Detecting Recombination and Gene Conversion in DNA Multiple Alignments with Bayesian Hidden Markov Models and Markov Chain Monte Carlo

Dirk Husmeier
Biomathematics and Statistics Scotland (BioSS)
SCRi, Invergowrie, Dundee DD2 5DA, United Kingdom
Email: dirk@bioss.ac.uk

Gráinne McGuire
Work carried out at:
Department of Applied Statistics
University of Reading
Reading RG6 6FN, United Kingdom
Email: grainnemcguire@yahoo.com

December 2, 2001

Both authors contributed equally to this work, and the order of names was chosen alphabetically.

Keywords
Statistical phylogenetics, recombination, gene conversion, hidden Markov models, Bayesian statistics, Markov chain Monte Carlo.
This article presents a statistical method for detecting recombination and gene conversion in DNA sequence alignments, which is based on combining two probabilistic graphical models: (1) a taxon graph (phylogenetic tree) representing the relationship between the taxa or genes, and (2) a site graph (hidden Markov model) representing interactions between different sites in the DNA sequence alignments. We adopt a Bayesian approach and sample the parameters of the model from the posterior distribution with Markov chain Monte Carlo, using a Metropolis-Hastings and Gibbs within Gibbs scheme. A simulation study confirms that this gives better results than two earlier approaches, where the model parameters were estimated heuristically or with maximum likelihood.

1 Introduction

An underlying assumption of most phylogenetic tree reconstruction methods is that there is one set of hierarchical relationships among the taxa/genes. While this is a reasonable approach when applied to most DNA sequence alignments, it can be violated if a recombination or gene conversion event has taken place. The resulting transfer or exchange of DNA subsequences leads to a change of the branching order (topology) in the affected region, which results in conflicting phylogenetic information from different regions of the alignment. If undetected, the presence of these so-called mosaic sequences can lead to systematic errors in phylogenetic tree estimation. Their detection, therefore, is a crucial prerequisite for consistently inferring the evolutionary history of a set of DNA sequences.

In the last few years, a plethora of methods for detecting recombination and gene con-
version have been developed – following up on the seminal paper by Smith (1992) – and it is beyond the scope of this article to present a comprehensive overview. Many detection methods for identifying the nature and the breakpoints of the resulting mosaic structure are based on moving a window along the alignment and computing a phylogenetic divergence score for each window position. Examples are the bootstrap support for the locally optimal topology (Salminen, Carr, Burke, and McCutchan, 1995), the likelihood ratio between the locally and globally optimal trees (Grassly and Holmes, 1997), and the difference in the fitting scores between two adjacent locally optimized trees (McGuire, Wright, and Prentice, 1997). The determination of the breakpoints of the mosaic structure is then based on an analysis of the signals thus obtained, using bootstrapping to estimate their significance. While these methods are useful for a preliminary scan of a DNA sequence alignment, the spatial resolution for the identification of the breakpoints is typically of the order of the window size and, consequently, rather poor.

This article discusses a different approach, which was first suggested by Hein (1993). The idea is to introduce a hidden state that represents the tree topology at a given site. A state transition from one topology into another corresponds to a recombination or gene conversion event. To introduce correlations between adjacent sites, a site graph is introduced, representing which nucleotides interact in determining the tree topology. Thus, the standard model of a phylogenetic tree is generalized by the combination of two graphical models: (1) a taxon graph (phylogenetic tree) representing the relationships between the taxa or genes, and (2) a site graph representing interactions between different sites in the DNA sequence alignments. To keep the mathematical model tractable and the computational costs limited, the latter are reduced to nearest-neighbour interactions. Breakpoints of mosaic segments are predicted by state transitions in the site graph. While this method can only deal with a small number of sequences simultaneously, it has, in principle, the potential to predict the locations and breakpoints
of recombinant regions more accurately than what can be achieved with most existing techniques.

The article is organized as follows. Section 2 introduces the mathematical method and discusses the shortcomings of existing parameter estimation techniques. Section 3 describes how this can be improved with a Bayesian approach using Markov chain Monte Carlo. Section 4 describes three DNA sequence alignments, on which the proposed scheme is tested. The simulation study is described in Section 5, and the results are discussed in Section 6. The paper finishes in Section 7 with a conclusion and an outlook on future work.

