

# A stochastic model of metabolic network evolution

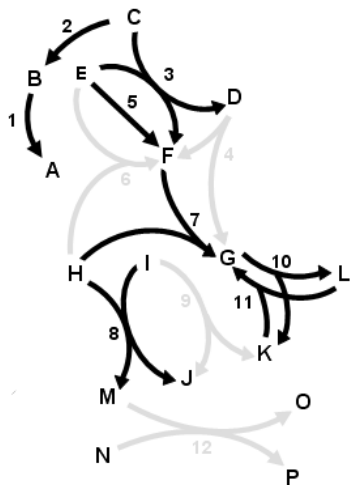
Aziz Mithani<sup>†</sup>, Gail Preston<sup>‡</sup> and Jotun Hein<sup>†</sup>  
[mithani@stats.ox.ac.uk](mailto:mithani@stats.ox.ac.uk)

<sup>†</sup>Department of Statistics and <sup>‡</sup>Department of Plant Sciences,  
University of Oxford, South Parks Road, Oxford, UK

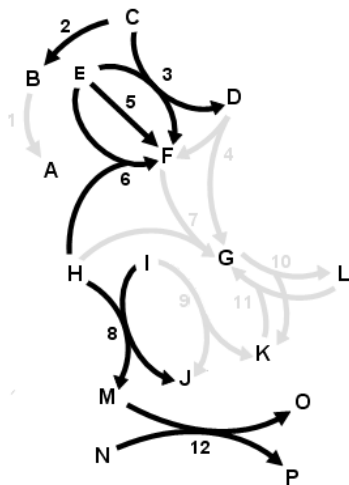
Network Biology 2008  
28 August 2008



# The Problem



Start Network ( $H_1$ )



End Network ( $H_2$ )

# Contents

- 1 Representation of networks as hypergraphs
- 2 Network evolution as a continuous time Markov process
- 3 Models of network evolution
- 4 Likelihood of network evolution
- 5 MCMC for sampling paths
- 6 Results
- 7 Summary

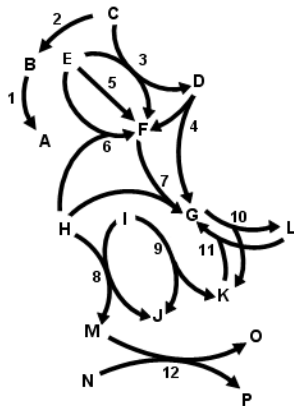
# Representation of networks as directed hypergraphs

## Components

- Nodes - sets of metabolites
- Edges - Reactions / Enzymes

## Advantages

- Biologically relevant - set of metabolites connected by a reaction
- Captures the relationship between multiple metabolites in a reaction
- Allows to think in terms of gain / loss of reactions



# Network evolution as a continuous time Markov process

**State:** The set  $G \subseteq \mathcal{E}$  of hyperedges present in the network

**Next State:** Characterised by the insertion of one edge or the deletion of one edge

**System Dynamics:** Described by following master equation

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

where

- $I(G)$ : set of networks reachable by insertion of a single hyperedge
- $D(G)$ : set of networks reachable by deletion of a single hyperedge
- $\lambda$ : rate of insertion of a hyperedge
- $\mu$ : rate of deletion of a hyperedge

## Models

- Independent Edge
  - Each edge changes independently.
- Neighbour-Dependent
  - Edge changes depending on the proportion of neighbours present or absent.

## Independent Edge Model

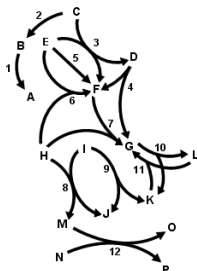
- State space size = 2 (an edge is either absent or present)
- Markov chain described by a  $2 \times 2$  rate matrix  $Q$  given as

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \quad (1)$$

where  $\lambda$  is the insertion rate and  $\mu$  is the deletion rate.

