BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 13 2005, pages 3034–3042
doi:10.1093/bioinformatics/bti459

Genetics and population analysis

Dual multiple change-point model leads to more accurate recombination detection

Vladimir N. Minin[1], Karin S. Dorman[2,3,4], Fang Fang[4] and Marc A. Suchard[1,*]

[1]Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095-1766, USA, [2]Department of Statistics and [3]Department of Genetics, Cell & Development Biology and [4]Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011, USA

Received on August 5, 2004; revised on April 14, 2005; accepted on April 18, 2005
Advance Access publication April 21, 2005

ABSTRACT

Motivation: We introduce a dual multiple change-point (MCP) model for recombination detection among aligned nucleotide sequences. The dual MCP model is an extension of the model introduced previously by Suchard and co-workers. In the original single MCP model, one change-point process is used to model spatial phylogenetic variation. Here, we show that using two change-point processes, one for spatial variation of tree topologies and the other for spatial variation of substitution process parameters, increases recombination detection accuracy. Statistical analysis is done in a Bayesian framework using reversible jump Markov chain Monte Carlo sampling to approximate the joint posterior distribution of all model parameters.

Results: We use primate mitochondrial DNA data with simulated recombination break-points at specific locations to compare the two models. We also analyze two real HIV sequences to identify recombination break-points using the dual MCP model.

Availability: A software program 'DualBrothers' implementing the dual MCP model is available in the form of a Java package at http://www.biomath.ucla.edu/msuchard/DualBrothers

Contact: msuchard@ucla.edu

Supplementary information: http://www.biomath.ucla.edu/msuchard/DualBrothers

1 INTRODUCTION

Recombination plays an important role in the evolution of almost all living organisms. In rapidly evolving viruses, homologous recombination is one way in which the viruses adapt quickly to changing environmental conditions (Worobey and Holmes, 1999). At least 10% of the circulating human immunodeficiency virus-1 (HIV-1) strains are believed to be recombinants containing genetic material from different viral subtypes (Robertson *et al*., 1995a,b). Recombination in HIV holds implications for vaccine development (Korber *et al*., 2001) and emerging drug resistance (Kellam and Larder, 1995). While point mutation processes have been used to study HIV immune response escape (Wei *et al*., 2003) and drug resistance development (Chen *et al*., 2004), the contributions of recombination are not well understood. Accurate knowledge of the frequency and location of recombination break-points may improve our understanding of these phenomena, but reliable and statistically rigorous methods are needed to provide this break-point information.

Most methods that can detect recombination from a multiple sequence alignment use statistical, phylogenetic procedures (Hein, 1990). These methods exploit the observation that if recombination occurred in the evolutionary history of a set of aligned sequences, then different segments of the alignment should support alternative phylogenies (Li *et al*., 1988). One of the most popular approaches to recombination detection is to slide a window along a sequence alignment and look for differences in the phylogenetic tree support within each window (Grassly and Holmes, 1997; McGuire *et al*., 1997; Husmeier and Wright, 2001). Although this approach can successfully detect recombination, it suffers from a multiple testing problem when assessing the significance of recombination (Suchard *et al*., 2002) and low resolution for locating recombination break-points, limited by the window size.

Husmeier and McGuire (2003) develop a Bayesian hidden Markov model (HMM), where the hidden states are phylogenetic trees and the observable states are consecutive nucleotide sites of a multiple sequence alignment. This method can predict recombination sites more accurately than sliding window methods as shown by the authors, but their current implementation is limited to only four sequences, because the dynamic programming, required for HMM inference, is computationally very expensive. This method also assumes that all regions of the alignment are under the same evolutionary pressure. This assumption is known to lead to false recombination identification under some circumstances (Dorman *et al*., 2002; Husmeier and McGuire, 2002). Finally, trees in the HMM are updated based only on the phylogenetic information from the neighboring sites. Therefore, noisy and sparse data can also reduce the accuracy of Bayesian HMM methods.

Suchard *et al*. (2003b) proposed a single multiple change-point (MCP) model that can capture spatial phylogenetic variation in both the trees and the evolutionary pressures. In this model, an alignment is partitioned into an unknown number of segments. Each segment has a vector of phylogenetic parameters associated with it, such as parameters describing the nucleotide substitution process and a bifurcating tree topology that specifies evolutionary relationships between sequences. End points between partitions are called change-points. Recombination is inferred if at least one change in topology is observed across a change-point. This model has been successfully applied to test recombination hypotheses in HIV strains (Suchard *et al*., 2002).

Modeling spatial variation of all parameters with a single change-point process results in prior correlation between sites where

[*]To whom correspondence should be addressed.

substitution parameters change and those where topologies change. This prior correlation can lead to biased break-point estimation when recombination occurs near the boundary of regions with different evolutionary pressures. When both substitution process parameters and topologies change at close, yet distinct, sites the single MCP will probably produce only one change-point in the neighborhood of the two true change-points. To overcome this difficulty, we develop a dual MCP model that decouples substitution process change-points from topology break-points by introducing two a priori independent change-point processes to describe spatial phylogenetic variation.

2 METHODS

2.1 Dual MCP model

We start with N aligned DNA or RNA sequences of length S. Columns of the alignment, also called sites, $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_S$ serve as observations of the evolutionary process. Each site $\mathbf{Y}_s = (Y_{s1}, Y_{s2}, \dots, Y_{sN})'$ contains a nucleotide or gap from each of the N sequences, such that $Y_{sn} \in \{\mathrm{A, G, C, T/U}, -\}$, for $s = 1, 2, \dots, S$, and $n = 1, 2, \dots, N$. We follow conventional likelihood-based phylogenetic approaches (Felsenstein, 2004) and model the evolutionary process in terms of an evolutionary tree $(\tau, \mathbf{T})$ and a rate matrix $\mathbf{Q}$, where τ is a bifurcating tree topology and $\mathbf{T} = (t_1, t_2, \dots, t_{2N-3})$ is a vector of branch lengths of τ. Matrix $\mathbf{Q} = \{q_{uv}\}$, $u, v \in \{\mathrm{A, G, C, T/U}\}$, defines the rates for a continuous-time Markovian nucleotide substitution process along each branch of τ. We follow the parameterization of $\mathbf{Q}$ by Hasegawa *et al.* (1985),

$$\mathbf{Q} = \begin{pmatrix} - & \alpha\pi_\mathrm{G} & \beta\pi_\mathrm{C} & \beta\pi_\mathrm{T} \\ \alpha\pi_\mathrm{A} & - & \beta\pi_\mathrm{C} & \beta\pi_\mathrm{T} \\ \beta\pi_\mathrm{A} & \beta\pi_\mathrm{G} & - & \alpha\pi_\mathrm{T} \\ \beta\pi_\mathrm{A} & \beta\pi_\mathrm{G} & \alpha\pi_\mathrm{C} & - \end{pmatrix}, \qquad (1)$$

