

Detecting interspecific recombination with a pruned probabilistic divergence measure

Dirk Husmeier¹ and Frank Wright²

Biomathematics and Statistics Scotland (BioSS)

¹JCMB, The King's Buildings, Edinburgh EH9 3JZ, United Kingdom

²SCRI, Invergowrie, Dundee DD2 5DA, United Kingdom

November 3, 2004

Keywords: Phylogenetics, interspecific recombination, sliding window methods, Markov chain Monte Carlo, probabilistic divergence measure, Robinson-Foulds distance.

MOTIVATION: A promising sliding-window method for the detection of interspecific recombination in DNA sequence alignments is based on the monitoring of changes in the posterior distribution of tree topologies with a probabilistic divergence measure. However, as the number of taxa in the alignment increases or the sliding window size decreases, the posterior distribution becomes increasingly diffuse. This diffusion blurs the probabilistic divergence signal and adversely affects the detection accuracy. The present study investigates how this shortcoming can be redeemed with a pruning method based on post-processing clustering, using the Robinson-Foulds distance as a metric in tree topology space.

RESULTS: An application of the proposed scheme to three synthetic and two real-world DNA sequence alignments illustrates the amount of improvement that can be obtained with the pruning method. The study also includes a comparison with two established recombination detection methods: Recpars, and the DSS method.

AVAILABILITY: Software and data are available from <http://www.bioss.sari.ac.uk/~dirk/Supplements/>. We are currently working on an integration of these programs into TOPALi, which is available from <http://www.bioss.ac.uk/~iainm/topali/>.

CONTACT: dirk@bioss.sari.ac.uk

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. Their detection, therefore, is a crucial prerequisite for consistently inferring the evolutionary history of a set of DNA sequences.

Recently, various methods for detecting evidence of interspecific recombination in DNA sequence alignments have been developed; see Posada, Crandall, and Holmes (2002) for a review. The objective of the present article is to propose an improved phylogenetic window method, which has been motivated by and is based on the ideas of PLATO (Partial Likelihood Assessed through Tree Optimization), the DSS (Difference of Sum of Squares) method, and the PDM (Probabilistic Divergence Measure) method.

PLATO, proposed by Grassly and Holmes (1997) and illustrated in the left panel of Figure 1, computes the likelihood of various subregions of the DNA sequence alignment from a reference tree and searches for subregions with significantly low likelihood scores. The idea is that if the reference tree is the “true” tree, then recombinant regions will show a low likelihood due to the change of the tree topology. However, the “true” tree is not known and is approximated by a tree estimated from the whole sequence alignment. This includes the recombinant regions, which perturb the parameter estimation for the reference tree. Consequently, the method becomes increasingly unreliable as the recombinant regions grow in length.

To overcome this shortcoming, the DSS method (McGuire, Wright, and Prentice, 1997; McGuire and Wright, 2000), illustrated in the middle of Figure 1, replaces the global by a local reference tree. A window of typically about 500 nucleotides is slid along the DNA sequence alignment. On the first half of the window, a distance matrix is calculated according to some Markov model of nucleotide substitution, and a phylogenetic tree is estimated using the least squares method. A distance matrix is then calculated for the second half of the window, and the topology estimated from the first half is fitted to it, again using least squares. Obviously, when the topology in the right window has changed as a result of recombination, the topology in the left window will be a poor fit to the distance matrix from the right, and the difference of the two sum-of-squares statistics – termed the ‘DSS’ statistic – will be large. The motivation for using a score based on pairwise sequence distances rather than the likelihood is the application of a rescaling method to distinguish between recombination and rate variation, as described in McGuire and Wright (2000). However, it is well known that DNA sequences contain more information than pairwise distances, and that estimation methods based on pairwise distances suffer from an inevitable information loss, which is aggravated by the fact that the reference tree is computed from a small section of the sequence alignment.

The PDM method (Husmeier and Wright, 2001b), illustrated in the right panel of Figure 1, is akin to the DSS method, but addresses some of its shortcomings. First,

by using a likelihood score, the intrinsic information loss associated with a score based on pairwise distances is prevented. Second, a single optimized reference tree is replaced by a distribution over trees, which captures the intrinsic uncertainty of tree estimation from short subsections of sequence alignments. Third, the method focuses on topology changes, thereby distinguishing true recombination events from the confounding effect of rate heterogeneity. An illustration of the concepts is given in the right panel and the caption of Figure 1.

For a formal introduction, consider a DNA sequence alignment, \mathcal{D} , where we follow the usual convention that rows represent DNA sequences, and columns represent sites. From \mathcal{D} , we select a subset \mathcal{D}_t of W consecutive columns (sites), centred on the t th site of the alignment; we refer to this as a ‘window’ of width W . Let S be an integer label for tree topologies, and consider the marginal posterior probability of tree topology S conditional on the ‘window’ \mathcal{D}_t ,

$$P(S|\mathcal{D}_t) = \iint P(S, \mathbf{w}, \boldsymbol{\theta}|\mathcal{D}_t) d\mathbf{w} d\boldsymbol{\theta} \quad (1)$$

From Bayes rule we have

$$P(S, \mathbf{w}, \boldsymbol{\theta}|\mathcal{D}_t) \propto P(\mathcal{D}_t|S, \mathbf{w}, \boldsymbol{\theta})P(S, \mathbf{w}, \boldsymbol{\theta}) \quad (2)$$

where $P(\mathcal{D}_t|S, \mathbf{w}, \boldsymbol{\theta})$ is the likelihood, which is computed with the pruning algorithm (Felsenstein, 1981), and $P(S, \mathbf{w}, \boldsymbol{\theta})$ is the prior, which is chosen uniform, as in Larget and Simon (1999). Note that (1) includes a marginalization over the branch lengths \mathbf{w} of the phylogenetic tree and the parameters of the nucleotide substitution model, $\boldsymbol{\theta}$. Since the integral in (1) is analytically intractable, it has to be solved numerically by means of a Markov chain Monte Carlo (MCMC) simulation, following, for instance, Larget and Simon (1999). This yields a sample of triples $\{S_{ti}, \mathbf{w}_{ti}, \boldsymbol{\theta}_{ti}\}_{i=1}^M$ simulated from the joint posterior distribution $P(S, \mathbf{w}, \boldsymbol{\theta}|\mathcal{D}_t)$. We then replace the true posterior distribution by the empirical distribution

$$P(S, \mathbf{w}, \boldsymbol{\theta}|\mathcal{D}_t) \approx \frac{1}{M} \sum_{i=1}^M \delta_{S, S_{ti}} \delta(\mathbf{w} - \mathbf{w}_{ti}) \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{ti}) \quad (3)$$

Here, $\delta_{S, S_{ti}}$ denotes the Kronecker delta function, which is 1 if the two arguments are identical, and 0 otherwise, and $\delta()$ is the delta distribution. Inserting (3) into (1) gives:

$$P(S|\mathcal{D}_t) = \frac{1}{M} \sum_{i=1}^M \delta_{S, S_{ti}} = \frac{M_t(S)}{M} \quad (4)$$

where M is the MCMC sample size, and $M_t(S)$ denotes the number of times the MCMC simulation applied to subalignment \mathcal{D}_t visits topology S . The basic idea of the PDM method for detecting recombinant regions is to move the window \mathcal{D}_t along the alignment, as illustrated in the right panel of Figures 1 and 5, 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 required. A standard divergence measure known from information theory is the Kullback-Leibler (KL) divergence:

$$KL(P, Q) = \sum_S P(S) \ln \frac{P(S)}{Q(S)} \quad (5)$$

in which P and Q denote probability distributions over tree topologies S . However, the KL divergence is asymmetric, and it may become singular unless the support of the target distribution P is a subset of the support of the reference distribution Q : $\text{Support}(P) \subseteq \text{Support}(Q)$ (where the support is the set of topologies with non-zero probability: $\text{Support}(P) = \{S|P(S) \neq 0\}$). To prevent these singularities and to symmetrize the measure, two separate KL divergences are combined, as in Krzanowski and Marriott (1995), chapter 14:

$$\begin{aligned} \tilde{D}(t) = & \frac{1}{2} \left[KL \left(P(S|\mathcal{D}_{t-\Delta t}), \frac{P(S|\mathcal{D}_{t-\Delta t}) + P(S|\mathcal{D}_{t+\Delta t})}{2} \right) \right. \\ & \left. + KL \left(P(S|\mathcal{D}_{t+\Delta t}), \frac{P(S|\mathcal{D}_{t-\Delta t}) + P(S|\mathcal{D}_{t+\Delta t})}{2} \right) \right] \end{aligned} \quad (6)$$

Note that $\text{Support}[P(S|\mathcal{D}_{t-\Delta t})], \text{Support}[P(S|\mathcal{D}_{t+\Delta t})] \subseteq \text{Support} \left[\frac{P(S|\mathcal{D}_{t-\Delta t}) + P(S|\mathcal{D}_{t+\Delta t})}{2} \right]$, which guarantees the non-singularity of $\tilde{D}(t)$. The final divergence measure $D(t)$ is obtained by averaging $\tilde{D}(t)$ in (6) over different degrees of window overlap, typically $0.1W \leq 2\Delta t \leq 0.9W$.

