Ann Inst Stat Math (2010) 62:145–159
DOI 10.1007/s10463-009-0259-8

Distribution of distances between topologies and its effect on detection of phylogenetic recombination

Leonardo de Oliveira Martins · Hirohisa Kishino

Received: 1 September 2008 / Revised: 3 April 2009 / Published online: 28 October 2009

Abstract Inferences about the evolutionary history of biological sequence data are greatly influenced by the presence of recombination, that tends to disrupt the phylogenetic signal. Current recombination detection procedures focus on the phylogenetic disagreement of the data along the aligned sequences, but only recently the link between the quantification of this disagreement and the strength of the recombination was realised. We previously described a hierarchical Bayesian procedure based on the distance between topologies of neighbouring sites and a Poisson-like prior for these distances. Here, we confirm the improvement provided by this topology distance and its prior over existing methods that neglect this information by analysing datasets simulated under a complex evolutionary model. We also show how to obtain a mosaic structure representative of the posterior sample based on a newly developed centroid method.

Keywords Viral recombination · SPR distance · Markov chain Monte Carlo · Phylogenetics

L. de Oliveira Martins (✉)
Department of Biochemistry, Genetics and Immunology, Faculty of Biology, University of Vigo, Campus Universitario, Vigo 36310, Spain
e-mail: leomrtns@uvigo.es

H. Kishino
Graduate School of Agriculture and Life Sciences, University of Tokyo, Bunkyo-ku Yayoi 1-1-1, Tokyo 113-8657, Japan
e-mail: kishino@lbm.ab.a.u-tokyo.ac.jp

1 Introduction

The main assumption of evolutionary theory is that the history of extant species can be traced back to one common ancestor in the past, through the process of speciation and extinction. In practice this usually amounts to representing the coalescence of individuals by a phylogenetic tree where leaves indicate the genotypes of extant taxa and internal nodes represent the ancestral forms. While this representation is certainly valid for one site, there are many cases where distinct sites do not share the same phylogeny. For example, in diploid populations each site came from one of the possible parents, and by crossing over distant sites may have distinct parents, in a manner dependent on the distance between the sites. Even haploid populations are subject to this disruptive force whereby distinct genomic regions support different phylogenetic trees. Examples include horizontal gene transfers, gene reassortment and viral recombination.

When the sequences are subjected to recombination, the amount of recombination (compared to the contribution of preserved mutations) will dictate if it is possible to reconstruct the phylogenetic history of the sequences. If recombination is rampant, then we should assume that each site has an independent evolutionary history, in which case the population genetic approach is more appropriate to describe the recombinational signal (Posada 2002; Awadalla 2003). Population genetic methods treat the individual phylogenetic trees as nuisance parameters, estimating population parameters that are averaged over all possible histories. The same applies whenever the recombination rate is higher than the substitution rate. On the other hand, if some phylogenetic signal is preserved we can detect recombination by a change in the underlying phylogenies. These phylogenies, nonetheless, have information about the recombinational process that most methods neglect, namely the minimum number of recombinations for a given break-point.

In this scenario the number of distinct topologies is not known in advance, and modelling directly the number of recombination breakpoints—the border between sequence regions supporting distinct topologies—generates a model with variable dimension. Previous attempts to infer the location of recombination breakpoints under a Bayesian framework have employed a reversible-jump MCMC strategy to cope with the variable number of phylogenetic segments and evolutionary model parameters, assuming a multiple change-point model (Suchard et al. 2003; Minin et al. 2005). Our strategy, instead, assumes that each site harbours a potentially distinct topology and model parameters independent of other sites. The dimensionality of the model (number of model parameters) is then constant and proportional to the number of sites, even when several of these parameters share the same value.

However, we employ a prior distribution on the distance between neighbouring sites that reflects the amount of recombination between neighbouring regions. These distances are analogous to latent variables, indicative of the homogeneity of the parameter values (Gelman 2004)—in our case, the topologies. This distance between the topologies facilitates the detection of recombination in the sense that fewer recombinations on the same break-point are more likely to occur, a priori, than many recombinations. This information was not given enough importance until recently (de Oliveira Martins et al. 2008). This same "parsimonious" recombination scenario

allows us to build an efficient Bayesian hierarchical procedure amenable to MCMC simulation.

In the present paper we make a systematic comparison between our most complex model and two of its special cases, namely neglecting the topological distance between segments and assuming a fixed penalty value for the prior over distances. For this comparison we use more complicated models than before, taking into account the variability in informativeness of sites. We also develop here a algorithm to summarise the distribution of break-points, in order to elect a centroid representative of the sample of mosaic structures. This sheds light on the question of how to compare samples exhibiting not only different break-point locations but also distinct number of break-points. Furthermore, we describe in detail the mini-sampler strategy adopted in the MCMC simulation to allow a better exploitation of the topological parameter space.

