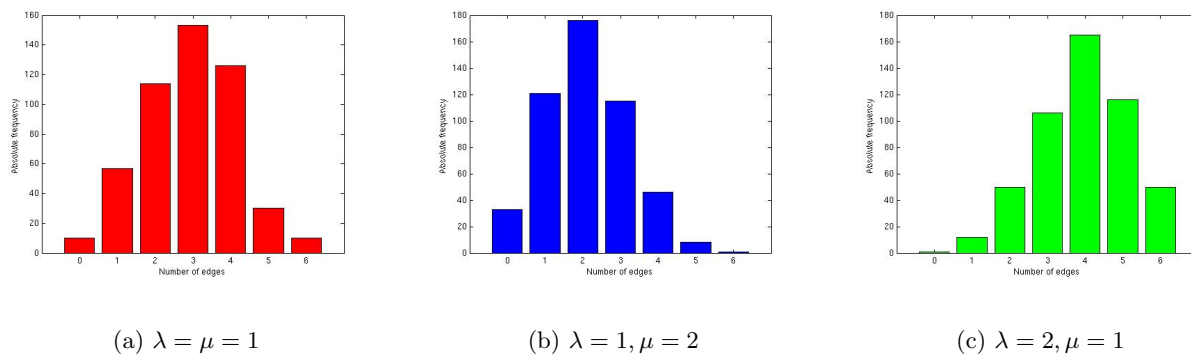


Figure 8: Frequency distribution. The number of observations of networks with different number of edges after evolution of an initial network of size 4 with no reactions. $t_m = 0$, $t_{m+1} = 100$.

3.2 Enumeration of all possible states

Since the evolution process is modelled as a continuous-time Markov chain, we need to calculate its finite state space. As we have mentioned in section 2, the problem this model faces is the large number of states it can have. Table 1 presents the number of all graphs with n vertices as well as the number of possible graphs for metabolisms with n metabolites. In the latter case, disconnected components in the graph are not allowed, while isolated vertices are. It is obvious that computation can get impractical for large networks. The connectedness requirement of the network over time does not improve the computational tractability of the problem. Therefore,

for the calculation of the likelihood of simple trees we use small metabolisms. We should note that the number of states is smaller when we take into account the core metabolism for a specific network, as not all the transitions are allowed.

n	Number of all graphs with n nodes	Number of states
1	1	1
2	2	2
3	8	8
4	64	61
5	1024	969
6	32768	31738
7	2097152	2069964
8	268435456	267270033
9	68719476736	68629753641
10	35184372088832	35171000942698

Table 1: Number of possible states of the Markov Chain for networks of size n .

3.2.1 Rate matrix and Stationary distribution

The equilibrium probability of a network N is used as the likelihood of the network under the evolutionary model, since we assume that evolution has been proceeding for a very long time. Therefore, for the calculation of likelihood of a phylogeny we need to have the stationary distributions of the networks. We use the balance equations 3 and 4 in order to calculate it. Since we enumerate all the states of the Markov chain we can create and store the rate matrix Q , when this is feasible (for small metabolisms). For example, for networks with 4 metabolites, the rate matrix is 61×61 and it has the form shown in figure 9. In this case, when $\lambda = \mu$ the stationary probabilities are equal for each state, $\pi_i = 0.0164$. When $\lambda > \mu$ the most probable state is the fully connected network, and all the highly connected ones have higher frequencies (Figure 10). Respectively, when $\lambda < \mu$, the most probable state in equilibrium is the network with no reactions. In this example, we have no core metabolism.

Network	Graph Structure
N_1	
N_2	
N_3	
N_4	

$Q(N_i \rightarrow N_j)$	N_1	N_2	N_3	N_4	...
N_1	$-\sum_j q_{1j}$	λ	λ	0	...
N_2	μ	$-\sum_j q_{2j}$	0	λ	...
N_3	μ	0	$-\sum_j q_{3j}$	λ	...
N_4	0	μ	μ	$-\sum_j q_{4j}$...
...

Figure 9: Examples of small metabolisms and the corresponding entries in the rate matrix. For metabolisms with 4 metabolites the matrix is 61×61 .