Method

The shortcoming of the PDM method is illustrated in Figure 2. A substantial change in the tree topology caused by a recombination event is not distinguished from changes in the clade configurations of closely related taxa. As the number of sequences in the alignment increases, so does the number of such within-clade reconfigurations. This is because for an increased number of taxa the posterior distribution over tree topologies, $P(S|\mathcal{D}_t)$, becomes increasingly disperse 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.

A simple pruning scheme

A possible remedy for this “dispersion” problem is to include information on the amount of change between different tree topologies and thereby to distinguish between the scenarios depicted in Figure 2. Since branch lengths have been marginalized over for the reason mentioned in the previous section, the difference has to be estimated solely on the basis of topological information. A possible metric in the space of tree topologies was introduced by Robinson and Foulds (1981). The idea is illustrated in Figure 3. Consider the two complimentary operations of merging two existing nodes into one, and splitting an existing node into two. The Robinson-Foulds (RF) distance between tree topologies S_1 and S_2 is defined as the minimum number of operations required to transform S_1 into S_2 (or S_2 into S_1). Robinson and Foulds (1981) showed that this definition is equivalent to the following, for practical implementations more relevant formulation. A branch induces a bi-partition of leaves (taxa). Denote by $E(S_1)$ the set of all bi-partitions induced in tree S_1 , and by $|E(S_1) - E(S_2)|$ the set of all bi-partitions induced by edges in tree S_1 that are not found in tree S_2 . The RF distance between

tree topologies S_1 and S_2 is then given by the following expression:

$$RF(S_1, S_2) = |E(S_1) - E(S_2)| + |E(S_2) - E(S_1)| \quad (7)$$

Define $\langle P(S|\mathcal{D}) \rangle$ to denote the average posterior distribution of tree topologies, averaged over all sliding window positions:

$$\langle P(S|\mathcal{D}) \rangle = \frac{1}{N_W} \sum_{t=1}^{N_W} P(S|\mathcal{D}_t) \quad (8)$$

in which N_W denotes the total number of window positions. Recall that the support of the average posterior distribution is the set of those tree topologies for which $\langle P(S|\mathcal{D}) \rangle \neq 0$, that is, the set of all tree topologies visited during the MCMC simulations. Now, an application of the PDM method to a DNA sequence alignment of ten strains of Hepatitis-B virus, discussed below, was found to lead to a support of 126 distinct tree topologies. Usually, we do not expect to find over hundred recombination events in an alignment of about 3000 nucleotides. Hence most of topology changes correspond to within-clade reconfigurations resulting from diffuse posterior distributions, which degrades the reliability and efficacy of the PDM method.

A way to proceed, then, is to reduce the dispersion of the posterior distribution by reducing the cardinality of the support of $\langle P(S|\mathcal{D}) \rangle$. This can be effected with a pruning scheme based on the RF distance. 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 3. The reduction in the dispersion 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 (Bishop, 1995). Also note that the pruning of the support of $\langle P(S|\mathcal{D}) \rangle$ has some similarity with a Bayesian approach in that it brings the data-based prediction in line with our prior assumption about the expected frequency of recombination events.

Improved pruning by clustering

The pruning scheme described in the previous subsection has an obvious disadvantage: when the recombinant region is short, the set of principal topologies may not contain any topology that reflects the recombination event. Also, the assignment of non-principal to principal topologies can be interpreted as the first step of an incomplete K-means clustering procedure. Such topology-based clustering of phylogenetic trees was proposed by Stockham, Wang, and Warnow (2002) to reduce the information loss incurred by a single consensus tree approach. It here offers the obvious method to be used for pruning. Based on the RF distance, the MCMC sample of tree topologies, $\{S_1, \dots, S_M\}$, is clustered. For a given number of clusters K , tree topologies are assigned to their respective cluster $\mathcal{C}_1, \dots, \mathcal{C}_K$:

$$I(S_i \in \mathcal{C}_k) = \begin{cases} 1 & \text{if } S_i \in \mathcal{C}_k \\ 0 & \text{if } S_i \notin \mathcal{C}_k \end{cases} \quad (9)$$

Now, define the posterior distribution over clusters,

$$P(\mathcal{C}_k|\mathcal{D}_t) = \sum_S I(S \in \mathcal{C}_k) P(S|\mathcal{D}_t) \quad (10)$$

which is computed from the MCMC sample of tree topologies, $\{S_1, \dots, S_M\}$:

$$P(\mathcal{C}_k | \mathcal{D}_t) = \frac{1}{M} \sum_{i=1}^M I(S_i \in \mathcal{C}_k) \quad (11)$$

An improved pruning scheme can be effected by using the posterior distribution over classes (11) instead of the original posterior distribution over topologies (4) in the computation of the divergence measure $D(t)$ in (6). This novel approach of computing $D(t)$ from $P(\mathcal{C}_k | \mathcal{D}_t)$ rather than $P(S | \mathcal{D}_t)$ will henceforth be referred to as the *pruned PDM* method.

Note that while this method addresses the shortcomings of the original PDM method and the simple pruning scheme of Figure 3, it is still heuristic in that the choice of the clustering algorithm is arbitrary. The present work is based on the findings of Stockham, Wang, and Warnow (2002). When applying MCMC to sample trees from the posterior distribution, as described in Larget and Simon (1999), one is faced with the problem of summarizing the information contained in the MCMC sample succinctly. A widely applied method is to resolve the conflicts within the obtained sample of tree topologies by computing a consensus tree, which, however, may incur a substantial information loss. Stockham, Wang, and Warnow (2002) investigated a post-processing alternative, by which trees are first divided into subsets with some clustering algorithm based on the RF distance, and each cluster is then characterized by its own consensus tree. The authors assessed various clustering algorithms with different measures of information loss, like the KL divergence between the original distribution over trees and the distribution induced by the clusters, as well as the specificity, that is, the normalized number of internal edges of the consensus tree (averaged over all clusters). They found that bottom-up average linkage clustering outperformed top-down K-means clustering and the method of phylogenetic islands (Maddison, 1991) in terms of complexity versus information content, and the former scheme will therefore be used in the present study.

The pruning-by-clustering procedure proposed in the present paper thus works as follows. First, generate an exhaustive list of all distinct tree topologies sampled in the complete set of MCMC simulations, that is, the MCMC simulations carried out for all window positions. Next, compute all pairwise RF distances between topologies, and infer a dendrogram of topologies with agglomerative average linkage clustering from these distances. Cut this dendrogram at a certain height such that it disintegrates into a pre-defined number of clusters. Finally, assign each tree topology to its respective cluster and use this assignment to compute the posterior distribution over clusters by application of (10). This posterior distribution over clusters supersedes the original posterior distribution over topologies in the computation of the divergence measure (6). A summary of the proposed algorithm is given in Figure 4.