2 Methods

In the standard evolutionary likelihood model, the nucleotide substitution process at a given site is described by a continuous-time Markov chain and a phylogenetic tree describing the ancestral relations between extant taxa (Felsenstein 1981). We assume that the infinitesimal substitution probability matrix Q for the Markov chain follows the Hasegawa-Kishino-Yano (HKY) model (Hasegawa et al. 1985):

$$Q = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} 1-\kappa\pi_C-\pi_R & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & 1-\kappa\pi_T-\pi_R & \pi_A & \pi_G \\ \pi_T & \pi_C & 1-\kappa\pi_G-\pi_Y & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & 1-\kappa\pi_A-\pi_Y \end{pmatrix} \end{array} \tag{1}$$

where the parameter κ is the transition:transversion ratio between purines and pyrimidines, $\boldsymbol{\pi} = (\pi_T, \pi_C, \pi_A, \pi_G)$ is the vector of equilibrium frequencies of each base and $\pi_Y = \pi_C + \pi_T$, $\pi_R = \pi_A + \pi_G$. This matrix satisfies the reversibility condition, since $\pi_i Q_{ij} = \pi_j Q_{ji}$. The probability $P(y \mid x, t, \kappa, \boldsymbol{\pi})$ of going from state x to state y in time t is given by

$$P(y \mid x, t, \kappa, \boldsymbol{\pi}) = \sum_{k=1}^{4} \exp(t\Psi_k) Z_{yk} Z_{kx}^{-1} \tag{2}$$

where $\boldsymbol{\Psi}$ is the matrix of eigenvalues and $\mathbf{Z}, \mathbf{Z^{-1}}$ are the matrices of eigenvectors (Hasegawa et al. 1985). The matrix Q in Eq. 1 is scaled to an overall rate of one, so that the time-scale is given in number of substitutions. This means that we work with $Q' = Q/u$, where the overall rate $u = -\sum_i \pi_i Q_{ii}$ is, for the HKY model,

$$u = 2\left[\pi_R\pi_Y + \kappa(\pi_T\pi_C + \pi_A\pi_G)\right]$$

The probability $P(X \mid T, \mathbf{t}, \kappa, \boldsymbol{\pi})$ of observing a site pattern X for a column of the alignment given the unrooted phylogenetic tree T with branch lengths vector $\mathbf{t}$ under the HKY evolutionary model is the likelihood of $(T, \mathbf{t}, \kappa, \boldsymbol{\pi})$. It can be found by recursion, if we observe that the partial likelihood $L_r(z \mid \mathbf{t}, \kappa, \boldsymbol{\pi})$ of state $z = T, C, A, G$ for the subtree rooted at an internal node r depends only on the partial likelihoods $L_a(x \mid \mathbf{t}, \kappa, \boldsymbol{\pi})$ and $L_b(y \mid \mathbf{t}, \kappa, \boldsymbol{\pi})$ of its descendant nodes a and b, assuming a binary tree, as

$$L_r(z \mid \mathbf{t}, \kappa, \boldsymbol{\pi}) = \left[\sum_x P(x \mid z, t_a, \kappa, \boldsymbol{\pi}) L_a(x \mid \mathbf{t}, \kappa, \boldsymbol{\pi})\right] \times \left[\sum_y P(y \mid z, t_b, \kappa, \boldsymbol{\pi}) L_b(y \mid \mathbf{t}, \kappa, \boldsymbol{\pi})\right] \tag{3}$$

for a subtree with branch lengths $t_a, t_b \in \mathbf{t}$ connecting a and b, respectively, to r. The states at the terminal nodes should fit to the observed data, and define their corresponding partial likelihoods trivially. For an unrooted tree, the likelihood at a site may then be obtained by choosing arbitrarily an edge connecting nodes r and r_0, and summing over the possible states as (Felsenstein 2004)

$$P(X \mid T, \mathbf{t}, \kappa, \boldsymbol{\pi}) = \sum_x \sum_y \left[\pi_x L_r(x \mid \mathbf{t}, \kappa, \boldsymbol{\pi}) P(y \mid x, t_r, \kappa, \boldsymbol{\pi}) L_{r_0}(y \mid \mathbf{t}, \kappa, \boldsymbol{\pi})\right]. \tag{4}$$

2.1 Segmentation model

We assume that the DNA alignment for L taxa can be divided into K contiguous segments, such that all sites within a segment share the same evolutionary parameters and phylogeny, but these are allowed to vary between segments. The alignment $\mathbf{X}$ can then be divided into K segments $\mathbf{X} = (X_1, \ldots, X_K)$, each segment j of arbitrary size n_j such that $X_j = (X_{j1}, \ldots, X_{jn_j})$ are the columns of the alignment belonging to segment j. Since all parameters are shared within a segment, we can write the likelihood of the segment X_j as

$$P(X_j \mid T_j, \mathbf{t}_j, \kappa_j, \boldsymbol{\pi}_j) = \prod_{k=1}^{n_j} P(X_{jk} \mid T_j, \mathbf{t}_j, \kappa_j, \boldsymbol{\pi}_j) \tag{5}$$

We assume that the equilibrium frequencies are the same across segments, leading to $\boldsymbol{\pi}_j = \boldsymbol{\pi}$. We assume further that the branch lengths $t_{j1}, \ldots, t_{j\,2L-3}$ are independent realisations of an exponential distribution of mean μ_j, so that Eq. 2 can be marginalised over t to generate

$$P(y \mid x, \mu_j, \kappa_j, \boldsymbol{\pi}) = \sum_{k=1}^{4} \frac{Z_{yk} Z_{kx}^{-1}}{1 - \psi_k \mu_j}.$$

which can then be used for all branches. This allows the model to account for heterotachy while avoiding overparametrization (de Oliveira Martins et al. 2008). Since we assume that the parameters are shared among sites belonging to the same segment, care should be taken so that rate heterogeneity is properly modelled, with the ideal case being one site per segment. In our hierarchical setting we assume that the ratios κ_j and average rates μ_j follow exponential distributions with exponentially distributed hyper-priors shared across segments:

$$\begin{aligned} P(\mu_j \mid \mu_0) &= (1/\mu_0)\ e^{-\mu_j/\mu_0} \quad j = (1, \ldots, K) \\ P(\mu_0 \mid \mathcal{M}) &= (1/\mathcal{M})\ e^{-\mu_0/\mathcal{M}} \\ P(\kappa_j \mid \kappa_0) &= (1/\kappa_0)\ e^{-\kappa_j/\kappa_0} \quad j = (1, \ldots, K) \\ P(\kappa_0 \mid \mathcal{K}) &= (1/\mathcal{K})\ e^{-\kappa_0/\mathcal{K}} \end{aligned}$$

