Addressing the shortcomings of three recent Bayesian methods for detecting interspecific recombination in DNA sequence alignments

Dirk Husmeier[a] and **Alexander Mantzaris**[a,b]
[a] Biomathematics and Statistics Scotland (BioSS)
Edinburgh, United Kingdom
[b]School of Informatics
University of Edinburgh, United Kingdom

Preprint of an article to appear in SAGMB (Statistical Applications in Molecular Biology and Evolution), October 2008

Keywords

Phylogenetics, recombination, Bayesian modelling, Markov chain Monte Carlo, change-point model, hidden Markov model, Felsenstein zone, consistency, incidental parameters.

Abstract

We address a potential shortcoming of three probabilistic models for detecting interspecific recombination in DNA sequence alignments: the multiple change-point model (MCP) of Suchard *et al.* (2003), the dual multiple change-point model (DMCP) of Minin *et al.* (2005), and the phylogenetic factorial hidden Markov model (PFHMM) of Husmeier (2005). These models are based on the Bayesian paradigm, which requires the solution of an integral over the space of branch lengths. To render this integration analytically tractable, all three models make the same assumption that the vectors of branch lengths of the phylogenetic tree are independent among sites. While this approximation reduces the computational complexity considerably, we show that it leads to the systematic prediction of spurious topology changes in the Felsenstein zone, that is, the area in the branch lengths configuration space where maximum parsimony consistently infers the wrong topology due to long-branch attraction. We apply two Bayesian hypothesis tests, based on an inter- and an intra-model approach to estimating the marginal likelihood. We then propose a revised model that addresses these shortcomings, and compare it with the aforementioned models on a set of synthetic DNA sequence alignments systematically generated around the Felsenstein zone.

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 sub-sequences 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 taxa.

The present work is related to three recent Bayesian methods for detecting recombination in DNA sequence alignments: the multiple change-point model (MCP) of Suchard *et al.* (2003), the dual multiple change-point model (DMCP) of Minin *et al.* (2005), and the phylogenetic factorial hidden Markov model (PFHMM) of Husmeier (2005). The idea underlying the MCP is to segment the DNA sequence alignment by the insertion of change points, and to infer different phylogenetic trees and nucleotide substitution rates for the separate segments thus obtained. Inference is carried out in a Bayesian way. Of particular interest are the number and locations of the change points, which mark putative recombination breakpoints. Starting from a truncated Poisson prior, the number of change points is sampled from the posterior distribution with reversible jump (RJ) Markov chain Monte Carlo (MCMC). A disadvantage of this approach is the inability of the model to distinguish between recombination and rate heterogeneity. This shortcoming is addressed in the DMCP, where two separate change-point processes associated with the phylogenetic tree topology and the nucleotide substitution rate are employed. A related but different modelling paradigm is provided by the PFHMM, where two a priori independent hidden Markov chains are introduced, whose states represent the tree topology and nucleotide substitution rate, respectively. The three models described above have one feature in common: different sites in the sequence alignment are associated with separate branch lengths, which allows the latter to be integrated out analytically. This is convenient, as the marginal likelihood of the tree topology, the nucleotide substitution rate, and further parameters of the nucleotide substitution model (like the transition- transversion ratio) can be computed in closed from. In this way, the computational complexity of sampling break points (MCP,DMCP) or hidden state sequences (PFHMM) from the posterior distribution with MCMC is substantially reduced. The subject of the present work is to investigate the effect of the approximation on which the analytic integration of the branch lengths is based. We will demonstrate that as a consequence of this approximation, the resulting model may predict spurious topology changes. A clearer analysis of the underlying approximation reveals that the resulting model exhibits a behaviour very similar to

maximum parsimony, and that it is intrinsically susceptible to the systematic failure in the Felsenstein zone (Felsenstein, 1978). We propose a modification of the PFHMM without the aforementioned distributional approximation for the branch lengths. This modification increases the computational complexity of the inference scheme, as the branch lengths have now to be numerically sampled from the posterior distribution. However, we demonstrate that the resulting model will avoid the prediction of spurious topology changes in the Felsenstein zone, and thereby increases the accuracy of detecting recombination in DNA sequence alignments.

2 Methods

Consider an alignment $\boldsymbol{\mathcal{D}}$ of m DNA sequences, N nucleotides long. Let each column in the alignment be represented by $\mathbf{y}_t$, where the subscript t represents the site, $1 \leq t \leq N$. Hence $\mathbf{y}_t$ is an m-dimensional column vector containing the nucleotides at the t site of the alignment, and $\boldsymbol{\mathcal{D}} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. Given a probabilistic model of nucleotide substitutions based on a homogeneous Markov chain with instantaneous rate matrix $\mathbf{Q}$, a phylogenetic tree topology S, and a vector of branch lengths $\mathbf{w}$, the probability of each column $\mathbf{y}_t$, $P(\mathbf{y}_t|S, \mathbf{w}, \boldsymbol{\theta})$, can be computed, as e.g. discussed in Husmeier *et al.* (2005a). Here, $\boldsymbol{\theta}$ denotes a (vector) of free nucleotide substitution parameters extracted from $\mathbf{Q}$. For instance, for the HKY85 model of Hasegawa *et al.* (1985), we have

$$\mathbf{Q} = \begin{pmatrix} . & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & . & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & . & \alpha\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha\pi_C & . \end{pmatrix} \tag{1}$$

where the dot in each row represents the additive inverse of the sum of the remaining elements in that row, $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$, with $\pi_i \in [0, 1]$ and $\sum_i \pi_i = 1$, is a vector of nucleotide equilibrium frequencies, and $\alpha, \beta \geq 0$ are separate nucleotide substitution rates for transitions and transversions. For identifiability between $\mathbf{w}$ and $\mathbf{Q}$, the constraint $\sum_i Q_{ii}\pi_i = -1$ is commonly introduced, which allows the branch lengths to be interpreted as expected numbers of mutations per site (see, e.g., Minin *et al.* (2005)). The normalization constraint on $\boldsymbol{\pi}$ further reduces the number of free parameters by one, so that without loss of generality we have $\boldsymbol{\theta} = (\pi_A, \pi_C, \pi_G, \tau)$, where $\tau = \alpha/\beta \geq 0$ is the transition-transversion ratio.

A Bayesian approach to phylogenetics without recombination was proposed and tested in Yang and Rannala (1997) and Larget and Simon (1999), where the objective is to sample the tree topology S, the branch lengths $\mathbf{w}$, and the parameters of the nucleotide substitution model, $\boldsymbol{\theta}$, from the posterior distribution $P(\mathbf{w}, S, \boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$ with MCMC. Generalizing this scheme to the presence of recombination requires replacing the single topology-indicating variable S by a sequence of topologies, $\mathbf{S} = (S_1, \ldots, S_N)$, where S_t (the 'state' at site t) represents the tree topology at site t. Each state $S_t \in \{1, \ldots, K\}$ can have a different vector of branch lengths,

$\mathbf{w}_{S_t}$, and nucleotide substitution parameters, $\boldsymbol{\theta}_{S_t}$. To simplify the notation, we introduce the accumulated vectors $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ and define: $P(\mathbf{y}_t|S_t, \mathbf{w}_{S_t}, \boldsymbol{\theta}_{S_t}) = P(\mathbf{y}_t|S_t, \mathbf{w}, \boldsymbol{\theta})$. This means that S_t indicates which subvectors of $\mathbf{w}$ and $\boldsymbol{\theta}$ apply.

Since a tree topology may change as a result of recombination, which corresponds to a transition into another state S_t at the breakpoint t of the affected region, our main objective is the prediction of the state sequence $\mathbf{S} = (S_1, \ldots, S_N)$. This prediction should be based on the posterior probability $P(S_t|\boldsymbol{\mathcal{D}})$, which requires a marginalization over the other states

$$P(S_t|\boldsymbol{\mathcal{D}}) \;=\; \sum_{S_1} \cdots \sum_{S_{t-1}} \sum_{S_{t+1}} \cdots \sum_{S_N} P(\mathbf{S}|\boldsymbol{\mathcal{D}}) \tag{2}$$

and the remaining parameters to be integrating out:

$$P(\mathbf{S}|\boldsymbol{\mathcal{D}}) = \int P(\mathbf{S}, \mathbf{w}, \boldsymbol{\theta}|\boldsymbol{\mathcal{D}}) d\mathbf{w} d\boldsymbol{\theta} \tag{3}$$

Alternatively, if the objective is to detect only the location of recombination breakpoints without explicitly inferring the tree topologies in the different regions of the alignment, then the state sequences become nuisance parameters that have to be marginalized over. In practice this is effected by the introduction of a breakpoint detection operator, $\mathcal{B}$, which is a function of the state sequence, $\mathbf{S}$, and then obtaining the posterior probabilities of the breakpoints by summing over the state sequences:

$$P(\mathcal{B}|\boldsymbol{\mathcal{D}}) \;=\; \sum_{\mathbf{S}} P(\mathcal{B}|\mathbf{S}) P(\mathbf{S}|\boldsymbol{\mathcal{D}})$$

The assumption made for all three models discussed in Section 1 – MCP, DMCP and PFHMM – is that the integral over the branch lengths $\mathbf{w}$ can be solved analytically. We will revisit this point in Section 2.4, after briefly summarizing the main ideas behind the three methods first.

2.1 Multiple change-point model (MCP)

In the MCP model, each state $S_t \in \{1, \ldots, K\}$ in $\mathbf{S} = (S_1, \ldots, S_N)$ represents a different tree topology. A separate vector of nucleotide substitution parameters $\boldsymbol{\theta}_k, k \in \{1, \ldots, K\}$, and an overall divergence hyperparameter $\rho_k, k \in \{1, \ldots, K\}$, is associated with each state. As we will show later, in equation (13) and Section 2.4, the hyperparameter ρ_k defines the prior distribution of the branch lengths. The posterior probability is obtained from Bayes rule

$$P(\mathbf{S}, \boldsymbol{\theta}, \boldsymbol{\rho}, K|\boldsymbol{\mathcal{D}}) \;\propto\; P(\boldsymbol{\mathcal{D}}|\mathbf{S}, \boldsymbol{\theta}, \boldsymbol{\rho}, K) P(\mathbf{S}) P(\boldsymbol{\theta}) P(\boldsymbol{\rho}) P(K) \tag{4}$$

and requires the specification of various prior distributions. Note that the branch lengths $\mathbf{w}$ have been integrated out analytically. The prior on the number of states

K is chosen to be a truncated Poisson distribution. For $P(\boldsymbol{\rho})$ a factorizable prior $P(\boldsymbol{\rho}) = \prod_k P(\rho_k)$ is assumed, where each $P(\rho_k)$ is taken to be an exponential distribution. For $P(\boldsymbol{\theta})$ a similar factorization is made: $P(\boldsymbol{\theta}) = \prod_k P(\boldsymbol{\theta}_k)$. The nucleotide substitution model chosen in Suchard *et al.* (2003) is the HKY85 model of (Hasegawa *et al.*, 1985), where the nucleotide equilibrium frequencies are kept fixed, estimated from the whole DNA sequence alignment. Hence each $\boldsymbol{\theta}_k$ corresponds to a single parameter, the transition-transversion ratio, and $P(\boldsymbol{\theta}_k)$ is chosen to be the exponential distribution again. Finally, a change-point process is chosen as the prior on $P(\mathbf{S})$. The posterior probability over the state assignments is, in principle, obtained by marginalization

$$P(\mathbf{S}|\boldsymbol{\mathcal{D}}) \;=\; \sum_K \int\int P(\mathbf{S}, \boldsymbol{\theta}, \boldsymbol{\rho}, K|\boldsymbol{\mathcal{D}}) d\boldsymbol{\theta} d\boldsymbol{\rho} \tag{5}$$

from which the prediction of topology changes is obtained by further marginalization, e.g. according to equation (2). In practice, the integral in (5) is intractable and is approximated by sampling state sequences $\mathbf{S}$ and model parameters $\boldsymbol{\theta}, \boldsymbol{\rho}$ and K approximately from the posterior distribution of equation (4) with reversible jump Markov chain Monte Carlo (RJMCMC).