# Neighbour-Dependent Model

**Neighbours:** Hyperedges sharing at least one node.



## Motivation

- Preferential attachment in metabolic networks (Light *et al.*, 2005)
- Popularity is attractive (Dorogovtsev and Mendes, 2003)
- Network tends to remain connected - metabolic pathways

## The Model

- State Space Size:  $2^m$
- The rate from network  $x$  to  $x'$  depends on  $x_i$ ,  $x'_i$  and the neighbouring hyperedges  $\Psi(x_i)$ , and is given as follows.

$$\gamma(x'_i; x_i, \Psi(x_i)) = q(x_i, x'_i)F(x_i, \Psi(x_i)) \quad (2)$$

where

- $q(x_i, x'_i)$ : rate from the independent edge model
- $F(x_i, \Psi(x_i))$ : neighbourhood component

# Neighbour-Dependent Model (3)

## Neighbourhood Component

$$F(x_i, \Psi(x_i)) = \begin{cases} \frac{|\Psi^+(x_i)|}{\sum_{i \neq j} x_j} & \text{insertion} \\ \frac{|\Psi^-(x_i)|}{\sum_{i \neq j} (1-x_j)} & \text{deletion} \end{cases} \quad (3)$$

## Example

$$F(x_7, \Psi^+(x_7)) = \frac{5}{7} = 0.714$$

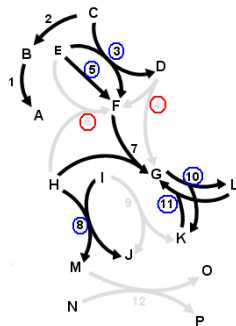
$$F(x_7, \Psi^-(x_7)) = \frac{2}{4} = 0.500$$

Let  $(\lambda, \mu) = (0.05, 0.03)$ , then

$$Q = \begin{bmatrix} -0.05 & 0.05 \\ 0.03 & -0.03 \end{bmatrix}$$

and

$$\gamma(x'_7; x_7, \Psi(x_7)) = 0.03 \times 0.5 = 0.015$$

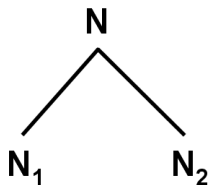


Start Network ( $H_1$ )

# Likelihood of network evolution

- The likelihood of a pair of networks,  $N_1$  and  $N_2$ , separated from a common ancestor  $N$  by divergence time  $t$  is:

$$\begin{aligned} P_t(N_1, N_2) &= \sum_N P_\infty(N) P_t(N_1|N) P_t(N_2|N) \\ &= P_\infty(N_1) P_{2t}(N_2|N_1) \quad (\text{assuming reversibility}) \end{aligned}$$



- The probability  $P_t(N_2|N_1)$  is calculated from the transition matrix  $P(t)$  given as

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

where  $Q$  is the rate matrix.

# Transition Probability $P_t(N_2|N_1)$

## Problem

- Enumeration of the rate matrix  $Q$  not feasible for larger systems
- Matrix exponentiation computationally too expensive

## Possible Solutions

- Corner cutting - **still not feasible**
- Markov Chain Monte Carlo (MCMC) method

## Estimating $P_t(N_2|N_1)$ using MCMC

$$P_t(N_2|N_1) = \sum_{z:N_1 \rightarrow N_2} P_t(z)$$

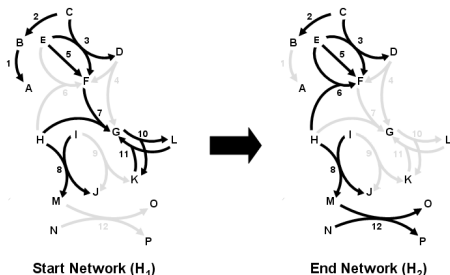
where  $z$  represents a path from  $N_1$  to  $N_2$

# MCMC for sampling paths

**Path:** Sequence of events transforming  $H_1$  into  $H_2$

## Example Paths

- ① 6, 12, 1, 7, 10, 11
- ② 12, 7, 10, 6, 11, 1
- ③ 12, 7, 4, 10, 6, 11, 4, 1



## Path Proposal

- Add Events

12, 7, 4, 7, 10, 7, 6, 11, 4, 1  $(k' = k + 2)$

- Delete Events

12, 7, 4, 10, 6, 11, 4, 1  $(k' = k - 2)$

- Permute Events

4, 7, 12, 6, 11, 1, 4, 10  $(k' = k)$

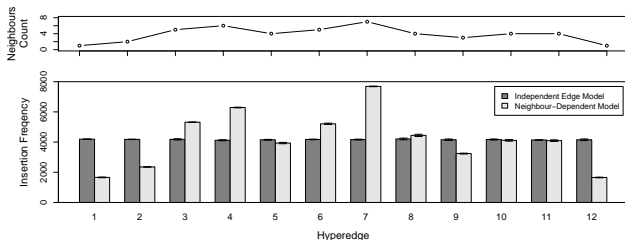
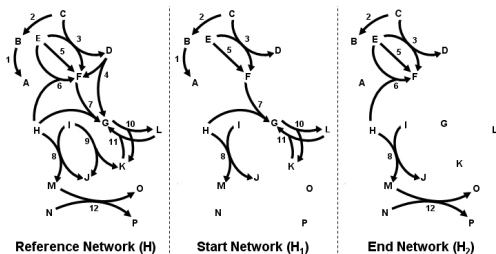
## Proposal Probability