2 Method: Background and earlier approaches

Consider an alignment $\mathcal{D}$ of $m$ DNA sequences, $N$ nucleotides long. Let each column in the alignment be represented by $y_t$, where the subscript $t$ represents the site, $1 \leq t \leq N$. Hence $y_t$ is an $m$-dimensional column vector containing the nucleotides at the $t$th site of the alignment, and $\mathcal{D} = (y_1, \ldots, y_N)$. Attached to each site is a hidden state variable $S_t$, which represents the tree topology at site $t$. For $m$ taxa, there are $K = (2^m - 5)!!$ distinct unrooted topologies, hence $S_t \in \{1, \ldots, K\}$. If a recombination or gene conversion event has occurred, then there will be a change in topology in this region, corresponding to a transition into another hidden state at the breakpoint of this region. Our objective is to predict the ‘optimal’ sequence of hidden states

$$S = (S_1, \ldots, S_N)$$  \hspace{1cm} (1)

given the sequence alignment $\mathcal{D}$ and some optimality criterion to be discussed below.
Obviously, this optimization problem is, in general, intractable. Firstly, the number of possible topologies at a given site increases super-exponentially with the number of sequences $m$. Secondly, there are $K^N$ different state sequences, which prevents an exhaustive search even for small values of $K$. Consequently, the introduction of approximations and restrictions is inevitable.

To deal with the second source of computational complexity, interactions between sites are limited to nearest neighbour interactions. This allows the application of a dynamic programming scheme, which reduces the computational complexity to $O(K^2N)$. To deal with the first source of complexity, the scheme has to be restricted to alignments with small numbers of sequences, $m$. This calls for the application of a fast low-resolution method that identifies a sufficiently small set of putative recombinant sequences. We will discuss this again in the last section.

**RecPars**

Hein (1993) defined optimality in a parsimony sense. His algorithm, RecPars, searches for the most parsimonious state sequence $S$, that is, the one that minimizes a given parsimony cost function $E(S)$. Interactions between sites are restricted to nearest-neighbour interactions, as discussed above, and the search is carried out with dynamic programming. While RecPars is faster than the methods to be discussed below, it suffers from the shortcomings inherent to parsimony, as discussed by Felsenstein (1988). Also, the cost function $E(S)$ depends on certain parameters - the transition cost, the transversion cost, and the recombination cost - which have to be chosen in advance and are not optimized by the algorithm.

*Figure 1 approximately here.*
Detecting recombination with hidden Markov models (HMMs)

Adopting a statistical approach to phylogenetics, illustrated in Figure 1, the probabilistic equivalent to RecPars is a hidden Markov model (HMM), whose application to the detection of recombination was first suggested by McGuire, Wright, and Prentice (2000). Figure 2, left shows the corresponding probabilistic graphical model. White nodes represent hidden states, $S_t$, which have direct interactions only with the states at adjacent sites, $S_{t-1}$ and $S_{t+1}$. Black nodes show columns in the DNA sequence alignment, $y_t$. The joint probability of the DNA sequence alignment, $D$, and the sequences of hidden states, $S$, factorizes:

$$P(D, S) = P(y_1, \ldots, y_N, S_1, \ldots, S_N)$$

$$= \prod_{t=1}^{N} P(y_t|S_t) \prod_{t=2}^{N} P(S_t|S_{t-1})P(S_1)$$ (2)

The optimal state sequence $\hat{S}$ is the one most supported by the data, that is, the mode of $P(S|D)$:

$$\hat{S} = \arg\max_S P(S|D)$$ (3)

While, in general, this problem would be intractable due to the exponential increase in the number of state sequences, see above, the reduction to nearest-neighbour interactions between hidden states and the resulting factorization (2) allows the application of a dynamic programming technique, the so-called Viterbi algorithm (Rabiner, 1989), to find the mode $\hat{S}$ with computational complexity $O(N)$. The factorization (2) contains three terms: $P(y_t|S_t)$, $P(S_t|S_{t-1})$, and $P(S_1)$. The transition probabilities $P(S_t|S_{t-1})$ correspond to recombination events (if $S_t \neq S_{t-1}$). Let $\nu$ denote the probability that the tree topology remains unchanged as we move from a given site in the alignment, $t$,
to an adjacent site, \( t + 1 \) or \( t - 1 \). We then obtain for the state transition probabilities:

\[
P(S_t|S_{t-1}, \nu) = \nu \delta(S_t, S_{t-1}) + \frac{1 - \nu}{K - 1} [1 - \delta(S_t, S_{t-1})]
\] (4)

where \( \delta(S_t, S_{t-1}) \) denotes the Kronecker delta function, which is 1 when \( S_t = S_{t-1} \), and 0 otherwise. It is easily checked that this satisfies the normalization constraint \( \sum_{S_t} P(S_t|S_{t-1}) = 1 \). The emission probabilities \( P(y_t|S_t) \) can easily be computed with the pruning algorithm (Felsenstein, 1981) if the branch lengths corresponding to the topology \( S_t \), \( w_{S_t} \), and the parameters of the nucleotide substitution model, \( \theta_{S_t} \), are known. So, more precisely, we have \( P(y_t|S_t) = P(y_t|S_t, w_{S_t}, \theta_{S_t}) \). To simplify the notation, define the accumulated vectors \( w = (w_1, \ldots, w_K) \) and \( \theta = (\theta_1, \ldots, \theta_K) \) and define: \( P(y_t|S_t, w_{S_t}, \theta_{S_t}) = P(y_t|S_t, w, \theta) \). This means that \( S_t \) indicates which subvectors of \( w \) and \( \theta \) apply. We can depict the dependence of the probability distribution on the parameters \( w \) and \( \nu \) in an extended graphical model, shown in Figure 2, left.

