

# A stochastic model for the evolution of metabolic networks with neighbor dependence

Aziz Mithani<sup>1,\*</sup>, Gail M. Preston<sup>2</sup> and Jotun Hein<sup>1</sup>

<sup>1</sup>Department of Statistics and <sup>2</sup>Department of Plant Sciences, University of Oxford, Oxford, UK

Associate Editor Prof. Martin Bishop

## ABSTRACT

**Motivation:** Most current research in network evolution focuses on networks that follow a Duplication Attachment model where the network is only allowed to grow. The evolution of metabolic networks, however, is characterized by gain as well as loss of reactions. It would be desirable to have a biologically relevant model of network evolution that could be used to calculate the likelihood of homologous metabolic networks.

**Results:** We describe metabolic network evolution as a discrete space continuous time Markov process and introduce a neighbor-dependent model for the evolution of metabolic networks where the rates with which reactions are added or removed depend on the fraction of neighboring reactions present in the network. We also present a Gibbs sampler for estimating the parameters of evolution without exploring the whole search space by iteratively sampling from the conditional distributions of the paths and parameters. A Metropolis-Hastings algorithm for sampling paths between two networks and calculating the likelihood of evolution is also presented. The sampler is used to estimate the parameters of evolution of metabolic networks in the genus *Pseudomonas*.

**Availability:** An implementation of the Gibbs sampler in Java is available at <http://www.stats.ox.ac.uk/~mithani/networkGibbs/>

**Contact:** mithani@stats.ox.ac.uk

**Supplementary information:** Supplementary data are available at the journal's website.

## 1 INTRODUCTION

With an increasing emphasis towards studying system processes as a whole, biological networks such as protein interaction networks, metabolic networks, and gene regulatory networks have gained much attention in recent years. This has led to the development of computational and mathematical techniques allowing modeling of biological networks. These networks evolve over time and their evolution is one of the major areas of research today. A number of models have been proposed in the literature to study evolution of biological networks (Dorogovtsev and Mendes, 2003; Berg *et al.*, 2004; Ueda and Hogenesch, 2005; Boccaletti *et al.*, 2006; Wiuf *et al.*, 2006).

In this work, we focus on metabolic networks. The evolution of metabolic networks is characterized by gain and loss of reactions (or enzymes) connecting two or more metabolites. However, most of the current research focuses on networks that follow a Duplication

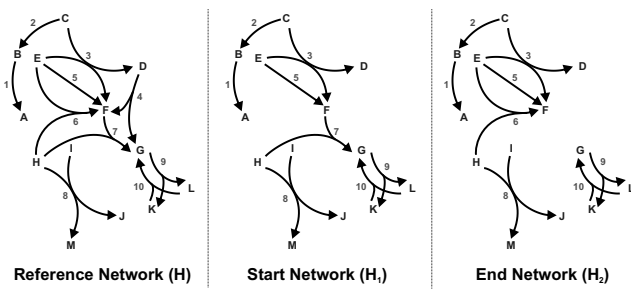
Attachment (DA) model (Chung *et al.*, 2003) where new nodes and edges are probabilistically added depending on a set of parameters, and the evolution of the network is considered as an evolutionary stochastic process where the number of nodes increases through a series of events relating to insertion of nodes and edges. Deletions of nodes or edges are not allowed in a DA model. This is clearly not realistic as the evolution of metabolic networks is characterized by both insertion and deletion events.

Moreover, the current models represent networks as directed or undirected graphs where nodes represent biological entities, e.g. proteins or metabolites, and the presence of an edge between any two nodes indicates the presence of some sort of relation between the nodes. However, the drawback of using ordinary graphs for representing biological networks is that they cannot capture the relationships between more than two nodes, for example multiple metabolites in a reaction.

We describe metabolic network evolution as a discrete space continuous time Markov process allowing both insertion and deletion of reactions. We represent metabolic networks as directed hypergraphs (Yeung *et al.*, 2007), where an edge (reaction) may connect any number of vertices (metabolites). Representing metabolic networks as hypergraphs not only captures the relationship between multiple metabolites involved in a reaction but also provides an intuitive approach to study evolution since loss or gain of reactions can be regarded as loss or gain of hyperedges.

We also introduce a neighbor-dependent model for the evolution of metabolic networks where the rates with which reactions are added or removed depend on the fraction of neighboring reactions present in the network. Two reactions are considered to be neighbors if they share at least one metabolite. The likelihood of evolution from one network to another for fixed parameter values can be calculated by integrating over the paths between the two networks. For this, we present a Metropolis-Hastings algorithm that allows moving in the space of paths between two networks. The likelihood of evolution can then be approximated by summing over the unique paths visited by the sampler. A Gibbs sampler for estimating the evolution parameters is also presented. The Gibbs sampler estimates the parameters without exploring the whole search space by iteratively sampling from the conditional distributions of the paths and parameters. This is particularly useful as exhaustive maximum likelihood or parsimony analysis is not feasible for estimating the rates for models with dependence between edges since the search space becomes intractable even for moderately sized networks (Snijders, 2005).

\*to whom correspondence should be addressed



**Fig. 1.** Toy networks consisting of 16 nodes. The nodes are labeled from A to M and the hyperedges are labeled from 1 to 10. The reference network consists of all allowed hyperedges for this example system. Networks  $H_1$  and  $H_2$  consists of subsets of the hyperedges from the reference network.

## 2 METHODS

### 2.1 Preliminaries

It is assumed that the number of nodes  $N$  in a network remains fixed and there is a set  $\mathcal{E}$  such that  $|\mathcal{E}| = M$  of hyperedges connecting these nodes. Let there be a network called *Reference Network* containing all these hyperedges. Assuming that hyperedges in the reference network are labeled 1 to  $M$ , a network  $x$  can be represented as a sequence of 0s and 1s such that  $i$ -th entry ( $0 < i \leq M$ ) in the sequence is 1 if and only if hyperedge labeled  $i$  is present in  $x$ , and 0 otherwise.