Significance estimation

Having obtained an improved probabilistic divergence signal, the next question is whether the observed peaks are significant. Following a classical statistical approach, this requires the computation of the distribution of peak heights under the null hypothesis of no recombination. In order to avoid the bias from asymptotic distributions, we would like to follow a bootstrap procedure. The idea is to repeat the analysis several (typically about a hundred) times under the null hypothesis of no recombination, either

after randomizing the columns of the DNA sequence alignment (nonparametric bootstrapping) or after generating DNA sequence alignments synthetically from a model fitted to the actual DNA sequence alignment under the null hypothesis of no recombination (parametric bootstrapping). The disadvantage of this approach, however, is the need for repeated MCMC simulations, resulting in unreasonably high computational costs. The approach proposed in the present paper is based on the bootstrap procedure just described, but without the need for repeated MCMC simulations. Recall that the PDM method computes a posterior distribution $P(S|\mathcal{D}_t)$ of tree topologies S for each sliding window position t . Under the null hypothesis of no recombination, these probability distributions are identical except for sampling-induced variations; obviously, when estimating a probability distribution from a finite sample size, variations result as a mere consequence of random fluctuations in the selected subalignment \mathcal{D}_t . The objective of bootstrapping is to distinguish the effect of these random fluctuations from systematic changes in the posterior distributions that reflect true recombination-induced topology changes. Now, averaging over all “local” distributions $P(S|\mathcal{D}_t)$ gives a “global” distribution $\langle P(S|\mathcal{D}) \rangle$, defined in (8), which is characteristic of the given DNA sequence alignment. We could therefore think of simulating the variation under the null hypothesis of no recombination by repeatedly drawing M samples from $\langle P(S|\mathcal{D}) \rangle$. In this way we could obtain a set of empirical bootstrap replica distributions, from which the probabilistic divergence signals under the null hypothesis could be computed. However, direct sampling from $\langle P(S|\mathcal{D}) \rangle$ leads to samples that are independent, whereas the actual MCMC sample results from a Markov chain. To model the sampling process under the null hypothesis correctly, we therefore generate tree topologies from a Markov model with the following transition probability:

$$P(S_{n+1} = i | S_n = k) = \lambda \delta_{ik} + (1 - \lambda) \pi_i \quad (12)$$

Here, S_n and S_{n+1} denote the tree topologies at subsequent sampling steps, δ_{ik} is the Kronecker delta symbol, which is 1 if $i = k$ and 0 otherwise, $\pi_i \in [0, 1]$ is the equilibrium distribution of tree topologies, which implies $\sum_i \pi_i = 1$, and $\lambda \in [0, 1]$ is the probability of not changing the tree topology from step n to step $n + 1$; this reflects the rejection of a move in the MCMC simulation. The value of λ can be estimated in a maximum likelihood sense, using the EM algorithm (Dempster, Laird, and Rubin, 1977) or a line search. The whole procedure can thus be summarized as follows. First, fit the Markov model of (12) to the complete set of tree topologies obtained from the MCMC simulations. Second, draw N_W bootstrap samples of M tree topologies from the Markov model (12), where M is the MCMC sample size, and N_W is the number of window positions in the original PDM application. Third, compute the empirical distribution for each of the N_W bootstrap samples. Fourth, compute the PDM signal and determine its highest peak. The whole procedure is repeated N_B times (for instance, $N_B = 100$), starting from different random number generator seeds. The resulting distribution of maximal peak heights is the distribution under the null hypothesis of no recombination. An illustration of this procedure is given in Figure 5.

Software

The original PDM method has been implemented in the software package TOPALi (Milne, Wright, Rowe, Marshall, Husmeier, and McGuire, 2004), which is freely available from <http://www.bioss.ac.uk/software.html>.

The pruned PDM method proposed in the present article has been implemented using a combination of four pieces of software: (1) the PDM code extracted from TOPALi; (2) BAMBE (Larget and Simon, 1999) for carrying out the MCMC simulations; (3) the program TREEDIST of the PHYLIP package (Felsenstein, 1996) for computing the Robinson-Foulds distance; and (4) MATLAB code for the hierarchical clustering and pruning steps. While these programs can already be freely obtained from <http://www.bioss.sari.ac.uk/~dirk/Supplements/>, we are currently improving their user-friendliness and will produce a pruned PDM option in TOPALi.

Data and simulations

The evaluation of a recombination detection method is best carried out through a combined analysis of real-world and simulated DNA sequence alignments. For simulated data, the true location of the recombinant regions and their breakpoints is known; hence a straightforward assessment of the detection accuracy is possible. The disadvantage, however, is that the models used in the simulation study are simplifications of reality. Real-world DNA sequence alignments are obviously not affected by this shortcoming. However, the assessment of the detection accuracy is impeded by the fact that the true location of the recombinant regions and their breakpoints is ultimately unknown. We therefore carried out the present evaluation on three synthetic data sets, for which the level of detection difficulty varied substantially, and two real-world DNA sequence alignments. These sequence alignments can be obtained from the accompanying website: <http://www.bioss.sari.ac.uk/~dirk/Supplements/>.

Simulated recombination

We tested the proposed detection method on two of the simulated data sets used in Husmeier and Wright (2001b). Here, evolution in a population of 8 taxa was simulated with the Kimura model (Kimura, 1980), from which a DNA sequence alignment of 5500 nucleotides was generated. Two recombination events were simulated: an ancient event, affecting the region between sites 1000 and 1500, and a recent event, affecting the region between sites 2500 and 3000. A mutational hotzone of the same length, located between sites 4000 and 4500, was evolved at an increased nucleotide substitution rate (factor 3); this is to test whether the detection method can successfully distinguish between recombination and rate variation.

For the present study, we used the data sets B1 and B3 in Husmeier and Wright (2001b). For B1, henceforth referred to as *simulated alignment 1*, the average branch length of the underlying phylogenetic tree was $w = 0.1$. For B3, henceforth referred to as *simulated alignment 2*, this value was reduced to $w = 0.01$. The ensuing reduction in the number of polymorphic sites renders the second detection problem harder, because recombination tends to be easier to detect with increasing levels of divergence (see also Posada (2002), p.714).

We also generated a new sequence alignment, which followed the simulation procedure described in Husmeier and Wright (2001b), but reduced the lengths of the tree branches even further, drawing them from a uniform distribution on the interval $[0.003, 0.005]$. The motivation for this data set, henceforth referred to as *simulated alignment 3*, was to increase the detection difficulty deliberately to a level where all alternative detection methods investigated in this study were found to fail.

Hepatitis B Virus

Hepatitis B is caused by a DNA virus with a short genome of 3200 nucleotides. Evidence for recombination was found in Bollyky, Rambaut, Harvey, and Holmes (1996). The present study investigates a subset of two recombinant strains (Genbank accession numbers D00329 and X68292) and eight nonrecombinant strains (V00866, M57663, D00330, M54923, X01587, D00630, M32138 and L27106). The sequences were aligned with ClustalW (Thompson, Higgins, and Gibson, 1994), using the default parameters. Columns with gaps were discarded, giving a total alignment length of 3049 nucleotides.

Maize actin genes

Gene conversion is a process equivalent to recombination. It occurs in multigene families, where a DNA subsequence of one gene can be replaced by the DNA subsequence from another. Indication of gene conversion between two pairs of maize actin genes has been reported in Moniz de Sa and Drouin (1996). For the present evaluation, we used the eight maize sequences studied by Moniz de Sa and Drouin (1996) (GenBank/EMBL accession numbers are in brackets): Mac1 (J01238), Maz56 (U60514), Maz63 (U60513), Maz81 (U60511), Maz83 (U60510), Maz87 (U60509), Maz89 (U60508), and Maz95 (U60507). The sequences were aligned with ClustalW, using the default parameter settings. Columns with more than three gaps were discarded, resulting in a total alignment length of 1281 nucleotides.

Methods

The PDM and pruned PDM methods were applied as follows. Windows of different lengths were moved along the alignment with a fixed step size of $\Delta t = 10$ nucleotides. The MCMC simulations were carried out with BAMBE¹, described in Larget and Simon (1999). For each new window position, the system was equilibrated over $M_{eq} = 100,000$ Metropolis-Hastings steps, starting from an initial tree obtained with Neighbour Joining (Saitou and Nei, 1987) and using global² tree manipulations for the proposal moves. This was followed by a sampling period of $M = 100,000$ Metropolis-Hastings steps, using local tree manipulations and sampling tree configurations in intervals of $\Delta M = 200$ Metropolis-Hastings steps. Full details of the MCMC simulations are available from the accompanying web site (see abstract), which includes the input file for BAMBE. Note that rather large values for M_{eq} and M were chosen to ensure sufficient mixing and convergence of the Markov chains. When applied sequentially, this leads to a total execution time of several hours, and one may therefore want to reduce the size of M and M_{eq} . In fact, rerunning the algorithm with $M = 10,000$, $M_{eq} = 10,000$, and $\Delta M = 20$

¹Available from <http://www.mathcs.duq.edu/larget/bambe.html>.

²In BAMBE, two different kinds of proposal moves – local and global – are used, which are described in Larget and Simon (1999).

was found to lead to results very similar to those presented in this article. However, as discussed in the Discussion section, the algorithm can easily be parallelized, in which case the total execution time is reduced to the order of minutes without having to compromise the convergence and mixing of the Markov chains.

The DSS signals were computed as described in McGuire and Wright (2000), using the same step size and window size as for the PDM and pruned PDM methods. We used the default options of the program, except for the computation of the pairwise distances, where we replaced the Jukes-Cantor model of nucleotide substitution by the Kimura 2-parameter model.