For reasons discussed by Husmeier and Wright (2001), the initial probabilities \( P(S_1) \) are not treated as parameters, but are kept fixed at \( P(S_1) = \frac{1}{K} \forall S_1 \in \{1, \ldots, K\} \). The prediction task is to find the most likely hidden state sequence conditional on the observations (that is, the DNA sequence alignment) and the parameters \( w, \theta, \) and \( \nu \):

\[
\arg\max_S P(S|\mathcal{D}, w, \theta, \nu) = \arg\max_{S_1, \ldots, S_N} P(S_1, \ldots, S_N|y_1, \ldots, y_N, w, \theta, \nu)
\] (5)

The parameters \( w, \theta, \) and \( \nu \) need to be estimated.

**HMM-heuristic**

McGuire, Wright, and Prentice (2000) estimated the branch lengths \( w \) for each tree topology separately with maximum likelihood. This approach is obviously suboptimal. For a proper estimation of the branch lengths of a recombinant tree, one would have to
base the parameter estimation on the recombinant region. The location of this region, however, is not known in advance. Estimating the branch lengths from the whole DNA sequence alignment leads to seriously distorted values, as demonstrated by Husmeier and Wright (2001), since the estimation includes data for which the tree topology is incorrect. A heuristic way to address this problem, suggested by McGuire, Wright, and Prentice (2000), is to estimate the branch lengths from a subregion of the alignment. The length of this region should be matched to the length of the recombinant region, which, however, is not known in advance. Also, this approach does not offer a way to estimate the recombination parameter $\nu$.

**HMM-ML**

A solution to this problem, proposed by Husmeier and Wright (2001), is a proper maximum likelihood estimation of the parameters so as to maximize

$$L(w, \nu) = \ln P(D|w, \nu) = \ln \sum_S P(D, S|w, \theta, \nu)$$

with respect to the vector of branch lengths $w$, the parameters of the nucleotide substitution model $\theta$, and the recombination parameter $\nu$. This requires a summation over all state sequences $S = (S_1, \ldots, S_N)$, that is, over $K^N$ terms, and seems to be intractable for all but very short sequence lengths $N$. However, Husmeier and Wright (2001) showed that by applying the expectation maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977), the sparseness of the connectivity in the HMM could be exploited to reduce the computational complexity to the order of $K$ separate tree optimizations. While the application of this scheme outperformed the heuristic approach of McGuire, Wright, and Prentice (2000), it still suffers from certain shortcomings. First, Husmeier and Wright (2001) did not optimize the nucleotide substitution parameters $\theta$: 
The Kimura model was used, with a fixed transition-transversion ratio of $\tau = 2$. Second, when optimizing the parameters in a maximum likelihood sense, the prediction of the optimal state sequence according to (5) might be susceptible to over-fitting. Husmeier and Wright (2001) addressed this problem by resorting to hypothesis testing with parametric bootstrapping, but this approach is computational demanding and suffers from the inherent problems of multiple testing. The authors also tried model averaging, but pointed out that this was a crude, heuristic approximation to a Bayesian scheme. The present article overcomes these shortcomings by developing a proper Bayesian approach, whereby all model parameters are sampled from the proper posterior distribution with Markov chain Monte Carlo.

3 Method: A Bayesian approach

A Bayesian approach to phylogenetics without recombination was proposed and tested by Yang and Rannala (1997) and Larget and Simon (1999). Generalizing this scheme to the presence of recombination requires replacing the single topology-indicating variable by the state sequence $S$, as discussed in the previous section. The prediction of this state sequence should be based on the posterior probability $P(S|\mathcal{D})$, which requires integrating out the remaining parameters:

$$P(S|\mathcal{D}) = \int P(S, w, \theta, \nu|\mathcal{D}) dw d\theta d\nu$$  \hspace{1cm} (7)

In principle this avoids the over-fitting problem mentioned above and removes the need for a separate hypothesis test. The difficulty, however, is that the integral in (7) is analytically intractable, which calls for the application of a numerical approximation, using Markov chain Monte Carlo (MCMC). The practical viability of the Bayesian framework thus hinges on the performance of this scheme. In the subsections below, we will dis-
cuss the following issues: (1) the choice of prior probabilities, (2) the chosen Markov chain Monte Carlo method, which has the form of a Metropolis Hastings and Gibbs within Gibbs sampling scheme, and a (3) simulated annealing scheme for accelerating convergence. We will then test this approach on various DNA sequence alignments.