In the case of metabolic networks, it is further assumed for computational convenience that a reversible reaction is represented by two hyperedges. As a result a reversible reaction is gained or lost in two steps. The total number of hyperedges  $M$  in a metabolic network then corresponds to  $R$  reactions out of which  $K$  are reversible, i.e.  $M = R + K$ .

### 2.2 Datasets

**2.2.1 Toy Networks** Figure 1 shows toy networks used in this paper for example purposes. The network in the left most panel is the reference network  $H$  containing all allowed hyperedges for this example system and the middle and right panels show two networks,  $H_1$  and  $H_2$ , acting as starting and ending networks while studying evolution. The two networks differ by 2 hyperedges corresponding to 1 insertion (hyperedge 6) and 1 deletion (hyperedge 7).

**2.2.2 Metabolic Networks** Metabolic network data was extracted from the KEGG database (Kanehisa et al., 2006) for three different pathway maps across four species belonging to the genus *Pseudomonas* using the Rahnuma tool (Mithani et al., submitted). These pathway maps and the organism names are listed in Table 1 and the basic information for each network is given in Supplementary Table 1. For example, the pentose phosphate pathway map in *P. fluorescens* Pf0-1 contains 26 out of a possible 43 reactions, of which 17 are reversible, requiring 43 hyperedges in the reconstructed network. Compounds such as  $H_2O$ ,  $CO_2$ , ATP and other ubiquitous metabolites are called *current metabolites* (Ma and Zeng, 2003) and were ignored when reconstructing the networks, for reasons explained in Section 2.4.2. A complete list of metabolites which were ignored is given in (Mithani et al., in preparation). The number of non-current metabolites present in a metabolic network is shown in the last column in Supplementary Table 1.

### 2.3 Network evolution as a continuous time Markov process

The evolution of metabolic networks can be described as a discrete space continuous time Markov process. Markov models have been widely used for modeling the evolution of DNA sequences in order to estimate parameters

such as substitution rates and tree topology as well as for inferring ancestral sequences and aligning sequences. When applied to metabolic networks, models similar to those for DNA sequences can be used to understand the changes occurring in these networks.

Consider a metabolic network represented as a hypergraph. At each step of the network evolution a hyperedge is either added or deleted until the desired network is obtained. Modeling network evolution as a Markov process implies that the future dynamics of the system is determined by the current state of the network where a *state* is defined as the current configuration of the network, i.e. the set of hyperedges present in the network. The waiting times between the events are exponentially distributed and the average waiting time of the network before an insertion or deletion event takes place is a function of the entire network and depends on the number of edges that can be inserted or deleted from the network and the rates at which the hyperedges can be changed.

Using standard Markov process theory, the rate of change of the probability of being in a specific state can be calculated. Let the current state of the system be denoted by  $G \subseteq \mathcal{E}$  and the probability of a state by  $P(G)$ . The next state is characterized by insertion or deletion of a single hyperedge from the current network. Let the rate with which a hyperedge is inserted be  $\lambda$  and the rate with which it is deleted be  $\mu$ . Given a set of networks reachable by insertion of a single hyperedge from state  $G$ ,  $I(G)$ , and a set of networks reachable by deletion of a single hyperedge,  $D(G)$ , the dynamics of the system can be described by the following master equation.

$$\frac{dP(G)}{dt} = \mu \sum_{G' \in I(G)} P(G') + \lambda \sum_{G'' \in D(G)} P(G'') - P(G) \left( \lambda |I(G)| + \mu |D(G)| \right) \quad (1)$$

The first two terms on the right hand side of (1) correspond to the gain in probability due to the incoming transitions to state  $G$ , whereas the last term gives the total probability value to be subtracted due to outward transitions from state  $G$ .

**2.3.1 Core and Prohibited Hyperedges** To make our model biologically relevant, we allow some hyperedges to be specified as *core edges*, hyperedges that cannot be deleted during the course of evolution, and *prohibited edges*, hyperedges that cannot be added to a network. The former correspond to reactions or enzymes whose deletion can be lethal for the survival of the organism while the latter are the ones that, for example, are predicted to be absent in the current lineage. We denote the set of alterable edges by  $\mathcal{E}'$ . Designating certain hyperedges as core and prohibited edges implies that the sets  $I(G)$  and  $D(G)$  contain only those networks which are reachable by an alterable edge, i.e. the difference between states  $G$  and  $G'$  where  $G' \in I(G)$ ,  $D(G)$  is in the set of alterable edges  $\mathcal{E}'$ . Defining core and prohibited edges in the model not only reduces the size of the state space but also allows the model to be focused on reactions that are likely to change during the course of evolution.

### 2.4 Models of network evolution

In this section we describe two models pertaining to the insertion and deletion of an edge from a network. We first introduce a very simple model where each edge evolves independent of each other and then describe in detail a neighbor-dependent model of the evolution of metabolic networks where the rates with which hyperedges are added or removed depend on the fraction of neighboring hyperedges present in the network.

**2.4.1 Independent Edge Model** The simplest model is perhaps to assume that each hyperedge evolves separately. In this case, the size of the state space is 2 (a hyperedge is either absent or present). The corresponding Markov chain is described by a  $2 \times 2$  rate matrix  $Q$ , which is given as

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \quad (2)$$

where  $\lambda$  is the insertion rate and  $\mu$  is the deletion rate. The transition probability matrix is given as  $P(t) = \exp(tQ)$  where an entry  $P_{ij}(t)$  corresponds to the probability that the corresponding edge changes from

state  $i$  to state  $j$  in time  $t$ . For example,  $P_{00}(t)$  corresponds to the probability that the edge remains absent from the network after time  $t$ ,  $P_{01}(t)$  corresponds to the probability that the edge which was initially absent is present in the network after time  $t$ , and so on. It can be seen that (2) is an efficient way of describing the model presented in (1) when the edges are independent of each other since under the independence assumption the dynamics of the network can be obtained by considering the dynamics of the individual edges and taking the product of the corresponding probabilities (Snijders, 2005; Ross, 2007 Ch. 6).