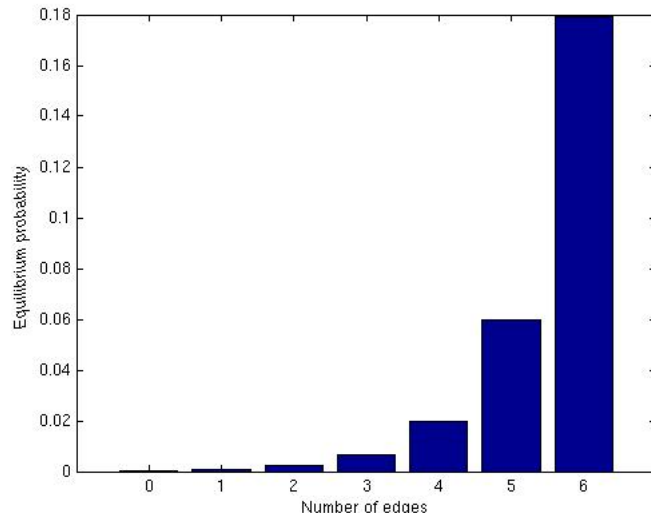


Figure 10: Stationary distribution for networks with 4 metabolites when $\lambda = 3 \cdot 10^{-3}$ and $\mu = 10^{-3}$. The networks with the same number of edges are only represented once in this graph.

3.3 Maximum Likelihood Estimates of a small phylogeny of networks

In the case of phylogenetic analysis of sequences, as Felsenstein (1981) suggested, the aim is to find the phylogeny that gives the maximum likelihood. This is done by searching in a space of trees with branch lengths. Specifically, we have to find the optimum branch lengths for each given tree topology and search the space of tree topologies for the one that has a set of branch lengths that gives it the highest likelihood. The likelihood approach can also be used to infer unspecified parameters such as substitution rates.

In the case of metabolic networks, the phylogenetic tree between organisms can be given by the sequence analysis, therefore it can be assumed. The branch lengths, i.e. the time of evolution between species diversification, can be known, so the likelihood approach is used to estimate parameters concerning the gain and loss of enzymatic interactions. Specifically, we present here an example of how the suggested stochastic model of evolution can be used to phylogenetically estimate the rates of insertion and deletion of interactions between metabolites. Since the computation can get really hard for large networks and phylogenies of more than 2 metabolisms, we describe here a simple theoretical example.

Let N_1 and N_2 be two theoretical metabolic networks as shown in figure 11. No core metabolism is used for this example, however the restriction for connectedness holds. The two networks have evolved from a common ancestor (Figure 12), however since the model is time reversible we do not need to sum over all possible ancestral networks. Instead, it is sufficient to treat one modern network as if it were the ancestor and the other modern network as if it were the descendant. Therefore, for the computation of the joint probability of both networks we use equation 7.

For the specific example we used 100 (million years) as the divergence time between each network and the unknown common ancestor. The maximum likelihood estimates for the insertion and deletion rates are: $\lambda \approx \mu \approx 0.003$. Figure 13 shows the likelihood curve where we can see the estimated values of the parameters of interest. In addition, in figure 14 we can see how the equilibrium probabilities of N_1 and N_2 and the transition probabilities $P(N_2|N_1)$ and $P(N_1|N_2)$ (used for the calculation of the likelihood function) change for different values of these rates. We should note that because of the time reversibility of the process we have $L = P(N_1)P(N_2|N_1) = P(N_2)P(N_1|N_2)$.