**Prior probabilities**

Inherent to the Bayesian framework is the choice of prior probabilities for all model parameters, as illustrated in Figure 2, right. We make the usual assumption of parameter independence (see, e.g., Heckermann (1999)), $P(\nu, w, \theta) = P(\nu)P(w)P(\theta)$, and choose rather vague priors to reflect the absence of true prior knowledge. The prior probabilities will either be conjugate, where possible, or uniform, but proper (that is, restricted to a finite interval).

The recombination parameter $\nu$ is a binomial random variable, for which the conjugate prior is a beta distribution:

$$P(\nu) = B(\nu|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\nu^{\alpha-1}(1 - \nu)^{\beta-1}$$

(8)

The hyperparameters $\alpha$ and $\beta$ determine the shape of the distribution, as shown in Figure 3, and $\Gamma$ denotes the gamma function (required for normalization).

**Figure 3 approximately here.**

The branch lengths $w$ are defined in the usual way, that is, they represent the average number of nucleotide substitutions per site. *A priori*, they are assumed to be uniformly distributed in the interval $[0, 1]$. Fixing an upper bound on the branch lengths is necessary to avoid the use of an improper prior, for which the MCMC scheme might
not converge. Since for real DNA sequence alignments branch lengths are unlikely to approach 1, this restriction should not cause any difficulties.

The prior on $\theta$ depends on the model of nucleotide substitution. In the present study, the Felsenstein 84 model (Felsenstein and Churchill, 1996) is used, which has four free parameters: The nucleotide frequencies $\pi_A$, $\pi_C$, $\pi_G$, and $\pi_T$, and the transition bias $\rho$. Note that because of the normalization constraint $\pi_A + \pi_C + \pi_G + \pi_T = 1$, there are 4 rather than 5 free parameters. Also note that the sixth parameter of the Felsenstein 84 model, the overall nucleotide substitution rate, is redundant and has been arbitrarily set to one. (Therefore it does not appear here. Note that this implies that the branch lengths correspond to the number of substitutions per site).

In our approach, each tree is allowed to have a different value of $\rho$. However, the nucleotide frequencies are assumed to be the same for all trees. This is for estimation reasons: allowing each tree to have a different set of frequencies means that some frequencies might be estimated from a small amount of data, thus leading to vague posterior distributions. The total vector of nucleotide substitution parameters is thus of the form

$$\theta = (\rho_1, \ldots, \rho_K, \pi_A, \pi_C, \pi_G, \pi_T)$$

A natural prior for the nucleotide frequencies $\pi_i$ is a Dirichlet distribution, which, as a multivariate generalization of the beta distribution (8), satisfies the normalization constraint. We here choose a Dirichlet($1,1,1,1$) distribution, which is a uniform distribution subject to the normalization constraint and thus maximally non-informative. For the transition biases $\rho_k$, we choose a uniform prior over the interval $[0, 2]$. Again, an upper bound is needed to prevent the prior from becoming improper. Allowing $\rho_k$ to be as large as 2 will account for extreme cases of transition bias, which should not impose any serious restrictions in practice.
The joint distribution of the DNA sequence alignment, the state sequences, and the model parameters, is given by

$$P(\mathbf{D}, \mathbf{S}, \mathbf{w}, \theta, \nu) = \prod_{t=1}^{N} P(y_t|S_t, \mathbf{w}, \theta) \prod_{t=2}^{N} P(S_t|S_{t-1}, \nu) P(S_1) P(\mathbf{w}) P(\theta) P(\nu)$$

where $P(y_t|S_t, \mathbf{w}, \theta)$ is the probability of the $t$th column of nucleotides in the alignment, which is computed with the pruning algorithm (Felsenstein, 1981), $P(S_t|S_{t-1}, \nu)$ is the probability of transitions between states, given by (4), $P(S_1)$ is fixed, as discussed above, and $P(\mathbf{w})$, $P(\theta)$, and $P(\nu)$ are the prior probabilities, also discussed above.

**Markov chain Monte Carlo (MCMC) sampling**

Ultimately, we are interested in the marginal posterior probability of the state sequences, $P(\mathbf{S}|\mathbf{D})$, which requires a marginalization over the model parameters according to (7). The numerical approximation is to sample from the joint posterior distribution

$$P(\mathbf{S}, \mathbf{w}, \theta, \nu|\mathbf{D})$$

and then to discard the model parameters. To sample from the joint posterior probability, we follow a Gibbs sampling procedure (see, e.g., (Casella and George, 1992)), and sample each parameter group separately conditional on the others. So if the superscript $(i)$ denotes the $i$th sample of the Markov chain, we obtain the $(i+1)$th sample as follows:

$$S^{(i+1)} \sim P(\cdot|\mathbf{w}^{(i)}, \theta^{(i)}, \nu^{(i)}, \mathbf{D})$$

$$\mathbf{w}^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)}, \theta^{(i)}, \nu^{(i)}, \mathbf{D})$$

$$\theta^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{w}^{(i+1)}, \nu^{(i)}, \mathbf{D})$$

$$\nu^{(i+1)} \sim P(\cdot|\mathbf{S}^{(i+1)}, \mathbf{w}^{(i+1)}, \theta^{(i+1)}, \mathbf{D})$$
The order of these sampling steps, which will be discussed in the remainder of this subsection, is arbitrary.