The problem with this model, however, is that the reactions in metabolic networks do not evolve independently. Selection tends to favor the gain of reactions corresponding to a functionality useful for an organism and the loss of reactions that are either redundant or are not critical for the survival of an organism. Besides this, metabolic networks are small world networks (Dorogovtsev and Mendes, 2003) suggesting that highly connected nodes called hubs change their connections more often than the poorly connected nodes. As shown in Section 3.1, the independent edge model fails to capture this property of metabolic networks. A better way of modeling evolution is, therefore, to take neighbor-dependence between hyperedges into account as discussed next.

**2.4.2 Neighbor-Dependent Model** The Neighbor-dependent model is an extension of the independent edge model where hyperedges are inserted or deleted depending on their neighbors. Neighbor dependence in the context of DNA sequence evolution has been well characterized (Jensen and Pedersen, 2000) but not much explored in the context of metabolic networks. Two hyperedges are considered to be neighbors if they share at least one node (Yeung *et al.*, 2007). The neighbor-dependent model relates to the preferential attachment property reported for metabolic network evolution (Light *et al.*, 2005) and produces a behavior such that highly connected hyperedges are added or removed more frequently than those hyperedges which have very few neighbors.

*The Model* Consider a network  $x = (x_1, \dots, x_M)$ . Our model allows only one hyperedge to be inserted or deleted at a time (see Section 2.3) and, therefore, the networks reachable from  $x$  differ by only one hyperedge. Let  $x'$  be a network that differs from  $x$  at position  $i$ . The value at position  $x'_i$  is given as  $1 - x_i$  and the rate from  $x$  to  $x'$  depends on  $x_i$ ,  $x'_i$  and the neighboring hyperedges  $\Psi(x_i)$  present in the network  $x$ , and is given as

$$\gamma(x'_i; x_i, \Psi(x_i)) = q(x_i, x'_i)F(x_i, \Psi(x_i)) \quad (3)$$

where  $q(x_i, x'_i)$  is the appropriate entry from the rate matrix  $Q$  given in (2) and the function  $F$  corresponds to the neighborhood component. Since hyperedges no longer evolve independently the resulting rate matrix is a  $2^M \times 2^M$  matrix. Let this matrix be denoted by  $\Gamma$ . The rate  $\gamma(x'_i; x_i, \Psi(x_i))$  is then an entry in this matrix.

The neighborhood component  $F(x_i, \Psi(x_i))$  weights the insertion and deletion rates by the proportion of neighbors present in the network and is given as follows:

$$F(x_i, \Psi(x_i)) = \frac{|\Psi(x_i)|}{\sum_{i \neq j} x_j} \quad (4)$$

The denominator  $\sum_{i \neq j} x_j$  in (4) gives the number of hyperedges present in the current network. A subtlety arises when all the neighbors of a hyperedge are either absent or a hyperedge does not have any neighbors resulting in zero weight. In such situations, the neighborhood component is calculated as  $1/(M + 1)$ . The idea behind this is to assign a very low weight but non-zero to the isolated hyperedges. It must be noted that the neighbor-dependent model described above satisfies the time-reversibility criterion, the proof of which is included in the Supplementary Material (Section S1).

When using the neighbor-dependent model, an important factor affecting the efficacy of the model is the role of current metabolites (see Section 2.2.2). It has been argued in the literature that using current metabolites in an analysis may lead to biologically incorrect results (Ma and Zeng, 2003). In our case, the inclusion of current metabolites in a network affects the neighborhood factor and, therefore, the rate of change of a hyperedge.

Supplementary Table 2 shows the hyperedge in and out-degrees for the networks introduced in Section 2.2.2 with and without current metabolites. The in-degree of a hyperedge  $e$  is defined as the number of hyperedges that share at least one of their products with substrates of the hyperedge  $e$  whereas the out-degree of a hyperedge  $e$  is defined as the number of hyperedges sharing at least one of their substrates with products of the hyperedge  $e$  (Yeung *et al.*, 2007). It can be seen from the table that keeping current metabolites in the network overestimates the hyperedge degrees, and therefore, the neighborhood effect. For example, consider the Lysine degradation networks in *P. aeruginosa* PAO1 and *P. syringae* pv. tomato DC3000. The in and out degrees for the filtered networks are zero suggesting that there are no connected reactions in the pathway map in contrast to the full network approach which results in non-zero degrees for these networks.

## 2.5 Evolutionary path

From Section 2.3 it can be seen that successive networks during the course of evolution differ only by a single edge. At each step, an edge is either inserted or deleted from the network. We refer to the addition or deletion of an edge as an *event* and define a path between two networks as a sequence of events that transform the first network into the second. Formally, for a given pair of networks  $x$  and  $x'$ , which differ by  $d$  number of hyperedges, a *path* of length  $k$  ( $k \geq d$ ) denoted by  $z_k$  is a sequence

$$[e_1, e_2, \dots, e_k]; e_i \in \mathcal{E}'$$

consisting of edge-toggle events that lead the network  $x$  to the network  $x'$ . For example,  $[6, 7]$ ,  $[7, 6]$ ,  $[7, 4, 6, 4]$  and  $[4, 6, 7, 4]$  are some example paths leading the network  $H_1$  to the network  $H_2$  shown in Figure 1.