2.2 Dual multiple change-point model (DMCP)

A disadvantage of the MCP model is its inability to distinguish between recombination and rate heterogeneity. This shortcoming is addressed in the DMCP model of Minin *et al.* (2005), where two separate change point processes associated with the phylogenetic tree topology and the nucleotide substitution rate are employed. Let $\mathbf{S} = (S_1, \ldots, S_N)$ denote, as before, a hidden state sequence in which each state $S_t \in \{1, \ldots, K\}$ is associated with a phylogenetic tree topology. Denote by $\mathbf{R} = (R_1, \ldots, R_N)$ a separate hidden state sequence in which each hidden state $R_t \in \{1, \ldots, K'\}$ is associated with a divergence hyperparameter $\rho_{k'}, k' \in \{1, \ldots, K'\}$, and a (vector of) nucleotide substitution parameter(s) $\boldsymbol{\theta}_{k'}, k' \in \{1, \ldots, K'\}$. Like $P(\mathbf{S})$, the prior on $\mathbf{R}$, $P(\mathbf{R})$, is chosen to be a change-point process, and both change-point processes are elected to be a priori independent: $P(\mathbf{S}, \mathbf{R}) = P(\mathbf{S})P(\mathbf{R})$. The objective of Bayesian inference is to sample both hidden state sequences from the posterior distribution

$$P(\mathbf{S}, \mathbf{R}|\boldsymbol{\mathcal{D}}) \;=\; \sum_K \sum_{K'} \int\int P(\mathbf{S}, \mathbf{R}, K, K', \boldsymbol{\theta}, \boldsymbol{\rho}|\boldsymbol{\mathcal{D}}) d\boldsymbol{\theta} d\boldsymbol{\rho} \tag{6}$$

which is approximately effected with RJMCMC.

2.3 Phylogenetic factorial hidden Markov model (PFHMM)

The concept of the PFHMM of Husmeier (2005) is similar to the DMCP model. The main difference is the choice of the prior distribution on the hidden state sequences,

$P(\mathbf{S}, \mathbf{R}) = P(\mathbf{S})P(\mathbf{R})$. Rather than using two a priori independent change-point processes, two a priori independent homogeneous Markov chains are used:

$$P(\mathbf{S}) = P(S_1) \prod_{t=1}^{N-1} P(S_{t+1}|S_t) \quad (7)$$

$$P(\mathbf{R}) = P(R_1) \prod_{t=1}^{N-1} P(R_{t+1}|R_t) \quad (8)$$

where the transition probabilities are given by

$$P(S_{t+1}|S_t) = \nu_S \delta_{S_{t+1},S_t} + \frac{1-\nu_S}{K-1}(1 - \delta_{S_{t+1},S_t}) \quad (9)$$

$$P(R_{t+1}|R_t) = \nu_R \delta_{R_{t+1},R_t} + \frac{1-\nu_R}{K'-1}(1 - \delta_{R_{t+1},R_t}) \quad (10)$$

Here, $\delta_{i,j}$ is the Kronecker delta symbol, which is one if $i = j$, and zero otherwise.

Note that the change-point process is a special case of a Markov chain, in which a state can only be visited once, without the possibility of a state reoccurring. This is an unnatural assumption in the context of recombination. When a recombination event has occurred in the central segment of a sequence alignment, then the evolutionary history of this central segment will be different from the flanking regions of the alignment. However, the two flanking regions share the same evolutionary history. This can be modelled with a Markov chain of two states and two transitions: from state 1 into state 2, and back from state 2 into state 1. However, a change-point process does not provide a mechanism to combine the two flanking regions into the same state. To rephrase this in terms of Markov chains: a change-point process corresponds to a Markov chain with two separate states for the two flanking regions, as the re-occurrence of a previously visited state is impossible. Consequently, the model has to infer the identity of the two states from the data. This is suboptimal, and it leads to an increased inference uncertainty (especially for short sequence alignments); see Lehrach (2008) for further details.

There are various differences in the detailed implementation of the methods. For the PFHMM described in Husmeier (2005), the parameters K, K', $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$ are fixed. This allows the computationally expensive RJMCMC simulations to be replaced by a much faster Gibbs sampling procedure. However, this difference is not essential to the PFHMM. In fact, the constraints on the parameters have been relaxed in Lehrach (2008) and Lehrach and Husmeier (2009), where – similarly to the work of Minin *et al.* (2005) – RJMCMC was used.

2.4 Analytic integration over the branch lengths

Consider a phylogenetic tree with topology S and branch lengths $\mathbf{w}$, denote the nucleotide substitution parameters by $\boldsymbol{\theta}$, and assume we are given a single column

$\mathbf{y}$ from a DNA sequence alignment. The probability of this column, $\mathbf{y}$, is given by the following standard form (see, e.g., Husmeier *et al.* (2005a)):

$$P(\mathbf{y}|\mathbf{w}, S, \boldsymbol{\theta}) \;=\; \sum_{hidden} P(\tilde{y}_r) \prod_n P(\tilde{y}_n|\tilde{y}_{pa(n)}, w^{pa(n)\to n}, \boldsymbol{\theta}) \tag{11}$$

Here, $\tilde{y}_n = y_n$ if the node n in the phylogenetic tree is observed (usually a leaf node). Otherwise, $\tilde{y}_n$ is a hidden variable (usually an ancestral node corresponding to a speciation point) that is marginalized over in the sum. The subscript r represents the root note, which for a reversible nucleotide substitution model can be chosen arbitrarily without affecting the probability of $\mathbf{y}$. The length of the branch connecting node n to its parent $pa(n)$ is denoted by $w^{pa(n)\to n}$. The factorization in the expansion of equation (11) is defined by the phylogenetic tree topology S. We are interested in integrating out the branch lengths $\mathbf{w}$ according to

$$P(\mathbf{y}|S, \boldsymbol{\theta}) \;=\; \int P(\mathbf{y}|\mathbf{w}, S, \boldsymbol{\theta}) P(\mathbf{w}) d\mathbf{w} \tag{12}$$

We follow Suchard *et al.* (2003) and put a completely factorizable prior on the vector of branch lengths:

$$P(\mathbf{w}) \;=\; \prod_i P(w^i) \;=\; \frac{1}{\rho} \exp\left(-\frac{w^i}{\rho}\right) \tag{13}$$

where w^i is a single element of $\mathbf{w}$ representing the length of an individual branch connecting two nodes in the phylogenetic tree. Inserting this expression and equation (11) into equation (12) gives:

$$\begin{aligned} P(\mathbf{y}|S, \boldsymbol{\theta}) \;&=\; \int \sum_{hidden} P(\tilde{y}_r) \prod_n P(\tilde{y}_n|\tilde{y}_{pa(n)}, w^{pa(n)\to n}, \boldsymbol{\theta}) P(w^{pa(n)\to n}) d\mathbf{w} \qquad (14) \\ &=\; \sum_{hidden} P(\tilde{y}_r) \prod_n \int P(\tilde{y}_n|\tilde{y}_{pa(n)}, w^{pa(n)\to n}, \boldsymbol{\theta}) P(w^{pa(n)\to n}) dw^{pa(n)\to n} \end{aligned}$$

Recall that $\tilde{y}_n$ and $\tilde{y}_{pa(n)}$ in $P(\tilde{y}_n|\tilde{y}_{pa(n)}, w^{pa(n)\to n}, \boldsymbol{\theta})$ represent nucleotides. The probability of nucleotide X mutating into Z along a branch of length w is of the following general form (Suchard *et al.*, 2003):

$$P(Z|X, w, \boldsymbol{\theta}) \;=\; A_{ZX} \exp(-B_{ZX} w) + C_{ZX} \exp(-D_{XZ} w) + \pi_Z \tag{15}$$

Here, $A_{ZX}, B_{ZX}, C_{ZX}, D_{XZ}$ are nucleotide-dependent constants that are determined by the eigensystem of the instantaneous rate matrix $\mathbf{Q}$ and, thus, depend on the chosen nucleotide substitution model. For the HKY85 model, for instance, the particular expressions can be found in Hasegawa *et al.* (1985). The last term, π_Z, represents the equilibrium frequency of nucleotide Z, which is a parameter of the

nucleotide substitution model. Hence, $A_{ZX}, B_{ZX}, C_{ZX}, D_{XZ}$ and π_Z are determined by $\boldsymbol{\theta}$.

Combining equations (13) and (15) allows the branch length to be integrated out analytically:

$$\begin{aligned} P(Z|X,\rho) &= \int P(Z|X,w)P(w|\rho)dw \\ &= \pi_Z + \frac{A_{ZX}}{\rho}\int \exp\left(-\left[B_{ZX}+\frac{1}{\rho}\right]w\right)dw \\ &\quad + \frac{C_{ZX}}{\rho}\int \exp\left(-\left[D_{XZ}+\frac{1}{\rho}\right]w\right)dw \\ &= \pi_Z + \frac{A_{ZX}}{1+B_{ZX}\rho} + \frac{C_{ZX}}{1+D_{XZ}\rho} \end{aligned} \tag{16}$$

Inserting eq. (16) into eq. (14) gives the following closed-form solution:

$$P(\mathbf{y}|S,\boldsymbol{\theta}) = \sum_{hidden} P(\tilde{y}_r)\prod_n \left(\pi_{\tilde{y}_r} + \frac{A_{\tilde{y}_n,\tilde{y}_{pa(n)}}}{1+B_{\tilde{y}_n,\tilde{y}_{pa(n)}}\rho} + \frac{C_{\tilde{y}_n,\tilde{y}_{pa(n)}}}{1+D_{\tilde{y}_n,\tilde{y}_{pa(n)}}\rho}\right) \tag{17}$$

Let us now consider a whole DNA sequence alignment $\boldsymbol{\mathcal{D}} = (\mathbf{y}_1,\ldots,\mathbf{y}_N)$:

$$P(\boldsymbol{\mathcal{D}}|S,\boldsymbol{\theta}) \;=\; \int P(\boldsymbol{\mathcal{D}}|\mathbf{w},S,\boldsymbol{\theta})P(\mathbf{w})d\mathbf{w} \;=\; \int \prod_{t=1}^{N} P(\mathbf{y}_t|\mathbf{w},S,\boldsymbol{\theta})P(\mathbf{w})d\mathbf{w} \tag{18}$$

It is seen that the independence assumption of equation (13), $P(\mathbf{w}) = \prod_i P(w^i)$, does not yet allow this integral to be solved in closed form. What is needed is the expansion of the parameter space

$$\mathbf{w} \rightarrow (\mathbf{w}_1,\ldots,\mathbf{w}_t,\ldots,\mathbf{w}_N) \tag{19}$$

and the further independence assumption:

$$P(\mathbf{w}_1,\ldots,\mathbf{w}_t,\ldots,\mathbf{w}_N) \;=\; \prod_{t=1}^{N} P(\mathbf{w}_t) \tag{20}$$

Inserting this prior into eq. (18) gives

$$P(\boldsymbol{\mathcal{D}}|S,\boldsymbol{\theta}) \;=\; \prod_{t=1}^{N}\int P(\mathbf{y}_t|\mathbf{w}_t,S,\boldsymbol{\theta})P(\mathbf{w}_t)d\mathbf{w}_t \;=\; \prod_{t=1}^{N} P(\mathbf{y}_t|S,\boldsymbol{\theta}) \tag{21}$$

where $P(\mathbf{y}_t|S,\boldsymbol{\theta})$ is given by (17). The commutation of the integral and the product, which is a direct consequence of equations (19) and (20), allows the integral to be solved in closed form, according to equations (14) and (17). The upshot is that for

the branch lengths to be integrated out analytically, the model has to be modified so as to associate a separate branch length vector $\mathbf{w}_t$ with each position t in the DNA sequence alignment. This model is equivalent to the no-common-mechanism model proposed by Tuffley and Steel (1997). It is important to note that it is not the independence assumption of eq. (13) alone that leads to this simplification, a conclusion one might erroneously draw from Suchard *et al.* (2003). Rather, the more restrictive independence assumption of eq. (20) is needed. As a consequence of the latter independence assumption and the parameter expansion of eq. (19) the model is over-complex, though, with no information sharing between different sites with respect to the branch length estimation. In terms of statistical terminology, the expansion of eq. (19) turns the structural parameters $\mathbf{w}$ into a set of incidental parameters[1]. As discussed in Goldman (1990), this implies that maximum likelihood is no longer guaranteed to provide a consistent estimator. This aspect, which has not been considered in any of the three methods discussed in Section 1 – MCP, DMCP and PFHMM – causes inconsistency problems that are related to those found in maximum parsimony. We will investigate them more closely in the subsequent sections.

3 Data

We suspect that the assumption of independent site-specific branch lengths, as discussed in Section 2.4, could lead to inconsistency problems akin to those that affect maximum parsimony (Felsenstein, 1978). To test this conjecture, we tested the models on synthetic DNA sequence alignments. We used two different programs for generating these alignments: SEQGEN (Rambaut, 1996) and the MATLAB programs used in Husmeier (2005). In both cases we simulated the nucleotide substitution processes with the HKY model (Hasegawa *et al.*, 1985), using a transition-transversion ratio of 2 and uniform nucleotide equilibrium frequencies. For SEQGEN, we used the implementation available via the web service provided by the Pasteur Institute, available from

$$http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=seqgen.$$

The MATLAB programs used in Husmeier (2005) are available from

$$http://www.bioss.ac.uk/staff/dirk/Supplements/,$$

and were preferred when running a large number of jobs in batch mode. We generated two different types of alignments: homogeneous alignments, and alignments with mosaic structures.

[1] Structural parameters are parameters that appear in the probability distributions of all the observations, whereas an incidental parameter appears in the probability distributions of only a subset of the observations. See Goldman (1990) for further details.