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

where

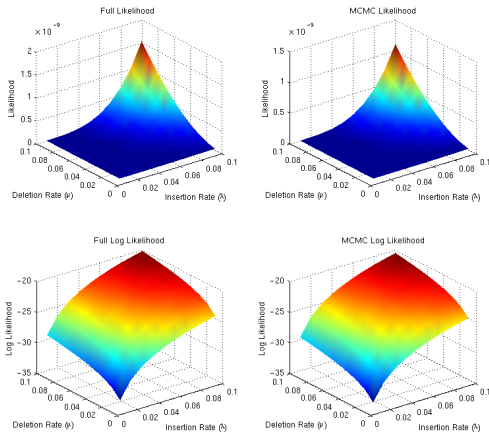
- $p(k' | k)$  is the probability of proposing new path length
- $q(z_{k'}' | z_k, k')$  is the probability of a new path length given the current path and new path length

# Results: Simulation of Network Evolution Models



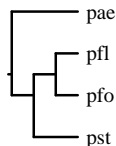
**Figure:** Insertion Frequency for toy network  $H_1$ . Top panel shows the number of neighbours for each hyperedge based on reference network  $H$ .

# Results: Likelihood Calculation for Toy Networks



**Figure:** Likelihood and log likelihood surfaces calculated by matrix exponentiation and by using MCMC for toy networks

# Results: Likelihood Calculation for Metabolic Networks

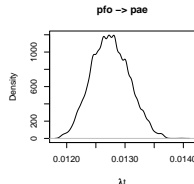


(*P. aeruginosa* PAO1)

(*P. fluorescens* Pf-5)

(*P. fluorescens* PfO-1)

(*P. syringae* pv. tomato DC3000)



Network	Org.	Differences	With Core and Prohibited Edges		
			Likelihood	$\lambda t$	$\mu t$
Lysine Degradation	pae	2 (D:2, I:0)	2.228E-03 ( $\pm 0.000E+00$ )	0.04361	0.92575
	pfl	2 (D:0, I:2)	3.394E-04 ( $\pm 1.263E-09$ )	0.19059	0.10928
	pst	3 (D:2, I:1)	8.865E-07 ( $\pm 6.759E-12$ )	0.06178	0.67177
Methionine Metabolism	pae	9 (D:6, I:3)	1.384E-18 ( $\pm 4.650E-21$ )	0.03656	0.08686
	pfl	4 (D:4, I:0)	3.143E-07 ( $\pm 2.226E-12$ )	0.01167	0.13972
	pst	10 (D:9, I:1)	9.914E-18 ( $\pm 4.019E-20$ )	0.01098	0.14392
Selenoamino acid Metabolism	pae	6 (D:3, I:3)	5.515E-10 ( $\pm 1.186E-16$ )	0.16541	0.08817
	pfl	3 (D:1, I:2)	8.756E-07 ( $\pm 1.239E-10$ )	0.20647	0.08133
	pst	5 (D:4, I:1)	7.465E-09 ( $\pm 5.177E-15$ )	0.04236	0.19960

I: number of insertions, D: number of deletions

## Summary

- A stochastic model of network evolution using neighbour dependence
  - hypergraph based - intuitive
  - biologically relevant
- Markov Chain Monte Carlo method for sampling path

## Further Work

- Calculation of equilibrium probabilities  $P_\infty(N_1)$ 
  - $P_\infty(N_1)P_t(N_2|N_1)$
- Extensions to the model
  - Reaction structure
    - number of metabolites
    - chemical efficiency
  - Ortholog data
    - status of the reaction in closely related species
- Likelihood calculation for phylogeny of networks

# Acknowledgements



**Higher Education Commission,  
Government of Pakistan**



**Department of Statistics,  
University of Oxford**

wellcome trust



***Cold Spring Harbor Laboratory/  
Wellcome Trust***



**St. Anne's College,  
University of Oxford**

# Rahnuma: Hypergraph based tool for metabolic pathway prediction and network comparison

<http://portal.stats.ox.ac.uk:8080/rahnuma>