where α is a transition rate, β is a transversion rate, $\boldsymbol{\pi} = (\pi_\mathrm{A}, \pi_\mathrm{G}, \pi_\mathrm{C}, \pi_\mathrm{T})$ is the stationary distribution of the nucleotides and the diagonal elements of $\mathbf{Q}$ are set such that the rows of $\mathbf{Q}$ sum to 0. The resulting finite-time transition matrix for substitutions is $\mathbf{P}(t_b) = \mathrm{e}^{t_b\mathbf{Q}} = \{p(u, v|t_b)\}$, where $p(u, v|t_b)$ is the probability of nucleotide u mutating to v along branch b, $b = 1, 2, \dots, 2N-3$. For identifiability between t_b and $\mathbf{Q}$, we fix β such that $\sum q_{uu}\pi_u = -1$ and branch lengths are expressed in terms of the expected number of substitutions per site. The transition/transversion ratio $\kappa = \alpha/\beta$ remains a free parameter ranging from 0 to ∞. This parameterization differs from the parameterization used by Suchard *et al.* (2003b), where transition rate $\alpha \in [0, 1]$ plays the role of the free parameter. Although $\boldsymbol{\pi}$ can be estimated simultaneously with other model parameters, the resulting estimates normally differ little from the observed frequencies in the alignment. Therefore, we fix the stationary distribution $\boldsymbol{\pi}$ to the observed nucleotide frequencies in the alignment to avoid unnecessary computations (Li *et al.*, 2000). Felsenstein (1981) provides an efficient algorithm for integrating out gaps and computing the site likelihood $f(\mathbf{Y}_s|\tau(s), \mathbf{T}(s), \mathbf{Q}(s))$, where $(\tau(s), \mathbf{T}(s), \mathbf{Q}(s))$ are the evolutionary parameters associated with site s. To reduce computations further, we assume a coalescent-like prior on branch lengths, such that $p(t_b(s)) \propto \exp(-t_b(s)/\mu(s))$ (Sinsheimer *et al.*, 2003). We refer to the prior mean of branch lengths $\mu(s)$ as the average divergence at site s. Then $\mathbf{T}(s)$ can be integrated out of the likelihood as shown in Suchard *et al.* (2003b), producing marginal probabilities of mutation $p(u, v|\mu(s))$ and marginal likelihoods $f(\mathbf{Y}_s|\tau(s), \mu(s), \mathbf{Q}(s))$. Assuming independence across sites conditional on the evolutionary parameters of the model, the total likelihood of the alignment becomes

$$f = \prod_{s=1}^{S} f(\mathbf{Y}_s|\tau(s), \mu(s), \mathbf{Q}(s)). \qquad (2)$$

The functional dependence of τ, μ and $\mathbf{Q}$ on s is introduced for a convenient representation of the total likelihood (2).

For statistical inference, it is very important to maintain a balance between the sample size of the data and the number of model parameters. This is the well-known bias-variance trade-off paradigm. Allowing independent parameters $\tau(s), \mu(s)$ and $\mathbf{Q}(s)$ for each site s makes the unrealistic evolutionary assumption of site independence and leads to unreasonably large uncertainty in parameter estimation. On the other hand, fixing all parameters to be equal across sites ignores the natural spatial variation in the data, especially in the presence of recombination (Grassly and Holmes, 1997). Suchard *et al.* (2003b) introduce an MCP model that serves as a middle ground between these two extremes. In their model, the authors divide an alignment into an unknown number of contiguous segments, allowing parameters $(\tau, \mu, \mathbf{Q})$ to differ across segments, but keeping them equal inside each segment. We further extend this model by decoupling tree topologies τ and substitution process parameters $(\mu, \mathbf{Q})$ into two separate change-point processes.

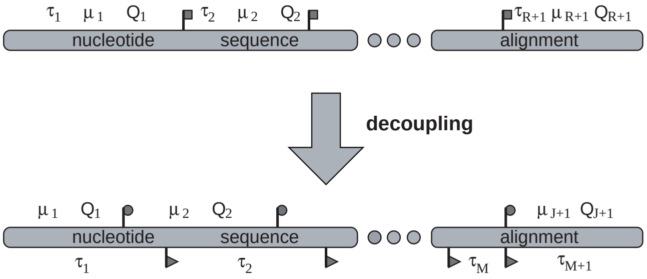

Fig. 1. Decoupling of substitution process change-points from recombination break-points. Square flags on the top alignment represent R random change-points of a single MCP model. The dual MCP model is illustrated on the bottom alignment with J circular flags denoting random substitution process $(\mu, \mathbf{Q})$ change-points and M triangular flags denoting random topology (τ) break-points.

To define the unknown segments for both processes, let ρ_j for $j = 0, \dots, J+1$ be the substitution process change-points that divide an alignment into $J+1$ non-overlapping intervals, subject to the constraint $1 = \rho_0 < \rho_1 < \cdots < \rho_J < \rho_{J+1} = S+1$. We allow substitution parameters to vary only across change-points, keeping them constant inside each segment. Specifically, $(\mu(s), \mathbf{Q}(s)) = (\mu_j, \mathbf{Q}_j)$ for all $s \in [\rho_{j-1}, \rho_j)$. Similarly, we introduce topology break-points ξ_m, for $m = 0, \dots, M+1$, subject to the constraint $1 = \xi_0 < \xi_1 < \cdots < \xi_M < \xi_{M+1} = S+1$ with $\tau(s) = \tau_m$, for all $s \in [\xi_{m-1}, \xi_m)$. To ensure ξ_m is truly an identifiable break-point, we place the additional constraint $\tau_m \neq \tau_{m+1}$, for all $m \in \{1, \dots, M\}$. A similar constraint on $(\mu_j, \mathbf{Q}_j)$ is unnecessary because the probability that two independent continuous random variables are equal across a change-point is zero. The random change-point $\boldsymbol{\rho} = (\rho_1, \rho_2, \dots, \rho_J)$ and break-point $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_M)$ positions are independent of each other and may coincide. We illustrate this decoupling process in Figure 1. In this figure, a possible realization of change-points from the single MCP model is represented as a sequence of square flags on a multiple sequence alignment. The two change-point processes of the dual MCP model are shown on the bottom, with circular and triangular flags denoting substitution process change-points and topology break-points, respectively. After decoupling, substitution process change-points and topology break-points do not necessarily align with the original single MCP change-points.

2.2 Priors

To complete our model specification, we assume independent truncated Poisson priors on both the number of random topology break-points M and the number of random substitution process change-points J, such that

$$p(M) \propto \frac{\delta^M}{M!} 1\{M < S\} \quad \text{and} \quad p(J) \propto \frac{\lambda^J}{J!} 1\{J < S\}, \qquad (3)$$

where $1\{\cdot\}$ is the indicator function and δ and λ are fixed hyperprior constants. These hyperpriors can be interpreted as the prior expected number of topology break-points and substitution process change-points, respectively. Motivation for the truncated Poisson stems from previous MCP studies (Green, 1995; DiMatteo *et al.*, 2001). Given J and M, locations of break-points $\boldsymbol{\xi}$ and change-points $\boldsymbol{\rho}$ are a priori uniform over all possible unordered selections of M or J locations from $S-1$ choices, such that

$$p(\boldsymbol{\xi}|M) = \frac{M!(S-M-1)!}{(S-1)!} \quad \text{and} \quad p(\boldsymbol{\rho}|J) = \frac{J!(S-J-1)!}{(S-1)!}. \tag{4}$$

For a small number of sequences N, we include all possible $E_N = (2N-5)!/2^{N-3}/(N-3)!$ unrooted choices (Felsenstein, 1978) in the space of tree topologies $\boldsymbol{\Omega}$ considered by the model. However, when handling a larger number of sequences, we suggest two ways to reduce the size of $\boldsymbol{\Omega}$. First, if the phylogenetic relationship between the potential parental sequences that could have recombined to produce the putative recombinant is known, we restrict $\boldsymbol{\Omega}$ to the $2N-5$ topologies that could result by adding a new leaf to any branch in the fixed parental tree with $N-1$ taxa (Suchard *et al.*, 2002). The second method exploits the fact that only a few tree topologies are supported by the data. All other topologies can be excluded from the analysis. To identify this subset of topologies, we follow Haake *et al.* (2004) and use MrBayes (Huelsenbeck and Ronquist, 2001; Paraskevis *et al.*, 2003) to precalculate the posterior probability distribution of tree topologies for several small alignments created by breaking the full alignment into subsections. If a tree garners a posterior probability greater than some fixed threshold for at least one of the alignment subsections, it is included in the tree space $\boldsymbol{\Omega}$. In all three of the above approaches, we assume τ_m is drawn from a uniform prior distribution over $\boldsymbol{\Omega}$.

