

Understanding the evolution of protein interaction networks.

John W. Pinney*¹, Magnus Rattray² & David L. Robertson¹

¹ Faculty of Life Sciences, University of Manchester

² School of Computer Science, University of Manchester

1 Introduction

As whole-genome protein interaction network datasets become available for a wide range of species, evolutionary biologists have the opportunity to address some of the unanswered questions surrounding the evolution of these complex systems (Sharan and Ideker, 2006). Given multiple observed protein interaction networks from divergent organisms, we can reconcile these data to investigate how gene duplication, deletion and ‘re-wiring’ processes may have shaped their evolution to their contemporary forms.

We are currently investigating how probabilistic modeling using Bayesian approaches can provide a platform for the quantitative analysis of multiple protein interaction networks, including the reconstruction of ancestral networks for families of transcription factors (Amoutzias *et al.*, 2004) and quality control for noisy high-throughput datasets.

2 Methods

Starting with a phylogeny for a gene family and a separate species tree, a reconciled tree may be derived using the NOTUNG 2.0 program (Durand *et al.*, 2006). The reconciled tree shows speciation, gene duplication and gene loss events during the evolution of this family (Figure 1). Although many such reconciled trees may be consistent with the original phylogeny, NOTUNG 2.0 uses a parsimony approach to produce trees that minimize the cost of gene duplication and loss, and can hence be considered to be the most likely explanations for the data.

Using a reconciled tree and a protein-protein interaction dataset for each extant species as input, our software constructs a Bayesian network model for the evolutionary history of the protein interactions (Figure 2). The reconciled tree specifies the protein complement at each common ancestor, but the ordering of gene duplication and loss events between two species nodes is in general unknown and must be represented by multiple paths in the model.

The protein-protein interaction networks of the ancestral species may be estimated from the model using a Markov Chain Monte Carlo (MCMC) approach (Gilks *et al.*, 1996). Our method assumes that gene sequence mutations are much more likely to result in the loss of protein-protein interactions than in the creation of novel interactions.

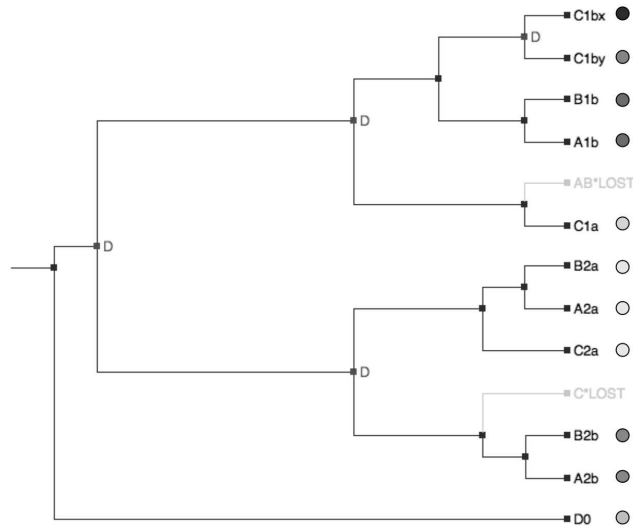


Figure 1: A reconciled gene family phylogeny created using NOTUNG 2.0 (Durand *et al.*, 2006). Unmarked internal nodes show speciation events, nodes marked **D** show gene duplications and light grey terminal nodes represent gene loss events.

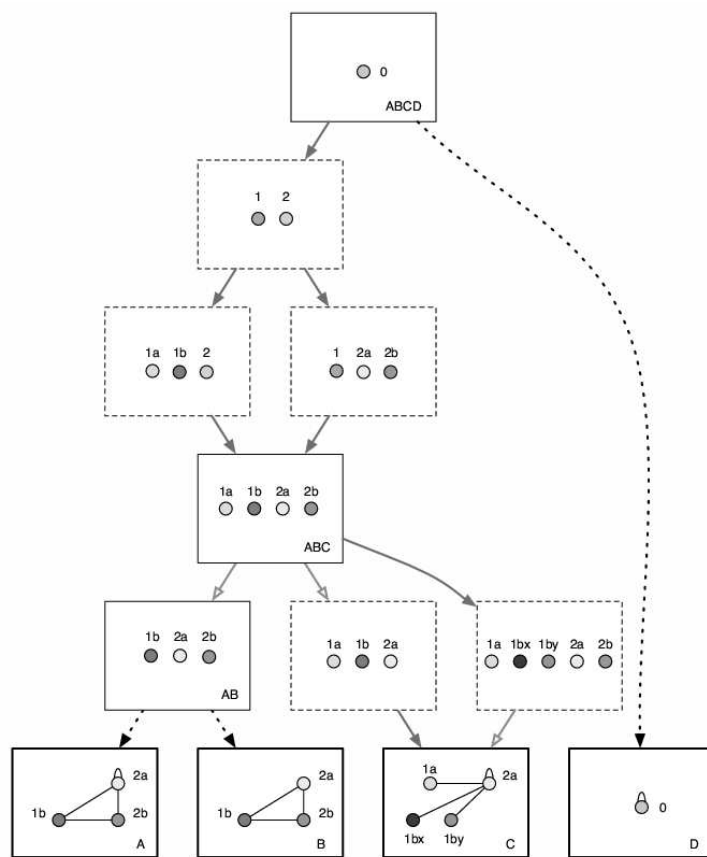


Figure 2: A Bayesian network model constructed from the reconciled tree in Figure 1. Boxes with thick solid borders represent extant species, thin solid borders, last common ancestors of those species and dashed borders, possible intermediate species. Solid arrows with filled heads show gene duplication events and those with unfilled heads, gene loss events. Evolution with no change in protein complement is shown by dotted arrows. Observed protein-protein interaction data is shown for extant species.