In de Oliveira Martins et al. (2008) we described a distance between topologies $\hat{d}_{SPR}(T_{j_1}, T_{j_2})$ that approximates the minimum number of subtree prune-and-regraft (SPR) moves separating the unrooted topologies T_{j_1} and T_{j_2} (Allen and Steel 2001). The rooted SPR distance equals the minimum number of recombinations between two rooted (time-directed) topologies, whose unrooted equivalents always have a smaller distance (Song 2003; Beiko and Hamilton 2006), and thus our $\hat{d}_{SPR}$ is a conservative quantification of recombination. Our strategy to correlate neighbouring segments j and $j+1$ ($j = 1, \ldots, K-1$) is to incorporate a prior on the distance $d_j = \hat{d}_{SPR}(T_j, T_{j+1})$ between their topologies T_j and T_{j+1}. This prior takes the form of a modified truncated Poisson:

$$\begin{aligned} P(d_j \mid \lambda_j, w_j, m) &= \frac{e^{-\lambda_j(w_j+1)}\lambda_j^{d_j(w_j+1)}}{\eta(\lambda_j, w_j, m) d_j!^{(w_j+1)}}; \\ \eta(\lambda_j, w_j, m) &= \sum_{d=0}^{m} \frac{e^{-\lambda_j(w_j+1)}\lambda_j^{d(w_j+1)}}{d!^{(w_j+1)}} \end{aligned} \tag{6}$$

with $m \leq L-3$ since the maximum distance between two trees on L taxa is $L-3$ (Allen and Steel 2001). We assume that topologies with a distance larger than m are prohibited, so care should be taken if deciding for $m < L-3$. If the number of segments is much larger than the number of recombination break-points, we expect the Poisson to be too lax for regions free from recombination. Therefore, this prior differs from a Poisson by the parameter w_j, which accommodates for underdispersion. The parameters λ_j and w_j are independent realisations of common gamma-distributed hyper-priors:

$$P(\lambda_j \mid \alpha_\lambda, \beta_\lambda) = \frac{\beta_\lambda^{\alpha_\lambda}}{\Gamma(\alpha_\lambda)} \lambda_j^{\alpha_\lambda - 1} e^{-\beta_\lambda \lambda_j} \quad j = (1, \ldots, K-1)$$

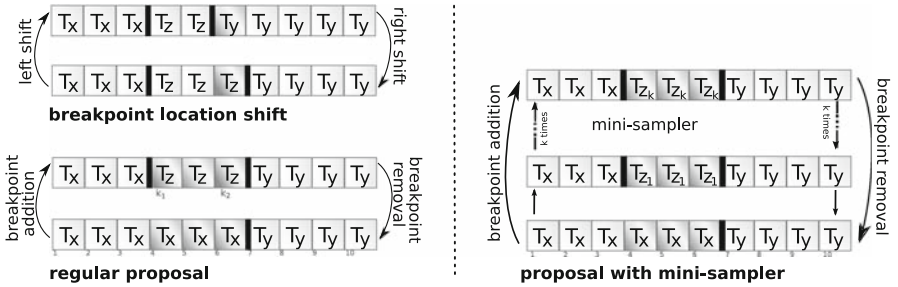

Fig. 1 Diagram representing the proposal of an addition, removal and location shift of a breakpoint. The *boxes* represent the alignment segments, labelled by their topologies (represented by T_x, T_y, etc.). The *thick vertical bars* represent the break-points, since the topologies surrounding them are different, and the *shaded boxes* indicate the segments that are affected by the proposal. For the addition and removal of a break-point, the left panel represents the simple proposal mechanism while in the right panel we have the mini-sampler strategy. Proposing a shift in the location of the break-point is the same for both strategies

$$P(w_j \mid \alpha_w, \beta_w) = \frac{\beta_w^{\alpha_w}}{\Gamma(\alpha_w)} w_j^{\alpha_w - 1} e^{-\beta_w w_j} \quad j = (1, \ldots, K-1)$$

where $\Gamma(\alpha)$ is the gamma function (Gelman et al. 2003).

2.2 Markov chain Monte Carlo (MCMC) scheme

The posterior distribution is simulated through a Gibbs sampler where each parameter is independently updated by a Metropolis-Hastings step, and two chains at different temperatures are run concurrently, with occasional swap of states (de Oliveira Martins et al. 2008). The continuous parameters θ_i are updated by a random perturbation where the new state θ_i^* is sampled through $\theta_i^* = \theta_i e^{\xi_{\theta_i}(u-0.5)}$ where $u \sim \text{uniform}(0, 1)$and ξ_{θ_i} is an arbitrary constant. The update of topologies is done in blocks of segments sharing the same topology, and the changes in the number and location of recombination break-points are proposed according to Fig. 1. In order to ensure detailed balance, it is enough to set up the frequency of a break-point addition f ($f \leq 0.5$) as being equal to the frequency of break-point removal. With probability $(1 - 2f)$ a change in break-point location (shift) is tried (Dimatteo et al. 2001). The topology distances between segments d_j are updated indirectly as a consequence of updating the topologies. Further details can be found in de Oliveira Martins et al. (2008).