For a further comparison, Recpars (Hein, 1993) was applied. This method requires the prior specification of a recombination and a nucleotide substitution penalty parameter, for which various ratios Ψ were chosen.

Results

Figure 6, top left panel, shows the unpruned PDM signal obtained from the first simulated sequence alignment, using a window size of $W = 500$. All four recombinant breakpoints are clearly detected, and the differently diverged region is successfully suppressed. No pruning is required, although it improves the spatial resolution slightly (Figure 6, top panel). The resolution is more dramatically improved by reducing the window size to $W = 200$; see the bottom panel of Figure 6. However, without pruning this reduction in W substantially increases the signal stemming from the confounding, differently diverged region. This shortcoming is avoided when pruning is used (Figure 6, bottom panel).

Figure 7, top left panel, shows the unpruned PDM signal obtained from the second simulated sequence alignment, again using a window size of $W = 500$. The four recombinant breakpoints are still detected, but at a poor spatial resolution. This deterioration is a consequence of the significantly decreased sequence divergence, as discussed above. Pruning improves the resolution slightly (top panel of Figure 7). However, a decrease of the window size to $W = 200$ no longer seems to be a viable option: the bottom left panel of Figure 7 demonstrates that the relevant peaks get lost in a noisy, erratic signal. Interestingly, when applying the pruning method, the true peaks of the recombinant breakpoints re-emerge, resulting in a spatial resolution that is equivalent to the one obtained from the first simulated sequence alignment (bottom panel of Figure 7).

Figures 8 and 9 show the performance of both the DSS method and Recpars on the first two simulated sequence alignments. With one exception, the DSS method predicts all recombinant breakpoints correctly when a window size of $W = 500$ is used, while successfully suppressing the confounding, differently diverged region. However, when the window size is decreased to $W = 300$ and $W = 200$, the true breakpoints get lost among spurious peaks. The prediction with Recpars depends critically on the recombination versus nucleotide substitution cost ratio, Ψ . Note that this parameter cannot be inferred from the data, but has to be selected rather arbitrarily by the user. For the first simulated sequence alignment, the correct mosaic structure is predicted when setting $\Psi = 10$ (Figure 8, left panel). For the second sequence alignment, $\Psi = 2$ leads to a satisfactory prediction (Figure 9, left panel). However, setting the value of Ψ with the benefit of hindsight after inspecting the outcome of the prediction is

methodologically incorrect, and for suboptimal values of Ψ poor predictions with too many or too few breakpoints are obtained.

On the third simulated sequence alignment, none of the alternative methods achieved a satisfactory performance; see the two subfigures on the left of Figure 11 (DSS method and Recpars), and the left panel of Figure 10 (unpruned PDM method). Interestingly, all four breakpoints are correctly identified when the pruning method with a sufficiently small number of clusters K is used (Figure 10).

The centre and right panels of Figure 13 show two DSS signals obtained from the maize actin gene sequence alignment, using the window sizes of $W = 200$ and $W = 300$. A clear breakpoint is detected around site 1000. This breakpoint is also found with Recpars for $\Psi = 10$ (see the left panel of Figure 13), and it concurs with the findings reported in Moniz de Sa and Drouin (1996). The unpruned PDM method also detects this breakpoint; however, it is obscured by various spurious peaks (Figure 12, left panel). This poor performance is rectified with the pruning method: Figure 12 shows that out of six predictions resulting from combining two window sizes ($W = 200$ and $W = 300$) with three threshold values ($K = 3, 5, 7$), five concur with the DSS signals.

On the Hepatitis-B virus DNA sequence alignment, the PDM method was applied with two window sizes: $W = 300$ and $W = 500$. Without pruning, the probabilistic divergence signals, shown in the left column of Figure 14, contain erratic oscillations that obscure any clear patterns. This is dramatically improved with the pruning method. The top row of Figure 14 shows the results for the larger window size. Three breakpoints emerge, for all chosen values of the threshold K . These breakpoints are also clearly predicted with the DSS method (Figure 11, right). When decreasing the window size, these breakpoints are predicted at a higher spatial resolution (bottom row of Figure 14), and two further breakpoints occur (at positions 1000 and 1250). Interestingly, these breakpoints also appear in the DSS signal – at the border of the significance threshold (Figure 11, right). The predictions with Recpars (Figure 11, centre right) vary too much to be conclusive.

Discussion

The present paper has proposed a novel phylogenetic window method for detecting interspecific recombination in DNA sequence alignments. The new approach combines the probabilistic divergence measure (PDM) method of Husmeier and Wright (2001b) with the post-processing clustering procedure of Stockham, Wang, and Warnow (2002). The idea is to reduce the dispersion of the posterior distribution of tree topologies, on which the computation of the PDM signal is based, by a pruning procedure. Pruning is effected by clustering trees with similar topologies together, and replacing the posterior distribution over tree topologies by the posterior distribution over topology clusters. The scheme requires the prior specification of the number of topology clusters, K , which reflects our prior assumption about the maximum permissible number of recombination events. For sufficiently small values of K the prediction accuracy was found to increase considerably. Also, for sufficiently small values of K , the results are rather robust with respect to a variation of K ; hence this variation is far less critical than the variation of, say, Ψ in Recpars. The pruning scheme may also allow for more flexibility in the choice of the window size. While a smaller window increases the spatial

resolution of the breakpoint detection, it also increases the dispersion of the posterior distribution, reflected by the total number of tree topologies sampled in the MCMC simulations. This dispersion aggravates the inherent shortcoming of the PDM method, as discussed above. By projecting the total topology space onto a small set of topology clusters, the dispersion of the posterior distribution is drastically reduced. The simulation studies discussed in the previous section have demonstrated that a decrease of the window size may lead to erratic and noisy PDM signals when no pruning is used, while the application of the novel pruning method was found to recover the true recombinant breakpoints at an improved spatial resolution. This finding suggests that the pruning scheme proposed in the present article may lead to a substantial improvement on the PDM method, rendering it a more viable tool for the detection of interspecific recombination.

Like the PDM method, the pruned PDM method requires several MCMC simulations to be carried out – one for each window position. When applied sequentially, this procedure is slow and may take several hours to complete. However, each individual MCMC simulation is independent of all the other MCMC simulations, and the method therefore lends itself to a straightforward parallelization. Note that each individual MCMC simulation is fast due to the small length of the selected sequence alignment (selected with the sliding window). Consequently, a parallelization of the algorithm can reduce the total execution time to the order of minutes.

An obvious disadvantage of the proposed scheme is the use of a sliding window, which gives the method an intrinsically heuristic flavour. The simulation studies have demonstrated that the performance of the pruned PDM scheme seems to depend less critically on the window size than the DSS method. This finding reflects the fact that the PDM method is a likelihood method and, consequently, extracts more information from the data than the DSS method, which is based on pairwise sequence distances. However, a method like Recpars, which avoids the use of a window altogether, is in principle superior. The principal shortcoming of Recpars is that, as a parsimony method, it depends on the arbitrary parameter Ψ , which cannot be estimated from the data. Husmeier and Wright (2001a) and Husmeier and McGuire (2003) have developed a likelihood method equivalent to Recpars, in which all parameters are estimated from the data, and in which the recombinant breakpoints can be determined at a higher resolution than with a window method. However, this approach is restricted to small sequence alignments of typically only four taxa. This restriction calls for the prior identification of putative recombinant sequences, using, for example, window-based methods – like the one discussed in the present article.

Future research may try to improve the method for significance estimation. Note that the proposed Markov process of (12) converges to the equilibrium distribution at a uniform rate. This approach is too restrictive for the unpruned PDM method, due to the scenario depicted in Figure 2. A more sophisticated approach would be the employment of a more complex transition matrix, for which the transition probabilities between different tree topologies depend on the respective RF distance. However, this extra complexity would lead to substantial computational overheads; hence, the current study resorted to the simple rectification of always estimating the transition probability λ from the *pruned* distributions. In fact, this estimation was found to consistently lead to values of $0.90 < \lambda < 0.99$. Since minor variations in the height of the significance

line do not seem to be particularly critical for the interpretation of the results, one may well decide to use a fixed *a priori* choice of $\lambda = 0.95$.