Evolutionary parameters κ_j and μ_j for individual segments are a priori independent and come from log–normal distributions:

$$p(\kappa_j) \propto \frac{1}{\kappa_j} \mathrm{e}^{(-(\log \kappa_j - \nu_\kappa)^2)/2\sigma_\kappa^2} \quad \text{and} \quad p(\mu_j) \propto \frac{1}{\mu_j} \mathrm{e}^{(-(\log \mu_j - \nu_\mu)^2)/2\sigma_\mu^2}. \tag{5}$$

We aim for a hierarchical prior over the evolutionary parameters across segments. Hierarchical frameworks improve estimation precision by pooling information across exchangeable parameters. To form a hierarchical prior, we must assume $\nu_\kappa, \sigma_\kappa^2, \nu_\mu$ and σ_μ^2 are unknown with their own hyperprior distributions. When the number of substitution process segments is less than or equal to 3, it becomes difficult to construct relatively uninformative distributions over σ_κ^2 and σ_μ^2 such that the posterior distribution remains proper (Gelman, 2004). Therefore, when we expect to see fewer than four distinct partitions with different evolutionary pressures, we fix $(\nu_\kappa = 2, \sigma_\kappa^2 = 1)$ and $(\nu_\mu = -2, \sigma_\mu^2 = 2)$. This choice leads to vague and independent prior distributions on κ_j and μ_j. In particular, the prior median of κ_j is 7 and $\kappa_j \in (1, 50)$ with 95% probability; the prior median μ_j is 0.1 and $\mu_j \in (0.003, 7)$ with 95% probability. If we expect four or more segments with varying evolutionary pressures, we continue the hierarchical construction by assuming diffuse, conjugate hyperpriors on $\nu_\kappa, \sigma_\kappa^2, \nu_\mu$ and σ_μ^2:

$$\begin{aligned} p(\nu_\kappa) &\propto \mathrm{e}^{-\nu_\kappa^2/(2\times 100^2)}, \\ p(\nu_\mu) &\propto \mathrm{e}^{-\nu_\mu^2/(2\times 100^2)}, \\ p(\sigma_\kappa^{-2}) &\propto \sigma_\kappa^{-2(0.01-1)} \mathrm{e}^{-\sigma_\kappa^{-2}\times 0.01}, \\ p(\sigma_\mu^{-2}) &\propto \sigma_\mu^{-2(0.01-1)} \mathrm{e}^{-\sigma_\mu^{-2}\times 0.01}. \end{aligned} \tag{6}$$

We estimate these hyperparameters simultaneously with other model parameters.

2.3 Sampling algorithm

In Bayesian analysis, one attempts to describe the posterior distribution of all model parameters given the observed data and then use this distribution for estimation or hypothesis testing. The parameters in our dual MCP model can be represented as

$$\boldsymbol{\theta} = (M, J, \boldsymbol{\xi}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\kappa}, \boldsymbol{\mu}, \boldsymbol{\phi}), \tag{7}$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{M+1}), \boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_{J+1}), \boldsymbol{\mu} = (\mu_1, \ldots, \mu_{J+1})$ and $\boldsymbol{\phi} = (\nu_\kappa, \sigma_\kappa^2, \nu_\mu, \sigma_\mu^2)$ is the vector of hyperparameters. Our objective is to approximate the posterior distribution of the dual MCP model

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{Y}) &\propto \prod_{s=1}^{S} f(\mathbf{Y}_s|\boldsymbol{\tau}(s), \mathbf{Q}(s), \mu(s)) \\ &\times p(\boldsymbol{\tau}) \prod_{j=1}^{J+1} p(\kappa_j|\boldsymbol{\phi}) p(\mu_j|\boldsymbol{\phi}) \\ &\times p(\boldsymbol{\phi}) p(\boldsymbol{\xi}|M) p(\boldsymbol{\rho}|J) p(M) p(J). \end{aligned} \tag{8}$$

Since analytic expression of our dual MCP model posterior is intractable, we employ Markov chain Monte Carlo (MCMC) to generate random samples from the posterior (Tierney, 1994). When simulating from a posterior distribution via MCMC, the states of the Markov chain are points in the parameter space of the model and the proportion of time the chain spends at each state approximates the posterior probability (density) of this state. In an MCP model, the dimension of the parameter space is not fixed, but depends on the number of change-points in the model, necessitating a means of transitioning between states with different numbers of components. Green (1995) has developed a reversible jump MCMC (rjMCMC) algorithm by extending the Metropolis–Hastings (MH) sampling scheme (Hastings, 1970) to allow moves between spaces of different dimensions. This method introduces auxiliary variables to construct a bijective map between parameter states with unequal numbers of components. This dimension-matching procedure is concluded by adjusting MH acceptance probabilities with the Jacobian of the transformation.

In developing an rjMCMC sampler for Equation (8), we follow the scheme introduced by Suchard *et al.* (2003b) adjusting it when necessary. In particular, steps involving adding or removing change-points are modified to accommodate two kinds of change-points, and the proposal distributions for continuous parameters are altered to improve convergence of the chain. We describe our general sampling scheme first and then focus the attention of the readers on the differences between the MCMC implementations for the single and the dual MCP models.

Our sampler achieves mobility between spaces with different dimensions by proposing birth and death steps for both substitution process change-points and topology break-points. At each step, one of the following moves is attempted: inserting or deleting topology break-point(s), inserting or deleting a substitution process change-point or updating all model parameters conditional on the current values of J and M. The probabilities of choosing a particular move follow with slight modification from the single MCP sampler. Proposals for changes in the number of substitution process change-points replicate the equivalent proposals for the single MCP sampler, but changes in the number of break-points necessitate novel proposals in order to preserve inequality between adjacent topologies. We demonstrate this requirement with a simple example. Suppose the current state $\boldsymbol{\theta}$ in the Markov chain has three break-points separating four topologies: $(\tau_1, \tau_2, \tau_3, \tau_4) = $ (A,B,A,B), where A $\neq$ B. If the sampler proposes to remove the single break-point ξ_2 that separates topologies τ_2 and τ_3, then two segments collapse into one. The new segment should reasonably inherit its topology from either of the original segments with some probability, but both possible realizations (A,B,B) or (A,A,B) from this proposal violate the identifiability restriction: $\tau_m \neq \tau_{m+1}$, for all m. To avoid this problem, we propose to add or delete two break-points in a single step. When the sampler encounters the situation above, it also removes the additional break-point that is producing the violation. To preserve detailed balance, we introduce a complementary birth step that adds two topology break-points at a time. A tuning parameter c is employed to adjust the fraction of time that the sampler uses the double rather than single break-point birth step. In our experience, setting $c = 1/E_N$ results in consistent and satisfactory mixing of topology break-points.

