

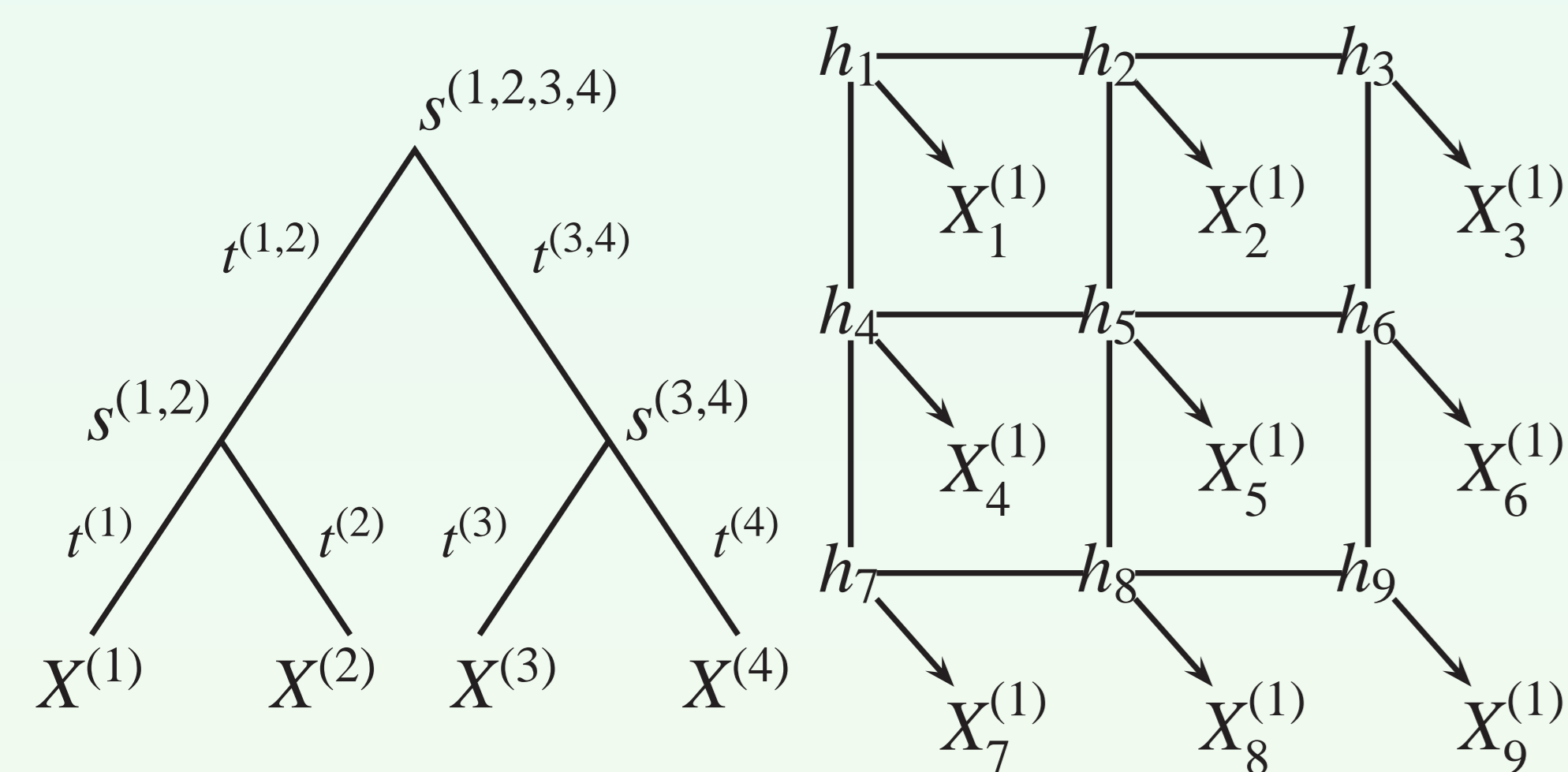
## Abstract

With ever-increasing computational power, models of biological networks have been extensively developed and studied in recent years. In this project, we focus on the evolution of metabolic networks, comprising the sets of highly regulated and correlated reactions that take place between metabolites within a cell.

We describe this evolution as a continuous time discrete space Markov process on a phylogenetic tree, where reactions can flip between being present or absent. The aim of this work was to explore whether there exists a canonical way of classing the reactions within a metabolic network in terms of the rates at which reactions are added and removed, by simultaneously inferring the parameters, and classifying the reactions into different groups.

Previous work has been carried out to perform inference of metabolic networks evolving on a phylogeny, incorporating dependencies between reactions if they share reactants or products [3], but this approach becomes infeasible as the number of metabolisms increases.

## Introduction



Each variable  $X^{(i)}$  represents the metabolism of a particular organism, related by the phylogenetic tree on the left. In order to make summing over ancestral metabolisms feasible for the purposes of calculating the likelihood, we instead consider  $X_k^{(i)}$  for each reaction  $k$  as independent conditional on a hidden variable  $h_k$ . The dependencies between the reactions are then modelled through a hidden Markov random field (HMRF) on  $h$ .