**2.5.1 Likelihood of a Path** Consider the path shown in Figure 2(a) consisting of two events that lead the network  $H_1$  to  $H_2$  shown in Figure 1 with jump times  $0 \leq t_1 \leq t_2 \leq 1$  visiting  $H'_1$  (Supplementary Figure 1) as the intermediary network. The plus and minus signs with each edge labels in the figure indicate whether the corresponding edge is added or removed from the current network. Let the network obtained after  $i$ -th event be denoted by  $x^i$ , i.e. the starting network is denoted by  $x^0$ , the network  $H'_1$  obtained after first event by  $x^1$  and the network  $H_2$  by  $x^2$ . Setting  $t_0 = 0$  and  $t_3 = 1$ , the likelihood density of the path conditioned on the starting network,  $x_0$ , is given as

$$f([6, 7]|x^0) = e^{-(t_1-t_0)v_0} r(x^0, 6) e^{-(t_2-t_1)v_1} r(x^1, 7) e^{-(t_3-t_2)v_2} \quad (5)$$

where  $r(x^i, e)$  is the rate of transition from the network  $x^i$  by toggling the edge labeled  $e$  and  $v_i$  is the total exit rate from the network  $x^i$ . In general, given a path  $z_k$  of length  $k$  consisting of events that lead the starting network  $x(0)$  to the ending network  $x(1)$ , the likelihood density of the path conditioned on the starting network is given as

$$f(z_k|x(0)) = \exp\left(-\sum_{i=0}^{k-1} (t_{i+1} - t_i)v_i\right) \times \prod_{i=1}^k r(x_{i-1}, z_k(i)) \quad (6)$$

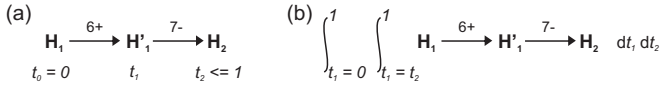
where  $r(x^{i-1}, z_k(i))$  is the rate of transition from the network  $x^{i-1}$  at step  $i-1$  by toggling the edge  $z_k(i)$ ,  $v_i$  is the total exit rate from the network  $x^i$  at step  $i$  and  $t_i$  are the exponentially distributed jump times with  $t_0 = 0$  and  $t_{k+1} = 1$ . Readers interested in details of (5) and (6) are referred to Baier *et al.* (2003) and Miklos *et al.* (2004).

To calculate the likelihood independent of the jump times in (5), the likelihood density  $f([6, 7]|x^0)$  can be integrated over variables  $t_1$  and  $t_2$  (Figure 2(b)) as follows.

$$P([6, 7]|x^0) = \int_{t_1=0}^1 \int_{t_2=t_1}^1 f([6, 7]|x^0) dt_1 dt_2$$

In general, the likelihood value can be obtained by integrating over variables  $t_i$ ,  $1 \leq i \leq k$  such that  $0 \leq t_1 \leq \dots \leq t_k \leq 1$  as follows.

$$P(z_k|x(0)) = \int_{t_1=t_0}^1 \dots \int_{t_k=t_{k-1}}^1 f(z_k|x(0)) dt_1 \dots dt_k \quad (7)$$



**Fig. 2.** An example path consisting of two events that lead the network  $H_1$  to  $H_2$  shown in Figure 1 visiting  $H'_1$  (Supplementary Figure 1) as the intermediary network with plus and minus signs indicating whether the corresponding edge is added or removed from the current network. The paths are shown (a) with jump times  $0 \leq t_1 \leq t_2 \leq 1$  and (b) independent of jump times by integrating over time variables.

Here  $f(z_k|x(0))$  is given by (6). The likelihood in (7) can be calculated using the trajectory likelihood algorithm described by Miklos *et al.* (2004).

## 2.6 Likelihood calculation

As mentioned in Section 2.4.2, when neighbor dependence is taken into account the rate matrix is a  $2^M \times 2^M$  matrix. The dimensions of the rate matrix grow exponentially and, therefore, enumerating the matrix is not possible for large networks. As a result, the exact likelihood of network evolution can not be calculated for large networks. Markov Chain Monte Carlo (MCMC) methods can be used to calculate the likelihood of evolution for given parameter values when enumeration of the rate matrix is not feasible.

Let  $L(\lambda t, \mu t)$  be the likelihood value to be calculated and  $z$  denote a path that a network  $x(0)$  can follow to evolve to network  $x(1)$ . Then we have

$$L(\lambda t, \mu t|x(0)) = \sum_{z:x(0) \rightarrow x(1)} P_{(\lambda t, \mu t)}(z|x(0)) \quad (8)$$

where  $P_{(\lambda t, \mu t)}(z|x(0))$  is calculated using (7). The above equation states that for given parameter values the likelihood of evolution of  $x(0)$  to  $x(1)$  conditioned on  $x(0)$  can be calculated by summing over the probabilities of all possible paths that transform  $x(0)$  into  $x(1)$ . Note that we use parameters  $(\lambda t, \mu t)$  instead of  $(\lambda, \mu)$ , as originally defined in (1). This is because, like DNA sequence evolution, it is impossible to separate the rate parameter from time without having complete knowledge about one or the other.

To calculate the summation in (8) we need a method to move in the space of paths between the two networks. For this we construct a Markov chain where each state is a path between the two given networks,  $x(0)$  and  $x(1)$  and use a Metropolis-Hastings algorithm to explore the state space of paths. Since the Metropolis-Hastings algorithm is a well-established method, it suffices here to give details about how a proposal for new path can be generated from the current path. Readers unfamiliar with MCMC methods are referred to Chapter 1 of Gilks *et al.* (1996). The performance of the algorithm is discussed in the Supplementary Material (Section S7).

## 2.7 Path proposal

A path between two networks is a sequence of events that transform the first network into the second. Since an edge can change multiple times during the course of evolution, a path between two networks, therefore, may contain events corresponding to an edge more than once. However, it is easy to see that if an edge is present an even number of times in a path then its overall effect is cancelled. Edges occurring an even number of times are redundant with respect to the most parsimonious path and can be removed without affecting the final network. An edge that occurs an odd number of times in a path, on the other hand, has all events redundant except one. Moreover, the order of the events does not affect the final network and, therefore, multiple sequences of events from the starting network can lead to the desired network. These facts can be used to define different operations on a path that can be used to generate a path proposal. These operations are described below.