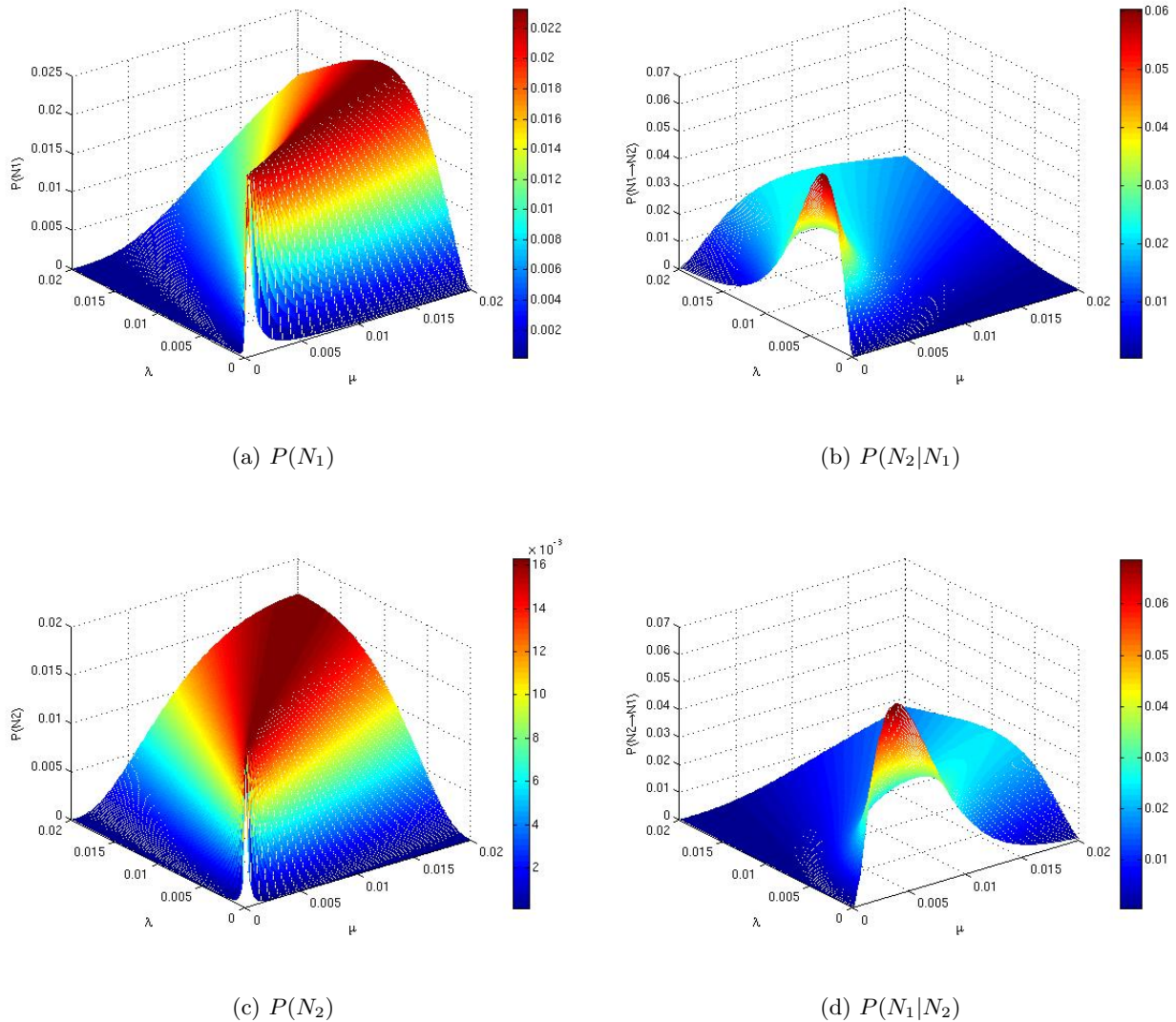


Figure 14: The equilibrium probabilities of N_1 (a) and N_2 (c) and the transition probabilities from N_1 to N_2 (b) and from N_2 to N_1 (d) as a function of the insertion and deletion rates.

3.3.1 Estimation using parsimony

Using the exact likelihood calculation we were able to estimate parameters for networks with 4 metabolites in phylogenies with up to three different species. However, for networks with more than 4 and up to 6 metabolites the same computation is not feasible due to limiting computational resources. This is because the internal nodes in the phylogenetic tree have to be inferred and 3 pairs of evolutionary model rates have to be estimated from maximising the likelihood function. Therefore, for networks of this size, estimates of evolutionary rates in a phylogeny with only two metabolites could be made.

We used parsimony methods, as described in the methods section in order to estimate parameters for these networks. Comparing the results of the parsimony analysis for the small networks, where the exact likelihood calculation is possible, we can say that the parsimony analysis performed quite well. The estimated rates of insertion and deletion of edges for the networks shown in figure 11 were accurate to three decimal digits.

For networks with more than 6 metabolites, the exact likelihood computation is not possible even for phylogenies of two metabolisms. In that case, the parsimony reasoning can be applied.