3.1 Homogeneous DNA sequence alignment

A homogeneous DNA sequence alignment is an alignment where one single phylogenetic tree with a specified branch length vector is used in the data generating process. We generated alignments from the 4-taxa tree depicted in Figure 1 for different settings of the branch length configurations, specified by the values $d2$ and $d3$. This corresponds to a study originally carried out by Felsenstein for investigating potential shortcomings and inconsistencies of maximum parsimony (Felsenstein, 1978). We varied the parameters $d2$ and $d3$, defined in Figure 1, over a large range that included the so-called Felsenstein zone, in which maximum parsimony systematically fails. The data thus generated were used for the studies reported in Figures 4, 5, 6, and 8. All sequence alignments were 1000 nucleotides long.

3.2 DNA sequence alignment with mosaic structure

A mosaic structure is a DNA sequence alignment subject to recombination and/or rate heterogeneity, where a segment in the DNA sequence alignment was generated from a tree with a different tree topology, or with different branch lengths. We generated DNA sequence alignments from the 4-taxa tree shown in Figure 1. The alignments were 1500 nucleotides long. They contained a central segment of 500 nucleotides, which was generated from a tree with the same topology as for the flanking regions, but with a different branch length configuration. The objective of our study was to investigate if spurious tree topology changes were inferred with the recombination detection methods described in Section 2 if the branch length configurations for the central and flanking regions were on different sides of the Felsenstein boundary. The alignments thus generated were used in the study described in the caption of Figure 9.

4 Bayesian model selection

As discussed in Section 2.4, the integration over the branch lengths, on which the three methods MCP, DCMP and PFHMM rely, is based on the choice of independent site-specific branch lengths, that is, the vector of branch lengths is allowed to be different at each site. Since separate branch length vectors $\mathbf{w}_t$ are associated with different positions t in the alignment, there is no longer a mechanism in place to over-rule a posteriori the prior independence assumption of equation (13). We suspect that MCP, DCMP and PFHMM might therefore be susceptible to the same inconsistency problems as the method of maximum parsimony, which could result in the prediction of spurious topology changes. Before investigating this conjecture in direct simulation studies, to be discussed in Section 5.2, we carried out systematic Bayesian model selection along the Felsenstein zone (introduced in Felsenstein (1978)). To this end, we generated data synthetically from the four-taxa tree of Figure 1 with two types of branches, $d2$ and $d3$, as described in Section 3.

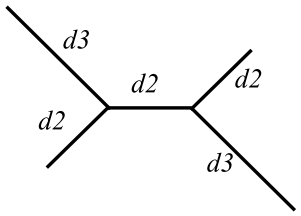


Figure 1: Phylogenetic tree of four taxa
The figure shows a phylogenetic tree of four taxa, which was used for generating the synthetic DNA sequence alignments, as described in Section 3. The tree contains two types of branch lengths, denoted by $d2$ and $d3$, as in Felsenstein (1978). For configurations with large branch lengths $d3$ and small branch lengths $d2$, the method of maximum parsimony is known to systematically infer the wrong tree topology.

For the different $d2/d3$ ratios, we estimated the marginal likelihood $P(\boldsymbol{\mathcal{D}}|S, \mathcal{H}_i)$ for each of the three possible tree topologies, $S \in \{\Psi_1, \Psi_2, \Psi_3\}$, under the two hypotheses or modelling approaches: independent site-specific branch length vectors $\mathbf{w}_t$ ($\mathcal{H}_0$), and a common vector of branch lengths $\mathbf{w}$ for the whole alignment ($\mathcal{H}_1$). Under the assumption of a uniform prior on the tree topologies, we estimated the posterior probability for the correct tree topology

$$P(S = true|\boldsymbol{\mathcal{D}}, \mathcal{H}_i) \;=\; \frac{P(\boldsymbol{\mathcal{D}}|S = true, \mathcal{H}_i)}{\sum_{k=1}^{3} P(\boldsymbol{\mathcal{D}}|S = \Psi_k, \mathcal{H}_i)} \tag{22}$$

We investigated the behaviour of $P(S = true|\boldsymbol{\mathcal{D}}, \mathcal{H}_i)$ in $d2/d3$ space, especially around the Felsenstein zone. For estimating the marginal likelihood, we pursued two approaches: an inter-model approach, using MCMC, and an intra-model approach, using the method of annealed importance sampling (AIS), as proposed in Neal (2001). Below, in Sections 4.1 and 4.2, we will first define the exact form of the probabilistic models associated with $\mathcal{H}_0$ and $\mathcal{H}_1$. We will then, in Sections 4.3 and 4.4, briefly describe the way we computed the marginal likelihoods. Finally, we will present the results and investigate the behaviour of the two modelling frameworks around the Felsenstein zone.

4.1 Independent site-specific branch-length model $\mathcal{H}_0$

To investigate whether there are potential inconsistency problems inherent in the recombination detection methods MCP, DMCP and PFHMM, we considered the standard phylogenetic model (without recombination) subject to the same independence assumptions of equations (13) and (20) on which MCP, DMCP and PFHMM are based. We refer to this modelling concept as $\mathcal{H}_0$. The corresponding graphical model is shown in Figure 2. The right panel depicts the site-independence of the branch lengths, inherent in equation (20). As a consequence of the analytic

integration over the branch lengths, discussed in Section 2.4, the model simplifies. The resulting probabilistic graphical model is shown in Figure 2a, which defines the following factorization:

$$P(\boldsymbol{\mathcal{D}}, S, \rho, \alpha) = P(\boldsymbol{\mathcal{D}}|S, \rho)P(S)P(\rho|\alpha) \tag{23}$$

Here, $\boldsymbol{\mathcal{D}}$ is the DNA sequence alignment, S is the tree topology, ρ (defined in equation (13)) represents the average mutational divergence, and α is a hyperparameter that defines the prior distribution of ρ:

$$P(\rho|\alpha) \;=\; \frac{1}{\alpha}e^{-\frac{\rho}{\alpha}} \tag{24}$$

The prior distribution over tree topologies, $P(S)$, is chosen to be uniform. The objective of Bayesian model selection for learning the best tree topology S is to estimate the marginal likelihood

$$P(\boldsymbol{\mathcal{D}}|S, \alpha) = \int P(\boldsymbol{\mathcal{D}}|S, \rho)P(\rho|\alpha)d\rho, \tag{25}$$

which is the numerator in the model selection equation (22). The term $P(\boldsymbol{\mathcal{D}}|S, \rho)$ is given in (21), where the explicit reference to $\mathcal{H}_0$ and the nucleotide substitution parameters $\boldsymbol{\theta}$ has been left out to reduce the notational complexity.[2]

4.2 Standard phylogenetic model $\mathcal{H}_1$

For comparison with $\mathcal{H}_0$, we consider the standard phylogenetic model, in which a common vector of branch lengths $\mathbf{w}$ is used for the whole DNA sequence alignment, as depicted in Figure 3b. We refer to this modelling concept as $\mathcal{H}_1$. The essential difference from $\mathcal{H}_0$ is that the independence assumption of equation (20), on which MCP, DMCP and PFHMM are based, is no longer valid. The consequence is that the elimination of the branch lengths, as described in Section 2.4 and represented in Figure 2a, is no longer feasible, resulting in the more complex probabilistic dependence model of Figure 3a. The structure of the model incorporates the average mutational divergence ρ, and the branch length vector $\mathbf{w}$, and the joint probability factorizes as follows:

$$P(\boldsymbol{\mathcal{D}}, S, \mathbf{w}, \rho, \alpha) = P(\boldsymbol{\mathcal{D}}|S, \mathbf{w})P(\mathbf{w}|\rho)P(S)P(\rho|\alpha) \tag{26}$$

$P(S)$ is the prior distribution over tree topologies, which we keep uniform. The prior distribution over branch lengths, $P(\mathbf{w}|\rho)$, is defined in equation (13). This distribution depends on the hyperparameter ρ, which is given the prior distribution

[2]Recall that the free nucleotide substitution parameters $\boldsymbol{\theta}$ of the HKY model are the equilibrium frequencies and the transition-transversion ratio. In our simulations, we chose a uniform distribution for the equilibrium frequencies, and a fixed transition-transversion ratio of 2. Also, note that in (21) the explicit reference to the hyperparameter ρ has been dropped for notational convenience.

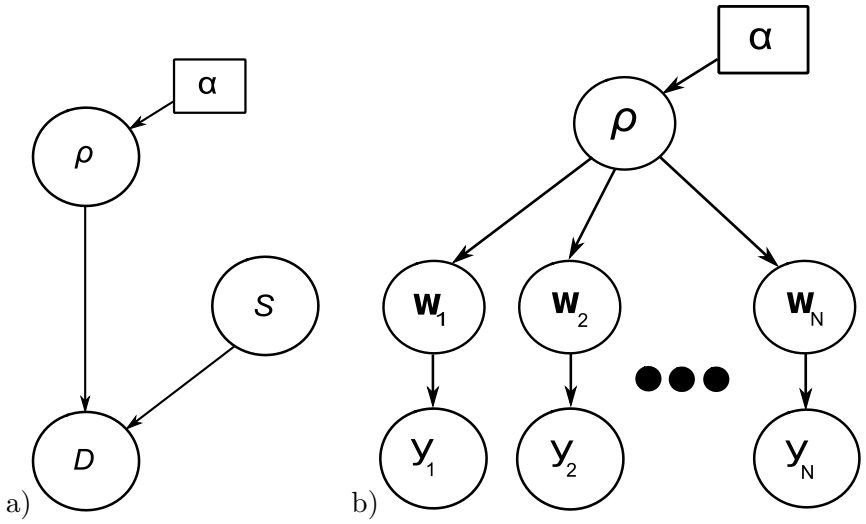


Figure 2: Graphical model for hypothesis $\mathcal{H}_0$

Hypothesis $\mathcal{H}_0$ is based on the independence assumption of equation (20), which is depicted in Panel b). Here, the $\mathbf{y}_t$ represent the columns in the DNA sequence alignment, the $\mathbf{w}_t$'s are separate independent vectors of branch lengths, associated with the sites t in the alignment, and ρ is a hyperparameter determining the prior distribution over the branch lengths, via equation (13). As a consequence of the independence assumptions inherent in this model, the branch lengths can be integrated out, as described in Section 2.4. The resulting probabilistic graphical model is shown in Panel a). Here D (which is equal to $\boldsymbol{\mathcal{D}}$ in the text) is the DNA sequence alignment, S is the tree topology, ρ (defined in equation (13)) represents the average mutational divergence, and α is a hyperparameter that defines the prior distribution of ρ; see equation (24). Further details are given in Section 4.1.

of equation (24). The objective of Bayesian model selection is to estimate the marginal likelihood

$$P(\boldsymbol{\mathcal{D}}|S,\alpha) = \int P(\boldsymbol{\mathcal{D}}|S,\mathbf{w})P(\mathbf{w}|\rho)P(\rho|\alpha)d\rho d\mathbf{w} \tag{27}$$

Here, $P(\boldsymbol{\mathcal{D}}|S,\mathbf{w})$ is the (non-marginal) likelihood, which is obtained from $P(\mathbf{y}_t|S,\mathbf{w})$ defined in equation (11) as follows:

$$P(\boldsymbol{\mathcal{D}}|S,\mathbf{w}) \;=\; \prod_{t=1}^{N} P(\mathbf{y}_t|S,\mathbf{w}) \tag{28}$$

Note that in order to simplify the notation and the graphical presentation, we have not made the dependence on the nucleotide substitution parameters $\boldsymbol{\theta}$ explicit in equation (28) and Figures 2-3.

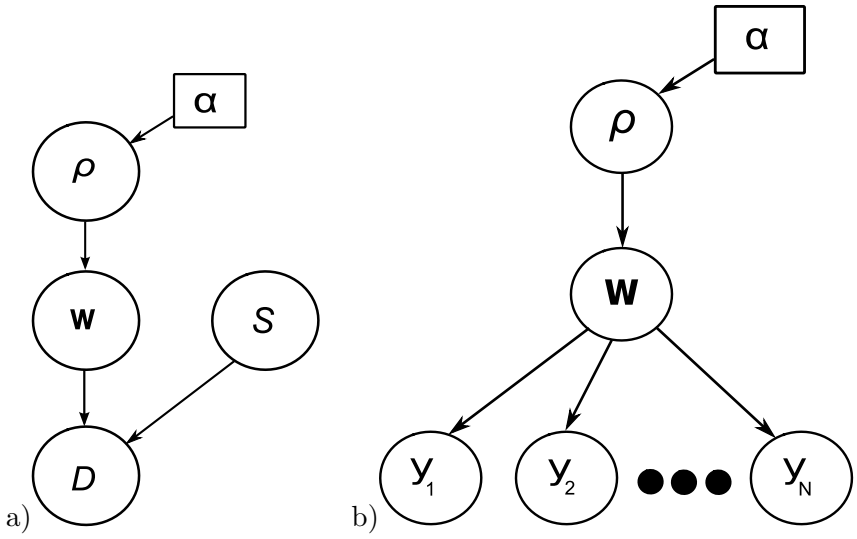


Figure 3: Graphical model for hypothesis $\mathcal{H}_1$
The symbols are the same as those defined in Figure 2. Panel b) shows that a common vector of branch lengths $\mathbf{w}$ is used to describe the whole DNA sequence alignment, rather than independent site-specific vectors, as in Figure 2b. The consequence is that the elimination of the branch lengths, as described in Section 2.4 and represented in Figure 2a, is no longer feasible, resulting in the more complex probabilistic dependence model of Panel a). Further details are given in Section 4.2.