A regular breakpoint addition proposal (displayed in the left panel of Fig. 1) corresponds to applying one SPR to the current topology. Thus, by design, the proposed topology will always have $\hat{d}_{SPR} = 1$ to one of its neighbouring segments (for example, T_x and T_z in the left panel of Fig. 1). To avoid favouring this scenario of neighbouring segments harbouring only similar topologies, we developed a mini-sampler strategy inspired by the reversible-jump MCMC problem of proposing a change in the number of dimensions (Al-Awadhi et al. 2004). The proposal is represented in the right panel of Fig. 1, and consists of running a few iterations where the topologies within a non-recombinant block are updated according to a heated posterior distribution $[P(x)]^h$ before deciding for the acceptance/rejection of the new state. Here x represents an

arbitrary state and $h > 0$ is the temperature (usually smaller than one). When proposing a break-point addition, the mini-sampler under the heated distribution is simulated after increasing the number of break-points, and when proposing a reduction in the number of break-points the mini-sampler is run prior to removing the break-point, to guarantee detailed balance. If we follow the right panel of Fig. 1 in representing the states before the break-point addition, just after the addition and after the mini-sampler as x, z_1 and z_k, respectively, the acceptance probability for the addition proposal will be given by $\min(1, A(z_k \mid x))$ where

$$A(z_k \mid x) = \frac{P(z_k)}{P(x)} \left[\frac{P(z_1)}{P(z_k)}\right]^h \frac{q(x \mid z_1)}{q(z_1 \mid x)} \tag{7}$$

The definition of the proposal ratio $q(x \mid z_1)/q(z_1 \mid x)$ is given in de Oliveira Martins et al. (2008). In the case of proposing a break-point removal, the inverse of Eq. 7 should be used. For a bad choice of h, the typical value of z_k will be z_1 and we recover the original acceptance probability, while if $h = 1$ then the final state z_k will always be accepted—since all intermediate states within the mini-sampler were subjected already to an acceptance-rejection. We realise that this mini-sampler is equivalent to a within-chain Metropolis-coupled MCMC (MC-MCMC) by replacing z_1 by x in Eq. 7 and noticing that it now reduces to the probability of swap between chains in the MC-MCMC simulation (Altekar et al. 2004; de Oliveira Martins et al. 2008).

2.3 Posterior samples

All relevant information about the parameters can be retrieved by storing the sampled values along the MCMC simulation, and the ensemble of sampled points will form the posterior distribution of each variable. We are mainly interested in the distribution of $\hat{d}_{SPR}$ distances and sampled topologies for each segment, which can be represented as matrices of dimension $(K-1) \times N$ for the distances and $K \times N$ for the topologies, where K is the number of segments and N is the number of samples from the posterior distribution. The matrix of sampled topologies $\mathbf{T} = \{T_{ij}\}$ has the information about the topologies T_{ij} sampled at iteration i ($i = 1, \ldots, N$) for segment j ($j = 1, \ldots, K$), and the element d_{ij} of the matrix of distances $\mathbf{d} = \{d_{ij}\}$ holds the $\hat{d}_{SPR}$ between topologies T_{ij} and $T_{i\ j+1}$.

We have previously shown (de Oliveira Martins et al. 2008) how we can summarise these posterior samples to make inferences about the alignment regions more likely to be the result of recombination by looking at the at the posterior mean distances $\overline{d_{\cdot j}}$, where the average is over all samples for each segment. Values of $\overline{d_{\cdot j}}$ larger than one might indicate a potential hotspot for region j, since the $\hat{d}_{SPR}$ is a lower bound on the number of recombinations.

On the other hand, more often than not, the researcher is interested in a point estimate of recombination break-points, the so-called mosaic structure. This is not trivial since $\overline{d_{\cdot j}}$ is usually a multimodal distribution. One alternative is to look at the distribution of the average frequency of recombination $\bar{I}_{d_{\cdot j}>0}$ for each segment, where I_x is the indicator function. This way, given the number of break-points (number of

modes, ideally) we can divide the alignment into unimodal regions and obtain a point estimate for the break-point location for each region. Credible sets can be constructed based on these posterior mean frequencies, and the number of modes can be estimated through the distribution of break-points $s_i = \sum_{j=1}^{K-1} I_{d_{ij}>0}$ over the samples i, or simply by $\sum_{j=1}^{K-1} \bar{I}_{d_{\cdot j}>0}$. A point estimate could also be found by looking at the most frequent (maximum *a posteriori*—MAP) topologies $map(T_{\cdot j})$ for each segment and a break-point j^* inferred whenever $map(T_{\cdot j^*}) \neq map(T_{\cdot j^*+1})$, but it can lead to overestimates due to random fluctuations (de Oliveira Martins et al. 2008).