However, problems might still occur if the networks differ significantly from each other. Then many states have to be enumerated again, even though the evolutionary paths between the two metabolisms are the shortest ones.

3.4 Model with Fitness-dependent link dynamics — Simulation results

As we described in section 2, we assume that a species mutates to other species by insertions or deletions of edges in their metabolism. Moreover, the formation of the required end product from an intermediate metabolism increases the fitness of the organism. Therefore, the concentration of key metabolites, the rates of insertion and deletion and the structure of the network determine the dynamics of network evolution. For the simulation of this model we used 7 small metabolisms as presented in table 2. These networks have two key metabolites (A and B) in their intermediate metabolism and an end product (C).

According to the structure of each network, the fitness function is chosen to be the concentrations of the key metabolites that affect the end product. Therefore, we have $f(s_{G_1}) = [A] + [B]$, which means that the fitness function of graph G_1 is equal to the sum of the concentrations of both metabolites A and B. This is because for this graph both of them affect the concentration of the end product. Similarly for the rest of the networks we have: $f(s_{G_2}) = [A]$, $f(s_{G_3}) = [B]$, $f(s_{G_4}) = [B]$, $f(s_{G_5}) = [A]$, $f(s_{G_6}) = [A]$ and $f(s_{G_7}) = [B]$. We should note that the stoichiometries of these two key metabolites are considered to be 1 : 1 in this example.

Furthermore, the mutation rate from one organism to the other depends on the structure of the two networks and the insertion and deletion rates. Therefore, the organism with the network G_1 can mutate to G_2 , G_3 , G_4 or G_5 with mutation rate equal to the deletion rate of an edge (μ), while these four networks can mutate back to G_1 with mutation rate equal to the insertion rate λ . Similarly, network G_2 can mutate to G_6 with rate μ , while G_6 can mutate to G_2 with rate λ and so on.

In order to calculate the concentration of the 7 species that have these metabolic structures, we numerically solved a set of seven linear differential equations that have the form of equation 10. In figure 15 we present some characteristic results. For the results in figure 15(a) we used $A = 0.001$ and $B = 0.002$ for the concentration of the metabolites and for the rates $\lambda = 0.01$, $\mu = 0.1$. Moreover the initial values for the relative concentrations were $s_{G_1} = 1$, $s_{G_2} = 0$, $s_{G_3} = 0$, $s_{G_4} = 0$, $s_{G_5} = 0$, $s_{G_6} = 0$ and $s_{G_7} = 0$. Since the deletion rate is much larger than the insertion rate the relative concentration of G_1 decreases since it loses some of the interactions between the metabolites. Therefore, the relative concentration of the networks with one less interaction starts to increase. However, the increase is not the same for all of them since the fitness function affects the result. Hence, the concentration of G_4 is larger than the one of G_5 , since we used higher concentration of the B metabolite than A metabolite. We should note that two pairs of graphs have exactly the same relative concentration since they have the same fitness function. These are graph G_2 with G_5 , and graph G_3 with G_4 . As time evolves, we see a large increase of the concentration of G_7 . This is because this network, is the least connected one, which is the result of the high deletion rate, and most importantly its key metabolite is B which has the highest concentration. In the stationary point, the relative concentration of species with G_7 is clearly the highest, while the concentration of G_4 follows since the fitness of these species depends on B. The rest of the species remain in very small concentrations.

For the results in figure 15 (b), we used the same initial values and insertion and deletion rates, but we made the concentration of metabolite A increase over time. Starting from zero it changes proportional to time. Initially, as the concentration of metabolite A is really small we get the same behaviour as in 15 (a) and in a faster rate. The relative concentration of G_1 starts to decrease rapidly while the rest start to increase. The concentration of G_7 increases very fast since it is one of the least connected networks. However, as the concentration of A increases the concentration of organisms with G_7 starts to decay. This is because the fittest species is now the one with G_6 metabolic network, which is the least connected one with metabolite A affecting the end product. It is worth mentioning the temporal increase of G_1 concentration, as

Network	Graph Structure
G_1	
G_2	
G_3	
G_4	
G_5	
G_6	
G_7	

Table 2: Small metabolisms used for simulation.