4.3 Inter-model approach: Markov chain Monte Carlo (MCMC)

The objective of the inter-model approach is to sample tree topologies from the posterior distribution of equation (22).

4.3.1 MCMC framework for hypothesis $\mathcal{H}_0$

Recall that for a DNA sequence alignment with four sequences, there are three different unrooted tree topologies. Our proposal distribution for proposing a new tree topology S^* from the current topology S is just the uniform distribution over the tree topology space. For proposing a new rate[3] ρ^*, we sample a value $\rho^\sharp$ from the uniform interval of length W centred on the current value ρ, using reflection to ensure the proposed value is positive: $\rho^* = \rho^\sharp$ if $\rho^\sharp \geq 0$; otherwise $\rho^* = -\rho^\sharp$. This proposal distribution depends on a tuning parameter W, which is adjusted during the burn-in period to achieve a target acceptance rate between 30% and 70%. The Metropolis-Hastings acceptance probability for this move is:

$$a(S^*, \rho^*|S, \rho) = \min\left\{1, \frac{Q\left(\rho|\rho^*\right) P(\rho^*|\alpha) Q(S|S^*) P(S^*) P(\boldsymbol{\mathcal{D}}|S^*, \rho^*)}{Q\left(\rho^*|\rho\right) P(\rho|\alpha) Q(S^*|S) P(S) P(\boldsymbol{\mathcal{D}}|S, \rho)}\right\} \qquad (29)$$

[3]In a slight abuse of terminology, we henceforth refer to the hyperparameter ρ as the "rate".

where $P(\boldsymbol{\mathcal{D}}|S,\rho)$ is the likelihood, defined in equation (21), $P(S)$ and $P(\rho|\alpha)$ are the prior distributions for the tree topology and the rate, defined in Section 4.1, and $Q(S^*|S)$ and $Q(\rho^*|\rho)$ are the proposal distributions, as discussed above. It is straightforward to show that the latter distributions are symmetric and thus cancel out. The prior distribution in tree topology space, $P(S)$, is uniform. Equation (29) thus simplifies as follows:

$$a(S^*,\rho^*|S,\rho) = \min\left\{1, \frac{P(\rho^*|\alpha)P(\boldsymbol{\mathcal{D}}|S^*,\rho^*)}{P(\rho|\alpha)P(\boldsymbol{\mathcal{D}}|S,\rho)}\right\} \tag{30}$$

To increase the acceptance probabilities, we de-couple the proposal step into two separate steps for proposing a new tree topology and a new rate, with the following acceptance probabilities:

$$a(S^*|S) = \min\left\{1, \frac{P(\boldsymbol{\mathcal{D}}|S^*,\rho)}{P(\boldsymbol{\mathcal{D}}|S,\rho)}\right\}, \tag{31}$$

$$a(\rho^*|\rho) = \min\left\{1, \frac{P(\rho^*|\alpha)P(\boldsymbol{\mathcal{D}}|S,\rho^*)}{P(\rho|\alpha)P(\boldsymbol{\mathcal{D}}|S,\rho)}\right\}. \tag{32}$$

4.3.2 MCMC framework for hypothesis $\mathcal{H}_1$

Recall that for Hypothesis $\mathcal{H}_1$ the analytic integration over the branch lengths $\mathbf{w}$ is no longer tractable; hence, the sampling of a new vector of branch lengths $\mathbf{w}^*$ from the existing branch lengths $\mathbf{w}$ has to be incorporated into the MCMC scheme. We elected to propose new values $w_i^\sharp$ independently from a Cauchy distribution centred on the current value w_i

$$Q(w_i^\sharp|w_i,\gamma) = \frac{1}{\pi\gamma\left(1+\left(\frac{w_i^\sharp - w_i}{\gamma}\right)\right)} \tag{33}$$

subject to the constraint that the proposed new branch length w_i^* must be non-negative. Again, this constraint is achieved by reflection: $w_i^* = w_i^\sharp$ if $w_i^\sharp \geq 0$; otherwise $w_i^* = -w_i^\sharp$. The spread of the proposal distribution is defined by the tuning parameter γ, which is adjusted during the burn-in phase to achieve an average acceptance rate between 30% and 70%. The proposal distributions for the tree topology S and the rate ρ are the same as discussed in the previous subsection. The Metropolis-Hastings acceptance probability is given by

$$a(S^*,\rho^*,\mathbf{w}^*|S,\rho,\mathbf{w}) = \min\{1,r\} \tag{34}$$

$$r = \frac{Q(\rho|\rho^*)\,P(\rho^*|\alpha)\left(\prod_{i=1}^K Q(w_i|w_i^*)P(w_i^*|\rho)\right)Q(S|S^*)P(S^*)P(\boldsymbol{\mathcal{D}}|S^*,\mathbf{w}^*)}{Q(\rho^*|\rho)\,P(\rho|\alpha)\left(\prod_{i=1}^K Q(w_i^*|w_i)P(w_i|\rho)\right)Q(S^*|S)P(S)P(\boldsymbol{\mathcal{D}}|S,\mathbf{w})}$$

where $K = \dim\{\mathbf{w}\}$, $Q(w_i^*|w_i)$ is the proposal distribution for a new branch length, which is straightforward to compute from equation (33) and the condition of reflection, $P(w_i^*|\rho)$ is the prior distribution of the branch lengths, defined in equation (13). $P(\boldsymbol{\mathcal{D}}|S, \mathbf{w})$ is defined in equation (28). The other expressions are the same as defined below equation (29) in the previous subsection.

It is straightforward to show that the proposal distribution for the branch lengths, $Q(w_i^*|w_i)$, is symmetric and thus cancels out. Together with the simplifications discussed below equation (29) we get the following simplified expression:

$$a(S^*, \rho^*|S, \rho) = \min\left\{1, \frac{P(\rho^*|\alpha)\prod_{i=1}^{K} P(w_i^*|\rho)P(\boldsymbol{\mathcal{D}}|S^*, \mathbf{w}^*)}{P(\rho|\alpha)\prod_{i=1}^{K} P(w_i|\rho)P(\boldsymbol{\mathcal{D}}|S, \mathbf{w})}\right\} \tag{35}$$

As with the model discussed in the previous subsection, we de-couple the individual update steps so as to increase the acceptance probability:

$$a(\mathbf{w}^*|\mathbf{w}) = \min\left\{1, \frac{\prod_{i=1}^{K} P(w_i^*|\rho)P(\boldsymbol{\mathcal{D}}|S, w_i^*)}{\prod_{i=1}^{K} P(w_i|\rho)P(\boldsymbol{\mathcal{D}}|S, w_i)}\right\} \tag{36}$$

$$a(\rho^*|\rho) = \min\left\{1, \frac{P(\rho^*|\alpha)\prod_{i=1}^{K} P(w_i|\rho^*)}{P(\rho|\alpha)\prod_{i=1}^{K} P(w_i|\rho)}\right\} \tag{37}$$

$$a(S^*|S) = \min\left\{1, \frac{P(\boldsymbol{\mathcal{D}}|\mathbf{w}, S^*)}{P(\boldsymbol{\mathcal{D}}|\mathbf{w}, S)}\right\} \tag{38}$$

4.3.3 Convergence diagnostics

The Gelman and Rubin diagnostic test (Gelman and Rubin, 1992) was used in the simulations for both models to investigate whether the chains have converged. The tests output a (set of) so-called potential scale reduction factor(s) (PSRF), where a value close to 1 provides a strong indication of convergence. We computed the PSRF from the branch lengths and the rate hyperparameter ρ, and chose burn-in and simulation lengths that led to PSRFs below 1.1. This was effected with the following settings. For $\mathcal{H}_0$, we carried out 2K burn-in and 5K sampling steps. For $\mathcal{H}_1$, these values had to be slightly increased (owing to the larger dimension of the parameter space), to 5K burn-in and 10K sampling steps.

4.4 Intra-model approach: Annealed importance sampling (AIS)

As an alternative to the inter-model MCMC sampling scheme discussed in the previous section, we consider an intra-model approach, where the objective is a direct (approximate) computation of the marginal likelihood

$$P(\boldsymbol{\mathcal{D}}|S) = \int P(\boldsymbol{\mathcal{D}}|\boldsymbol{\phi}, S)P(\boldsymbol{\phi}|S)d\boldsymbol{\phi} \tag{39}$$

where $\boldsymbol{\phi}$ is the vector of all parameters associated with the respective hypothesis: $\boldsymbol{\phi} = \rho$ under the site-independent branch length hypothesis $\mathcal{H}_0$, and $\boldsymbol{\phi} = (\mathbf{w}, \rho)$ for $\mathcal{H}_1$. In principle one could approximate the marginal likelihood by

$$P(\boldsymbol{\mathcal{D}}|S) \;\approx\; \frac{1}{N}\sum_{t=1}^{N} P(\boldsymbol{\mathcal{D}}|\boldsymbol{\phi}_t, S) \tag{40}$$

where $\{\boldsymbol{\phi}_t\}$ is a sample from the prior distribution $P(\boldsymbol{\phi}|S)$. However, the convergence of this estimator is known to be poor unless the prior and posterior distributions are very similar (Raftery, 1996). Alternatively, one could exploit the Bayesian identity $P(\boldsymbol{\mathcal{D}}|\boldsymbol{\phi}, S)P(\boldsymbol{\phi}|S) = P(\boldsymbol{\phi}|\boldsymbol{\mathcal{D}}, S)P(\boldsymbol{\mathcal{D}}|S)$ and compute the marginal likelihood from the so-called harmonic mean estimator (Raftery, 1996)

$$\frac{1}{P(\boldsymbol{\mathcal{D}}|S)} \;\approx\; \frac{1}{N}\sum_{t=1}^{N} \frac{1}{P(\boldsymbol{\mathcal{D}}|\boldsymbol{\phi}_t, S)} \tag{41}$$

using a sample $\{\boldsymbol{\phi}_t\}$ from the posterior distribution $P(\boldsymbol{\phi}|\boldsymbol{\mathcal{D}}, S)$. This estimator is known to be numerically unstable, since for modestly informative priors the main contributions to the sum on the right-hand side of equation (41) come from the tail rather than the bulk of the posterior distribution. The standard approach to deal with these problems is to use importance sampling. Define some (possibly unnormalized) distribution $Q(\boldsymbol{\phi})$, and rewrite equation (39) in the form:

$$\frac{P(\boldsymbol{\mathcal{D}}|S)}{Z_Q} \;=\; \int \frac{P(\boldsymbol{\mathcal{D}}|\boldsymbol{\phi}, S)P(\boldsymbol{\phi}|S)}{Q(\boldsymbol{\phi})} \frac{Q(\boldsymbol{\phi})}{Z_Q} d\boldsymbol{\phi} \tag{42}$$

where $Z_Q = \int Q(\boldsymbol{\phi})d\boldsymbol{\phi}$. Provided $Q(\boldsymbol{\phi}) \neq 0$ whenever $P(\boldsymbol{\mathcal{D}}|\boldsymbol{\phi}, S)P(\boldsymbol{\phi}|S) \neq 0$, we get the following unbiased and consistent estimator of the marginal likelihood (Neal, 2001):

$$\frac{P(\boldsymbol{\mathcal{D}}|S)}{Z_Q} \;\longleftarrow\; \frac{1}{N}\sum_{t=1}^{N} c_t \tag{43}$$

where $\{\boldsymbol{\phi}_t\}$ is a sample drawn from $\frac{Q(\boldsymbol{\phi})}{Z_Q}$, and the weights c_t are defined as $c_t = \frac{P(\boldsymbol{\mathcal{D}}|\boldsymbol{\phi}_t, S)P(\boldsymbol{\phi}_t|S)}{Q(\boldsymbol{\phi}_t)}$. Rather than using some fixed distribution $Q(\boldsymbol{\phi})$ as a compromise between the prior and the posterior distribution, as in Raftery (1996), we follow the annealed importance sampling (AIS) scheme proposed in Neal (2001), where the idea is to propose new values $\{\boldsymbol{\phi}_t\}$ by gradually transforming the prior into the posterior distribution. Define

$$Q_m(\boldsymbol{\phi}) \;=\; P(\boldsymbol{\phi}|S)^{[1-\beta_m]} P(\boldsymbol{\phi}|\boldsymbol{\mathcal{D}}, S)^{\beta_m} \tag{44}$$

where $1 = \beta_0 > \beta_1 > \ldots > \beta_M = 0$. That is, Q_M is equal to the prior, and Q_0 is equal to the posterior distribution. AIS produces a sample of parameter vectors $\{\boldsymbol{\phi}_t\}$ and associated weights $\{c_t\}$ according to the following procedure. Consider a Markov chain transition defined by $T_m(\mathbf{x}'|\mathbf{x})$ giving the probability of moving from the current state $\mathbf{x}$ to the new state $\mathbf{x}'$. The choice of T_m is decided

by the requirement that it must leave the corresponding probability distribution Q_m in equation (44) invariant, e.g. by satisfying the equation of detailed balance: $T_m(\mathbf{x}'|\mathbf{x})Q_m(\mathbf{x}) = T_m(\mathbf{x}|\mathbf{x}')Q_m(\mathbf{x}')$. Next, a sequence of points is generated as follows:

$$\begin{array}{l} \text{Generate } \mathbf{x}_{M-1} \text{ from } Q_M \\ \text{Generate } \mathbf{x}_{M-2} \text{ from } \mathbf{x}_{M-1} \text{ using } T_{M-1} \\ \ldots \\ \text{Generate } \mathbf{x}_1 \text{ from } \mathbf{x}_2 \text{ using } T_2 \\ \text{Generate } \mathbf{x}_0 \text{ from } \mathbf{x}_1 \text{ using } T_1 \end{array} \tag{45}$$