3 Results

Results are shown in Figure 3 for our illustrative dataset. The MCMC simulation was run for 10,000 iterations, with parameters $P(\text{loss of protein interaction}) = 0.1$ and $P(\text{gain of protein interaction}) = 0.01$. For each potential protein-protein interaction within an ancestral species, the program estimates the probability that it was present at that point in evolution.

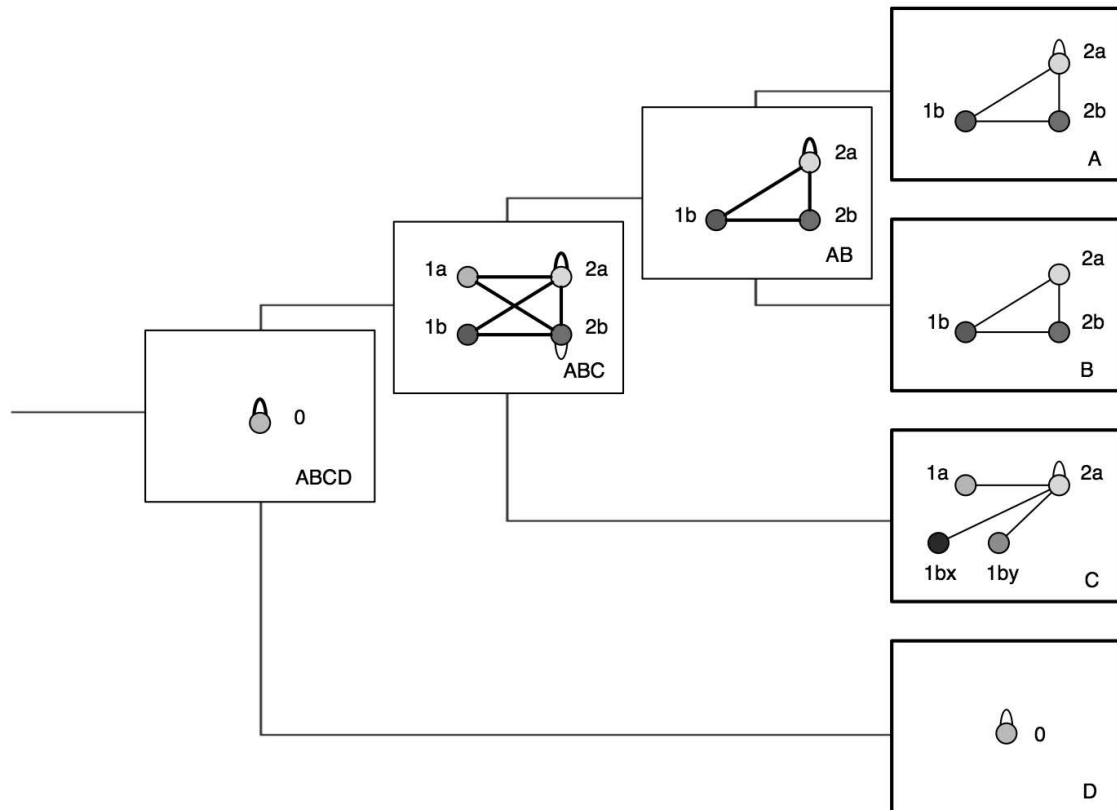


Figure 3: Results from running our MCMC simulation for 10,000 iterations on the model shown in Figure 2. The weights of the edges in the ancestral species \tilde{O} networks represent the probability of their existence as estimated by the simulation. Bold edges show $P > 0.75$ and normal lines, $P > 0.5$. Note that the ultimate ancestor of this family (ABCD0) is overwhelmingly likely to have been a self-interacting protein.

4 Discussion

This Bayesian approach to reconciling protein interaction networks with gene phylogenies has great potential as a quantitative tool for the study of how specific gene families have evolved.

In addition, the same type of model may be used to infer genuine protein-protein interactions from noisy high-throughput datasets for multiple divergent species, given estimates of false-negative and false-positive rates for each experimental protocol used.

References

- Amoutzias, G.D., Robertson, D.L., Oliver, S.G. and Bornberg-Bauer, E. (2004). Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Rep*, **5**, 274-279.
- Durand, D., Halldorsson, B.V. and Vernet, B. (2006). A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol*, **13**, 320-335.
- Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (eds). (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
- Sharan, R. and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, **24**, 427-433.