## Model and Methods

We simulated the evolution of three metabolic pathways with the following model:

$\lambda, \mu$ : Gamma random variables that define rates of addition and deletion of reactions in the network. These define a rate matrix  $Q$  describing a CTMC with "present" and "absent" states.

$P(t) = e^{tQ}$ : the matrix of transition probabilities after time  $t$ . We parametrize it with a Gamma random variable  $r = \lambda + \mu$  and a Beta random variable  $p = \lambda/(\lambda + \mu)$

$$P(t) = \begin{pmatrix} 1 - p(1 - e^{-tr}) & p(1 - e^{-tr}) \\ (1 - p)(1 - e^{-tr}) & p + (1 - p)e^{-tr} \end{pmatrix}$$

**Potts model:** Given a set of data  $\mathcal{X}$ , a subset  $\mathcal{X}_i \subset \mathcal{X}$  corresponding to the observations for reaction  $i$ , and a set of hidden nodes  $\{h_1, h_2, \dots, h_n\}$ , then the posterior probability of the hidden state  $i$  conditional on each of its neighbours  $j \sim i$  is written as

$$p(h_i | h_{j \sim i}, \mathcal{X}, \beta) \propto p(\mathcal{X}_i | h_i) \exp \left\{ \beta \sum_{j \sim i} J(h_i, h_j) \right\}$$

where  $J(h_i, h_j)$  is 1 if  $h_i = h_j$  and 0 otherwise.

$\beta$ : neighbourhood interaction strength of reactions. It determines probability of changing states.

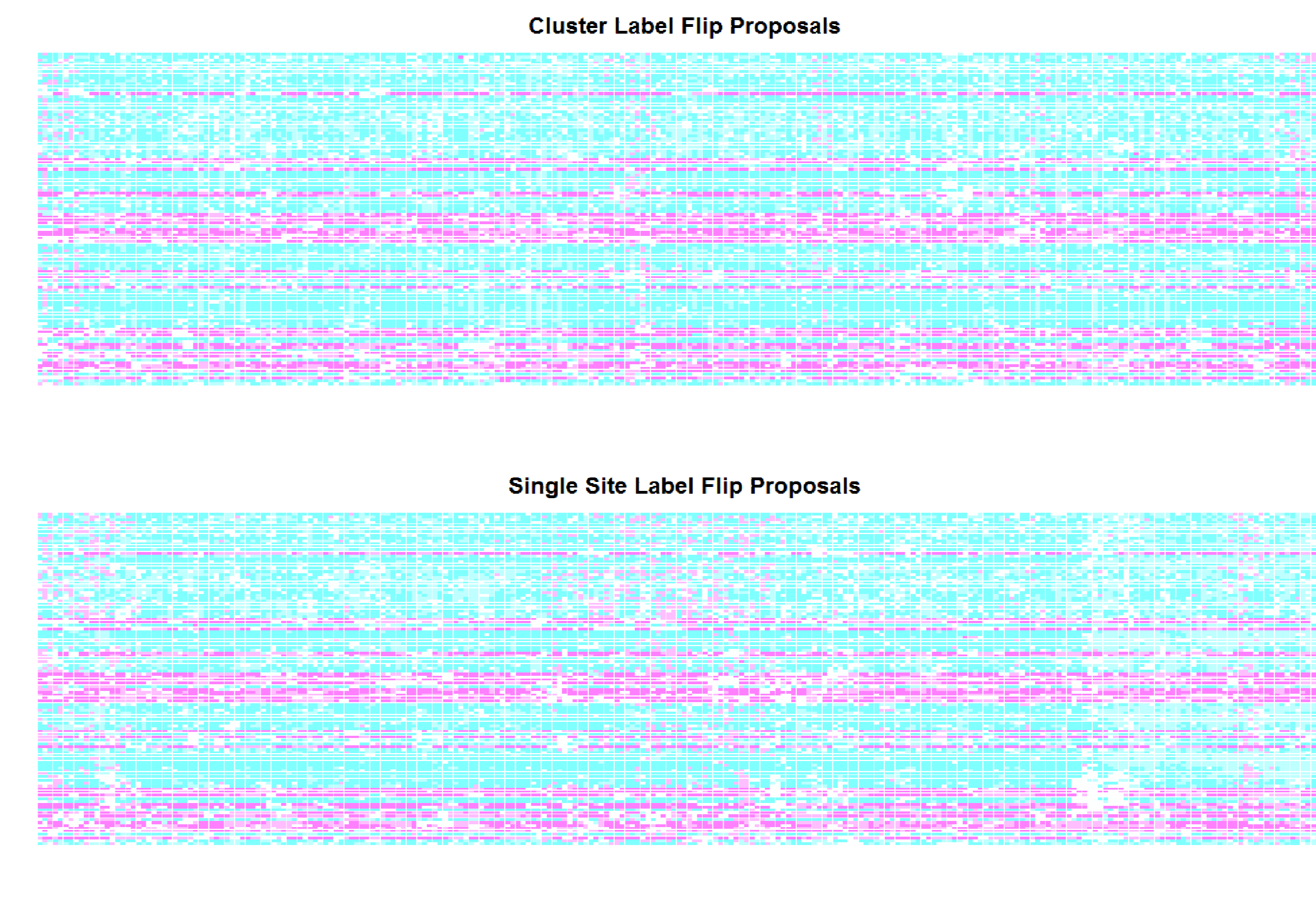
$K$ : the number of possible values that can be taken by each hidden state.