We update $(\boldsymbol{\xi}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\kappa}, \boldsymbol{\mu}, \boldsymbol{\phi})$ conditional on M and J in a Metropolis-within-Gibbs cycle. During this step, we sequentially propose new values for tree topologies and substitution model parameters for each partition, accepting or rejecting them according to the MH rule. All updates are as in Suchard *et al.* (2003b), except new values of κ_j and μ_j are proposed by generating a uniform random variable $U \in [0, 1]$ and multiplying current values of κ_j and μ_j by $e^{U-0.5}$. Under this scheme, our sampler can take bigger jumps as current values of the variables get larger. These long jumps allow for faster exploration of the parameter state space. After the dual MCP sampler updates model parameters for all partitions, the locations of any change-points and break-points are updated. If we let hyperparameters $\boldsymbol{\phi}$ vary, they are also updated during this round. Since full conditional distributions of $\nu_\kappa, \sigma_\kappa^2, \nu_\mu$ and σ_μ^2 are available (Gelfand and Smith, 1990; Suchard *et al*., 2003a), we use Gibbs sampling to update them.

We generate MCMC chains of length 2 100 000. Each chain starts with one partition (i.e. no change-points or break-points). Values for κ_1 and μ_1 are generated uniformly from the intervals $(0, 100)$ and $(0, 1)$, τ_1 is drawn uniformly from Ω. The first 100 000 iterations of each chain are discarded as burn-in and every 200th iteration thereafter is saved, resulting in posterior samples containing 10 000 draws.

2.4 Bayes factors

Bayesian model selection is often accomplished by comparing prior and posterior probabilities of competing hypotheses via Bayes factors (Kass and Raftery, 1995). We adopt this approach to test for the presence of recombination in the evolutionary history of a putative recombinant. Let $H_1 = \{M > 0\}$ be a hypothesis postulating that there is at least one site in an alignment where recombination has occurred, and let $H_2 = \{M = 0\}$ be the alternative hypothesis in which the evolutionary history under investigation does not contain recombination. Then the Bayes factor in favor of recombination is

$$B_{12} = \frac{\Pr(M > 0|\boldsymbol{Y})}{\Pr(M = 0|\boldsymbol{Y})} \Bigg/ \frac{\Pr(M > 0)}{\Pr(M = 0)}. \qquad (9)$$

Recalling that a priori M approximately follows a Poisson distribution with mean δ, we obtain $\Pr(M > 0)/\Pr(M = 0) \approx e^\delta - 1$. The posterior probability of recombination $\Pr(M > 0|\boldsymbol{Y})$ can be directly estimated as the fraction of MCMC simulants satisfying the condition $M > 0$. The simplicity of this procedure is deceptive, because the standard error of such an estimator can be quite large when $\Pr(M > 0|\boldsymbol{Y})$ approaches 0 or 1 (Weiss *et al*., 1999). To minimize this standard error, Carlin and Chib (1995) propose adjusting prior probabilities of competing hypotheses so that a posteriori the hypotheses are approximately equiprobable. In line with this idea, we employ the logistic regression model for Bayes factor estimation introduced by Suchard *et al*. (2005). This approach benefits from averaging across several values of prior odds to arrive at a more efficient estimate of the Bayes factor.

3 DATA

To demonstrate the advantages of the dual MCP sampler over the single MCP sampler, we examine three datasets containing either simulated data or real sequence alignments including HIV recombinants.

We start with a previously used test example involving mitochondrial DNA (mtDNA) coding subunits 4 and 5 of the NADH-dehydrogenase enzyme and three transfer RNAs (tRNAs) from humans (H), orangutans (O), squirrel monkeys (S) and lemurs (L) (Hayasaka *et al*., 1988). Previous phylogenetic studies report the consensus tree (H,O,(S,L)) as the evolutionary relationship among these taxa (Yang and Rannala, 1997; Larget and Simon, 1999; Suchard *et al*., 2001). We construct an artificial alignment from these data, where sites are rearranged to form four distinct partitions: the first three partitions comprise the three codon positions from the protein-coding region of the alignment, while the last partition consists of tRNA sites. Larget and Simon (1999) have demonstrated that these partitions differ greatly in their evolutionary pressures. The greatest evolutionary divergence can be observed in the third (codon) partition where silent mutations are common. We therefore expect, among others, a substitution process change-point around site 464, the starting site for partition three. To investigate the effect of distance between change-points and break-points on the accuracy of recombination detection, we generate 18 alignments, each with a single simulated recombination break-point. We simulate the recombination event by permuting the nucleotides from the H and the L sequences starting at a fixed site in each alignment. Such permutations change the inferred relationship between taxa to the alternative topology (L,O,(S,H)) after the chosen site. The artificial break-points are placed every 10 sites starting at 405 and ending at 575.

We also apply the dual MCP model to two real HIV datasets. The first dataset consists of a portion of the *gag* gene from HIV-1 isolate RW024 (accession number U86548) aligned with the eight HIV-1 subtype consensus sequences A, B, C, D, F, G, H and J from the Los Alamos HIV Database. The alignment contains 729 sites encoding *gag* proteins p17 and p24. We assume that the phylogenetic tree describing evolutionary relationships among consensus subtype sequences is fixed and equals ((((((A,H),G),J),C),F),B,D) as reported by Robertson *et al*. (1999). This assumption allows us to consider only those topologies that can be obtained by adding sequence RW024 to any branch in the subtype tree, thus reducing the size of the tree space $\boldsymbol{\Omega}$ to 13 possible topologies. The isolate RW024 has been analyzed by Cornelissen *et al*. (1996), who found that different portions of the *gag* gene (p17, p24) support different parental heritage (A, H) of the isolate. Dorman *et al*. (2002) and Suchard *et al*. (2002) find support in favor of recombination generating this isolate with a P-value < 0.0001 and a Bayes factor of 10^{19}, respectively.

In the third example, we analyze the full genome of HIV-1 isolate KAL153 from Kaliningrad, Russia (accession number AF193276). Liitsola *et al*. (1998) show that the *gag* and *env* genes of this isolate originated from subtypes A and B, respectively. The resulting A/B recombinant viral strain is suspected to be the causative agent of an explosive HIV-1 epidemic in Kaliningrad among intravenous drug users. We use an alignment of the KAL153 isolate along with subtypes A, B and F consensus sequences.

Both of the HIV putative recombinants have been successfully analyzed in a single MCP framework, but correlation between the inferred locations of substitution process change-points and topology break-points is observed (Suchard *et al*., 2002, 2003b). A dual MCP model analysis should reveal if this correlation is supported by the data or is an artifact of the a priori correlation between change-points and break-points from which the single MCP model suffers.

Although all the data used in our examples can be found in public databases, the multiple sequence alignments may not be easily reproduced without the knowledge of original alignment algorithm settings. Therefore, we make the unpermuted alignment of the mtDNA and alignments of the putative HIV recombinants available from http://www.biomath.ucla.edu/msuchard/DualBrothers as Supplementary information.

4 RESULTS

4.1 Prior choice and sensitivity analysis

Most parameters in the dual MCP model receive non-informative prior distributions, requiring little if any a priori knowledge to specify. However, the numbers of change-points and break-points follow Poisson distributions that can be tuned to incorporate prior information. Suchard *et al*. (2003b) outline one procedure of choosing a prior mean for the number of change-points in the single MCP model.