Another point estimate can be obtained by finding the centroid mosaic structure S^*, that minimises its distance to all other mosaic structures S_i ($i = 1, \cdots, N$) (Ding et al. 2005; Webb-Robertson et al. 2008; Carvalho and Lawrence 2008). The mosaic structure S_i of sample i can be represented by a vector $S_i = \{S_{i0}, \ldots, S_{is_i}, S_{i\,s_i+1}\}$ where S_{is} is the position of the last site of the s-th break-point—that is, the position of the last site belonging to segment j^* such that $\sum_{j=1}^{j^*} I_{d_{ij}>0} = s$. There are two special "break-points" in the boundaries of the alignment, the position $S_{i0} = 0$ and the position $S_{i\,s_i+1}$ which equals the sequence length, for all samples. The distance $D(S_{i_1}, S_{i_2})$ between two mosaics S_{i_1} and S_{i_2} can then be estimated by the *ad-hoc* function

$$D(S_{i_1}, S_{i_2}) = D(S_{i_2}, S_{i_1}) = \sum_{s=1}^{s_{i_1}} \min_{k=0}^{s_{i_2}+1} (|\, S_{i_1 s} - S_{i_2 k}\,|) + \sum_{s=1}^{s_{i_2}} \min_{k=0}^{s_{i_1}+1} (|\, S_{i_2 k} - S_{i_1 s}\,|). \tag{8}$$

This distance is based on a mapping between each break-point in mosaic S_{i_1} and its closest equivalent in mosaic S_{i_2}, with the property of being less influenced by the number of break-points than by their locations. Our centroid estimate will then be the mosaic structure S_{i*} such that the total distance $\sum_{i=1}^{N} D(S_{i*}, S_i)$ from other mosaics is minimal. Here we make the simplifying assumption that the centroid can be found among the samples, but there may be other mosaics, not present in the posterior samples, with a total distance even smaller. Given a point estimate of the mosaic structure, the underlying topologies can also be found through the modal value of the MAP topologies $map(T_{\cdot j})$ for the segments between the break-points.

3 Results

We simulated 100 alignments with 8 taxa, supporting distinct evolutionary histories at every 64 base pairs (bp), comprising a total alignment of 256 bp. For each dataset we ran the MC-MCMC for 10^5 iterations, sampling at each 500 iterations (200 posterior samples in total), after a burn-in period of 5×10^3 iterations. We assumed that the alignment could be divided into 128 segments composed of two sites each, and unless otherwise stated the truncation term m was set to five, and the hyperparameters for the penalty w_j were set at $\alpha_w = \beta_w = 1$. For all simulations we further assumed that $\alpha_\lambda = 1$ and $\beta_\lambda = 128$, and the number of steps of the mini-sampler was set at two with a temperature $h = 0.6$. We call this scenario the "unrestricted

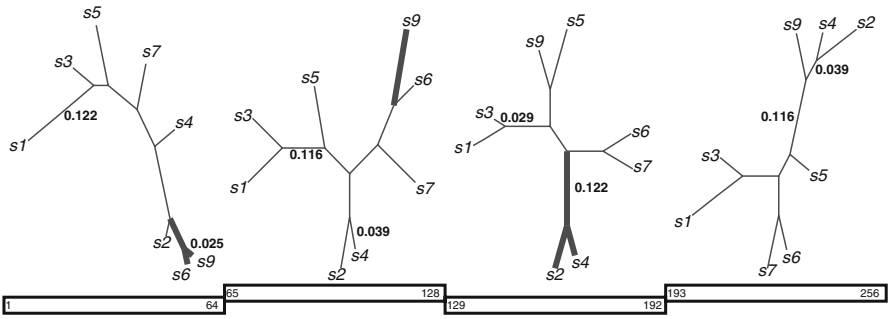

Fig. 2 Trees used to simulate the recombinant alignment of 256 base pairs. The branch lengths represented in each tree are proportional to the amount of evolution, with the total branch lengths summing up to one substitution per site. The labels on the branches are the smallest and largest branch lengths for each tree, and the branches highlighted in *bold* represent one possible recombination scenario. On the bottom we have the alignment regions where each tree was used on the simulation

model", since we make no further assumptions, and then compare it with two simplified models.

3.1 Simulation scenario

We simulated alignments evolving under evolutionary scenarios where we had heterogeneity at three levels: of branch lengths, model parameters and evolutionary rates. The alignment was simulated by making each site evolve according to the topologies displayed in Fig. 2. In this figure we realise that there are branches were substitutions are more prone to occur, like for instance the long branch leading to sequence *s1* in the first segment. Since we expect most substitutions to occurs there, we might need many sites to have information about a particular branch, like for example the branches around sequences *s6* and *s9*.

The sequences evolved according to a general time reversible (GTR) Markov model of nucleotide substitution (Yang 1994a; Tavaré 1986) and under a discretized gamma model of rate heterogeneity between sites. The GTR model can be described by the rate matrix

$$Q = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{pmatrix} \end{array} \tag{9}$$

where the diagonals are such that the row sums equal zero. In our simulations the rate ratios were fixed at $a = 0.1$, $b = 0.3$, $c = 0.5$, $d = 0.7$, $e = 0.9$ and $f = 1$, while the equilibrium frequencies were set at $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$ and $\pi_G = 0.4$. This is a very heterogeneous model, more complex than the HKY model we include in our hierarchical procedure which assumes that $a = f = \kappa$ and $b = c = d = e = 1$.

