

Phylogenetic Factorial Hidden Markov Models

Dirk Husmeier¹

Keywords: phylogenetics, rate heterogeneity, interspecific recombination, hidden Markov models, Markov chain Monte Carlo.

Introduction

A recently proposed method for detecting recombination in DNA sequence alignments is based on the combination of hidden Markov models (HMMs) with phylogenetic trees (1). The idea is to introduce a hidden state that represents the tree topology at a given site. A state transition from one topology into another corresponds to a recombination event. To introduce correlations between adjacent sites, the hidden states are given a Markovian dependence structure. Thus, the standard model of a phylogenetic tree is generalized by the combination of two probabilistic models: (1) a taxon graph (phylogenetic tree) representing the relationships among the taxa, and (2) a site graph (HMM) representing dependencies between different sites in the DNA sequence alignments. Breakpoints of mosaic segments in the alignment are predicted by state transitions in the site graph. Inference is done in a Bayesian way by sampling the model parameters and hidden state sequences from the posterior distribution with Markov chain Monte Carlo (MCMC). While this method was found to detect breakpoints of recombinant regions more accurately than most existing techniques, it inherently fails to distinguish between recombination and rate variation. Hence, genomic regions under different selective pressure tend to be erroneously predicted as recombination events.

Method

To distinguish between recombination and rate heterogeneity, we marry the phylogenetic tree to a factorial HMM. The states of the first hidden chain represent tree topologies, as before, and transitions between these states are indicative of recombination events. The states of the second independent hidden chain represent different global scaling factors of the weights, and transitions between these rate states indicate variations in the selective pressure. Inference is done in terms of a hierarchical Bayesian model. Parameters are divided into groups, and parameter groups are sampled from the posterior distribution with Gibbs sampling, that is, one group is sampled conditional on fixed settings of the other groups; see (1) for details. The model is illustrated in Figure 1.

Results

We chose a subset of the 787-nucleotide *Neisseria argF* DNA multiple alignment studied in (2), where we selected the four strains *N.gonorrhoeae* (X64860), *N.meningitidis* (X64866), *N.cinera* (X64869), and *N.mucosa* (X64873) (GenBank/EMBL accession numbers are in brackets). Zhou et al. (2) found two anomalous, or more diverged regions in the DNA sequence alignment, which occur at positions $t = 1 - 202$ and $t = 507 - 538$. In the rest of the alignment, *N.meningitidis* clusters with *N.gonorrhoeae*, while between sites $t = 1$ and $t = 202$, they found that it is grouped with *N.cinera*. Zhou et al. (2) suggested that the region $t = 507 - 538$ is the result of rate variation. The phylogenetic HMM proposed in (1) predicted the first recombination event correctly. However, the differently diverged region was erroneously predicted as a recombinant region. Figure 1 shows the results obtained

¹Biomathematics & Statistics Scotland, JCMB, The King's Buildings. Edinburgh EH9 3JZ, UK.
E-mail: dirk@bioass.ac.uk

with the proposed phylogenetic factorial HMM. The model clearly distinguishes between recombination and rate heterogeneity and avoids the prediction of spurious recombinant regions.

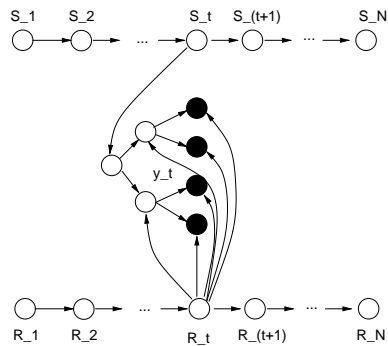


Figure 1: Phylogenetic factorial HMM, drawn as a Bayesian network. Black nodes represent observations; these are columns of nucleotides in the DNA sequence alignment. Empty circles represent hidden nodes, which can be separated into three classes: (1) ancestral nodes in the phylogenetic tree; (2) nodes of the first (top) hidden Markov chain, which represent different tree topologies; and (3) nodes of the second (bottom) hidden Markov chain, which are associated with different global scaling factors and represent different selective pressures.

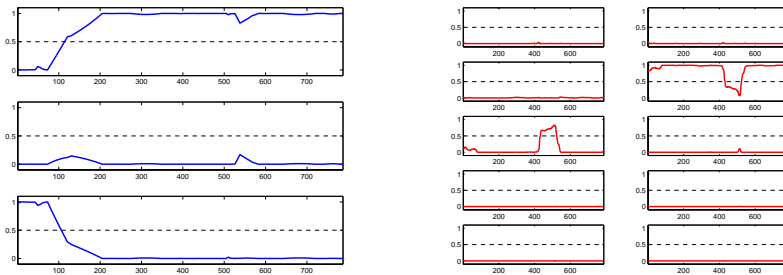


Figure 2: Left panel: Posterior probability of topology states. The three subfigures show the predicted posterior probabilities of the three topologies, plotted against the site t in the DNA sequence alignment. Right panel: Posterior probability of rate states. Ten rate states were used. The associated rate factors (increasing from top to bottom and from left to right) were varied between 0.001 and 100 with an approximately uniform spacing on a log scale.

References

- [1] Husmeier, D. and McGuire, G. (2003) Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution*, **20**, 315–337.
- [2] Zhou, J. and Spratt, B. G. (1992) Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Molecular Microbiology*, **6**, 2135–2146.