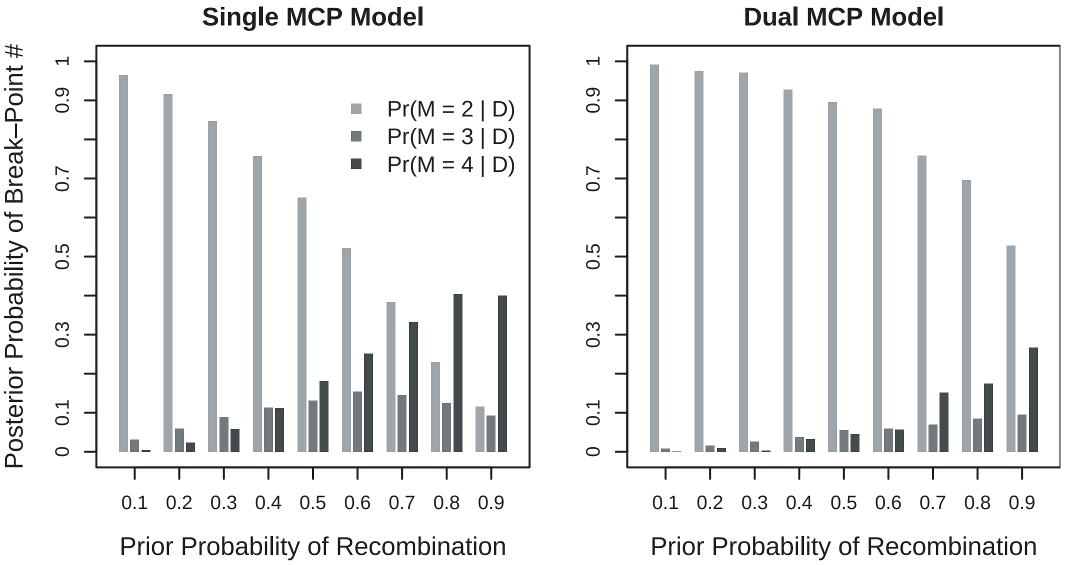

Fig. 2. Sensitivity to the choice of prior probability of recombination in the KAL153 example. Posterior probabilities of the number of topology break-points attaining values 2, 3 and 4 (vertical bars) are plotted for different values of the prior probability of at least one recombination break-point (x-axis).

Since most of the change-points in the single MCP model correspond to substitution change-points in the dual MCP model, we readily apply these guidelines to specify the prior mean number of substitution change-points λ. Following the suggestions of Suchard *et al.* (2003b) we assign λ roughly to the number of boundaries between genes or gene products in the alignment. In the primate mtDNA example we set $\lambda = 3$ corresponding to the four artificial partitions of the nucleotide sites. For the RW024 HIV example, we expect a priori $\lambda = 1$ substitution process change-point somewhere near the boundary between the two *gag* gene products, and for the KAL153 analysis, we recall that the alignment contains all 10 HIV genes, suggesting $\lambda = 9$ divisions. Usually, no prior information about recombination is available, especially for newly sequenced strains. Therefore, in both our synthetic and real examples, we set the prior mean number of break-points $\delta = \log 2$; this translates into a prior probability of at least one recombination point $\Pr(M > 0) \approx 1 - \mathrm{e}^{-\delta} = \frac{1}{2}$, generating a fair test a priori.

Naturally, the results of any Bayesian analysis depend on the choice of prior distributions (O'Hagan, 2003). The more informative priors, $p(J)$ and $p(M)$ in our case, tend to have the greatest impact on results. After perturbing the prior mean number of substitution change-points in the primate example, we find that despite the differences in the posterior distributions of J, posterior profiles of the evolutionary parameters remain robust to the choice of λ (see Supplementary information at http://www.biomath.ucla.edu/msuchard/DualBrothers). Since the number of topology break-points is the most important parameter for recombination detection, we examined its sensitivity to the choice of prior in more detail. We apply the single and dual MCP models to the KAL153 example nine times by varying the prior probability of at least one recombination break-point from 0.1 to 0.9. Manipulating this prior probability is trivial in the dual MCP model since $\Pr(M > 0) \approx 1 - \mathrm{e}^{-\delta}$. On the contrary, controlling the prior probability of recombination is rather challenging in the single MCP model (Suchard *et al.*, 2002) as this probability is a function of both the prior mean number of change-points and the probability that two adjacent segments share the same topology. In our sensitivity analysis we set the former to 9 and vary the latter. Vertical bars in Figure 2 denote the posterior probability that the number of topology break-points attains values 2, 3 and 4. Other values are not shown as they do not gain substantial posterior support under either single or dual MCP models. Both models support two topology break-points under conservative priors on recombination. As the prior probability of recombination increases, the single MCP model more quickly favors the next most probable configuration with four topology break-points than the dual MCP model. The two additional break-points are located near the boundary of *pol* and *vif* genes as can be seen in Figure 5. This region contains very few sparsely distributed informative sites. This sparseness explains the sensitivity of the MCP models to the prior in this part of the alignment. We conclude that the contribution of the prior to posterior estimates of the number of topology break-points is minimal in the presence of sufficient phylogenetic information, but can be significant when the data are sparse or noisy. The fact that even for high values of the prior probability of recombination the posterior mode of M under the dual MCP model remains at 2 in contrast to the results of the single MCP analysis indicates that the dual MCP model is more robust to misspecification of the prior probability of recombination.

4.2 Results of posterior simulations

We first discuss the simulation study aimed at demonstrating the improved accuracy of the dual MCP model. In Figure 3, we plot the inferred against simulated recombination sites using the single and dual MCP models (open circles). Ideally one expects all plotted points to lie on the line $y = x$, indicating perfect estimation of recombination sites. The dashed horizontal line in each plot marks site 464, where a substitution process change-point is inferred by the dual MCP. The single MCP model shows a strong attraction between the inferred recombination sites and the substitution process change-point, with inferred recombination sites clustering along