The proposed parameter vector of the tth iteration is set to $\boldsymbol{\phi}_t = \mathbf{x}_0$, and the associated weight is set to

$$c_t = \frac{Q_{M-1}(\mathbf{x}_{M-1})}{Q_M(\mathbf{x}_{M-1})} \frac{Q_{M-2}(\mathbf{x}_{M-2})}{Q_{M-1}(\mathbf{x}_{M-2})} \cdots \frac{Q_1(\mathbf{x}_1)}{Q_2(\mathbf{x}_1)} \frac{Q_0(\mathbf{x}_0)}{Q_1(\mathbf{x}_0)} \tag{46}$$

The scheme is continued to generate a sample of weights $\{c_t\}$. It can be shown that for the sample of weights thus obtained, equation (43) provides a consistent and unbiased estimator of the marginal likelihood (Neal, 2001). The individual steps of (45) can be constructed by applying the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953) to the respective transition probability T_m, as in MCMC. Note that as opposed to MCMC, the respective Markov chains do not need to be run to convergence, though.

4.4.1 Details of the simulations

In our simulations, we carried out for each step in (45) 10 Metropolis-Hastings steps according to the description in Section 4.3.1, for $\mathcal{H}_0$, and Section 4.3.2, for $\mathcal{H}_1$. A "temperature" ladder of $M = 10$ equidistant β_m values for defining the intermediate distributions Q_m in (44) was selected, and we chose a total sample size of $N = 400$ for computing the marginal likelihood according to (40). We experimented with a polynomial rather than an equidistant cooling scheme for β, but did not find any noticeable differences in the results.

As a heuristic indicator of how accurate the estimation with AIS is, we followed Neal (2001) and computed the variance of $c_t^* = c_t / \frac{1}{N}\sum_{t=1}^{N} c_t$. The term $\psi = 1/[1 + Var(c_t^*)]$ gives a rough indication of the factor by which the sample size is effectively reduced when drawing samples according to the procedure (45) rather than from the correct posterior distribution. In our simulations, we typically found values of $\psi \leq 1.3$, indicating a sufficient degree of convergence.

5 Results

5.1 Investigating the behaviour around the Felsenstein zone

We generated synthetic DNA sequence alignments from 4-taxa trees with different branch lengths. In the vein of Felsenstein's seminal study for demonstrating the inconsistency of maximum parsimony (Felsenstein, 1978), we systematically varied the parameters $d2$ and $d3$ in Figure 1, and generated DNA sequence alignments with SEQGEN (Rambaut, 1996), as described in Section 3. For each branch length configuration $[d2, d3]$, we estimated the posterior probabilities for the three possible tree topologies under the two different models discussed above: the site-specific branch length model $\mathcal{H}_0$, described in Section 4.1, and the standard phylogenetic model $\mathcal{H}_1$, described in Section 4.2. We repeated the estimation of the posterior probabilities with two different methods: MCMC, as described in Section 4.3, and annealed importance sampling, as described in Section 4.4. The results are shown in Figures 4 and 5. In both figures, subfigure a) shows the results for the site-specific branch length model $\mathcal{H}_0$, while subfigure b) shows the results for the standard phylogenetic model $\mathcal{H}_1$. The axes represent the values of the parameters $d2$ and $d3$; hence each grid location defines a phylogenetic tree with a specific branch length configuration. The estimated posterior probabilities are indicated with a grey shading ranging from 0 (black) to 1 (white) and the values in between are indicated by the legend in subfigure c). It is clearly seen that the independent branch length model $\mathcal{H}_0$ leads to a systematic failure in the Felsenstein zone (characterized by a small value of $d2$ and a large value of $d3$) in that the posterior probability of the correct tree is consistently close to zero. In fact, the tree topology with the highest posterior probability was found to be the one in which the two longer branches were grouped together. This suggests that the independent branch length model $\mathcal{H}_0$ has the same problem with long-branch attraction as maximum parsimony. This failure was avoided with the standard phylogenetic model $\mathcal{H}_1$, whose posterior probability of the correct tree topology was consistently above 0.5 (and mostly close to 1) for the whole branch length configurations space. The results obtained with MCMC and AIS were largely consistent, although the difference between the posterior probabilities for $\mathcal{H}_0$ and $\mathcal{H}_1$ in the Felsenstein zone was slightly larger with MCMC than with AIS.

5.2 Evaluation of the performance of DMCP and PFHMM

The previous section has shown that for the model of independent, site-specific branch lengths ($\mathcal{H}_0$), there is a systematic failure in the Felsenstein zone, which is avoided with the standard phylogenetic model of common branch lengths, $\mathcal{H}_1$. Since the recombination detection methods PFHMM and DMCP are based on $\mathcal{H}_0$, we suspect that they are susceptible to the same systematic failure. We tested this conjecture by applying both methods, DMCP and PFHMM, to the same synthetic

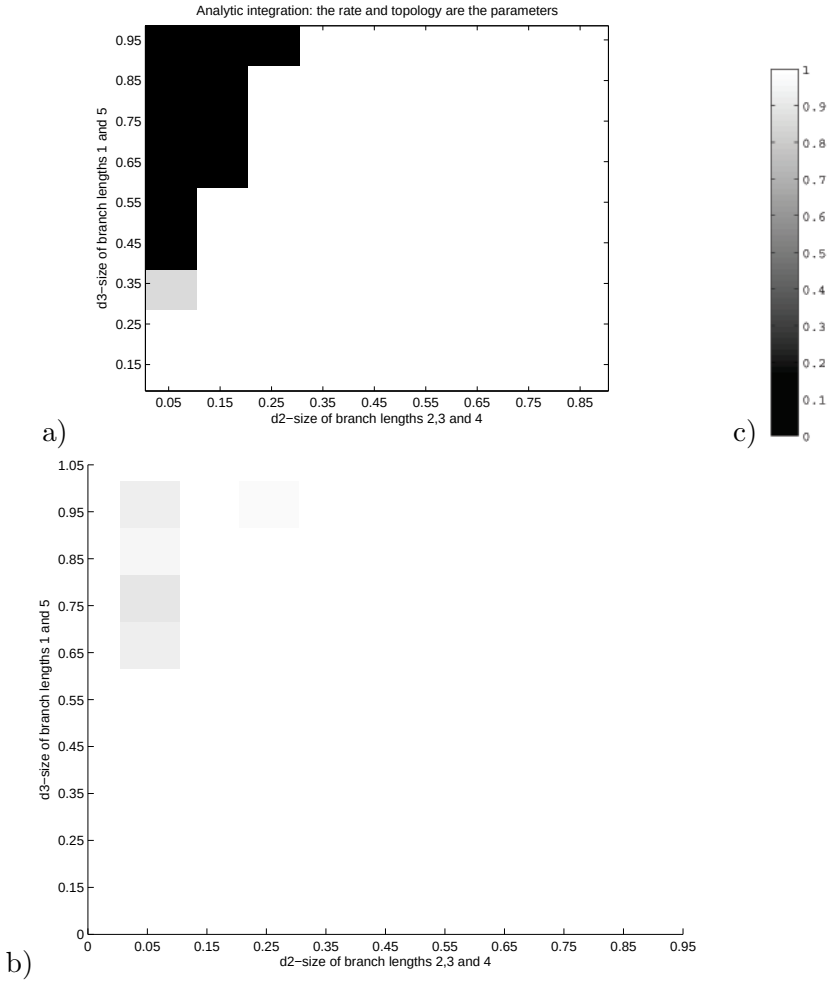


Figure 4: Posterior probabilities estimated with MCMC.
The two figures show the posterior probability of the correct tree topology for different branch length configurations. These configurations are determined by the values of $d2$ and $d3$, as defined in Figure 1. In each subfigure, the horizontal axis refers to $d2$, and the vertical axis refers to $d3$. The grey shading indicates the value of the inferred posterior probability, as indicated in the legend on the right, ranging from 0 (black) to 1 (white). Subfigure a) shows the results obtained for Model $\mathcal{H}_0$, represented in Figure 2. Subfigure b) shows the results obtained for Model $\mathcal{H}_1$, represented in Figure 3. The results shown are those obtained from a specific set of DNA sequence alignments generated from trees with the indicated $[d2, d3]$ configurations, as described in Section 3.1. Repeating the simulations for different sequences generated from the same trees was found to give nearly identical results. It is clearly observed that Model $\mathcal{H}_0$, which is shown in Subfigure a), leads to the systematic prediction of the wrong tree topology in the Felsenstein zone.

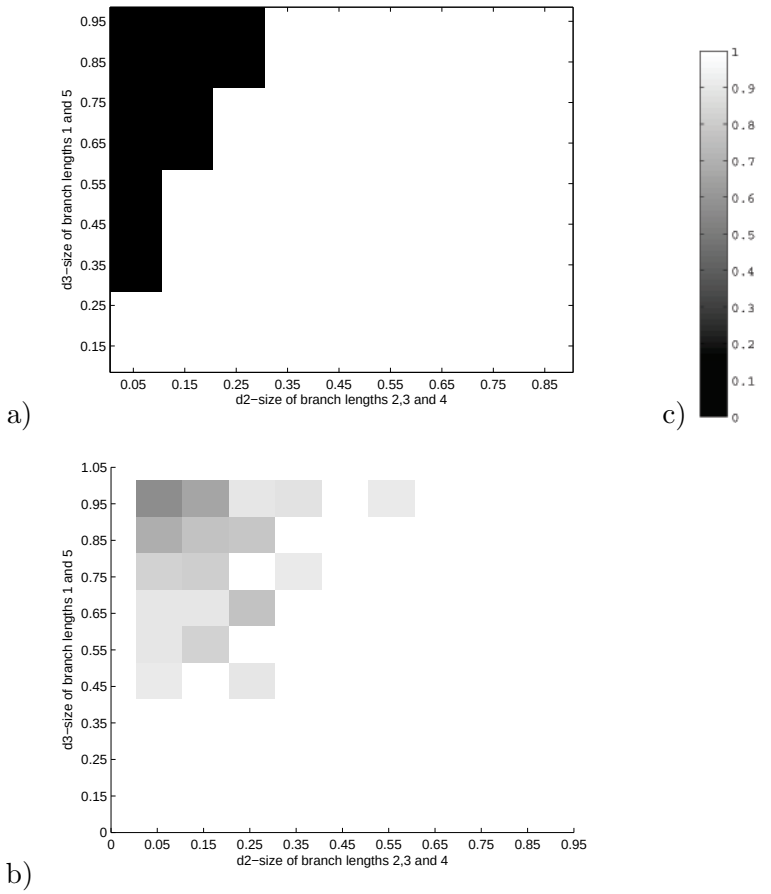


Figure 5: Posterior probabilities estimated with Annealed Importance Sampling. As in Figure 4, Subfigures a) and b) show the posterior probability of the correct tree topology for different branch length configurations. The results were obtained with annealed importance sampling rather than MCMC and show an average over five DNA sequence alignments independently generated for each branch length configuration. These configurations are determined by the values of $d2$ and $d3$, as defined in Figure 1. In each subfigure, the horizontal axis refers to $d2$, and the vertical axis refers to $d3$. The grey shading indicates the value of the inferred posterior probability, as indicated in the legend on the right, ranging from 0 (black) to 1 (white). Subfigure a) shows the results obtained for Model $\mathcal{H}_0$, represented in Figure 2. Subfigure b) shows the results obtained for Model $\mathcal{H}_1$, represented in Figure 3. Like in Figure 4, it is clearly observed that Model $\mathcal{H}_0$, which is shown in Subfigure a), leads to the systematic prediction of the wrong tree topology in the Felsenstein zone.