Define $\Psi = \sum_{t=1}^{N-1} \delta(S_t, S_{t+1})$. From (4) and (8) it is seen that writing the joint probability (10) as a function of $\nu$ gives:

$$P(\mathcal{D}, S, w, \theta, \nu) \propto \nu^{\Psi+\alpha-1}(1-\nu)^{N-\Psi+\beta-2} \tag{13}$$

On normalization this gives

$$P(\nu|\mathcal{D}, S, w, \theta) = B(\nu|\Psi + \alpha, N - 1 - \Psi + \beta) \tag{14}$$

where $B$ is the beta distribution (8), from which sampling is straightforward (see, e.g., Rubinstein (1981)).

For sampling the state sequences $S$, we adopt the approach suggested by Robert, Celeux, and Diebolt (1993) and sample each state $S_t$ separately conditional on the others, that is, with a Gibbs-within-Gibbs scheme:

$$S_1^{(i+1)} \sim P(.|S_2^{(i)}, S_3^{(i)}, \ldots, S_N^{(i)}, \mathcal{D}, w^{(i)}, \theta^{(i)}, \nu^{(i)})$$

$$S_2^{(i+1)} \sim P(.|S_1^{(i+1)}, S_3^{(i)}, \ldots, S_N^{(i)}, \mathcal{D}, w^{(i)}, \theta^{(i)}, \nu^{(i)})$$

$$\vdots$$

$$S_N^{(i+1)} \sim P(.|S_2^{(i+1)}, S_3^{(i+1)}, \ldots, S_{N-1}^{(i+1)}, \mathcal{D}, w^{(i)}, \theta^{(i)}, \nu^{(i)}) \tag{15}$$

The computational complexity of this scheme is reduced considerably by the sparseness of the connectivity in the HMM. Form the theory of graphical models it is known (see, e.g., Heckermann (1999)) that a node in the graph is only dependent on the Markov
blanket, that is, the set of parents, children, and coparents. This implies that

\[ P(S_t|S_1, \ldots, S_{t-1}, S_{t+1}, \ldots, S_N, D, w, \theta, \nu) = P(S_t|S_{t-1}, S_{t+1}, y_t, w, \theta, \nu) \]

\[ \propto P(S_{t+1}|S_t, \nu)P(S_t|S_{t-1}, \nu)P(y_t|S_t, w, \theta) \]

where \( P(S_t|S_{t-1}, \nu) \) and \( P(S_{t+1}|S_t, \nu) \) are given by (4). Note that the second expression on the right is easily normalized to give a proper probability, from which sampling is straightforward (since \( S_t \in \{1, \ldots, K\} \) is discrete).

For sampling the remaining parameters, \( w \) and \( \theta \), we apply the Metropolis-Hastings algorithm (see, e.g., (Chib and Greenberg, 1995)). Let \( z^{(i)} \) denote the parameter configuration in the \( i \)th sampling step. A new parameter configuration \( \tilde{z} \) is sampled from a proposal distribution \( Q(\tilde{z}|z^{(i)}) \), and then accepted with probability

\[ A(\tilde{z}) = \min \left\{ \frac{P(\tilde{z})Q(z^{(i)}|\tilde{z})}{P(z^{(i)})Q(\tilde{z}|z^{(i)})}, 1 \right\} \]

in which case \( z^{(i+1)} = \tilde{z} \). Otherwise, \( z^{(i+1)} = z^{(i)} \). The distribution \( P \) is given by (10). In theory the algorithm converges to the posterior distribution (11) irrespective of the choice of the proposal distribution (assuming ergodicity). In practice, a ‘good’ choice of \( Q(.,.) \) is crucial to achieve convergence within a reasonable amount of time, and will be discussed next.

For the components \( w_l \) of the vector of branch lengths \( w \) and for the transition biases \( \rho_k \), a new value is selected from a uniform interval centred around the existing value. This is a symmetric proposal distribution, so the terms \( Q(.,.) \) cancel out in (17). For the nucleotide frequencies \( \pi_A, \pi_C, \pi_G, \pi_T \), new values are sampled from a Dirichlet distribution. This ensures that the normalization constraint \( \pi_A + \pi_C + \pi_G + \pi_T = 1 \) is satisfied. The parameters of the Dirichlet distribution are chosen proportional to the current values of the nucleotide frequencies, thereby proposing new values close to
the current ones, which in turn makes it more likely that the proposed values will be accepted. This proposal distribution is not symmetric, so the $Q(\cdot|\cdot)$ terms must be calculated in (17).