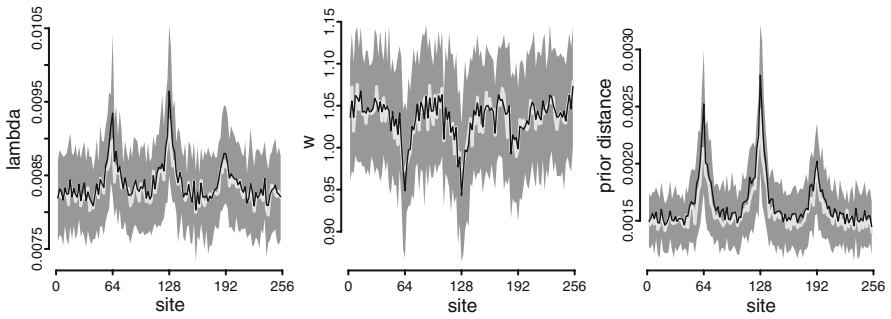

Fig. 3 Posterior distributions for 100 simulations under the unrestricted model. The *dark grey shadows* represent the interquartile range, the *light grey lines* represent the median and the *black lines* show the mean values over the simulated datasets. The *left panel* shows the distribution of the posterior mean of λ_j, while on the *middle panel* we have the posterior mean of the penalties w_j. On the *right panel* we have the mean of the modified Poisson distribution (Eq. 6) calculated through the method of moments

We assumed that the substitution rates are heterogeneous across sites, with the level of heterogeneity being described by a rate factor r following a gamma distribution with $\beta = \alpha$, such that $E[r] = 1$ and $Var[r] = 1/\alpha$ (Yang 1993, 1994b). The lower the value of α the higher the heterogeneity, with most sites having low information (low average number of substitutions) and a few sites being saturated (too many substitutions). We assumed a gamma distribution with $\alpha = 5$ discretized into four categories and an average substitution rate of one. The discretized rates are 0.50, 0.80, 1.06 and 1.61, which means that for 64 bp, on average ten sites will be monomorphic (no substitutions) and only 16 sites will harbour more than one substitution (Yang 1994b). It is worth noticing that on real datasets (like HIV-1) α is usually much smaller but the sequence length is much larger. The software PAML, described in Yang (2007), was used to simulate the alignments.

3.2 Unrestricted model

We were interested in observing the behaviour of the modified Poisson distribution, so for each λ_{ij} and w_{ij} sampled at iteration i we estimated its mean value o_{ij} as the first moment around zero:

$$o_{ij} = \sum_{d=0}^{m} dP(d \mid \lambda_{ij}, w_{ij}, m) \tag{10}$$

where $P(d \mid \lambda_{ij}, w_{ij}, m)$ is defined in Eq. 6. In Fig. 3 we have the distribution of the posterior means $\overline{\lambda_{\cdot j}}$, $\overline{w_{\cdot j}}$ and $\overline{o_{\cdot j}}$ calculated for each dataset. In this figure we can see that for regions away from the break-points not only λ_j is lower but w_j is higher, with the opposite being observed around the break-points (sites 64, 128 and 192). Their compound effect amounts to a sharper distinction, when we observe the distribution of their $\overline{o_{\cdot j}}$. The mean value $\overline{o_{\cdot j}}$ reflects in fact the mean prior for d_j.

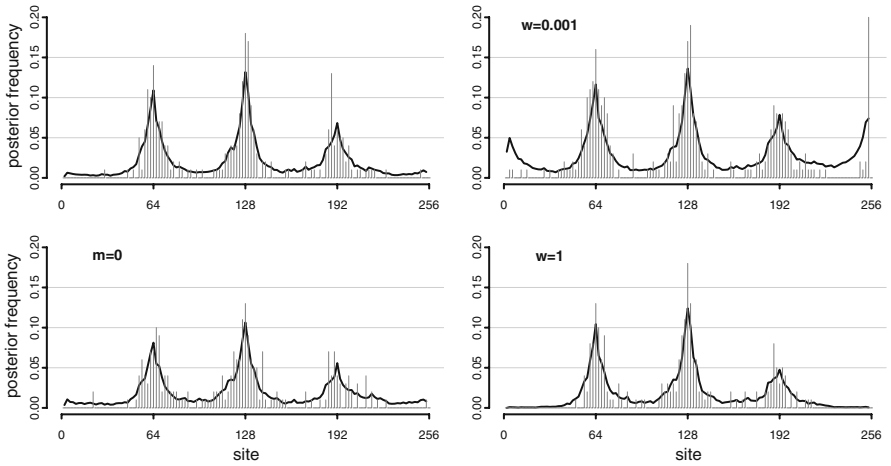

Fig. 4 Site-wise distribution of recombination break-points for different prior assumptions, over 100 simulated datasets. The *top-left panel* represents the unrestricted prior, with maximum topology distance $m = 5$ and penalty parameter w such that $E[w] = Var[w] = 1$. In the *left panel* at the bottom we assume that the recombination distance is not taken into account, keeping the hyper-prior on the penalty as in the unrestricted model. For the panels at the right, we enforce the penalty parameter to a fixed value by setting $Var[w] = 10^{-8}$ (with $E[w] = 10^{-3}$ for the *top panel* and $E[w] = 1$ for the *bottom panel*), while maintaining the maximum distance at $m = 5$. For each panel, we show the mean over 100 simulations of the posterior recombination distribution using two sample estimates: the black lines represent the site-wise average, over all samples, of the break-point frequency ($\hat{d}_{SPR} > 0$), and the *grey vertical bars* show the break-points belonging to the centroid sample of each dataset

3.3 Alternative prior: no distance information

To test whether the distance between topologies has any effect on the recombination detection, we used a modified model where the distance is not taken into account, and the prior described in Eq. 6 reduces to

$$\begin{aligned} P(T_j = T_{j+1} \mid \lambda_j, w_j) &= \frac{1}{1 + \lambda^{(w_j+1)}}; \\ P(T_j \neq T_{j+1} \mid \lambda_j, w_j) &= \frac{\lambda^{(w+1)}}{1 + \lambda^{(w_j+1)}} \end{aligned} \tag{11}$$