As a final remark, note that the objective of the PDM method is to detect changes in the tree topology, that is, systematic deviations from the assumption that a single phylogenetic tree is consistent with the whole DNA sequence alignment. The focus of this paper has been on interspecific recombination and gene conversion. Other processes that may lead to topology changes are gene duplication and lineage sorting. Conversely, certain recombination events may only affect the branch lengths of a phylogenetic tree while leaving the topology unchanged. Consequently, not all recombination events can be detected with the PDM method, while certain breakpoints detected with the PDM method may have been caused by events different from recombination. While, on the face of it, this looks like a shortcoming, we note that the motivation for the PDM method does not come from population genetics, where one is interested in recombination events per se, but from phylogenetics. Most standard phylogenetic methods are based on the implicit assumption that the given alignment is consistent with a single phylogenetic tree. When a recombination event only affects the branch lengths of a phylogenetic tree, this assumption is not particularly critical: the inferred topology will still be correct (as it has not been changed), and the inferred branch lengths will show some average value. However, when recombination (or any other event, like lineage sorting) affects the tree topology, the inference of an average tree is no longer reasonable: tree topologies are cardinal entities, for which an average value is *not* defined. In fact, it is well-known that in this case the application of a standard phylogenetic inference method may lead to a distorted tree that is unrepresentative of the underlying evolutionary history of the sequences. By detecting systematic changes of the tree topology along the sequence alignment, the pruned PDM method proposed in the present article promises to offer a useful pre-screening tool for improving the consistency of phylogenetic analysis.

Acknowledgements

This work was supported by the Scottish Executive Environmental and Rural Affairs Department (SEERAD). We would like to thank Chris Glasbey for helpful discussions.

References

- Bishop, C. M., 1995, *Neural Networks for Pattern Recognition*. (Oxford University Press New York) ISBN 0-19-853864-2.
- Bollyky, Paul L., Andrew Rambaut, Paul H. Harvey, and Edward C. Holmes, 1996, Recombination between Sequences of Hepatitis B Virus from Different Genotypes, *Journal of Molecular Evolution* 42, 97–102.
- Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society* B39, 1–38.
- Felsenstein, Joe, 1981, Evolution trees from DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution* 17, 368–376.

- Felsenstein, Joe, 1996, PHYLIP, Free package of programs for inferring phylogenies, available from <http://evolution.genetics.washington.edu/phylip.html>.
- Grassly, Nicholas C., and Edward C. Holmes, 1997, A Likelihood Method for the Detection of Selection and Recombination Using Nucleotide Sequences, *Molecular Biology and Evolution* 14, 239–247.
- Hein, Jotun, 1993, A Heuristic Method to Reconstruct the History of Sequences Subject to Recombination, *Journal of Molecular Evolution* 36, 396–405.
- Husmeier, Dirk, and Grainne McGuire, 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.
- Husmeier, Dirk, and Frank Wright, 2001a, Detection of Recombination in DNA Multiple Alignments with Hidden Markov Models, *Journal of Computational Biology* 8, 401–427.
- Husmeier, Dirk, and Frank Wright, 2001b, Probabilistic Divergence Measures for Detecting Interspecies Recombination, *Bioinformatics* 17, S123–S131.
- Kimura, M., 1980, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *Journal of Molecular Evolution* 16, 111–120.
- Krzanowski, W. J., and F. H. C. Marriott, 1995, *Multivariate Analysis* vol. 2. (Arnold) ISBN 0-340-59325-3.
- Larget, B., and D. L. Simon, 1999, Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees, *Molecular Biology and Evolution* 16, 750–759.
- Maddison, D, 1991, The discovery and importance of multiple islands of most parsimonious trees, *Systematic Zoology* 40, 315–328.
- McGuire, Grainne, and Frank Wright, 2000, TOPAL 2.0: improved detection of mosaic sequences within multiple alignments, *Bioinformatics* 16, 130–134.
- McGuire, G., F. Wright, and M.J. Prentice, 1997, A Graphical Method for Detecting Recombination in Phylogenetic Data Sets, *Molecular Biology and Evolution* 14, 1125–1131.
- Milne, Iain, Frank Wright, Glenn Rowe, David F. Marshall, Dirk Husmeier, and Grainne McGuire, 2004, TOPALi: Software for automatic identification of recombinant sequences within DNA multiple alignments, *Bioinformatics* 20, 1806–1807.
- Moniz de Sa, M., and G. Drouin, 1996, Phylogeny and substitution rates of angiosperm actin genes, *Molecular Biology and Evolution* 13, 1198–1212.
- Posada, David, 2002, Evaluation of Methods for Detecting Recombination from DNA Sequences: Empirical Data, *Molecular Biology and Evolution* 19, 708–717.

- Posada, David, Keith A. Crandall, and Edward C. Holmes, 2002, Recombination in Evolutionary Genomics, *Annual Review of Genetics* 36, 75–97.
- Robinson, D., and L. Foulds, 1981, Comparison of phylogenetic trees, *Mathematical Biosciences* 53, 131–147.
- Saitou, Naruya, and Masatoshi Nei, 1987, The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution* 4, 406–425.
- Stockham, Cara, Li-San Wang, and Tandy Warnow, 2002, Statistically based postprocessing of phylogenetic analysis by clustering, *Bioinformatics* 18, S285–S293.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson, 1994, CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice, *Nucleic Acids Research* 22, 4673–4680.

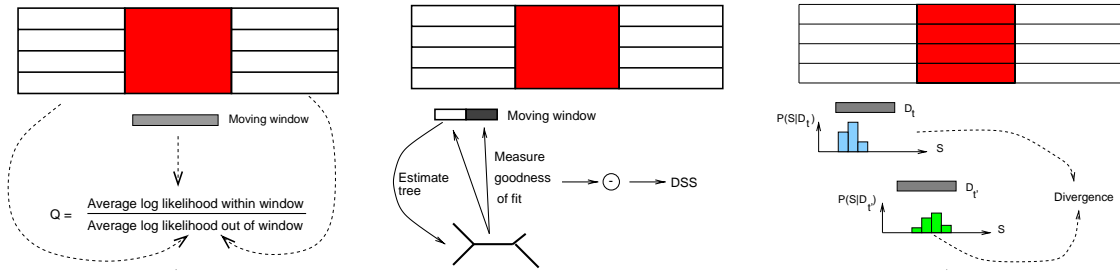


Figure 1: Different methods for detecting recombination. The figure illustrates three window-based phylogenetic methods for detecting recombination in DNA sequence alignments. *Left:* PLATO. A window of varying size is moved along the DNA sequence alignment. The average log likelihood is computed for both the window and the flanking regions of the sequence alignment, and the Q statistic is defined as the ratio of these values. Large Q values are taken as indications of recombinant regions. *Centre:* The DSS method. A window consisting of two subwindows is moved along the DNA sequence alignment. A tree is estimated from the left subwindow, and a goodness-of-fit score is computed for both subwindows. The DSS statistic is defined as the difference between these scores. When the window is centred on or near the breakpoint of a recombinant region, the tree estimated from the left subwindow is not an adequate model for the data on the right, which leads to a large DSS value. *Right:* The PDM method. The figure shows the posterior distribution $P(S|D_t)$ of tree topologies S conditional on two subsets of columns, 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 high probabilistic divergence score. Further details are given in the text.

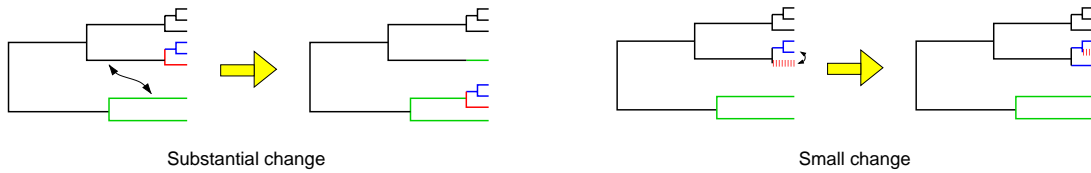


Figure 2: Shortcoming of the PDM method. A substantial change in the branching structure of a phylogenetic tree, illustrated on the left, is indicative of a recombination event. Changes of the branching structure that only involve closely related strains, shown on the right, may result as a mere consequence of statistical fluctuations. The PDM method of Husmeier and Wright (2001b) does not distinguish between these two types of changes in the tree topology, which renders the approach suboptimal.