If too few proposed values are accepted, the corresponding proposal distributions $Q(\cdot|\cdot)$ may be tuned to make acceptance more likely and thereby to accelerate convergence. For the branch lengths $w_l$ and the transition biases $\rho_k$, this is done by decreasing the width of the uniform interval from which the new value is sampled. For the nucleotide frequencies $\pi_A, \pi_C, \pi_G, \pi_T$, the constant of proportionality in the Dirichlet distribution is increased so that the proposed frequencies are more likely to be closer to the existing values.

The algorithm is started by firstly initializing the chain. The sequence of topologies $S$ is chosen randomly or from some initial estimation, e.g., using RecPars. The branch lengths are set to some plausible value, e.g. $w_l = 0.1$. Initial values for the transition biases $\rho_k$ and the nucleotide frequencies $\pi_A, \pi_C, \pi_G, \pi_T$ can be estimated from the data, as described below. The parameter groups are then updated in order according to (12) and the details described above. An initial equilibration or burn-in period must be run to allow the Markov chain to reach stationarity. In this part of the simulation, the parameters of the proposal distributions $Q(\cdot|\cdot)$ are tuned as described above. This is followed by the sampling phase of the simulation, in which the state sequences $S$ (and, if of interest, the model parameters) are saved for further analysis. Note that during the sampling phase, the parameters of the proposal distributions must not be tuned as this might lead to biased samples that do not represent the correct posterior probability (11).
Simulated annealing

The bottleneck of the presented Markov chain Monte Carlo scheme is the sampling of the state sequences $S$. Since recombination events are quite rare, the number of topology changes along the DNA sequence alignment is usually small and, consequently, the posterior distribution of $\nu$ concentrated on values close to 1. This discourages state transitions and may thus slow down the mixing and convergence of the Markov chain.

Now, it is well-known that mixing of a Markov chain generated with the Metropolis-Hastings algorithm may be improved with simulated annealing (Kirkpatrick, Gelatt, and Vecchi, 1983), where a modified acceptance probability is used, which is larger than (17) at the beginning of the equilibration phase, and then, during this phase, slowly converges to (17). However, since the recombination parameter $\nu$ is sampled with the Gibbs sampler, for which proposed new values are always accepted, this approach is not applicable.

We therefore use a modification of this scheme, which follows the same idea as simulated annealing, but is applicable to Gibbs sampling. Let $T$ denote the total length of the equilibration phase, and let $i \in \{1, \ldots, T\}$ denote the $i$th sample of the Markov chain during equilibration. Then, during equilibration, (14) is replaced by

$$ P^{(i)}(\nu|D, S, w, \theta) = B \left( \nu| \frac{i}{T} \Psi + \alpha, \frac{i}{T} (N - 1 - \Psi) + \beta \right) $$

For $i = 0$, this distribution is identical to the prior distribution (8), while for $i = T$ it is identical to the posterior distribution (14). For intermediate values, $0 < i < T$, the hyperparameters of (18) are a mixture of the prior and posterior hyperparameters, and the distribution (18) thus shows a gradual transition from the prior to the posterior distribution. Consequently, during and especially at the beginning of the equilibration
phase, small values of $\nu$ will be sampled with a higher likelihood than with the standard (unannealed) scheme; see Figure 9, top right (discussed in more detail below). This facilitates transitions between state sequences and can be expected to increase mixing of the Markov chain.

3.0.1 Implementation

The method discussed above has been implemented in the C++ program package BARCE, which is available upon request from the authors.

4 Data

We tested the viability of the proposed method on the following three DNA sequence alignments.

**Synthetic data.** DNA sequences, 1000 bases long, were evolved along a 4-species tree, using the Kimura model of nucleotide substitution (Kimura, 1980) with a transition-transversion ratio of 2. Two recombination or gene conversion events were simulated by exchanging the indicated lineages, as shown in Figure 4.

**Maize.** Indication of gene conversion between a pair of maize actin genes has been reported by Moniz de Sa and Drouin (1996), who showed that the Maz56 and Maz63 genes had a gene conversion covering the first 875 nucleotides of their coding regions. We applied our algorithm to a multiple alignment of the following four
maize sequences (1008 nucleotides long): Maz56 (GenBank/EMBL accession number U60514), Maz63 (U60513), Maz89 (U60508), and Maz95 (U60507). The sequences were aligned with CLUSTAL W (Thompson, Higgins, and Gibson, 1994), using the default parameter settings. We define the states of the HMM as follows: State 1: ((Maz56,Maz63),(Maz89,Maz95)); state 2: ((Maz56,Maz89),(Maz63,Maz95)); state 3: ((Maz56,Maz95),(Maz63,Maz89)).