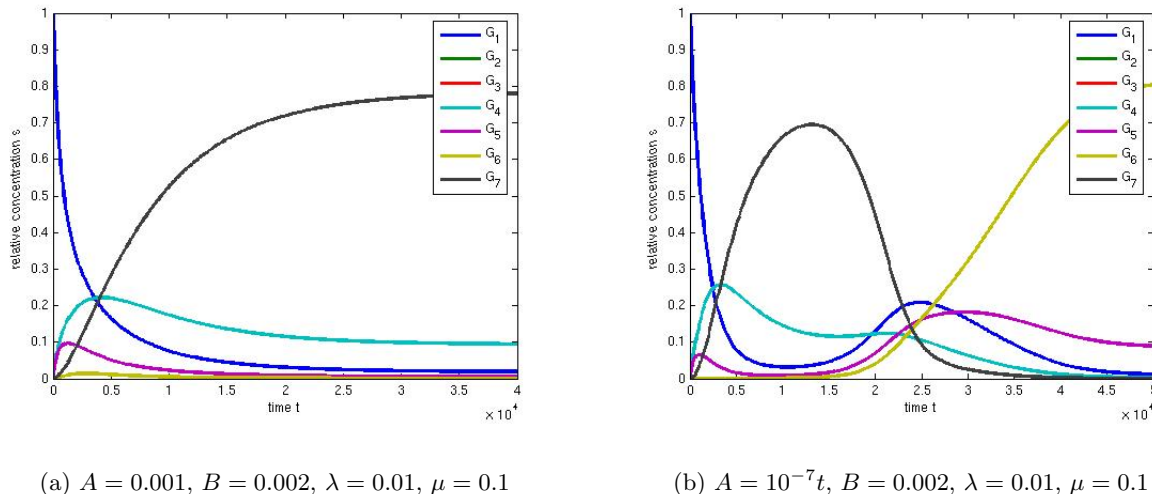


Figure 15: Relative concentration of species with different metabolisms over time. Simulation of model with fitness-dependent link dynamics.

the concentration of A and B become equal. At this time point ($t = 2 \times 10^4$), all 7 concentration of species tend to become close as the organisms have similar fitness values, with G_1 being the fittest. Since though the deletion rate is much higher than the insertion rate and as the fitness functions change due to the increase of $[A]$, the bottleneck effect takes place. In the long run, the relative concentration of organisms with G_6 metabolic network greatly outnumbers the rest.

4 Discussion and Further Work

The analysis of metabolic networks has so far mainly relied on ad hoc descriptions and summary statistics. Many properties of these networks have been revealed, although most of them depend on the way of representing and defining a metabolism or a metabolic pathway. The process of evolution of these networks can be used in order to justify their principal design. A number of theories have been developed that attempt to explain the evolution of the biochemical pathways, with each one finding applications in different pathways. Furthermore, studies have attempted to construct phylogenies based on metabolic networks using mainly distance based methods. Although models for network evolution have been introduced, they do not apply directly to metabolic networks where the representation can be different. There is clearly a need of modelling the evolutionary process of metabolic networks in order to be able to infer important parameters, such as the rate of gain or loss of enzymatic reactions.

In this work we have suggested a simplified stochastic model of evolution of metabolic networks in order to estimate these important parameters. This approach is computationally very demanding, so we have restricted the illustration to small theoretical metabolisms. The use of real data was not possible at this point as the real metabolic networks of organisms are really complex and large, while the small metabolic pathways do not show any substantial difference between species. However our work leaves room for improvement. Markov Chain Monte Carlo methods can be used to perform computation for large networks as suggested in Appendix A. The estimated evolutionary model parameters can be compared with the observed ones in order to evaluate the model.

For dealing with phylogenies of more than two networks, efficient algorithms need to be used in order to examine unknown internal states and to infer interesting parameters. In that case, Bayesian phylogenetic inference can be used where we can specify prior distributions for all the unknown parameters.

The model can be improved by incorporating preferential attachment in the context of metabolic networks. This can also be done if a different representation is used, where the enzymes are mapped as nodes in the graph and the metabolites are the edges. In that case, a possible biological explanation for preferential attachment growth of metabolic networks is that novel enzymes created through gene duplication maintain some of the compounds involved in the original reaction, throughout its future evolution. In addition, it has been found that enzymes which are candidates for horizontal gene transfer have a higher average connectivity than other enzymes (Light et al., 2005).