This equation is similar to Eq. 6 with $m = 1$, but since we assume that the equality can be inferred but the similarity can not be quantified we refer to this simplified model as the $m = 0$. Its effect can be seen in the left panels of Fig. 4. The plot at the top shows the original prior with distance information, showing that both the posterior $\bar{I}_{d_{.j}>0}$ and the centroid-estimated mosaic structures can infer the recombination break-points reasonably well, especially for the first two break-points. On the bottom panel, on the other hand, the information about the break-point locations is more sparse. The unrestricted analyses detected the correct number of break-points in 56% of the simulations, while underestimating it in 39% and overestimating it in 5% of the datasets. The analyses without topology distance information inferred correctly the

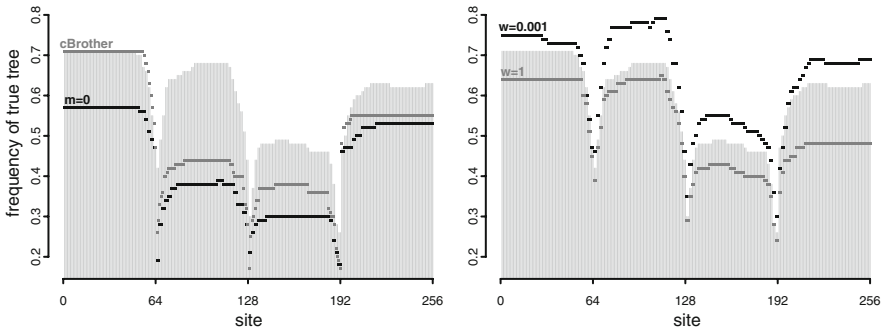

Fig. 5 Frequency, over 100 simulations, where the true trees (described in Fig. 2) were found using different methods. On both panels the performance of the unrestricted model is shown by the *light gray bars*. On the *left panel* we have two methods that do not take the topology distance into consideration: our modified model (*black points*) and the cBrother software (*dark gray points*). On the *right panel* we compare the original model and the penalty parameter fixed at 0.001 (*black points*) or fixed at one (*dark gray points*)

three recombination break-points in only 30% of the cases, while underestimated this number in 63% of the simulations.

Another Bayesian procedure for recombination detection has been described in Minin et al. (2005), which is based on a multiple change-point model where the number of break-points is explicitly taken into account (Suchard et al. 2003). This model (as most recombination detection procedures widely employed) is not able to quantify the difference between recombinant regions, and so we compared our procedure with the implementation of this change-point model developed by Fang et al. (2007) and present in the software cBrother. The comparison is summarised in the left panel of Fig. 5, where the performance of each method is measured by its ability in inferring the original topology, used to simulate the alignment. While the performance of cBrother is better than our simplified model without distance information, it is less accurate than the model using Eq. 6. The first 64 bp were the most informative according to any analysis, while the region between sites 128 and 192 was the less informative, suggesting that this is not an artefact of the algorithm. The cBrothers software detected three recombination break-points in 40% of the datasets, and detected less than three break-points in the other 60% of the simulations.

3.4 Alternative prior: constrained hierarchical model

To further explore the influence of the prior distribution on recombination detection, we analysed the same datasets by setting w_j for all segments at a fixed value. This was done, in practice, by choosing α_w and β_w such that the variance is very small (namely, 10^{-8}). The gamma-distributed prior for λ_j is a robust alternative to a Poisson distribution, and to compare our results with a prior for the topology distance closer to the negative-binomial (Poisson marginalised over λ), we set up w_j to be kept at a low value. We fixed $E[w_j] = 10^{-3}$ by setting $\alpha_w = 10^2$ and $\beta_w = 10^5$, which is not equivalent to the negative-binomial (we still have the truncation term and the exponent $w_j + 1 > 1$) but reflects a prior where the penalty parameter w_j

has a smaller influence. We also used a prior where w_j was fixed at one by choosing $\alpha_w = \beta_w = 10^8$, to compare with the previous settings.

The results are summarised in the right panels of Figs. 4 and 5. As expected (by looking at Fig. 3, e.g.) the low w_j had a poor performance on regions away from the "real" break-points—the exact location can not be found with arbitrary precision in general due to stochastic fluctuations—and the stringent $w_j = 1$ performed worse than the unrestricted model in the recombinant regions. This distinction is more evident on the borders, where there is no information from neighbouring sites. For $w_j = 10^{-3}$, the correct number of break-points was found in only 41% of the datasets, while for $w_j = 1$ in only 35% of the datasets the correct number was found. In 47% of the simulations the low-penalty setting overestimated the number of recombination break-points and in 64% of the datasets the high penalty assumption underestimated it. The topology inference itself was not so much affected, with the low penalty prior performing even better, on average, than the less informative prior of the unrestricted model. This might be due to a better exploration of the topology space, at the cost of inferring spurious recombination.

4 Discussion

Our present work shows that the power to detect recombinations is significantly improved by introducing a prior on the distance between the topologies of neighbouring segments. The performance of the prior depends on a hierarchical setting allowing for underdispersion as well as overdispersion of the distribution of distances. In addition, it was crucial to develop a distance measure which reflects well the number of recombination events required to explain the difference between the topologies. Our point estimates for the mosaic structure based on the sample closest to the centroid structure also reflected well the posterior distribution of break-points, since our proposed distance between mosaics is more influenced by the closeness between break-points than by the number of break-points alone.