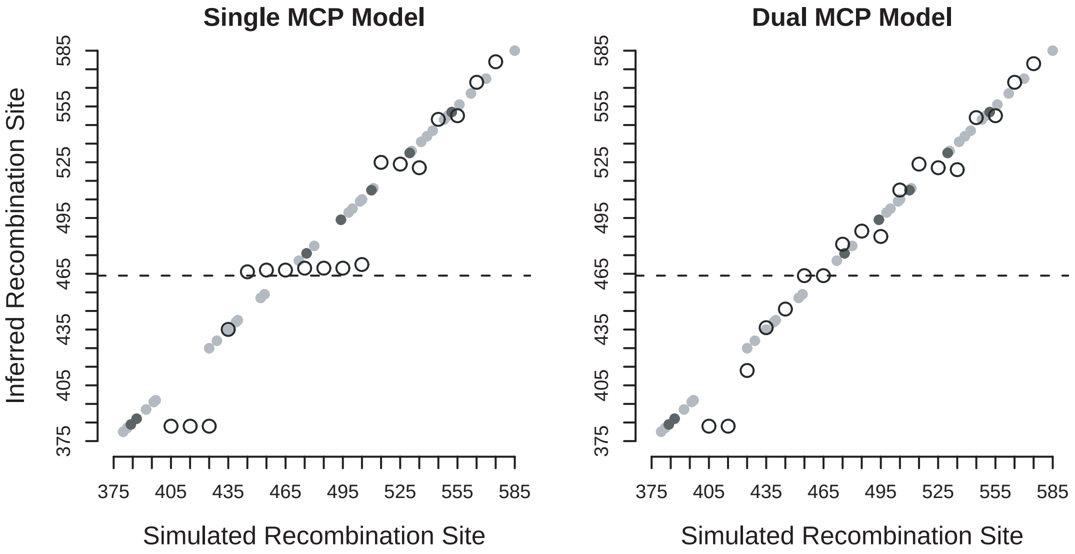

Fig. 3. Simulation study demonstrates improved accuracy of dual MCP sampler. For both the single and the dual MCP models, estimated locations of recombination are plotted against true break-points (open circles), simulated at various positions near site 464, where a substantial change in evolutionary pressures occurs (dashed line). Circles on the diagonal denote informative sites that support (light grey) or contradict (dark grey) the recombinant structure; blank diagonal gaps correspond to the uninformative regions. Posterior attraction of inferred recombination sites towards the substitution process change-point is greatly reduced in the dual as compared with the single MCP model.

the dashed horizontal line. The dual MCP model yields more accurate inference with small variation about the diagonal. To interpret the off-diagonal variation, we define two classes of topologically informative sites in the alignment, those supportive or contradictory of the simulated recombinant structure. A site is called supportive when it supports the consensus topology and sits left of a simulated break-point or when it supports the alternative topology and lies right of the break-point. Supportive sites provide both the single and the dual MCP algorithms information with which to infer the location of recombination. Sites contradictory to the recombinant structure are sites that support the alternative topology to the left of the break-point or support the consensus topology to the right of the break-point. Careful thought shows that all sites can be classified based solely on the topology supported by the unpermuted data and regardless of the actual break-point location. We mark supportive and contradictory sites in Figure 3 as light and dark grey circles, respectively. Sites that do not support either of the two competing topologies are not plotted. As expected, the greatest inaccuracies in detection occur when the simulated recombination events are located in uninformative regions, especially those bordered by contradictory sites.

We analyze the RW024 putative recombinant with the single and the dual MCP models and summarize the results of our posterior simulations in Figure 4. For each site in the alignment we plot the marginal posterior probabilities of the tree topologies as well as medians and 95% Bayesian credible intervals (BCIs) for κ and μ. In Figure 4, arrows mark distinctions in the results between the two models. As estimated by the single MCP model, changes in the evolutionary substitution process and the most probable topology effectively occur at the same location. Over 95% of the change-points located between nt 280 and 320 are both substitution process change-points and topology break-points. Decoupling of the MCP process results in a shift in the estimated break-point and change-point in opposite directions. Change-points and break-points located between sites 280 and 320 coincide only 0.3% of the time during MCMC simulation. Both models predict a higher κ in the 3′ end of the genome. The BCIs of κ calculated under the dual MCP model are larger than those of the single MCP model for sites in the middle of the alignment. Visual examination of this alignment region reveals that the dual MCP model correctly identifies a region with a relatively high transition/transversion ratio κ, but lack of information leads to great uncertainty about its actual value. The single MCP model places the substitution process change-point further 5′, near the topology break-point; therefore, averaging κ among sites with high and low transition/transversion ratio in the region downstream of the change-point.

We plot the results of posterior simulations for the KAL153 putative recombinant in Figure 5. This figure follows the same arrangement as in Figure 4. As before, differences in performance of the single and the dual MCP models are marked by arrows. The dual MCP model estimate of the first recombination site lies 5′ of the single MCP estimate. Both models identify a strongly supported (A,B,A) recombinant structure and a small region on the boundary of *pol* and *vif* with somewhat uncertain origin. A Bayes factor calculated under the dual MCP model amounts to 10^{77} providing decisive support for at least one recombination point in the KAL153 sequence. We find no evidence of recombination when we substitute KAL153 with a pure subtype A representative (see Supplementary information at http://www. biomath.ucla.edu/msuchard/DualBrothers).

Table 1 reports estimates of recombination break-points together with their 95% BCIs calculated under both models. In the KAL153 example, despite the increase in the number of parameters, the dual MCP break-point estimates have roughly the same size BCIs as the estimates calculated under the single MCP model. Therefore, the dual MCP model improves the accuracy of recombination detection while preserving the precision of the single MCP model. However,

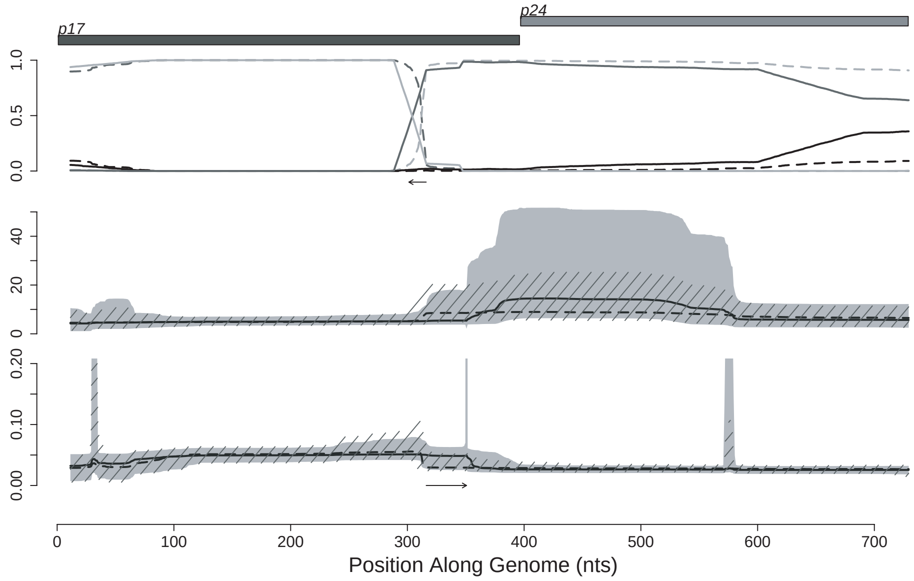

Fig. 4. Comparison of the single and the dual MCP models for HIV-1 RW024. The top plot shows the locations of gene products within the *gag* gene. The plot labeled τ shows marginal posterior probabilities of the two most probable tree topologies (light grey—subtype A heritage, medium grey—subtype H heritage) or the sum of the marginal posterior probabilities of all other topologies (dark grey line) for each site in the alignment, calculated under the single (dashed) and dual (solid) MCP model. The last two plots show medians (lines) and 95% BCIs (shading) of the transition/transversion ratio κ and average divergence μ. Single MCP model estimates are drawn with a dashed line and hatched shading, while dual MCP model estimates are drawn with a solid line and uniform shading. Arrows indicate regions where substantial differences in the results of the two models occur.