## Implementation

The simulation consists of alternatively sampling the hidden states  $h$ , and the model parameters. We developed Markov chain Monte Carlo (MCMC) methods in order to achieve this, and ran the model on a highly reliable Archaea phylogeny on three subnetworks of the whole metabolome taken from the KEGG database [2].

For different networks we observe similar convergence behaviour for different values of  $\beta$ . Interestingly, for the different networks the inference about the rate parameter is very similar (around one insertion/deletion event for an evolutionary distance of 0.07 mutations per site), suggesting that reactions in different parts of the metabolome are evolving in a similar fashion.

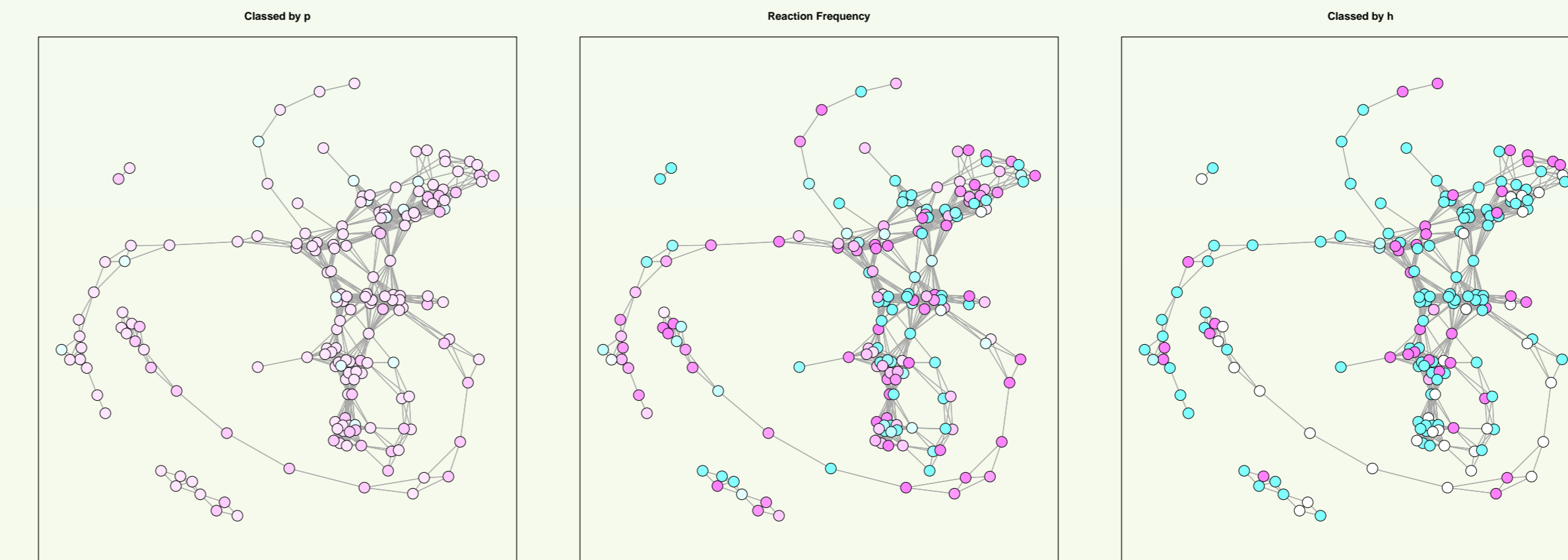
## Results



We ran the simulation with many parameter combinations. In all cases we can see a limited number of possible combinations of the rate  $r$  and proportion of additions  $p$ .

- **A:** high  $r$  and high  $p$ ; reactions that are gained easily.
- **B:** high  $r$  and low  $p$ ; reactions that evolve and are lost easily.
- **C:** Low  $r$  and high  $p$
- **D:** Low  $r$  and medium  $p$ ; reactions that are conserved or absent in most or all organisms.

We annotated the reactions nodes according to which class they fall into



These annotations were mapped back to the KEGG network, allowing us to investigate the biological significance of the four observed classes of reactions (**A**, **B**, **C** and **D**).

## References

- [1] J Felsenstein and G A Churchill. (1996) A hidden markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13(1):93?104, 1996.
- [2] M. Kanehisa, et al. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34:D354?D357.
- [3] A. Mithani, G.M. Preston, and J. Hein. (2010) A bayesian approach to the evolution of metabolic networks on a phylogeny. *PLoS Comput Biol*, 6(8):e1000868.

## Acknowledgements

This work was carried out as part of the Oxford Summer School in Computational Biology, 2011, in conjunction with the Department of Plant Sciences, and with support from the Department of Zoology. Funding was provided by EPSRC, BBSRC, and the EU COGANGS project. We thank Dr Steven Kelly for providing computational resources, and Aziz Mithani for helpful discussions.