When the number of parameters necessary to explain the data is not known in advance a strategy like reversible-jump MCMC can be employed to ensure that the dimensionality will not increase beyond a desirable level (Dimatteo et al. 2001; Suchard et al. 2003). Modelling and sampling between different dimensions, however, can become very complicated. An alternative strategy is to assume that the number of variables is constant (let us say, at its maximum dimension) and then work with an augmented data where latent variables will tell us whether there is a change in the values of other parameters (Mitchell and Beauchamp 1988). The number of "active" latent variables can be regarded as the number of necessary variables in the non-augmented model, since the block of variables sharing the same value in the augmented model can be equivalent to one variable in another model. Our model is an extension of this reasoning, where the latent variables not only indicate the minimum number of distinct variables (non-recombinant segments in our case) but also quantifies the difference between them. We believe that this procedure of having a latent variable interpretable as more than an indicator can be used as a replacement to variable

models, provided we can quantify the difference between variables and we are able to model this quantification through an adequate prior.

We have chosen the HKY model since it is the most complex model that can be solved analytically for arbitrary time intervals, and is a good compromise between accuracy and simplicity (Yang 1994a). Our model takes account of rate heterogeneity among sites, which is known to be indispensable for unbiased estimation of phylogenetic trees. Additionally, our marginal likelihood over the branch lengths allows for heterotachy, that is, the variability of the relative lengths of branches of the trees among sites in each of the segments. It also alleviates the computational burden of sampling individual branch lengths.

Therefore, the crucial factor that affects the power and the bias of the estimation of recombination in a phylogenetic context should incorporate the distance between topologies and the priors on the distances. The AIC (Akaike 1974) is widely used for the selection of evolutionary models in the inference of phylogenetic trees (Posada and Buckley 2004). We could have used it, for instance, to quantify the applicability of the HKY assumption of our model, since the sequences were simulated under a distinct model. It may be an interesting future study to select the most appropriate form of the prior based on the MCMC-based criteria like Bayes factor (Kass and Raftery 1995) or DIC (Spiegelhalter et al. 2002).

Acknowledgments We would like to thank Ziheng Yang and Mahendra Mariadassou for valuable discussions on our work. The careful comments of an anonymous reviewer are also much appreciated. This work was supported in part by a Grant-in-Aid for Scientific Research (B-19300094) from the Japan Society for the Promotion of Science (JSPS).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Al-Awadhi, F., Hurn, M., Jennison, C. (2004). Improving the acceptance rate of reversible jump MCMC proposals. *Statistics and Probability Letters, 69*(2), 189–198.

Allen, B., Steel, M. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics, 5*(1), 1–15.

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics, 20*(3), 407–415.

Awadalla, P. (2003). The evolutionary genomics of pathogen recombination. *Nature Reviews Genetics, 4*(1), 50–60.

Beiko, R. G., Hamilton, N. (2006). Phylogenetic identification of lateral genetic transfer events. *BMC Evolutionary Biology, 6*, 15.

Carvalho, L. E., Lawrence, C. E. (2008). Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proceedings of the National Academy of Sciences USA, 105*(9), 3209–3214.

Dimatteo, I., Genovese, C., Kass, R. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika, 88*(4), 1055–1071.

Ding, Y., Chan, C. Y., Lawrence, C. E. (2005). Rna secondary structure prediction by centroids in a boltzmann weighted ensemble. *RNA, 11*(8), 1157–1166.

Fang, F., Ding, J., Minin, V. N., Suchard, M. A., Dorman, K. S. (2007). cBrother: relaxing parental tree assumptions for Bayesian recombination detection. *Bioinformatics, 23*(4), 507–508.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution, 17*(6), 368–376.

Felsenstein, J. (2004). *Inferring phylogenies*. Sunderland, MA: Sinauer Associates.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association, 99*(466), 537–545.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed). Boca Raton, FL: Chapman & Hall/CRC.

Hasegawa, M., Kishino, H., Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution, 22*(2), 160–174.

Kass, R. E., Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773–795.

Minin, V. N., Dorman, K. S., Fang, F., Suchard, M. A. (2005). Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics, 21*(13), 3034–3042.

Mitchell, T. J., Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association, 83*(404), 1023–1032.

de Oliveira Martins, L., Leal, É., Kishino, H. (2008). Phylogenetic detection of recombination with a Bayesian prior on the distance between trees. *PLoS ONE, 3*(7), e2651.

Posada, D. (2002). Evaluation of methods for detecting recombination from dna sequences: empirical data. *Molecular Biology and Evolution, 19*, 708–717.

Posada, D., Buckley, T. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology, 53*(5), 793–808.

Song, Y. (2003). On the combinatorics of rooted binary phylogenetic trees. *Annals of Combinatorics, 7*(3), 365–379.

Spiegelhalter, D., Best, N., Carlin, B., van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B, 64*(4), 583–639.

Suchard, M., Weiss, R., Dorman, K., Sinsheimer, J. (2003). Inferring spatial phylogenetic variation along nucleotide sequences: a multiple changepoint model. *Journal of the American Statistical Association, 98*(462), 427–438.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In R.M. Miura (Ed.), *Some Mathematical Questions in Biology—DNA Sequence Analysis* (pp. 57–86). Providence: AMS Bookstore.

Webb-Robertson, B. J. M., McCue, L. A., Lawrence, C. E. (2008). Measuring global credibility with application to local sequence alignment. *PLoS Computational Biology, 4*(5), e1000077.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Molecular Biology and Evolution, 10*(6), 1396–1401.

Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution, 39*(1), 105–111.

Yang, Z. (1994b). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution, 39*(3), 306–314.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution, 24*(8), 1586–1591.