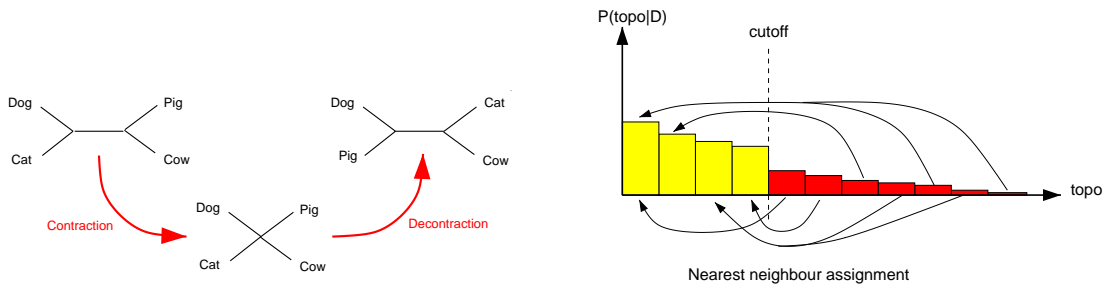


Figure 3: RF-distance based pruning. *Left:* The RF distance between tree topologies is defined as the minimum number of contraction and decontraction operations needed to transform one topology into the other. *Right:* On the average posterior distribution of tree topologies, averaged over all sliding 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 minimum RF distance. After this re-assignment, the posterior distribution is re-normalized.

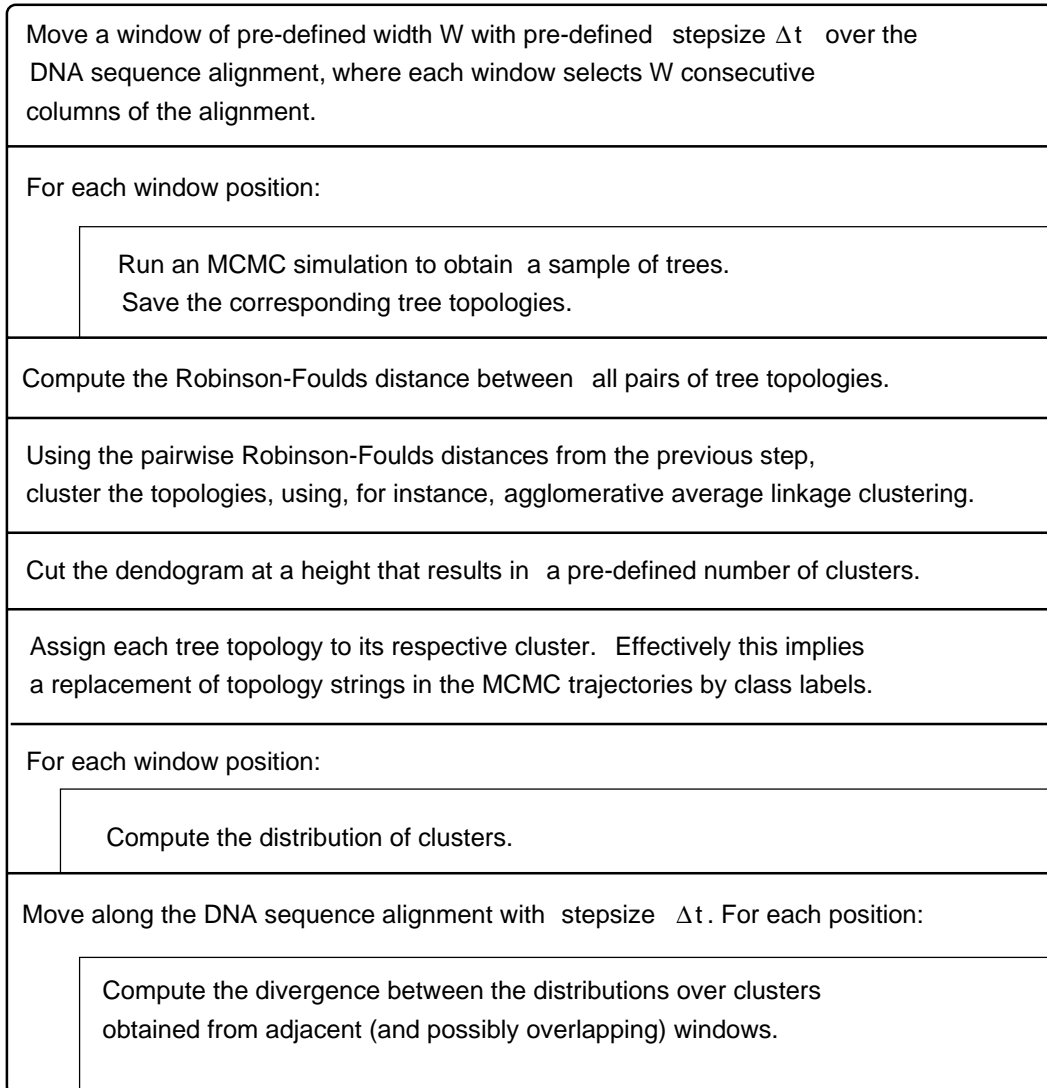


Figure 4: **Pruning by clustering.** The figure summarizes the various steps involved in the computation of the pruned probabilistic divergence signal.

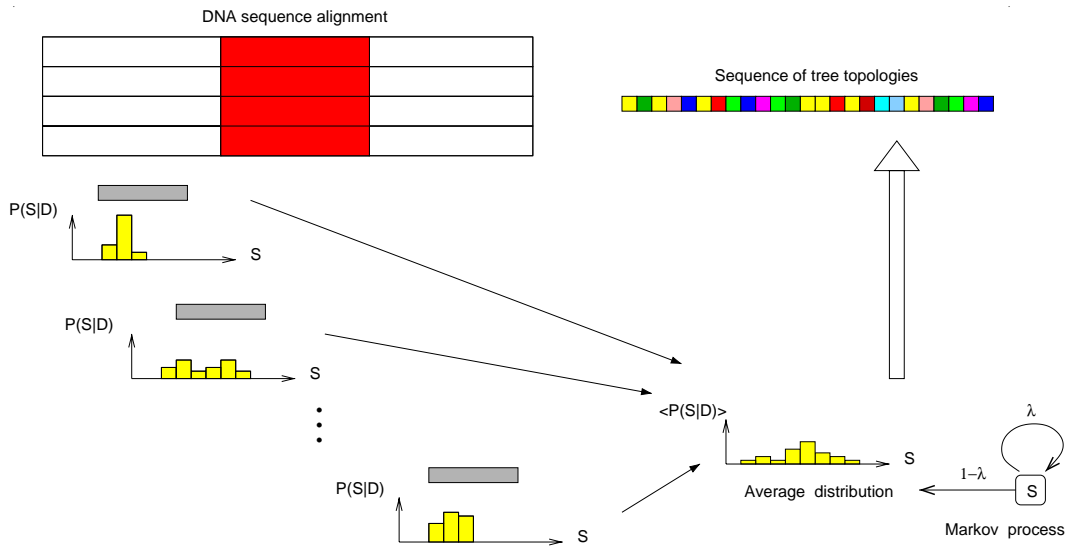


Figure 5: **PDM method and bootstrapping.** The PDM procedure computes the posterior distribution of tree topologies for each sliding window position. Averaging over all local posterior distributions leads to a “global” distribution. Sampling independently from this distribution ignores the fact that the tree topologies were generated from a Markov chain. In order to capture the dependence structure correctly, a Markov model is fitted to the set of tree topologies obtained from the MCMC simulations. Generating tree topologies from this Markov model simulates a sample under the null hypothesis of no recombination.

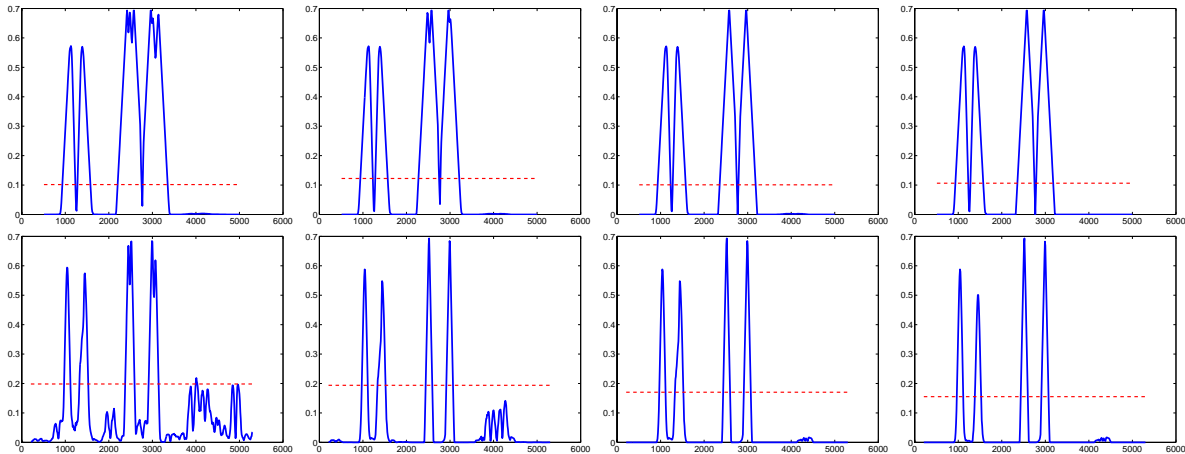


Figure 6: **Detection of recombination in the first synthetic DNA sequence alignment: PDM method.** The graphs show probabilistic divergence signals; the dashed horizontal lines show the 99 percentiles under the null hypothesis of no recombination, computed with parametric bootstrapping based on the Markov model of Figure 5. Top row: window size 500. Bottom row: window size 200. From left to right: no pruning (resulting in 16 tree topologies when the window size is 500, and 104 topologies for a window size of 200), pruning down to 7, 5, and 3 clusters. The true mosaic structure of the sequence alignment is as follows. Ancient recombination event: sites 1000–1500; recent recombination event: sites 2500–3000; differently diverged region: sites 4000–4500.