- **Add Events** Given a path  $z_k$ , a new path of length  $k + 2$  can be obtained by adding two events involving an alterable edge at any two positions in the path.
- **Delete Events** Given a path  $z_k$  containing redundant events, a new path of length  $k - 2$  can be obtained by deleting two events involving an edge from the path.
- **Permute Events** Given a path  $z_k$ , a new path of length  $k$  can be obtained by permuting the events in the current path.

Formal definitions of these operations along with examples of path proposal are given in the Supplementary Material (Section S4).

**2.7.1 Proposal Probability** The proposal probability of the new path  $z'_{k'}$  is given as

$$q(z'_{k'}|z_k) \propto p(k'|k) \cdot q(z'_{k'}|z_k, k')$$

where the first term on the right hand side is the probability of proposing a new path length or selecting one of the three operations described above, and the second term is the probability of selecting a new path given the current path and new path length. The details of how these probabilities are calculated are given in the Supplementary Material (Sections S5 and S6).

## 2.8 Parameter estimation

The Metropolis-Hastings algorithm described above calculates the likelihood for given parameter values. This can be extended to estimate the parameters  $(\lambda t, \mu t)$  of evolution. One way is to use a Gibbs sampler by iteratively sampling paths and parameters from the distributions  $P(z|\lambda t, \mu t)$  and  $P(\lambda t, \mu t|z)$  respectively. The general outline of the Gibbs sampler is as follows. Readers are referred to the Chapter 6 of Liu (2001) for general details of a Gibbs sampler.

- Choose initial values for the parameters, i.e.  $\lambda t^{(0)}, \mu t^{(0)}$ .
- Generate  $z^{(0)}$  by sampling from the distribution  $P(z|\lambda t, \mu t)$  using  $\lambda t^{(0)}, \mu t^{(0)}$ .
- Use  $z^{(0)}$  to generate  $\lambda t^{(1)}, \mu t^{(1)}$  by drawing from the distribution  $P(\lambda t, \mu t|z)$ .
- Repeat  $n$  times to get subset of points  $(z^{(i)}, \lambda t^{(i)}, \mu t^{(i)})$ , where  $1 \leq i \leq n$ , are the simulated estimates from the joint distribution  $P(z, \lambda t, \mu t)$ .

A Metropolis-Hastings algorithm for drawing samples from  $P(z|\lambda t, \mu t)$  was presented in Section 2.7 and the samples for parameters can be drawn in a similar fashion as described in the next section. The performance of the sampler is discussed in the Supplementary Material (Section S8). Related work has been done on ordinary graphs in the context of social networks (Koskinen, 2004; Koskinen and Snijders, 2007).

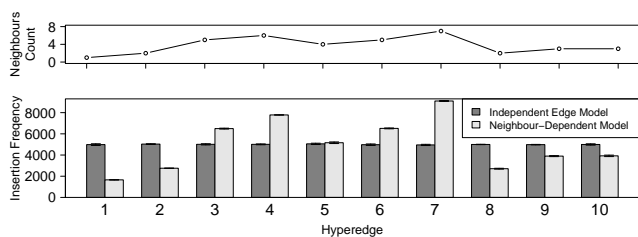
## 2.9 Rates proposal

Given a path and current set of rate parameters, a Metropolis-Hastings algorithm similar to the one described above can be used to sample new rate parameters. As before, we present the method to propose new parameter values. For a given path  $z_k$ , a proposal for the rates can be generated from a gamma distribution as follows.

$$\lambda t' \sim \Gamma\left(n_{ins} + 1, \frac{1}{k} \sum_{i=1}^k \eta_{ins}(x(z_k^i))\right)$$

$$\mu t' \sim \Gamma\left(n_{del} + 1, \frac{1}{k} \sum_{i=1}^k \eta_{del}(x(z_k^i))\right)$$

where  $n_{ins}$  and  $n_{del}$  are the number of insertion and deletion events respectively in the given path, and  $\eta_{ins}(x(z_k^i))$  and  $\eta_{del}(x(z_k^i))$  are the total number of possible insertion and deletion events at  $i$ -th step of the path.



**Fig. 3.** Simulation results for insertion frequencies for the toy network  $H_1$  shown in Figure 1 using independent edge and neighbor-dependent models. Also shown in the top panel are the number of neighbors for each hyperedge based on the reference network.

**2.9.1 Proposal Probability** The proposal probability  $q(\lambda t', \mu t' | \lambda t, \mu t)$  for the parameter vector is given as follows.

$$q(\lambda t', \mu t' | \lambda t, \mu t) = \prod_{\gamma=\lambda t, \mu t} q(\gamma' | \gamma)$$

where

$$q(\gamma' | \gamma) \propto \gamma^\alpha \exp\left(-\frac{\gamma}{k} \sum_{i=1}^k \eta_{\gamma'}(x(z_k(i)))\right).$$

Here  $\alpha$  corresponds to  $n_{ins}$  if  $\gamma = \lambda t$  and  $n_{del}$  otherwise.

## 3 RESULTS

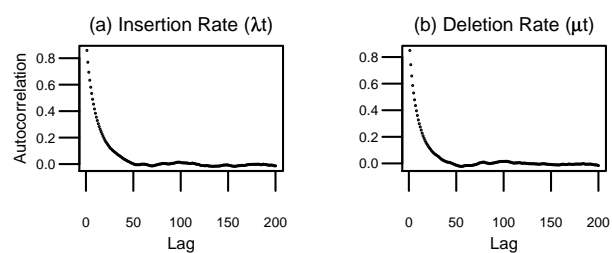
### 3.1 Simulating network evolution under different models

We simulated 100,000 iterations of the toy network  $H_1$  shown in Figure 1 with parameter  $(\lambda, \mu) = (0.05, 0.03)$  to study and compare the models of network evolution. At each step, the insertion and deletion rates were calculated using (2) or (3) depending on the model and were normalized to get probability values. An edge was then selected based on these probability values and was inserted if absent from the current network and deleted otherwise. The insertion frequencies for each hyperedge under the two models are shown in Figure 3 along with the number of neighbors in the top panel. The number of neighbors was calculated based on the reference network. It can be seen that with the neighbor-dependent model, the insertion frequency varies for different hyperedges based on their neighbor counts as compared to the independent edge model where all hyperedges have similar insertion frequencies.