**Neisseria.** One of the first indications for sporadic recombination was found in the bacterial genus *Neisseria* Smith (1992). We chose a subset of the 787-nucleotide *Neisseria argF* DNA multiple alignment studied by Zhou and Spratt (1992), where we selected the four strains (1) *N.gonorrhoeae* (X64860), (2) *N.meningitidis* (X64866), (3) *N.cinera* (X64869), and (4) *N.mucosa* (X64873) (GenBank/EMBL accession numbers are in brackets). Zhou and Spratt (1992) found two anomalous, or more diverged regions in the DNA alignment, which occur at positions $t = 1 - 202$ and $t = 507 - 538$ (Note that Zhou and Spratt (1992) used a different labelling scheme, with the first nucleotide at $t = 296$, and the last one at $t = 1082$.) In the rest of the alignment, *N.meningitidis* clusters with *N.gonorrhoeae* (defined as state $S_t = 1$ in our HMM), while between $t = 1$ and $t = 202$, they found that it is grouped with *N.cinera* (defined as state $S_t = 3$). Zhou and Spratt (1992) suggested that the region $t = 507 - 538$ was more diverged as a result of rate variation. The situation is illustrated in Figure 5.

**Figure 5 approximately here.**

Note that by restricting the alignments to $m = 4$ sequences we keep the dimension of the state space limited to $K = 3$ tree topologies: $S_t \in \{1, 2, 3\}$. 

18
5 Simulations

Application of the methods

We applied RecPars with the parameters used by Hein (1993), that is, with a transition cost of $C_{\text{transition}} = 2$, a transversion cost of $C_{\text{transversion}} = 5$, but different choices for the recombination cost $C_{\text{recomb}}$.

The application of HMM-Heu was similar to the study by McGuire, Wright, and Prentice (2000). We chose the Felsenstein 84 model of nucleotide substitution, estimating the transition-transversion ratio with maximum likelihood (using the program package PUZZLE by Strimmer and von Haeseler (1996)), and estimating the nucleotide frequencies $\pi_A, \pi_C, \pi_G, \pi_T$ from the data according to $\pi_X = N_X / X$, where $N_X$ is the number of occurrences of nucleotide $X \in \{A, C, G, T\}$. For each topology in turn, we optimized the branch lengths of the corresponding phylogenetic tree with maximum likelihood on the whole alignment, using the program DNAML of the PHYLIP package. As opposed to McGuire, Wright, and Prentice (2000), we did not restrict the optimization to subsets of the alignments, since the subset size is a parameter that cannot be properly optimized within the framework of this approach.

For training the HMM-ML, we followed Husmeier and Wright (2001) and optimized $\nu$ and all the branch lengths simultaneously in a maximum likelihood sense with the EM algorithm, using the MATLAB programs written by the authors.

Finally, the newly proposed method, HMM-Bayes, was applied as follows. We used the Felsenstein 84 model of nucleotide substitution, with a prior on the parameters as described in Section 3. For the prior on the recombination parameter $\nu$, we chose a beta distribution with hyperparameters $\alpha = 8$ and $\beta = 2$, shown in the bottom left
of Figure 3 (mean $\mu = \frac{\alpha}{\alpha + \beta} = 0.8$). This seems a reasonable choice in that the prior is both sufficiently vague and shifted towards larger, more plausible values of $\nu$. The initial nucleotide frequencies and the initial transition-transversion ratio were estimated from the data, as described above (see HMM-Heu). For the initialization of the hidden state sequences $S$, we tried three different options: (1) a uniform sequence that contains only the maximum likelihood topology – this was found to be equal to the sequence predicted with RecPars, $C_{recomb} = 100$; (2) the state sequence predicted with RecPars, $C_{recomb} = 10$; (3) a random sequence of nine uniform sections of equal length, three for each state. The parameters and state sequences were sampled with MCMC, as described in Section 3, using the annealing scheme of (18) during equilibration. Equilibration was carried out over $10^6$ MCMC steps. This was followed by a sampling phase over $10^6$ MCMC steps, during which parameters and state sequences were recorded every $10^3$ MCMC steps.

**Results**

From the recorded state sequences, we computed the marginal posterior probability for a state at the $t$th site of the alignment, $P(S_t|\mathcal{D})$. For each of the three possible tree topologies $S_t \in \{1, 2, 3\}$, we plotted $P(S_t|\mathcal{D})$ along the DNA sequence alignment, $1 \leq t \leq N$. The results are shown in Figures 6-8. The figures contain several subfigures, where each subfigure is composed of three graphs: $P(S_t = 1|\mathcal{D})$ (top), $P(S_t = 2|\mathcal{D})$ (middle), and $P(S_t = 3|\mathcal{D})$ (bottom).