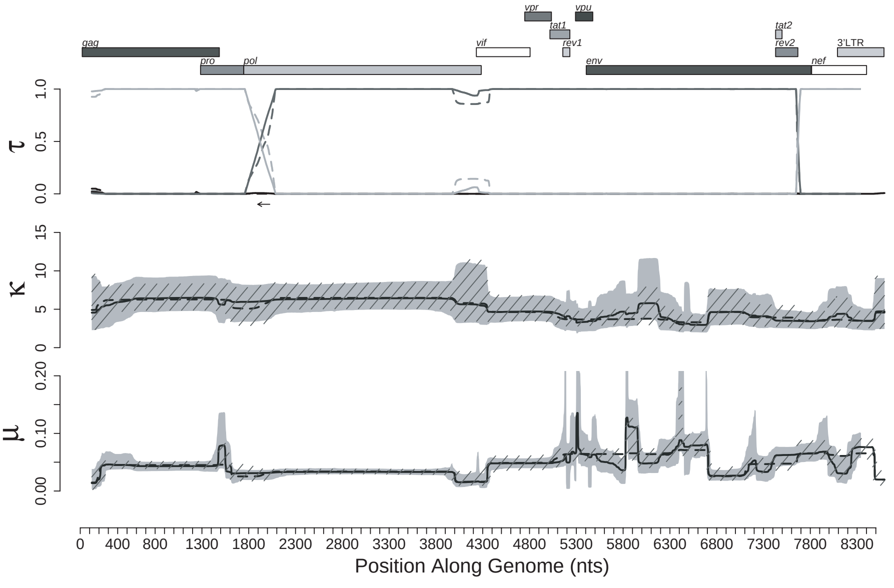

Fig. 5. Comparison of the single and the dual MCP models for HIV-1 KAL153. The top plot highlights positions of the open reading frames along the HIV-1 genome. The rest of the figure follows the format in Figure 4, with light grey representing subtype A heritage and medium grey representing subtype B heritage in the plot labeled τ. Inferred recombinant structure is (A,B,A) under the dual and the single MCP models.

the topology break-point in the RW024 has a noticeably larger BCI under the dual MCP model. The roots of the increased uncertainty lie not only in the different complexities of the two models, but also in differences in the utilization of sequence information. In this example, it is not surprising that the change-point inferred by the single MCP has smaller posterior variance since its location was deduced from both topological incongruence and heterogeneous sequence divergence. Pulling these events together may not be advantageous. We believe that recombination inference should be based solely on the discordance in tree topologies since the contribution of other sequence features to recombination has not been rigorously established.

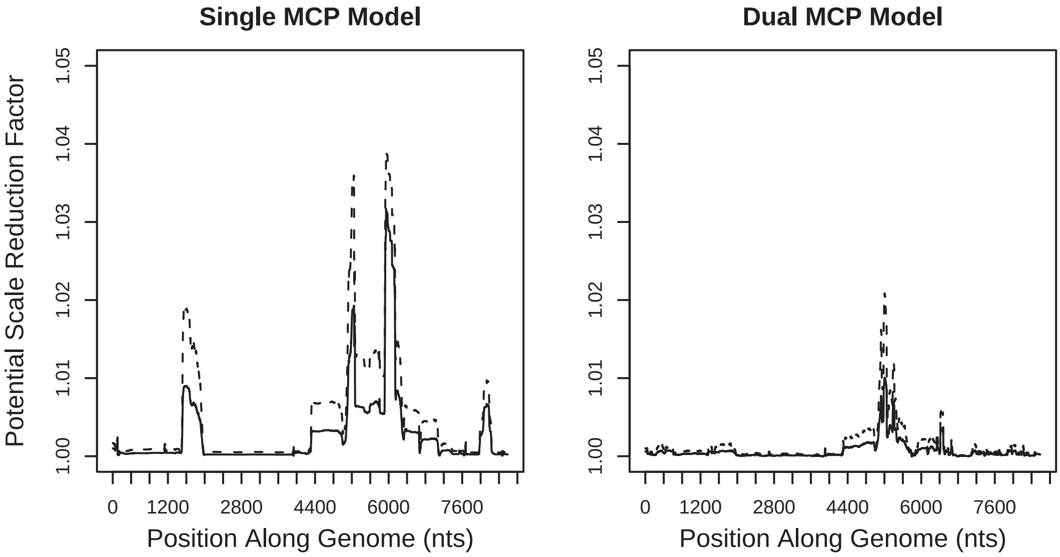

Fig. 6. Gelman–Rubin PSRFs and their corresponding 97.5% quantiles calculated for site-specific transition/transversion ratios $\kappa(s)$ along the genome. PSRFs represent the factor by which the posterior variance of a monitored parameter will be reduced as the number of iterations of MCMC sampling approaches infinity. The closeness of PSRFs to 1 indicates how well the sampler is converging to the target distribution. The dual MCP model exhibits better overall convergence suggesting that appropriate partitioning of the data may result in more efficient exploration of the posterior distribution of model parameters.

Table 1. Posterior medians and 95% BCIs of inferred recombination sites using single and dual MCP models

Single MCP Median	95% BCI	Dual MCP Median	95% BCI
KAL153 example			
1886	1666–1988	1825	1669–1985
7643	7631–7677	7652	7631–7678
RW024 example			
306	290–313	298	283–337

We report convergence of our MCMC sampler for the KAL153 example using the Gelman–Rubin potential scale reduction factor (PSRF), a convergence statistic based on a comparison of the within-chain and between-chain variances from multiple MCMC runs (Gelman and Rubin, 1992). We follow Brooks and Giudici (1998) and identify site-specific transition/transversion ratios $\kappa(1),\ldots,\kappa(S)$ as a set of parameters that retain their interpretation even as the dimension of the parameter space changes. We compute PSRFs using 10 independent MCMC chains with overdispersed starting values. Figure 6 plots point estimates of the PSRFs and the corresponding 97.5% quantiles for each site in the alignment. The closeness of the estimates to 1 indicates good convergence of the sampler. The dual MCP model shows better overall convergence of the site-specific transition/transversion ratios, in spite of the fact that both samplers use the same transition kernels.

5 DISCUSSION

In this paper, we improve the accuracy of recombination detection afforded by Bayesian phylogenetic methods. Accuracy is improved by modeling spatial phylogenetic variation with two independent change-point processes, an extension of the single change-point process used by Suchard *et al*. (2003b). One process models changes in the tree topology along a multiple sequence alignment; the other process allows nucleotide substitution pressures to vary. Decoupling these processes eliminates the prior correlation between locations of changes in topology and evolutionary pressure, yielding more accurate estimation of both change-point types. The dual MCP model inherits a major strength from the original MCP model in its realistic modeling of spatial phylogenetic variation using a parsimonious number of parameters. In addition, similar to the single MCP model, the dual model allows simultaneous estimation of regions with different evolutionary pressures, uncertainty in topologies and locations of recombination sites.

Analysis of alignments with recombination simulated in an area with a significant change in evolutionary pressure shows that dual MCP estimates of break-points are more accurate than single MCP results. The recombination sites inferred under the single MCP model are clearly attracted to the evolutionary change-point location owing to the prior correlation between the two kinds of change-points. In the RW024 example under the single MCP model, the estimates of change in evolutionary pressure and recombination site locations effectively coincide. In contrast, the dual model predicts the substitution change-points and topology break-points to occur at distinct and fairly distant positions along the genome. Similar separation of substitution change-points and topology break-points are observed in the KAL153 example within the *pol* coding region. We conclude that strong posterior correlation between locations of change in evolutionary pressure and recombination is an artifact of the single MCP model. The dual MCP model should result in more accurate estimation of locations of substitution change-points and topology break-points. Moreover, as the KAL153 examples show, the dual MCP model is less sensitive to the choice of prior on recombination in the presence of sparse and noisy data.