### 3.2 Likelihood calculation for toy networks

We ran the MCMC for sampling paths on the toy networks shown in Figure 1 using neighbor dependence for different parameters and calculated the likelihood conditional on the first network using (8) by summing over all distinct paths. A total of 100,000 iterations were run for each rate combination with additional 10,000 iterations as burn-in period. The exact likelihood (conditional on the first network) by matrix exponentiation was also calculated using all 1024 ( $= 2^{10}$ ) networks. The likelihood values estimated using MCMC were comparable to those obtained using matrix exponentiation. The true and estimated likelihood surfaces for a range of parameter values are shown in Supplementary Figure 2.

We also calculated the path length distribution for different rate combinations averaged over three runs (see Supplementary Figure 3). As expected, when rates were higher the sampler visited



**Fig. 4.** Autocorrelation of parameters estimated using the Gibbs sampler with 11,000 iterations for toy networks shown in Figure 1.

longer paths and explored a larger part of the search space than the case when rates were smaller. We also ran the MCMC for sampling paths on different networks which differed from the starting network  $H_1$  (Figure 1) by 1 to 5 hyperedges to test how well the MCMC estimated likelihood values. The results are presented in the Supplementary Material (Section S7).

### 3.3 Parameter estimation for toy networks

To test the Gibbs sampler for parameter estimation described in Section 2.8, we simulated the network evolution starting from the network  $H_1$  shown in Figure 1 as described in Section 3.1 using fixed parameter values until the system reached stationarity. We then used the sampler to estimate the parameters  $(\lambda t, \mu t)$  of evolution. The details of the test are presented in the Supplementary Material (Section S8 and Supplementary Figure 5).

In order to estimate the parameters of evolution for the toy networks shown in Figure 1, we ran the Gibbs sampler with random starting rates for 10,000 iterations with a further 1,000 iterations as burn-in period. The samples were collected every 10th iteration and the posterior mean of the parameters was calculated. The posterior means of the parameters averaged across three runs was calculated as  $1.116268 \pm 0.776611$  for  $\lambda t$  and  $0.366839 \pm 0.140616$  for  $\mu t$ . The effective sample sizes (ESS) (Liu, 2001 pg 125–127) used for parameter estimation were 423 and 531 for  $\lambda t$  and  $\mu t$  respectively. A sample MCMC trace for the first 1,000 iterations of the sampler for the rate parameters is shown in Supplementary Figure 4. The autocorrelation of parameters is plotted in Figure 4 suggesting an exponential decrease in the correlation as the lag between the samples increases. We also calculated the likelihood of evolution for different rate combinations visited by the sampler using (8). The maximum likelihood averaged over three runs was found to be  $0.020382 \pm 2.4938 \times 10^{-6}$  for parameters  $(\lambda t, \mu t) = (0.991056, 0.537419)$  which is very close to the true likelihood (see Supplementary Figure 2).

### 3.4 Parameter estimation for metabolic networks

We applied the Gibbs sampler on the metabolic networks introduced in Section 2.2.2. Evolution parameters were estimated for evolution between two independent pairs; from *P. fluorescens* Pf-5 to *P. fluorescens* Pf0-1 and from *P. aeruginosa* PAO1 to *P. syringae* pv. tomato DC3000. The phylogenetic relationship between these species is shown in Supplementary Figure 6.

To study the effects of core and prohibited hyperedges in the model (Section 2.3.1), we used the hyperedges corresponding to the reactions that were common to the nine genome sequence strains

**Table 1.** Posterior expectation and standard deviation (STD) of parameter values ( $\lambda t$ : insertion rate and  $\mu t$ : deletion rate) estimated using the Gibbs sampler with and without including core and prohibited hyperedges in the model. The values are averaged over three runs of 100,000 iterations each with additional 10,000 iterations as burn-in period. Samples were collected every 10th iteration. The codes MAPxxxx correspond to the respective KEGG pathway codes.

| Pathway Map                             | Start Organism             | End Organism                | Differences* | CnP**   | $\lambda t$ | STD     | $\mu t$ | STD     | $\lambda/\mu$ |
|---|----------------------------|-----------------------------|--------------|---------|-------------|---------|---------|---------|---------------|
| Pentose phosphate pathway<br>(MAP00030) | <i>P. aeruginosa</i> PAO1  | <i>P. syringae</i> DC3000   | 9 (I:8, D:1) | –       | 0.31694     | 0.00853 | 0.06168 | 0.00656 | 5.13849       |
|   |                            |                             | +            | 1.91912 | 0.26728     | 0.32402 | 0.12122 | 5.92285 |               |
|   | <i>P. fluorescens</i> Pf-5 | <i>P. fluorescens</i> Pf0-1 | 2 (I:2, D:0) | –       | 0.09724     | 0.00176 | 0.01852 | 0.00186 | 5.25147       |
|   |                            |                             | +            | 0.72834 | 0.11021     | 0.08162 | 0.04039 | 8.92367 |               |
| Lysine degradation<br>(MAP00310)        | <i>P. aeruginosa</i> PAO1  | <i>P. syringae</i> DC3000   | 1 (I:1, D:0) | –       | 0.02283     | 0.00058 | 0.06976 | 0.00576 | 0.32732       |
|   |                            |                             | +            | 0.38110 | 0.22654     | 3.25666 | 3.28194 | 0.11702 |               |
|   | <i>P. fluorescens</i> Pf-5 | <i>P. fluorescens</i> Pf0-1 | 2 (I:0, D:2) | –       | 0.00787     | 0.00074 | 0.29881 | 0.00534 | 0.02635       |
|   |                            |                             | +            | 0.12794 | 0.10561     | 1.31184 | 0.44755 | 0.09753 |               |
| Phenylalanine metabolism<br>(MAP00360)  | <i>P. aeruginosa</i> PAO1  | <i>P. syringae</i> DC3000   | 6 (I:4, D:2) | –       | 0.07867     | 0.00759 | 0.48296 | 0.08009 | 0.16289       |
|   |                            |                             | +            | 0.52749 | 0.13457     | 1.95439 | 0.71231 | 0.26990 |               |
|   | <i>P. fluorescens</i> Pf-5 | <i>P. fluorescens</i> Pf0-1 | 7 (I:2, D:5) | –       | 0.04486     | 0.00528 | 0.78573 | 0.06140 | 0.05709       |
|   |                            |                             | +            | 0.34472 | 0.10428     | 2.05643 | 0.46145 | 0.16763 |               |