Moreover, in order to make the model biologically more meaningful we need to consider introducing varying evolutionary rates. Specific parts of the network can have differences in the rates of insertion and deletion of the enzymatic reactions. Another important thing to consider is the addition of directionality to connections that would determine the nature of the reactions. In that case, we can also look at the evolution of metabolites.

Comparative studies will also help to better understand the evolutionary process in organisms and therefore create more realistic mathematical models. It is of interest to ask how the diversification of metabolic networks during the evolution affected phenotypic traits of organisms. Using the gene set predicted from the complete genome sequence, we can examine whether a particular gene existed in a particular species.

We have also introduced a mathematical framework that takes into account the fitness of organisms during evolution. It is useful mainly in providing a qualitative understanding of the evolutionary processes of metabolic networks, where gains and losses of enzymatic reactions can occur. As we have already mentioned, the process is similar to the description of Darwinian evolution of self-replicating entities, given by the quasi-species model. Although this is generally considered a good model for self-replicating macromolecules such as RNA or DNA or simple asexual organisms such as bacteria or viruses, it is not very descriptive for species where there is a high degree of error-correction in replication. However, we can get an understanding about the gene loss and gain in metabolism evolution; when a network starts with all its enzymatic reactions, but mutations aim to remove the genes, selection tries to keep the network connected in a favourable way. Hence, the model is more realistic in describing reductive evolution of genomes. If the deletion rate is much higher than the insertion rate, gene function is lost, as the specific gene is not needed in the environment of the organism. Generally, high mutation rates lead to high rate of initial evolution, but do not produce the maximum extent of evolution due to mutation/selection balance. Although we have not demonstrated that, we should note that in this model, networks with high fitness can lose out against networks with lower fitness that have better support from their mutational neighbors (Wilke, 2005). This effect has been termed survival of the flattest as these organisms although they occupy lower fitness peaks, were located in flatter regions of the fitness surface and were therefore more robust with respect to mutations (Wilke et al., 2001).

In this work, the *in silico* network evolution performed using the above model presented the effect of environmental changes in the gene content of the organisms. Specifically, by altering the concentration of key metabolites, the organisms adapt to the new environments by acquiring new enzymes or losing some of the existing. The fitness of the organisms as well as the structure of the network determines which of the enzymatic content will be lost or gained. A challenge related to this model is the inference of the environment of the different species according to their relative concentrations.

Acknowledgements

Many thanks to Prof. Jotun Hein, Dr. Thomas Mailund and Dr. Andrea Rocco for their invaluable help, guidance and support, to Dr. Gail Preston for providing essential biological advice, to the staff of the DTC for giving me the opportunity to work on this project, and to Robin Ryder for some mathematical comments.

Appendix

A Estimation of parameters

Monte Carlo Markov Chain methods can be used for the estimation of parameters such as the rates of insertion and deletion of edges when we are dealing with large networks. In this case, the enumeration of all the states that a network can go through (before it evolves to another), is not needed. Moreover, the summation over all the different paths that it can go through can be estimated using Gibbs sampling (Casella and George, 1992).

Specifically, let $L_t(\lambda, \mu)$ be the likelihood function (at time t) that we want to maximise, and ϱ the vector denoting a path that a network N_1 can follow to evolve to network N_2 . Then we have:

$$L_t(\lambda, \mu) = \sum_{\varrho: N_1 \rightarrow N_2} P_t(\varrho, \lambda, \mu)$$

In this case, we have (λ, μ) and (ϱ) and we wish to compute the marginal $L_t(\lambda, \mu)$. The idea behind the Gibbs sampler is that it is easier to consider the conditional distributions $P_t(\varrho|\lambda, \mu)$ and $P_t(\lambda, \mu|\varrho)$ than to obtain the marginal by the summation of the joint density. The sampler starts with some initial values (λ_0, μ_0) for (λ, μ) and generates ϱ_0 by sampling from the distribution $P_t(\varrho|\lambda, \mu)$. Consequently, the sampler uses ϱ_0 to generate (λ_1, μ_1) drawing from the distribution $P_t(\lambda, \mu|\varrho)$. Repeating this process n times, it generates a sequence of length n where a subset of points $(\lambda_i, \mu_i, \varrho_i)$ (where $1 < i < n$) are the simulated estimates from the joint distribution $P(\lambda, \mu, \varrho)$.