DNA sequence alignments as used in the previous section. For comparison, we also applied the phylogenetic hidden Markov model (PHMM) of Husmeier and McGuire (2003). Note that the latter model is based on $\mathcal{H}_1$ and should therefore not be susceptible to inferring wrong tree topologies in the Felsenstein zone.[4] We used the authors' own programs, available from the webpages referenced in Minin *et al.* (2005) (for DMCP), Husmeier (2005) (for PFHMM) and Husmeier and McGuire (2003) (for PHMM). All three methods sample parameters and hidden states from the posterior distribution with MCMC. To test for convergence of these simulations, we computed the potential scale reduction factor from different quantities, as in Gelman and Rubin (1992), taking values below 1.1 as an indication of sufficient convergence[5]. From the sampling phase of the MCMC simulations, we computed for each site t in the alignment the marginal posterior probabilities $P(S_t|\boldsymbol{\mathcal{D}})$ of the three possible tree topologies $S_t \in \{\Psi_1, \Psi_2, \Psi_3\}$[6]. The results were similar to those discussed in the previous section, with a clear systematic failure of PFHMM and DMCP in the Felsenstein zone. This failure was avoided when using PHMM. A specific example is presented in Figure 6, which shows the posterior probabilities $P(S_t|\boldsymbol{\mathcal{D}})$ for a DNA sequence alignment $\boldsymbol{\mathcal{D}}$ generated from the tree in Figure 1 with a branch length configuration $d2 = 0.15, d3 = 0.85$. A comparison with Figures 4 and 5 shows that this branch length configuration lies clearly in the Felsenstein zone. In support of our conjecture, both PFHMM and DMCP systematically show high posterior probabilities $P(S_t|\boldsymbol{\mathcal{D}})$ close to 1 for a wrong tree topology throughout the whole DNA sequence alignment. Incidentally, in the high-scoring tree topology the two long non adjacent branches $d3$ in Figure 1 are grouped together, suggesting that PFHMM and DMCP suffer from the same long branch attraction as the method of maximum parsimony (Felsenstein, 1978). There are no problems with PHMM, which consistently scored high posterior probabilities $P(S_t|\boldsymbol{\mathcal{D}})$ close to 1 for the correct tree topology throughout the whole alignment.

6 Improving the phylogenetic factorial HMM

Our study described in the previous sections has revealed that the phylogenetic factorial HMM (PFHMM) of Husmeier (2005) is susceptible to systematically predicting spurious topology changes in the Felsenstein zone. The objective of the present section is to describe a modification of the PFHMM that avoids this short-

[4]Note that as opposed to DMCP and PFHMM, PHMM cannot distinguish between recombination and rate heterogeneity, though.

[5]This was achieved with the following burn-in and sampling lengths. Burn-in: 1000 steps for DMPC, 250 steps for PFHMM, and 10K steps for PHMM. Sampling phase: 200 subsample steps (in intervals of 50 steps) for DMPC, 250 steps for PFHMM, and 1000 subsample steps (in intervals of 10 steps) for PHMM. PFHMM needs as input a set of fixed nucleotide substitution rates, corresponding to the hyperparameter ρ in eq 13. These values were selected as $\rho \in \{0.05, 0.1, 0.5, 1, 2, 4, 6, 8\}$.

[6] These tree topologies are $\Psi_1 = (1, 2, (3, 4)), \Psi_2 = (1, 3, (2, 4))$, and $\Psi_3 = (1, 4, (2, 3))$, where the numbers refer to the four taxa.

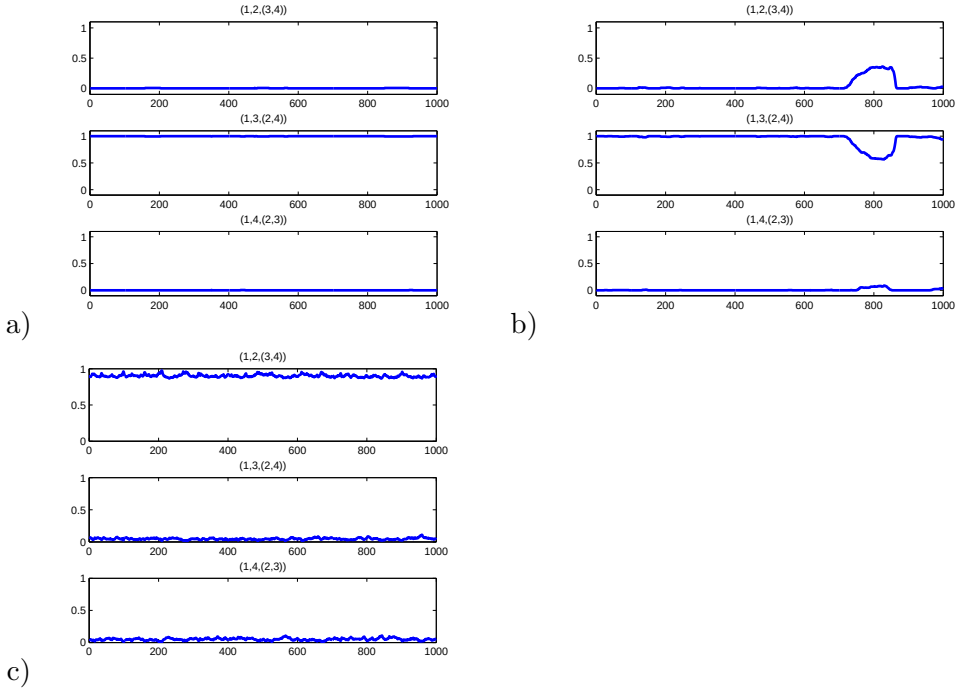


Figure 6: Failure of PFHMM and DMCP in the Felsenstein zone
Each figure shows a plot of the marginal posterior probability $P(S_t|\boldsymbol{\mathcal{D}})$ (vertical axes) of the three possible tree topologies $S_t \in \{\Psi_1, \Psi_2, \Psi_3\}$ for the 4-taxa tree of Figure 1, plotted against the position t in the DNA sequence alignment (horizontal axes). Each subfigure consists of three panels, where the top panel corresponds to the true tree topology, from which the data were generated. The middle panel corresponds to a wrong tree topology, in which the two long branches $d3$ in Figure 1 are grouped together (long branch attraction). The bottom panel corresponds to another wrong tree topology. The three subfigures show the results obtained for the three recombination detection methods investigated: DCMP (Subfigure a), PFHMM (Subfigure b), and PHMM (Subfigure c). PHMM predicts high posterior probabilities $P(S_t|\boldsymbol{\mathcal{D}})$ close to 1 for the true tree topology throughout the whole sequence alignment. However, both PFHMM and DMCP systematically show high posterior probabilities $P(S_t|\boldsymbol{\mathcal{D}})$ close to 1 for the wrong tree topology in which the two long non-adjacent branches are joined.

coming. A probabilistic graphical model representation of the PFHMM of Husmeier (2005) is shown in Figure 7 a. The model is essentially based on Model $\mathcal{H}_0$ of Figure 2 in that separate branch length vectors are associated with different sites of the alignment. This allows the branch lengths to be integrated out analytically, as described in Section 2.4, resulting in the simplified model depicted in Figure 7 b. The modified PFHMM is shown in Figure 7 c. Akin to Model $\mathcal{H}_1$ of Figure 3, a common vector of branch lengths is shared by all sites in the alignment[7]. The rate states $R_t \in \{\rho_1, \ldots, \rho_{k'}\}$, which in the original PFHMM of Husmeier (2005) are associated with the hyperparameter ρ of the prior distribution on the branch lengths, equation (13), are now associated with a global scaling factor by which the vector of branch lengths is multiplied. The hidden state sequences, $\mathbf{S}$ and $\mathbf{R}$, and the model parameters are sampled from the posterior distribution with a Gibbs sampling procedure:

$$\mathbf{S}^{(i+1)} \sim P\left(\cdot|\mathbf{R}^{(i)}, \nu_S^{(i)}, \nu_R^{(i)}, \mathbf{w}^{(i)}, \boldsymbol{\mathcal{D}}\right) \tag{47}$$

$$\mathbf{R}^{(i+1)} \sim P\left(\cdot|\mathbf{S}^{(i+1)}, \nu_S^{(i)}, \nu_R^{(i)}, \mathbf{w}^{(i)}, \boldsymbol{\mathcal{D}}\right) \tag{48}$$

$$\nu_S^{(i+1)} \sim P\left(\cdot|\mathbf{S}^{(i+1)}, \mathbf{R}^{(i+1)}, \nu_R^{(i)}, \mathbf{w}^{(i)}, \boldsymbol{\mathcal{D}}\right) \tag{49}$$

$$\nu_R^{(i+1)} \sim P\left(\cdot|\mathbf{S}^{(i+1)}, \mathbf{R}^{(i+1)}, \nu_S^{(i+1)}, \mathbf{w}^{(i)}, \boldsymbol{\mathcal{D}}\right) \tag{50}$$

$$\mathbf{w}^{(i+1)} \sim P\left(\cdot|S^{(i+1)}, \mathbf{R}^{(i+1)}, \nu_S^{(i+1)}, \nu_R^{(i+1)}, \boldsymbol{\mathcal{D}}\right) \tag{51}$$

where the superscript i denotes the iteration number. The first four steps are identical to those in Husmeier (2005): The hidden state sequences $\mathbf{S}$ and $\mathbf{R}$ are sampled with the stochastic forward-backward algorithm of Boys *et al.* (2000); the transition probabilities ν_S and ν_R, defined in equations (9) and (10), are sampled from beta distributions whose sufficient statistics are determined by $\mathbf{S}$ and $\mathbf{R}$. The new aspect of our algorithm is the sampling of the branch length vector $\mathbf{w}$. Since there is no closed-form expression for the distribution on the right-hand side of equation (51), we resort to a Metropolis-Hastings-within-Gibbs procedure. Note that $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3)$ is composed of three subvectors $\mathbf{w}_k$, $k \in \{1, 2, 3\}$, associated with the three tree topologies represented by the hidden state $S_t \in \{\Psi_1, \Psi_2, \Psi_3\}$. To ensure that the model is identifiable, we constrain the L1-norm of the branch length vectors to be equal to one: $||\mathbf{w}_k||_1 = 1$, $k \in \{1, 2, 3\}$; recall that the scaling of the branch lengths is effected by multiplication with a factor defined by the hidden states, $R_t \in \{\rho_1, \ldots, \rho_{K'}\}$. This constraint, as well as the positivity constraint $w_{ki} \geq 0$, is automatically guaranteed when proposing new branch length vectors $\mathbf{w}_k^*$

[7]More accurately, there are three vectors of branch lengths $\mathbf{w}_k$, $k \in \{1, 2, 3\}$, associated with the three different tree topologies. This can be modelled as a common vector composed of three sub-vectors, where the state variable S_t indicates which of these subvectors applies to site t.

from a Dirichlet distribution:

$$Q(\mathbf{w}_k^*|\mathbf{w}_k) \;\propto\; \prod_i [w_{ki}^*]^{\alpha w_{ki}-1} \tag{52}$$

whose mean and variance are given by

$$\mathrm{E}[w_{ki}^*|w_{ki}] \;=\; w_{ki}; \quad \mathrm{Var}[w_{ki}^*|w_{ki}] \;=\; \frac{w_{ki}(1-w_{ki})}{\alpha+1} \tag{53}$$

Hence, the mean of the proposal distribution is equal to the current branch length, while the variance depends on a scaling parameter α. In our simulations, α was automatically adjusted in the burn-in phase to achieve an average acceptance probability between 30% and 70%. The proposed vector of branch lengths $\mathbf{w}^*$ was accepted or rejected according to the standard Metropolis-Hastings criterion (Hastings, 1970), with the following acceptance probability:

$$A \;=\; \min\left\{1, \frac{L(\mathbf{w}_k^*)P(\mathbf{w}_k^*)Q(\mathbf{w}_k|\mathbf{w}_k^*)}{L(\mathbf{w}_k)P(\mathbf{w}_k)Q(\mathbf{w}_k^*|\mathbf{w}_k)}\right\} \tag{54}$$

where the proposal distribution $Q(\mathbf{w}_k^*|\mathbf{w}_k)$ is defined in equation (52), the prior distribution $P(\mathbf{w}_k)$ was chosen as defined in equation (13), with a fixed hyperparameter $\rho = 1$, and the likelihood $L(\mathbf{w}_k)$ depends on the hidden state sequences $\mathbf{S}$ and $\mathbf{R}$ as follows:

$$L(\mathbf{w}_k) \;=\; \prod_{t|S_t=\Psi_k} P(\mathbf{y}_t|R_t\mathbf{w}_k, \Psi_k, \boldsymbol{\theta}) \tag{55}$$

where the expression in the argument of the product is given by equation (11). The details of the Gibbs sampling scheme used in our simulations are summarized in the appendix.