\* I: number of insertions, and D: number of deletions going from the start organism to the end organism

\*\* CnP: Core and prohibited hyperedges

of *Pseudomonas* (Mithani et al., in preparation) as core edges and the hyperedges corresponding to the reactions not present in any of these nine species as prohibited edges and estimated the parameters for the evolution of metabolic networks with and without core and prohibited edges in the model using the Gibbs sampler. The sampler was run for 100,000 iterations with further 10,000 iterations as burn-in period. The samples were collected every 10th iteration to calculate the posterior expectation. The posterior expectations of the parameters averaged across three runs with random starting values are listed in Table 1 and the effective sample sizes for the parameters are given in Supplementary Table 6.

The parameter values are relatively higher when core and prohibited edges are added in the model than the case when core or prohibited edges are not defined. By defining the core and prohibited edges, the sampler does not spend time in toggling hyperedges that are defined as unlikely to change during the course of evolution and, therefore, only samples the networks that are more likely to be observed. The ratios between the parameter values ( $\lambda/\mu$ ) are generally higher when core and prohibited edges are defined in the model. An exception is the Lysine degradation pathway map where the parameter ratio decreases by adding the core and prohibited edges in the model for the evolution of *P. aeruginosa* PAO1 network to *P. syringae* pv. tomato DC3000. This is due to poor neighborhood effect arising as a result of disconnected hyperedges in these networks (Section 2.4.2).

We also obtained the maximum likelihood values for each pair by using (8) for different parameter combinations visited by the sampler. These values are listed in Supplementary Table 7. Like the parameters, the likelihood values are also higher when core and prohibited edges were defined in the model since the paths between network pairs now consists of those networks which are more likely to be seen during the course of evolution.

Table 1 also shows that the evolution parameters are higher for evolution from *P. aeruginosa* PAO1 to *P. syringae* pv. tomato DC3000 than those from *P. fluorescens* Pf-5 to *P. fluorescens* Pf0-1 for both the cases, i.e. with and without core and prohibited edges,

irrespective of the number of differences between the networks. Supplementary Table 7 shows that the likelihood of evolution from *P. fluorescens* Pf-5 to *P. fluorescens* Pf0-1 is higher than that from *P. aeruginosa* PAO1 to *P. syringae* pv. tomato DC3000 irrespective of the number of differences between the two networks. This suggests that the overall network structure as captured by reaction neighborhood is more similar between *P. fluorescens* Pf-5 and *P. fluorescens* Pf0-1 as compared to the *P. aeruginosa* PAO1 and *P. syringae* pv. tomato DC3000. This was expected since the evolutionary distance between the two strains of *P. fluorescens* is much smaller than the distance between *P. aeruginosa* PAO1 to *P. syringae* pv. tomato DC3000 (see Supplementary Figure 6).

## 4 DISCUSSION

The evolution of metabolic networks can be modeled as a continuous time Markov process. A model based on continuous time Markov chain was presented in this work which allows both insertion and deletion of edges. This is similar to the model presented by Snijders and Van Duijn (1997) for social networks, but allows, in addition, specification of unalterable hyperedges in the form of core and prohibited edges. Incorporating deletion of edges, unlike most of the current models, allows a more realistic model of evolution which is characterized by both gain and loss of enzymes.

A model for insertion and deletion of hyperedges from the network based on neighborhood was also presented. Using a reaction neighborhood while modeling metabolic evolution allows the dependence between reactions to be taken into account, thereby considering network functionality. The simulation results presented above suggest that the neighbor dependent model produces a behavior where highly connected nodes tend to go through a high number of insertions thus agreeing with the common notions of “popularity is attractive” and preferential attachment for biological networks (Dorogovtsev and Mendes, 2003; Light et al., 2005). In a metabolic context this model favors the insertion of edges that complete pathways and connect metabolites to hub metabolites,

and favors deletion of non-core edges that correspond to redundant connections in a network. There are, however, limitations associated with the neighbor-dependent model. In order to obtain useful results with this approach it is necessary to exclude current metabolites such as  $H_2O$ ,  $CO_2$  and ATP, which in turn can lead to exclusion of neighborhood for reactions for which current metabolites are the sole substrate or product (Mithani *et al.*, in preparation). Moreover, the evolution of metabolic networks is complex and involves other factors such as response to environmental, developmental or physiological stimuli as well as lateral gene transfers that are not addressed by this model.