For the application of the Gibbs sampler described, we need to determine the two conditional distributions $P_t(\varrho|\lambda, \mu)$ and $P_t(\lambda, \mu|\varrho)$. For sampling ϱ from $P_t(\varrho|\lambda, \mu)$, the Metropolis-Hastings algorithm can be used (Chib and Greenberg, 1995). It can generate a Markov chain whose stationary distribution is the target density $P_t(\varrho|\lambda, \mu)$. At each time, the next state of the chain is chosen by sampling a candidate point ϱ' from a proposal distribution $q(\varrho'|\varrho)$. The candidate ϱ' is then accepted with probability $\alpha(\varrho, \varrho')$ where

$$\alpha(\varrho, \varrho') = \min \left(1, \frac{P_t(\varrho'|\lambda, \mu)q(\varrho|\varrho')}{P_t(\varrho|\lambda, \mu)q(\varrho'|\varrho)} \right)$$

Subsequently, we want to sample (λ, μ) from $P_t(\lambda, \mu|\varrho)$. Since, in this case, the path between the two networks (during time $[t_0, t_n]$) is known, by associating each event with time we have:

$$P(\lambda, \mu|\varrho) = \prod_{0 < k < n} v_i(k) \exp(-v_i(k-1)[t_k - t_{k-1}])$$

where $v_i(k)$ is the total rate of events in state k , which is $v_i(k) = \lambda A(k) + \mu D(k)$, ($A(k)$ is the number of addable edges in state k , while $D(k)$ is the number of removable edges in state k). We have:

$$\begin{aligned} P(\lambda, \mu|\varrho) &= \prod_{0 < k < n} v_i(k) \exp\left(-\int_{t_0}^{t_n} v_i(k(t)) dt\right) \\ &= \prod_{k=\text{insertion}} v_i(k) \exp\left(-\int_{t_0}^{t_n} \lambda A(k(t)) dt\right) \prod_{k=\text{deletion}} v_i(k) \exp\left(-\int_{t_0}^{t_n} \mu D(k(t)) dt\right) \\ &\propto \lambda^{n_{ins}} \exp\left(-\lambda \int_{t_0}^{t_n} A(k(t)) dt\right) \mu^{n_{del}} \exp\left(-\mu \int_{t_0}^{t_n} D(k(t)) dt\right) \end{aligned}$$

where n_{ins} and n_{del} are the number of insertions and deletions of edges that happened in the path between the two networks. From the above we get that we can sample from the target distributions:

$$\lambda|\mu, \varrho \sim \Gamma\left(n_{ins} + 1, 1/\int_{t_0}^{t_n} A(k(t))dt\right)$$

$$\mu|\lambda, \varrho \sim \Gamma\left(n_{del} + 1, 1/\int_{t_0}^{t_n} D(k(t))dt\right)$$

Therefore, by using the conditional distributions defined above, the Gibbs sampler can estimate the likelihood of the insertion and deletion rates.

References

- R. Alves, R. Chaleil, and M. Sternberg. Evolution of enzymes in metabolism: a network perspective. *J Mol Biol*, 320(4):751–70, 2002.
- M. Arita. The metabolic world of Escherichia coli is not small. *Proceedings of the National Academy of Sciences*, 101(6):1543–1547, 2004.
- A. Barabasi and Z. Oltvai. Network biology: understanding the cell’s functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.
- J. Berg, M. Lässig, and A. Wagner. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology*, 4(1): 51–51, 2004.
- S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.
- G. Casella and E. George. Explaining the Gibbs sampler. *The American statistician*, 46(3): 167–174, 1992.
- S. Chib and E. Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335, 1995.
- S. Dorogovtsev and J. Mendes. Evolution of networks. *Adv Phys*, 51:1079–1146, 2002.
- M. Eigen, J. McCaskill, and P. Schuster. Molecular quasi-species. *The Journal of Physical Chemistry*, 92(24):6881–6891, 1988.
- E. Eisenberg and E. Levanon. Preferential Attachment in the Protein Network Evolution. *Physical Review Letters*, 91(13):138701, 2003.
- J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.
- J. Felsenstein. *Inferring phylogenies*. Sinauer Associates Sunderland, Mass., USA, 2004.
- C. Forst, C. Flamm, I. Hofacker, and P. Stadler. Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. *BMC Bioinformatics*, 7(1):67, 2006.
- C. Forst and K. Schulten. Evolution of Metabolisms: A New Method for the Comparison of Metabolic Pathways Using Genomics Information. *Journal of Computational Biology*, 6(3-4): 343–360, 1999.
- D. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.
- R. Guimera and L. A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, Feb 2005.