7 Results for the improved PFHMM

7.1 Simulation details

We tested the improved PFHMM on the two types of synthetic DNA sequence alignments described in Section 3. The homogeneous DNA sequence alignments were the same as those used in the previous studies. The DNA sequence alignment with the mosaic structure was generated as described in Section 3, setting $d2 = d3 = 0.25$ for the flanking segments, and $d2 = 0.15, d3 = 0.85$ for the central segment. Hence, the branch length configuration corresponding to the central segment lies clearly in the Felsenstein zone; compare with Figures 4 and 5. Note that the DNA sequence alignment does not contain any change of the tree topology, though. For both the original PFHMM of Husmeier (2005) and the improved PFHMM we sampled the state sequences $\mathbf{S}$ from the posterior distribution with MCMC, monitoring convergence with

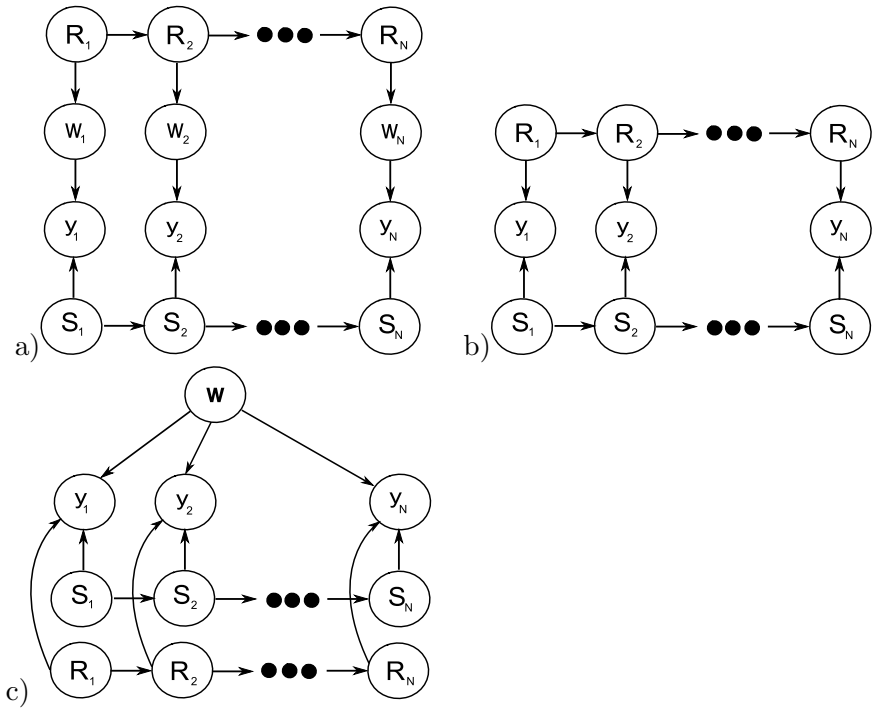
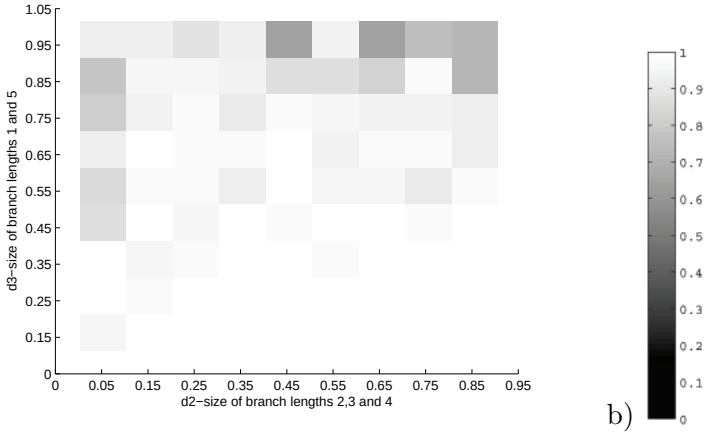


Figure 7: Graphical model of the PFHMM and the improved PFHMM
Subfigure a) shows the probabilistic graphical model representation of the phylogenetic factorial HMM of Husmeier (2005). The $\mathbf{y}_t$'s represent the columns in the DNA sequence alignment, where the subscript $t = 1, \ldots, N$ indicates the site in the alignment. Each site t is associated with a hidden state S_t that defines the tree topology, a vector of branch lengths $\mathbf{w}_t$, and a second hidden state R_t that defines the hyperparameter of the prior distribution on the branch lengths, as defined in equation (13). Both hidden states S_t and R_t have a Markovian dependence structure. The chosen form of the model allows the branch lengths to be integrated out analytically, as described in Section 2.4. This results in the simplified model depicted in Subfigure b). Note that this model is a phylogenetic factorial HMM, where one type of hidden states $(S_1, \ldots, S_N)$ defines the tree topology, and the other type of hidden states $(R_1, \ldots, R_N)$ defines the average amount of mutational divergence. Hence, the model presented here is a generalization of the model shown in Figure 2 so as to allow for recombination and rate heterogeneity. Subfigure c) shows the probabilistic graphical model representation of the improved phylogenetic factorial HMM proposed in the present article. The model is similar to the one presented in the previous subfigures with the difference that a common branch length vector $\mathbf{w}$ is shared among all sites. This is a generalization of the standard phylogenetic model of Figure 3 that allows for recombination and rate heterogeneity.



a) b)

Figure 8: Results of the improved PFHMM
The figure shows, for different branch length configurations $[d2, d3]$, as defined in Figure 1, the posterior probability $\overline{P}(S = \Psi_{true}|\boldsymbol{\mathcal{D}})$, defined in equation (56). The horizontal axis represents $d2$, and the vertical axis represents $d3$. Probabilities are represented by a grey shading, ranging from white (1) to black (0), as indicated by the legend on the right. The figure shows that as a consequence of the modification of the PFHMM, described in Section 6, the systematic failure in the Felsenstein zone, which was found in Subfigure a) of Figures 4 and 5, is avoided. These results are the averaging over 2 independent simulations.

the diagnostic test based on potential scale reduction factors (Gelman and Rubin, 1992); the details were given in Section 4.3.3. Note that both the original and the improved PFHMM need as input a set of fixed nucleotide substitution rates, corresponding to the hyperparameter ρ in equation (13). These values, which are associated with the rate states $\mathbf{R}$, were selected as follows: $\rho \in \{0.05, 0.1, 0.5, 1, 2, 4, 6, 8\}$.

7.2 Simulation results

Figure 8 shows the results obtained with the improved PFHMM on the homogeneous DNA sequence alignment. The figure shows, for different branch length configurations $[d2, d3]$, the average posterior probability of the correct tree topology Ψ_{true}, averaged over all positions in the alignment:

$$\overline{P}(S = \Psi_{true}|\boldsymbol{\mathcal{D}}) \;=\; \frac{1}{N}\sum_{t=1}^{N} P(S_t = \Psi_{true}|\boldsymbol{\mathcal{D}}) \tag{56}$$

It is clearly seen that the failure in the Felsenstein zone is avoided, and that $\overline{P}(S = \Psi_{true}|\boldsymbol{\mathcal{D}})$ is consistently greater than 0.5 (and close to 1 in most cases).

Figure 9 shows the results obtained on the DNA sequence alignment with the mosaic structure. Both subfigures show a plot of the predicted marginal posterior

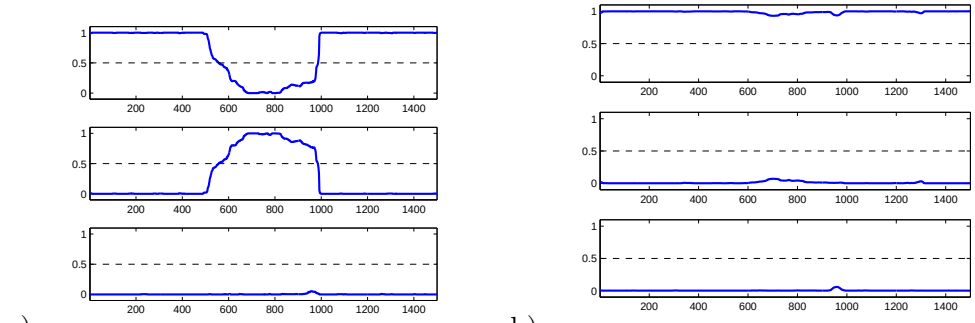


Figure 9: Mosaic DNA sequence alignment
The figure shows the predictions obtained with the original PFHMM of Husmeier (2005) versus the improved PFHMM proposed in the present paper. Both models were applied to a synthetic DNA sequence alignment with mosaic structure, where the branch length configuration in the central segment lies in the Felsenstein zone; see the description in Section 3. Each panel shows a plot of the predicted marginal posterior probabilities $P(S_t = \Psi_i|\boldsymbol{\mathcal{D}})$ for the three possible tree topologies $i \in \{1, 2, 3\}$, where the true topology corresponds to the panel in the top. The vertical axes show the marginal posterior probabilities, while the horizontal axes represent the site t in the alignment. The original PFHMM, shown in Subfigure a), predicts a spurious topology change in the central segment, which is avoided with the improved PFHMM, shown in Subfigure b). Subfigure b) is an average over 5 independent simulations showing that the spurious topology change is avoided consistently.

probabilities $P(S_t = \Psi_i|\boldsymbol{\mathcal{D}})$, for the three possible tree topologies $i \in \{1, 2, 3\}$, plotted against the position t in the alignment. The left subfigure shows the prediction obtained with the original PFHMM of Husmeier (2005). There is a clear transition into a different tree topology in the central region, where the branch length configuration $[d2, d3]$ lies in the Felsenstein zone. This confirms our conjecture that PFHMM is susceptible to the prediction of spurious topology changes. The right panel shows the prediction made with the improved PFHMM averaged over 5 independent simulations. The posterior probability for the correct tree topology, $P(S_t = \Psi_{true}|\boldsymbol{\mathcal{D}})$, is consistently close to 1, indicating that the prediction of spurious topology changes is avoided.

8 Discussion

In our paper, we have investigated a possible shortcoming of three recent Bayesian methods for detecting recombination in DNA sequence alignments: the multiple change-point (MCP) model of Suchard *et al.* (2003), the dual multiple change-point

(DMCP) model of Minin *et al.* (2005), and the phylogenetic factorial hidden Markov model (PFHMM) of Husmeier (2005). All three models assume separate branch lengths for different sites, which allows the branch lengths to be integrated out analytically. This reduces the computational complexity of the Bayesian inference scheme, which can now be formulated in terms of posterior distributions of the tree topologies and the nucleotide substitution parameters only. This makes the approach quite popular, and it has been applied in more recent works; see Lehrach (2008) and Lehrach and Husmeier (2009).

Note that the model of site-independent branch lengths, as expressed in eq. (20), was first introduced by Tuffley and Steel (1997), where it was called the "no-common-mechanism" model. In combination with the prior independence of the branch length components, expressed in eq. (13), the vector of branch lengths can be integrated out in the likelihood, as shown by Suchard *et al.* (2003), and discussed in Section 2.4. However, in the no-common-mechanism model, the branch lengths are incidental rather than structural parameters. As discussed in Goldman (1990), this implies that maximum likelihood is no longer guaranteed to provide a consistent estimator. In fact, Tuffley and Steel (1997) showed that under certain regularity conditions, maximum parsimony and maximum likelihood with no common mechanisms are equivalent. This suggests that maximum likelihood with no common mechanisms will be susceptible to the prediction of wrong tree topologies for certain branch length configurations (long branch attraction). To confirm this hypothesis, we have generated synthetic DNA sequence alignments with the HKY nucleotide substitution model (Hasegawa *et al.*, 1985) in the vein of Felsenstein's seminal study for demonstrating the inconsistency of maximum parsimony Felsenstein (1978), and we have estimated the marginal posterior probability for the tree topology in two different ways: using an inter-model approach, in which tree topologies are sampled from the posterior distribution with MCMC; and applying an intra-model approach, in which the marginal likelihood is estimated with annealed importance sampling. Both studies consistently reveal that as a consequence of the separate site-dependent branch lengths, the mode of the posterior distribution is systematically shifted to a wrong tree topology whenever the branch length configuration of the data-generating tree falls into the Felsenstein zone. The inferred tree topology with the highest posterior probability is the one in which the long branches are grouped together. This finding suggests that as a consequence of the aforementioned independence assumption (i.e., the "no-common-mechanism" model of separate site-dependent branch lengths), the resulting model suffers from the same inconsistency (long-branch attraction) as the method of maximum parsimony. We have further confirmed this conjecture by applying the recombination detection methods DMCP and PFHMM to the DNA sequence alignments generated in our study, using the authors' programs. Again, we found a systematic failure in the Felsenstein zone, where consistently the wrong tree topology was inferred. This suggests that these recombination detection methods are susceptible to predicting

spurious recombination events whenever branch-length configurations happen to fall near the boundary of the Felsenstein zone.