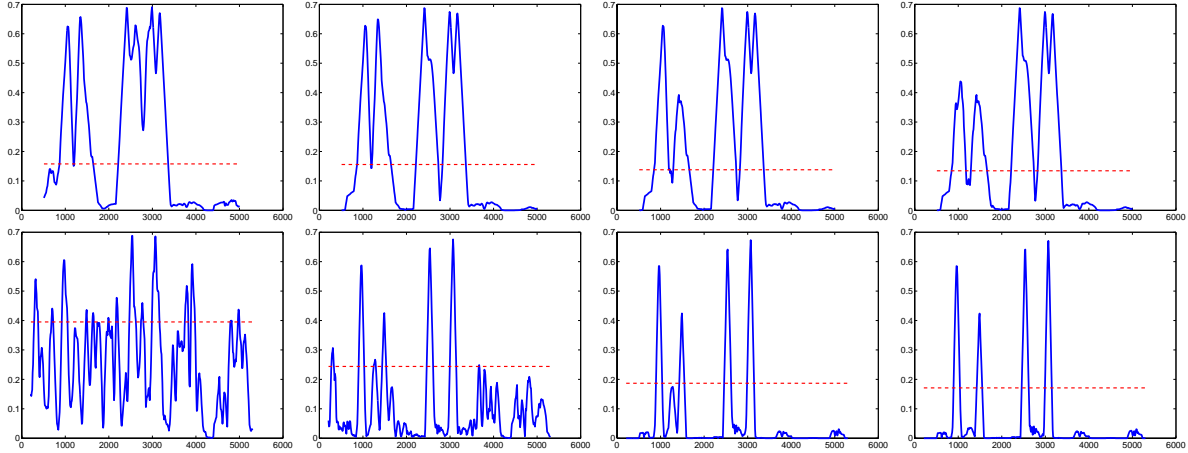


Figure 7: **Detection of recombination in the second synthetic DNA sequence alignment: PDM method.** The graphs show probabilistic divergence signals; the dashed horizontal lines show the 99 percentiles under the null hypothesis of no recombination, computed with parametric bootstrapping based on the Markov model of Figure 5. Top row: window size 500. Bottom row: window size 200. From left to right: no pruning (resulting in 50 tree topologies when the window size is 500, and 406 topologies for a window size of 200), pruning down to 7, 5, and 3 clusters. The true mosaic structure of the sequence alignment is as follows. Ancient recombination event: sites 1000–1500; recent recombination event: sites 2500–3000; differently diverged region: sites 4000–4500.

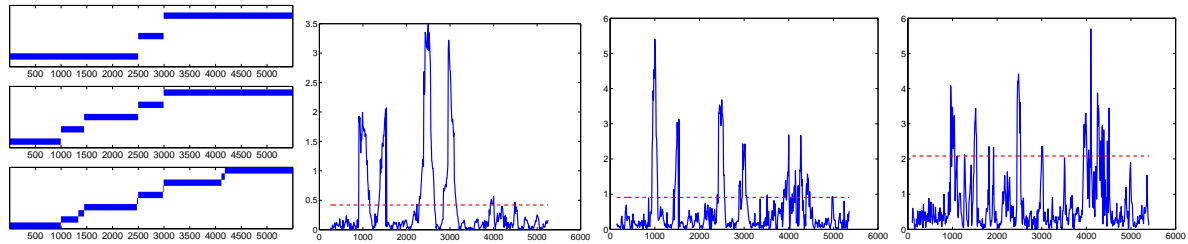


Figure 8: **Detection of recombination in the first synthetic DNA sequence alignment: alternative methods.** *Left:* Prediction obtained with Recpars. The three panels show segmentations predicted for different recombination versus nucleotide substitution cost ratios Ψ . Top: $\Psi = 20$; middle: $\Psi = 10$; bottom: $\Psi = 5$. *Centre left:* DSS signal obtained with a window size of 500. *Centre right:* DSS signal obtained with a window size of 300. *Right:* DSS signal obtained with a window size of 200. The horizontal dashed line shows the 99 percentile under the null hypothesis of no recombination, estimated with parametric bootstrapping. The true mosaic structure of the sequence alignment is as follows. Ancient recombination event: sites 1000–1500; recent recombination event: sites 2500–3000; differently diverged region: sites 4000–4500.

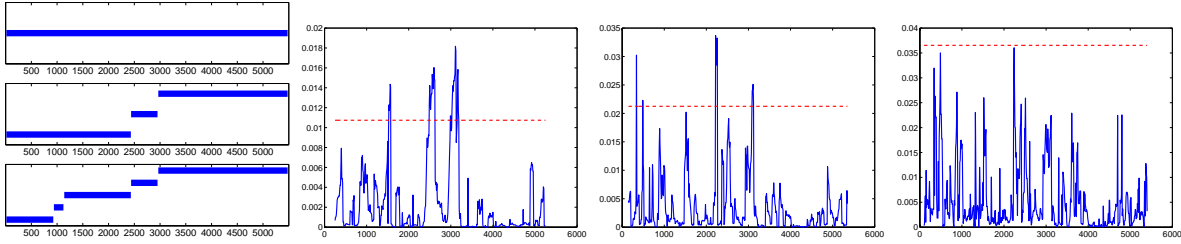


Figure 9: **Detection of recombination in the second synthetic DNA sequence alignment: alternative methods.** *Left:* Prediction obtained with Recpars. The three panels show segmentations predicted for different recombination versus nucleotide substitution cost ratios Ψ . Top: $\Psi = 20$; middle: $\Psi = 10$; bottom: $\Psi = 2$. *Centre left:* DSS signal obtained with a window size of 500. *Centre right:* DSS signal obtained with a window size of 300. *Right:* DSS signal obtained with a window size of 200. The horizontal dashed line shows the 99 percentile under the null hypothesis of no recombination, estimated with parametric bootstrapping. The true mosaic structure of the sequence alignment is as follows. Ancient recombination event: sites 1000–1500; recent recombination event: sites 2500–3000; differently diverged region: sites 4000–4500.

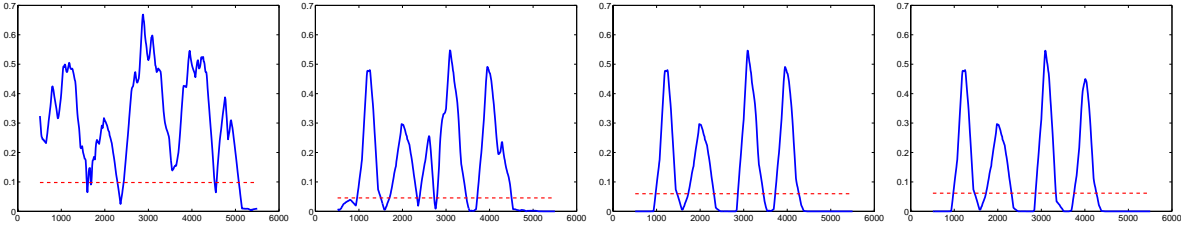


Figure 10: **Detection of recombination in the third synthetic DNA sequence alignment with the PDM method.** The graphs show probabilistic divergence signals; the dashed horizontal lines show the 99 percentiles under the null hypothesis of no recombination, computed with parametric bootstrapping based on the Markov model of Figure 5. The window size of the PDM method was set to 500. From left to right: no pruning (leading to 361 tree topologies in total), pruning down to 7, 5, and 3 clusters. The sequence alignment contains the following true mosaic structure: an ancient recombination event between sites 1000 and 2000, a recent recombination event between sites 3000 and 4000, and a differently diverged region (factor 3) affecting the last 1000 nucleotides.