- V. Hatzimanikatis, C. Li, J. Ionita, and L. Broadbelt. Metabolic networks: enzyme function and metabolite structure. *Curr Opin Struct Biol*, 14:300–306, 2004.
- S. Hong, T. Kim, and S. Lee. Phylogenetic analysis based on genome-scale metabolic pathway reaction content. *Applied Microbiology and Biotechnology*, 65(2):203–210, 2004.
- H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct 2000.
- M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.
- A. Lazcano. On the Origin of Metabolic Pathways. *Journal of Molecular Evolution*, 49(4):424–431, 1999.
- S. Light, P. Kraulis, and A. Elofsson. Preferential attachment in the evolution of metabolic networks. *BMC Genomics*, 6(1):159, 2005.
- H. Ma and A. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277, 2003a.
- H. Ma and A. Zeng. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11):1423–1430, 2003b.
- R. Mahadevan and B. Palsson. Properties of Metabolic Networks: Structure versus Function. *Biophysical Journal*, 88(1):7–9, 2005.
- C. Pal, B. Papp, M. Lercher, P. Csermely, S. Oliver, and L. Hurst. Chance and necessity in the evolution of minimal metabolic networks. *Nature*, 440(7084):667–70, 2006.
- C. Pal, B. Papp, and M. J. Lercher. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet*, 37(12):1372–1375, Dec 2005.
- J. Papin, N. Price, S. Wiback, D. Fell, and B. Palsson. Metabolic pathways in the post-genome era. *Trends Biochem Sci*, 28(5):250–258, 2003.
- B. Papp and C. Pál. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, 429:661–664, 2004.
- T. Pfeiffer, O. Soyer, and S. Bonhoeffer. The Evolution of Connectivity in Metabolic Networks. *Evolution*, 3(7), 2005.
- R. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401–3408, 2005.
- S. Rison and J. Thornton. Pathway evolution, structurally speaking. *Curr. Opin. Struct. Biol*, 12:374–382, 2002.
- C. Schilling, D. Letscher, and B. Palsson. Theory for the Systemic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective. *Journal of Theoretical Biology*, 203(3):229–248, 2000.
- S. Schmidt, S. Sunyaev, P. Bork, and T. Dandekar. Metabolites: a helping hand for pathway evolution. *Trends Biochem Sci*, 28(6):336–341, 2003.
- T. Snijders and M. Van Duijn. Simulation for statistical inference in dynamic network models. *Simulating Social Phenomena*, pages 493–512, 1997.
- K. Tun, P. K. Dhar, M. Palumbo, and A. Giuliani. Metabolic pathways variability and sequence/networks comparisons. *BMC Bioinformatics*, 7:24, 2006.

- H. Ueda and J. Hogenesch. Principles in the Evolution of Metabolic Networks. *Arxiv preprint q-bio.MN/0503038*, 2005.
- A. Wagner. The small world inside large metabolic networks. *Proceedings: Biological Sciences*, 268(1478):1803–1810, 2001.
- C. Wilke. Quasispecies theory in the context of population genetics. *BMC Evolutionary Biology*, 5(1):44, 2005.
- C. Wilke, J. Wang, C. Ofria, R. Lenski, and C. Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–3, 2001.
- C. Wiuf, M. Brameier, O. Hagberg, and M. Stumpf. A likelihood approach to analysis of network data. *Proceedings of the National Academy of Sciences*, 103(20):7566, 2006.
- Y. Zhang, S. Li, G. Skogerbo, Z. Zhang, X. Zhu, Z. Zhang, S. Sun, H. Lu, B. Shi, and R. Chen. Phylogenetic properties of metabolic pathway topologies as revealed by global analysis. *BMC Bioinformatics*, 7:252, 2006.