We observe improved convergence of the rjMCMC sampler after decoupling of change-points. Because the dual MCP model selects more accurate partitions of the data, it may produce better sampler mixing by regularizing the posterior landscape. For example, if the marginal posterior distribution of the transition/transversion ratio κ is bimodal in a given partition, splitting this partition by adding an appropriate change-point may result in two new partitions with unimodal posterior distributions of κ. Two unimodal distributions may be more efficiently explored by the rjMCMC sampler. Since the dual MCP model identifies partition boundaries more accurately, fewer multimodal distributions and faster exploration of the posterior distribution of model parameters are expected.

In addition to the increased accuracy in recombination site identification, using two change-point processes allows us to model recombination sites explicitly as model parameters. This explicit representation of recombination makes estimation and hypothesis testing more rigorous and the specification of the prior on recombination sites more flexible. For instance, it is now possible to use a site-specific recombination prior, where each site is explicitly assigned a prior probability of being a topology break-point. Such prior specification allows one to incorporate information from previous recombination detection studies, directing the algorithm to regions where recombination is more likely to occur. We believe that the dual MCP shows great promise for accurately detecting recombination and finding patterns in the spatial distribution of recombination sites.

ACKNOWLEDGEMENTS

We thank Benjamin Redelings for helpful discussions and Robert Rovetti for careful reading of the manuscript. We are also grateful to the two anonymous reviewers for their helpful comments and suggestions that greatly improved the manuscript. This work was supported by NIH grant GM068955, UCLA AIDS Institute and the James B. Pendleton Charitable Trust.

REFERENCES

Brooks,S. and Giudici,P. (1998) Convergence assessment for reversible jump MCMC simulations. *Bayesian Stat.*, **6**, 733–742.

Carlin,B. and Chib,S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B*, **57**, 473–484.

Chen,L. *et al*. (2004) Positive selection detection in 40 000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J. Virol.*, **78**, 3722–3732.

Cornelissen,M. *et al*. (1996) Human immunodeficiency virus type 1 subtypes defined by *env* show high frequency of recombinant *gag* genes. *J. Virol.*, **70**, 8209–8212.

DiMatteo,I. *et al*. (2001) Bayesian curve fitting with free-knot splines. *Biometrika*, **88**, 1055–1073.

Dorman,K. *et al*. (2002) Bootstrap confidence levels for HIV-1 recombinants. *J. Mol. Evol.*, **54**, 200–209.

Felsenstein,J. (1978) The number of evolutionary trees. *Syst. Zool.*, **27**, 27–33.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **13**, 93–104.

Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA.

Gelfand,A. and Smith,A. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, **85**, 398–409.

Gelman,A. (2004) Prior distributions for variance parameters in hierarchical models. *Technical report*, Columbia University NY.

Gelman,A. and Rubin,D. (1992) Inference from iterative simulation using multiple sequences. *Stat. Sci.*, **7**, 457–511.

Grassly,N. and Holmes,E. (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.*, **14**, 239–247.

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

Haake,D. *et al*. (2004) Molecular evolution and mosaicism of leptospiral outer membrane proteins involves horizontal DNA transfer. *J. Bacteriol.*, **186**, 2818–2828.

Hasegawa,M. *et al*. (1985) Dating the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.

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

Hayasaka,K. *et al*. (1988) Molecular phylogeny and evolution of primate mitochondrial DNA. *Mol. Biol. Evol.*, **5**, 626–644.

Hein,J. (1990) Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.*, **98**, 185–200.

Huelsenbeck,J. and Ronquist,F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Husmeier,D. and McGuire,G. (2002) Detecting recombination with MCMC. *Bioinformatics*, **18**, S345–S353.

Husmeier,D. and McGuire,G. (2003) Detecting recombination in 4-taxa DNA sequence alignment with Bayesian hidden Markov models and Markov chain Monte Carlo. *Mol. Biol. Evol.*, **20**, 315–337.

Husmeier,D. and Wright,F. (2001) Probabilistic divergence measures for detecting interspecies recombination. *Bioinformatics*, **17**, S123–S131.

Kass,R. and Raftery,A. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.

Kellam,P. and Larder,B. (1995) Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased zidovudine resistance. *J. Virol.*, **69**, 669–674.

Korber,B. *et al*. (2001) Evolutionary and immunological implications of contemporary HIV-1 variation. *Br. Med. Bull.*, **58**, 19–42.

Larget,B. and Simon,D. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.*, **16**, 750–759.

Li,S. *et al*. (2000) Phylogenetic tree construction using Markov chain Monte Carlo. *J. Am. Stat. Assoc.*, **95**, 493–508.

Li,W. *et al*. (1988) Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol. Biol. Evol.*, **5**, 313–330.

Liitsola,K. *et al*. (1998) HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. *AIDS*, **12**, 1907–1919.

McGuire,G. *et al*. (1997) A graphical method for detecting recombination in phylogenetic datasets. *Mol. Biol. Evol.*, **14**, 1125–1131.

O'Hagan,A. (2003) HSSS model criticism (with discussion). In Green,P., Hjort,N. and Richardson,S. (eds), *Highly Structured Stochastic Systems.*, Oxford University Press, Oxford, UK, pp. 423–453.

Paraskevis,D. *et al*. (2003) Analysis of the evolutionary relationships of HIV-1 and SIVcpz sequences using Bayesian inference: implications for the origin of HIV-1. *Mol. Biol. Evol.*, **20**, 1986–1996.

Robertson,D. *et al*. (1995a) Recombination in AIDS viruses. *J. Mol. Evol.*, **40**, 249–259.

Robertson,D. *et al*. (1995b) Recombination in HIV-1. *Nature*, **374**, 124–126.

Robertson,D. *et al*. (1999) HIV-1 nomenclature proposal: a reference guide to HIV-1 classification. In Korber,B., Kuiken,C., Foley,B., Hahn,B., McCutchan,F., Mellors,J. and Sodroski,J. (eds.), *Human Retroviruses and AIDS 1999*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, pp. 492–505.

Sinsheimer,J. *et al*. (2003) Are you my mother? Bayesian phylogenetic models to detect recombination among putative parental strains. *Appl. Bioinformatics*, **2**, 131–144.

Suchard,M. *et al*. (2001) Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.*, **18**, 1001–1013.

Suchard,M. *et al*. (2002) Oh brother, where art thou? a Bayes factor test for recombination with uncertain heritage. *Syst. Biol.*, **51**, 715–728.

Suchard,M. *et al*. (2003a) Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst. Biol.*, **52**, 649–664.

Suchard,M. *et al*. (2003b) Inferring spatial phylogenetic variation along nucleotide sequences: a multiple change-point model. *J. Am. Stat. Assoc.*, **98**, 427–437.

Suchard,M. *et al*. (2005) Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. *Biometrics*, in press.

Tierney,L. (1994) Markov chains for exploring posterior distributions (with discussion). *Ann. Stat.*, **22**, 1701–1762.

Wei,X. *et al*. (2003) Antibody neutralization and escape by HIV-1. *Nature*, **422**, 307–312.

Weiss,R. *et al*. (1999) On Bayesian calculations for mixture likelihoods and priors. *Stat. Med.*, **18**, 1555–1570.

Worobey,M. and Holmes,E. (1999) Evolutionary aspects of recombination in RNA viruses. *J. Gen. Virol.*, **80**, 2535–2543.

Yang,Z. and Rannala,B. (1997) Bayesian phylogenetic inference using DNA sequences. *Mol. Biol. Evol.*, **14**, 717–724.