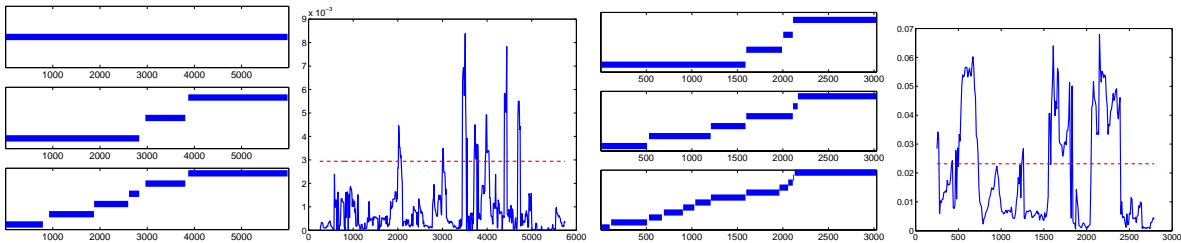


Figure 11: **Prediction of recombination in the third simulated sequence alignment and the Hepatitis B virus alignment with alternative methods.** *Left:* Recpars applied to the third simulated sequence alignment. The three panels show segmentations predicted for different recombination versus nucleotide substitution cost ratios Ψ . Top: $\Psi = 10$; middle: $\Psi = 5$; bottom: $\Psi = 3$. *Centre left:* DSS signal obtained from the third simulated sequence alignment. The horizontal dashed line shows the 99 percentile under the null hypothesis of no recombination, estimated with parametric bootstrapping. For a location of the true recombinant regions, see the caption of Figure 10. *Centre right:* Recpars applied to the Hepatitis-B virus sequence alignment. Top: $\Psi = 20$; middle: $\Psi = 10$; bottom: $\Psi = 5$. *Right:* DSS signal obtained from the Hepatitis-B virus sequence alignment, with the 99 percentile under the null hypothesis shown as a horizontal dashed line.

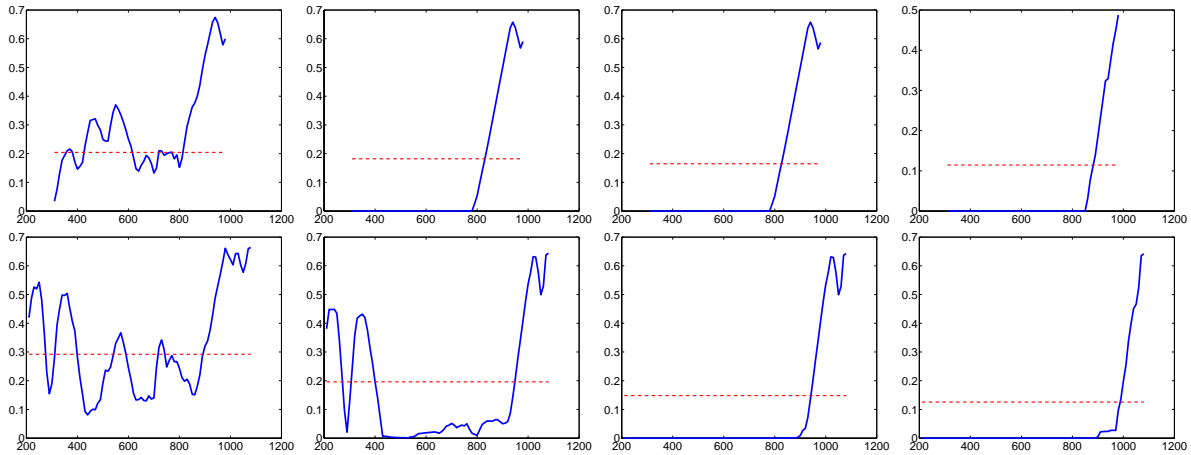


Figure 12: **Gene conversion in maize actin genes: prediction with the PDM method.** The graphs show probabilistic divergence signals; the dashed horizontal lines show the 99 percentiles under the null hypothesis of no recombination, computed with parametric bootstrapping based on the Markov model of Figure 5. *Top*: Window size 300. *Bottom*: Window size 200. The various columns refer to different pruning thresholds. From left to right: no pruning, pruning down to 7, 5, and 3 clusters. No pruning led to a support of 52 (window size 300) and 136 (window size 200) topologies.

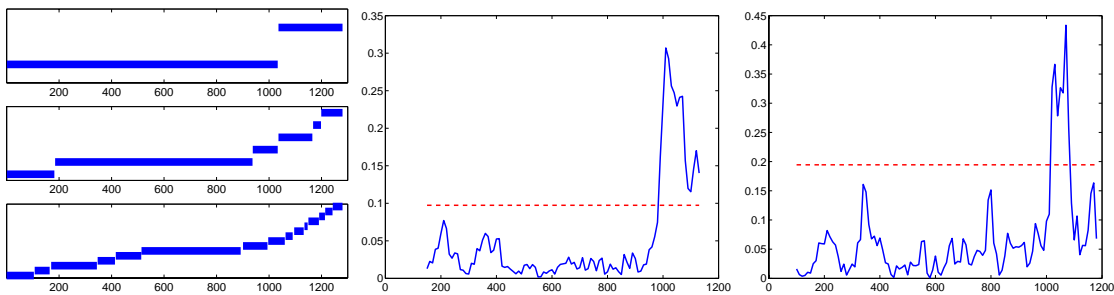


Figure 13: **Gene conversion in maize actin genes: prediction with alternative methods.** *Left*: Prediction with Recpars. The three panels show segmentations predicted for different recombination versus nucleotide substitution cost ratios Ψ . Top: $\Psi = 10$; middle: $\Psi = 5$; bottom: $\Psi = 3$. *Centre*: DSS signal obtained with a window size of 300. *Right*: DSS signal obtained with a window size of 200. The horizontal dashed line shows the 99 percentile under the null hypothesis of no recombination, estimated with parametric bootstrapping.

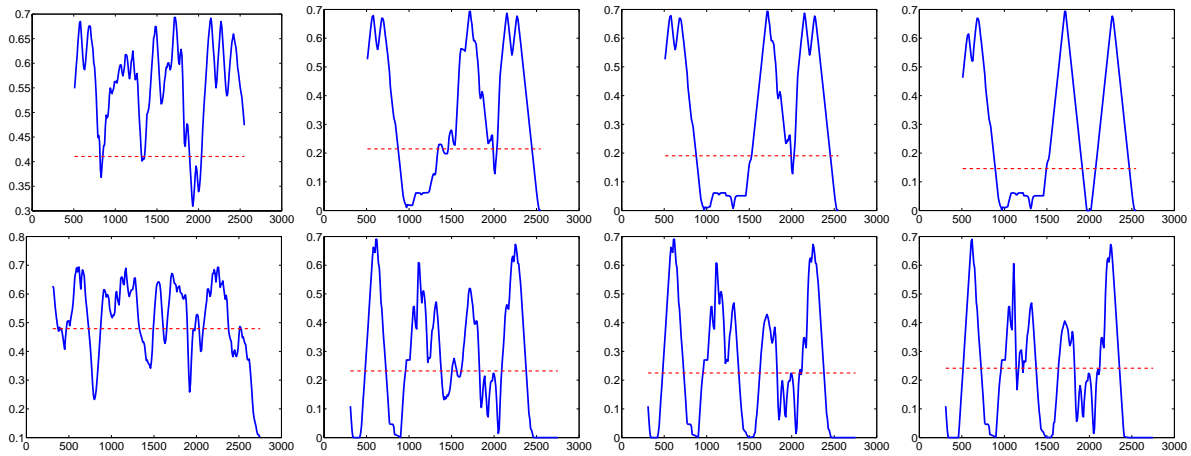


Figure 14: **Recombination in Hepatitis B virus: prediction with the PDM method.** The graphs show probabilistic divergence signals; the dashed horizontal lines show the 99 percentiles under the null hypothesis of no recombination, computed with parametric bootstrapping based on the Markov model of Figure 5. *Top:* Window size 500. *Bottom:* Window size 300. The various columns refer to different pruning thresholds. From left to right: no pruning, pruning down to 7, 5, and 3 clusters. No pruning led to a support of 126 (window size 500) and 439 (window size 300) topologies.