Pruned PDM Method for Detecting Recombination

Dirk Husmeier [1]

Keywords: phylogenetics, interspecific recombination, sliding window methods, Markov chain Monte Carlo, probabilistic divergence measure.

1 Introduction

The underlying assumption of most phylogenetic tree reconstruction methods is that there is one set of hierarchical relationships among the taxa. While this is a reasonable approach when applied to most DNA sequence alignments, it can be violated in certain bacteria and viruses due to interspecific recombination. The resulting transfer or exchange of DNA subsequences can lead to a change of the branching order (topology) in the affected region, which results in conflicting phylogenetic information from different regions of the alignment. If undetected, the presence of these so-called mosaic sequences can lead to systematic errors in phylogenetic tree estimation.

The idea of a recently proposed method for detecting evidence of recombination in DNA sequence alignments is illustrated in the left panel of Figure 1. Consider a given alignment of DNA sequences, $\mathcal{D}$, from which we select a consecutive subset $\mathcal{D}_t$ of predefined width W, centred on the tth site of the alignment. Let S be an integer label for tree topologies, and consider the marginal posterior probability of tree topologies S conditional on the 'window' $\mathcal{D}_t$, $P(S|\mathcal{D}_t)$, which is numerically computed with Markov chain Monte Carlo by marginalizing over the branch lengths of the phylogenetic tree and the parameters of the nucleotide substitution model. The basic idea of the probabilistic divergence method (PDM) for detecting recombinant regions is to move the window $\mathcal{D}_t$ along the alignment and to monitor the distribution $P(S|\mathcal{D}_t)$. We would then, obviously, expect a substantial change in the shape of this distribution as we move the window into a recombinant region. To quantify the degree of change, a probabilistic divergence measure is computed, as discussed in (2).

A shortcoming of the PDM method is that its performance deteriorates as the number of taxa increases. This is because for an increased number of taxa the posterior distribution over tree topologies, $P(S|\mathcal{D}_t)$, becomes more diffuse unless the size of the data set $\mathcal{D}_t$ is increased. An increased amount of data $\mathcal{D}_t$, however, corresponds to an increased length of the sliding window, which compromises the spatial resolution of the detection and is not an option for short alignments.

2 Method

A possible remedy to this problem is to reduce the vagueness of the posterior distribution by reducing the cardinality of the support of $\langle P(S|\mathcal{D})\rangle$, where $\langle . \rangle$ denotes an average over all window positions. This can be effected with a pruning scheme based on the Robinson-Foulds (RF) distance (3). First, identify a set of principal tree topologies, for instance, those that maximize $\langle P(S|\mathcal{D})\rangle$. Next, assign each non-principal tree topology to the principal topology with the minimum RF-distance. Finally, renormalize the posterior distributions $P(S|\mathcal{D}_t)$ and recompute the PDM signal. An illustration is given in Figure 1. The reduction in the vagueness of $P(S|\mathcal{D}_t)$ is likely to reduce the noise in the PDM signal. Note that similar pruning methods are used in machine learning to improve the generalization performance of a predictor. Also note that the pruning of the support of $\langle P(S|\mathcal{D})\rangle$ can be justified in a Bayesian way as bringing the data-based prediction in line with our prior assumptions about the expected frequency of recombination events.

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

3 Results

The method was applied to a DNA sequence alignment of ten strains of Hepatitis-B virus with the following Genbank accession numbers: D00329, X68292, V00866, M57663, D00330, M54923, X01587, D00630, M32138 and L27106. Without pruning, the probabilistic divergence signal, shown in the left panel of Figure 2, contains erratic oscillations that obscure any breakpoint patterns. This is dramatically improved with the pruning method (middle and left panels of Figure 2). Note that three clear breakpoints occur, which were also found in an independent earlier study (1). Also note that for sufficiently small values of the cutoff threshold K, the results are rather independent of K.

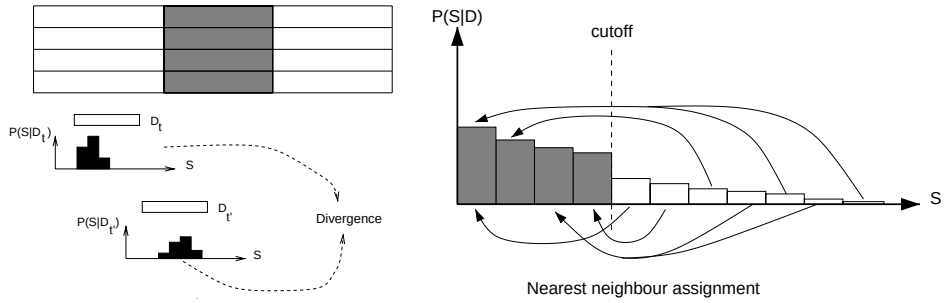

Figure 1: *Left:* PDM method. The figure shows the posterior distribution $P(S|D_t)$ of tree topologies S conditional on two subsets D_t and $D_{t'}$ selected by a moving window. When the window is moved into a recombinant region, the posterior distribution $P(S|D_t)$ can be expected to change significantly, which leads to a large probabilistic divergence score. *Right:* On the average posterior distribution of tree topologies, averaged over all sliding window positions, a cutoff threshold is defined. Tree topologies above this threshold are kept as "principal topologies". Tree topologies below the threshold are assigned to the principal topology with the minimal RF distance. After this re-assignment, the posterior distributions $P(S|D_t)$ are re-normalized.

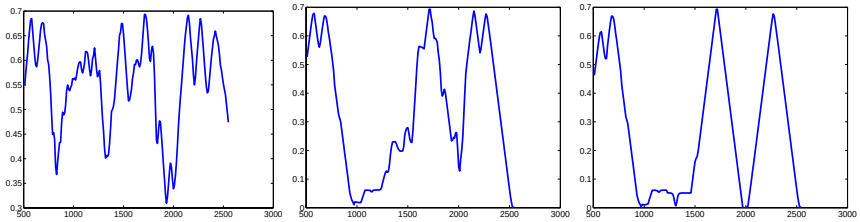

Figure 2: Detection of recombination in a DNA sequence alignment of ten strains of Hepatitis B virus. The graphs show the probabilistic divergence signals obtained with a window size of 500. From left to right: No pruning, pruning down to 7 and 3 tree topologies.

References

[1] Bollyky, P. L., Rambaut, A., Harvey, P. H., and Holmes, E. C. (1996). *Journal of Molecular Evolution*, 42:97–102.
[2] Husmeier, D. and Wright, F. (2001). *Bioinformatics*, 17(Suppl.1):S123–S131.
[3] Robinson, D. and Foulds, L. (1981). *Mathematical Biosciences*, 53:131–147.