We have concluded our study with a demonstration of how the PFHMM can be improved to avoid this shortcoming. In principle this can be achieved by removing the site-independence assumption for the branch lengths. As a consequence, however, the analytic integration over the branch lengths is no longer tractable, which requires them to be sampled approximately from the posterior distribution with MCMC. To avoid an identifiability problem resulting from the fact that the global scaling of the branch lengths (defined by one of the two types of hidden states) is an additional independent parameter of the model, we have imposed a normalization constraint on the branch lengths, which can easily be effected by the choice of a suitable proposal distribution in the MCMC scheme. We have tested the proposed method on the same DNA sequence alignments as for the other models, and found that it succeeded in avoiding the failure in the Felsenstein zone.

Note that in the proposed phylogenetic PFHMM, each hidden state is associated with a distinct tree topology. The number of tree topologies increases super-exponentially with the number of taxa; for this reason, we have applied our model to DNA sequence alignments of four sequences only, as in Husmeier and McGuire (2003). There are various heuristic simplifications one could adopt in order to apply the method to sequence alignments with more than four taxa. One method would be to apply a preliminary phylogenetic analysis to consecutive subsets of the DNA sequence alignment, effected for instance in the way described in Husmeier *et al.* (2005b). The phylogenetic FHMM would then include only those topology states that match one of the tree topologies inferred in the preliminary analysis. Another method would be to proceed in the way described in Minin *et al.* (2005). Here, the assumption is that we are given a sequence alignment composed of N-1 nonrecombinant and 1 putative recombinant strain. Additionally, it is assumed that the tree of the N-1 nonrecombinant sequences is known or that it can easily be inferred. The states of the phylogenetic FHMM are restricted to the set of those tree topologies that are obtained by adding a new leaf node to any branch in the fixed parental tree with $N-1$ nonrecombinant taxa. Both of these heuristic simplifications substantially restrict the set of permissible tree topologies, thereby rendering the application of the phylogenetic FHMM to larger alignments viable. Note, though, that in principle these restrictions are not necessary. Our method could in principle be implemented with a transdimensional MCMC scheme using reversible jumps associated with the birth and death of topology states, where each birth creates a new tree topology derived from the adjacent topology by some local modification, e.g. using nearest neighbour interchange. However, the computational costs of such an approach would be huge, and it would pose a challenging problem for novel high-performance distributed computing techniques.

It has been pointed out by one of the referees that our work is closely related to the work of Huelsenbeck *et al.* (2008). Like in our paper, the authors investigate

a Bayesian implementation of the no-common-mechanism model, and they empirically demonstrate that this model is not consistent and shows a systematic failure in the Felsenstein zone. There are various ways in which our study complements this work. Firstly, Huelsenbeck *et al.* (2008) use a fully symmetric nucleotide substitution model that makes no distinction between any character states (the Jukes-Cantor model). For this model, Tuffley and Steel (1997) showed that maximum parsimony and maximum likelihood with no common mechanism are equivalent in the sense that both choose the same tree. Hence, the work of Huelsenbeck *et al.* (2008) can be seen as an empirical corroboration of the theoretical findings in Tuffley and Steel (1997). Our study complements this work by using the HKY model (Hasegawa *et al.*, 1985) as a more general and more widely applied nucleotide substitution model, for which no theoretical proof was given in Tuffley and Steel (1997). Secondly, Huelsenbeck *et al.* (2008) use fixed parameters for the prior distribution on the branch lengths and find that these parameters "play an inordinately strong role in determining the probabilities of the trees". In our work on the homogeneous DNA sequence alignment, we use a hierarchical Bayesian model with an extra layer (a hyperprior) – see Figures 2 and 3 – and infer the parameters of the prior from the data. In our work on the DNA sequence alignment with mosaic structure, we use a phylogenetic FHMM, in which the parameters of the prior distribution are associated with different hidden states. The assignment of these hidden states to sites is inferred from the data. Thirdly, one has to appreciate that there is no sufficient criterion to prove that an MCMC simulation has converged. For this reason computing the posterior probabilities of tree topologies with an alternative paradigm, as we do in our intra-model approach based on annealed importance sampling, offers an independent corroboration of the findings. Fourthly and most importantly, however, there has been a completely different focus of our work. The motivation for the work of Huelsenbeck *et al.* (2008) has been the development of a new Bayesian MCMC scheme for learning tree topologies from sequence alignments that are adequately described by a single tree. The focus of our study is the prediction of recombination and mosaic structures in DNA sequence alignments, and it has been motivated by three recent detection methods that are based on the no-common-mechanism model. These models are more flexible than the single-tree model investigated by Huelsenbeck *et al.* (2008). In particular, they allow for breakpoints in the DNA sequence alignment at which the tree topology may change. While this mechanism provides the extra flexibility required for dealing with recombination, we have shown that in combination with site-independent branch lengths (the no-common-mechanism of Tuffley and Steel (1997)), the resulting model becomes susceptible to predicting spurious topology changes and recombination breakpoints.

9 Future work

The phylogenetic FHMM proposed in Section 6 of our paper provides a trade-off between two extreme scenarios: the homogeneous model, which employs the same branch lengths for the whole alignment, and the no-common-mechanism model. The first approach is too restrictive. In the second approach, the branch lengths are incidental rather than structural parameters, resulting in the inconsistency problems discussed in the present paper. The proposed phylogenetic FHMM contains a hidden factor by which the branch lengths are rescaled. This scaling factor is site dependent via its association with a hidden state of the FHMM. Since the number of hidden states is finite, and each hidden state can be revisited repeatedly when traversing along the alignment, all parameters of the model are structural (rather than incidental). In this way, the consistency of our model is guaranteed. However, while our model is appropriate to incorporate the effects of rate heterogeneity, it is too restrictive when dealing with certain recombination events that do not induce a tree topology change. This can happen when, in the coalescence tree, recombinant lineages coalesce before merging with any other lineage (Wiuf *et al.*, 2001). In certain scenarios, discussed in Wiuf *et al.* (2001), this can result in a more complex change of the branch lengths than can be modelled by a global rescaling. One way to proceed would be the following modification of our model. Rather than associating the second hidden state with a global scaling factor, we could associate it with a separate vector of branch lengths. In this model there would no longer be a common branch length vector, but the branch lengths would be site-dependent, as in the no-common-mechanism model. The substantial difference from the no-common-mechanism model would be that in the new model the site dependence is effected indirectly via a hidden state. Since the number of hidden states is finite, and the hidden states can be revisited (at least as long as all transition probabilities are non-zero), the new model contains structural rather than incidental parameters. In this way, its consistency is guaranteed. Note, however, that this model is more complex than the one proposed in our paper. In particular, it will require the number of hidden states and their associated parameters to be properly inferred from the data rather than chosen in advance. This calls for the development of a trans-dimensional MCMC scheme with RJMCMC (Green, 1995), as applied in the studies by Suchard *et al.* (2003), Minin *et al.* (2005), Lehrach (2008) and Lehrach and Husmeier (2009). We believe that this would be an important and stimulating topic for future research.

When extending the phylogenetic FHMM along the line discussed in the previous paragraph, one has to decide on the appropriate form of the prior distribution on tree topologies. It is a common approach in Bayesian analysis to use a uniform prior distribution. The intention is to reflect our prior level of ignorance, as especially promoted by the school of "objective Bayesianism". The question, then, is what exactly it is that we are ignorant about. A prior distribution that is uniform over

tree topologies is not uniform over labelled histories or clade formations, where the latter inconsistency has been used to (erroneously!) question the validity of the Bayesian approach per se (Pickett and Randle, 2005). As pointed out by Velasco (2008), the ignorance should be expressed in terms of the physical processes that generate the entities of interest. A phylogenetic tree is the result of the biological process of common ancestry and descent with modification, which can be modelled by a Yule random branching process (forward in time) or a coalescence process (backward in time). Kingman (1982) and Thompson (1975) showed that under certain regularity conditions, the Yule birth process, the Yule birth-death process and the coalescence process lead to the same distribution. This distribution is uniform over labelled histories (Edwards, 1970), which induces a prior distribution on tree topologies that is no longer uniform. In particular, Velasco (2008) showed that a tree topology that is more balanced (as opposed to pectinate) is consistent with more labelled histories and, consequently, has a higher prior probability. An early application of this approach can be found in Yang and Rannala (1997). However, the computational costs were found to be huge – about two orders of magnitude larger than those of the competing method of Larget and Simon (1999). It therefore will pose a substantial computational challenge for future work to render the approach based on labelled histories viable in the context of the model proposed in the present paper.

Appendix: Details of the Gibbs sampling scheme used for the improved phylogenetic FHMM

We briefly describe the Gibbs sampling procedure that we used for the improved phylogenetic FHMM described in Section 6.

We sampled the hidden state sequences and model parameters according to the Gibbs sampling scheme described in Sections 4 and 5. We carried out 200 Gibbs sampling steps in the burn-in phase, and 200 steps in the sampling phase. Recall that each Gibbs sampling step includes a set of Metropolis-Hastings (MH) steps for adapting the branch lengths, according to equations (51) and (54). Within each Gibbs step, we carried out 200 MH steps for the MH burn-in phase, and 1200 MH steps for the MH sampling phase. The final branch length vector was kept, and constituted the output of the Gibbs sampling step (51). During the MH burn-in phase, the parameter α of the proposal distribution (52) was adjusted, as described in Section 6. We used the MH sampling phase to compute, for all branch lengths, the potential scale reduction factor of Gelman and Rubin (1992).

For the simulations thus carried out, we found that the potential scale reduction factor was consistently smaller than 1.1, indicating a satisfactory degree of convergence. The marginal posterior probabilities of the topology states, $P(S_t = \Psi_k|\boldsymbol{\mathcal{D}})$, were computed straight from the state sequences $\{\mathbf{S}_i\}$ sampled during the sampling phase of the Gibbs sampling scheme by application of (47); the results are shown in

Figures 8 and 9.

Acknowledgement

This work was supported by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD). We are grateful to Wolfgang Lehrach for helpful and stimulating discussions. We are also indebted to Jeff Thorne for pointing out the literature on the no-common-mechanism model to us, and for constructive feedback that has led to a substantial revision of our original manuscript.

References

Boys, R. J., Henderson, D. A. and Wilkinson, D. J. (2000) Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Applied Statistics*, **49**, 269–285.

Edwards, A. (1970) Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society B*, **32:2**, 155–174.

Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **27**, 401–440.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **Vol. 7, No.4**, 457–472.

Goldman, G. (1990) Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Systematic Zoology*, **39**, 345–361.

Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Huelsenbeck, J., Ane, C., Larget, B. and Ronquist, F. (2008) A Bayesian perspective on a non-parsimonious parsimony model. *Systematic Biology*, **57:3**, 406–419.

Husmeier, D. (2005) Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics*, **21**, ii166–ii172.

Husmeier, D., Dybowski, R. and Roberts, S. (2005a) *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Advanced Information and Knowledge Processing. Springer, New York.

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.

Husmeier, D., Wright, F. and Milne, I. (2005b) Detecting interspecific recombination with a pruned probabilistic divergence measure. *Bioinformatics*, **21**, 1797–1806.

Kingman, J. F. C. (1982) The coalescent. *Stochastic Process Applications*, **13**, 235–248.

Larget, B. and Simon, D. L. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, **16**, 750–759.

Lehrach, W. and Husmeier, D. (2009) Segmenting bacterial and viral DNA sequence alignments with a transdimensional phylogenetic factorial hidden Markov model. *Applied Statistics*, **in press**.

Lehrach, W. P. (2008) *Bayesian machine learning methods for predicting protein-peptide interactions and detecting mosaic structures in DNA sequence alignments*. Ph.D. thesis, Univeristy of Edinburgh.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Minin, V. N., Dorman, K. S., Fang, F. and Suchard, M. A. (2005) Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, **21**, 3034–3042.

Neal, R. M. (2001) Annealed importance sampling. *Statistics and Computing*, **11**, 125–139.

Pickett, K. and Randle, C. (2005) Strange Bayes indeed: uniform topological priors imply non-uniform clade priors. *Molecular Phylogenetics and Evolution*, **34:1**, 203–211.

Raftery, A. E. (1996) Hypothesis testing and model selection. In Gilks, W. R., Richardson, S. and Spieglehalter, D. J. (eds.), *Markov Chain Monte Carlo in Practice*, pp. 163–187. Chapman & Hall, Suffolk. ISBN 0-412-05551-1.

Rambaut, N. C., A.; & Grassly (1996) Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*

Suchard, M. A., Weiss, R. E., Dorman, K. S. and Sinsheimer, J. S. (2003) Inferring spatial phylogenetic variation along nucleotide sequences: A multiple changepoint model. *Journal of the American Statistical Association*, **98**, 427–437.

Thompson, A. (1975) *Human evolutionary trees*. Cambridge University Press.

Tuffley, C. and Steel, M. (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, **59**, 581–607.

Velasco, J. (2008) The prior probabilities of phylogenetic trees. *Biology and Philosophy*, **23:4**, 455–473.

Wiuf, C., Christensen, T. and Hein, J. (2001) A simulation study of the reliability of recombination detection methods. *Molecular Biology and Evolution*, **18**, 1929–1939.

Yang, Z. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molecular Biology and Evolution*, **14**, 717–724.