Extensions to the neighbor-dependent model can be made, for example by taking into account reaction structure. A reaction which involves fewer metabolites or is chemically more efficient may be assigned higher weight than a reaction which uses a larger number of metabolites or is difficult to carry out. Note that this requires knowledge of the reaction mechanism and chemical structures of the metabolites involved. In this study we have made the assumption that reversible reactions are lost or added in two independent steps. It would be desirable to allow the deletion or insertion of reversible reactions in a single step with a certain probability as a single enzyme might be catalyzing a reversible reaction in both directions. In addition, ortholog and synteny data could be used to weight each hyperedge. The idea being that if a reaction is present in most of the species that are evolutionarily close to the one being considered then it has higher chances of being added, and if it is genetically linked to other reactions then they have a greater chance of being consecutively added or deleted. Yet another extension would be to include a parameter that corresponds to various environmental stimuli affecting the evolution.

We also presented a Metropolis-Hastings algorithm to move in the space of paths and to calculate the likelihood of evolution by summing over the paths between two networks. We demonstrated the method on toy networks where exact calculations were possible. The likelihood calculations presented in this text were conditioned on the starting network. The full likelihood of evolution is given as  $P(H_1)P(H_2|H_1)$ , i.e. the likelihood of observing the starting network  $H_1$  times the likelihood of the evolution of  $H_1$  into  $H_2$ . Assuming that evolution has been going on for a long time we can use the equilibrium probability  $\pi(H_1)$  to approximate  $P(H_1)$ . This requires solving the equation  $\pi Q = 0$ . However, as discussed earlier, enumerating the rate matrix  $Q$  is not feasible even for medium sized networks. We have developed a method to approximate the equilibrium probability by dividing the network into smaller components and solving the equation  $\pi Q = 0$  for each component. This is explained in the Supplementary Material (Section S3).

A Gibbs sampler to estimate the evolution parameters was also presented. When estimating the parameters, a uniform prior was used which assigns equal probability to each point in the parameter space. The elicitation of prior distributions by taking into account the relationship between parameters is an important area for further studies. It might be useful, for example, to consider the dependence between the number of insertions and deletions in a path to investigate other prior distributions.

Using the Gibbs sampler, we estimated the parameters of evolution between different pairs of bacteria belonging to the genus *Pseudomonas* based on neighbor dependence. The results

suggest that the network structure can be used to provide clues about the metabolic similarity of bacteria. The results show that network structure is more conserved in the closely related strains as compared to the distantly related strains. Furthermore, defining core and prohibited hyperedges in the model not only improves the likelihood but also allows biologically meaningful results by focusing on regions of the network which are under selection.

A logical extension to the work presented here would be to calculate the likelihood and estimate the parameters over a phylogeny, which is known through sequence analysis. Calculating the likelihood over a phylogeny requires a sum, over all possible networks that may have existed at the interior nodes of the tree, of the probabilities of each scenario of events. This is similar to the idea introduced by Felsenstein (1981) for observing DNA sequences over a phylogeny. Using phylogenetic information in conjunction with the ortholog data may also be useful for the inference of evolutionary histories over a phylogeny.

## ACKNOWLEDGEMENT

This work was supported by the grants from the Higher Education Commission, Government of Pakistan (A.M.), The Royal Society (G.P.) and the Biotechnology and Biological Sciences Research Council grant BB/E007872/1 (G.P.).

## REFERENCES

- Baier, C. *et al.* (2003) Model-checking algorithms for continuous-time Markov chains, *IEEE Trans. Software Eng.*, **29**, 524–541.
- Berg, J. *et al.* (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications, *BMC Evol. Biol.*, **4**, 51.
- Boccaletti, S. *et al.* (2006) Complex networks: Structure and dynamics, *Phys. Rep.*, **424**, 175–308.
- Chung, F. *et al.* (2003) Duplication models for biological networks, *J. Comp. Biol.*, **10**, 677–687.
- Dorogovtsev, S. and Mendes, J. (2003) *Evolution of networks: From biological nets to the Internet and WWW*, Oxford University Press.
- Felsenstein, J. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach, *J. Mol. Evol.*, **17**, 368–376.
- Gilks, W. *et al.* (1996) Markov chain Monte Carlo in practice, Chapman & Hall/CRC.
- Jensen, J. and Pedersen, A. (2000) Probabilistic models of DNA sequence evolution with context dependent rates of substitution, *Adv. Appl. Probab.*, **32**, 499–517.
- Kanehisa, M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Res.*, **34**, D354–357.
- Koskinen, J. (2004) Bayesian inference for longitudinal social networks. Research Report, number 2004:4, Stockholm University, Department of Statistics.
- Koskinen, J. and Snijders, T. (2007) Bayesian inference for dynamic social network data, *J. Statist. Plann. Inference*, **137**, 3930–3938.
- Light, S. *et al.* (2005) Preferential attachment in the evolution of metabolic networks, *BMC Genomics*, **6**, 159.
- Liu, J. (2001) Monte Carlo strategies in scientific computing, Springer.
- Ma, H. and Zeng, A. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms, *Bioinformatics*, **19**, 270–277.
- Miklos, I. *et al.* (2004) A “long indel” model for evolutionary sequence alignment, *Mol. Biol. Evol.*, **21**, 529–540.
- Ross, S. (2007) Introduction to probability models, Academic press.
- Snijders, T. (2005) Models for longitudinal network data. Chapter 11 in P. Carrington, J. Scott, and S. Wasserman (Eds.), *Models and methods in social network analysis*, New York: Cambridge University Press.
- Snijders, T. and Van Duijn, M. (1997) Simulation for statistical inference in dynamic network models, *Simulating Social Phenomena*, 493–512.
- Ueda, H. and Hogenesch, J. (2005) Principles in the evolution of metabolic networks, *Arxiv preprint q-bio.MN/0503038*.
- Wiuf, C. *et al.* (2006) A likelihood approach to analysis of network data, *Proc. Nat. Acad. Sci. U.S.A.*, **103**, 7566–7570.
- Yeung, M. *et al.* (2007) Estimation of the number of extreme pathways for metabolic networks., *BMC Bioinformatics*, **8**, 363.