Figure 6 approximately here.

Figure 7 approximately here.

Figure 8 approximately here.
Convergence of the Markov chain

In principle, the initialization of the hidden states is unimportant since the Markov chain will forget its initial configuration and converge to the equilibrium distribution irrespective of its starting point. In practice, however, extreme starting values could slow down the mixing of the chain and result in a very long burn-in, in which case the MCMC sampler may fail to converge towards the main support of the posterior distribution.

On the synthetic alignment, the final prediction was independent of the initialization, irrespective of whether or not the annealing scheme was used. The lengths of the equilibration and sampling periods could even be reduced by a factor of 10 without changing the result.

On the two real-world alignments, the final predictions were dependent on the initialization when no annealing scheme was used. When using the annealing scheme, the predictions turned out to be (almost) independent of the initialization, suggesting that the MCMC sampler had converged to the proper equilibrium posterior distribution (11).

The typical behaviour of the sampling process is shown in Figure 9. The subfigure in the top right shows the evolution of the recombination parameter $\nu$. Without annealing (circles), $\nu$ is always close to one, which discourages topology changes. The annealing scheme (crosses) reduces these values during equilibration as a result of mixing the posterior distribution with the prior. This makes topology changes more likely at the beginning of the sampling process and helps the Markov chain to leave the initial configuration of hidden states.

The subfigure in the top left shows the evolution of $L$, the logarithm of the unnormalized
posterior (10), during the MCMC sampling process with (thick line) and without (thin line) annealing. As a result of the annealing scheme, $\mathcal{L}$ is reduced at the early stage of the equilibration phase, but eventually takes on values that are larger than in the simulation without annealing. This suggests that annealing improves the mixing of the Markov chain and accelerates convergence.

The predictions obtained from the sampling phases of the two chains are shown in the bottom subfigures of Figure 9. The subfigure on the right shows the prediction obtained with annealing. This prediction is similar to that of Figure 8, bottom right, which was obtained by averaging over all the predictions resulting from the various initializations (always using annealing). The rather small difference between these predictions suggests that the dependence of the sampling scheme on the initialization is weak, which suggests that the Markov chain has (nearly) converged to the equilibrium posterior distribution. On the contrary, the subfigure on the bottom left shows a prediction that is very different, does not accord with the results from the literature (Zhou and Spratt, 1992), and suggests that without annealing, convergence of the Markov chain has not been achieved.

Figure 9 approximately here.

6 Discussion

Comparison with HMM-Heu: Synthetic sequence alignment

Figure 6 shows the results obtained on the synthetic DNA sequence alignment. The subfigure on the left shows the prediction of $P(S_i|\mathcal{D})$ with HMM-Heu. For this method,
the recombination parameter $\nu$ has to be specified in advance, and we have set it to the mean of the prior distribution: $\nu = 0.8$. It is seen that the overall pattern of the posterior probabilities is correct, showing an increase for state $S_t = 2$ in the region $200 < t < 400$, and an increase for state $S_t = 3$ in the region $600 < t < 800$. However, the signals are very noisy, and an automatic classification based on the mode of the posterior probability would incur a high proportion of erroneously predicted topology changes. The Bayesian scheme, HMM-Bayes, shown on the right of Figure 6, overcomes this shortcoming. The predicted state transitions coincide with the true breakpoints, and the tree topologies are predicted correctly. The posterior probabilities for the states, $P(S_t|D)$, are mostly close to zero or one. This indicates a high confidence in the prediction, which is reasonable: Since the DNA sequence alignment results from the simulation of a recombination process, the transitions between topologies are, in fact, well defined. The mean and the standard deviation of the posterior distribution $P(\nu|D)$ are $\langle \nu \rangle_{\text{posterior}} = 0.992$ and $\sigma_{\text{posterior}} = 0.004$. With four breakpoints in an alignment of length 1000 bases, the correct value for the recombination parameter is $\nu = 0.996$, which deviates from the prediction by only 0.4%.

Comparison with HMM-Heu: Maize

Figure 7 shows the prediction of $P(S_t|D)$ for the maize sequence alignment. The subfigures in the top row show predictions obtained with HMM-Heu, using different recombination parameters, $\nu = 0.8$ (left) and $\nu = 0.95$ (right). The overall pattern of the graphs captures the gene conversion event in that the final section shows a clear increase of the posterior probability for state $S_t = 3$. However, the signals are very noisy and unsuitable for an automatic detection of gene conversion without human intervention. The subfigure on the bottom left of Figure 7 shows the prediction with HMM-Heu when setting $\nu$ to the Bayesian posterior mean, $\nu = 0.997$, obtained with HMM-Bayes. This
leads to a considerable reduction of the noise and a qualitatively correct prediction of the gene conversion event. However, the breakpoint deviates considerably from that predicted by Moniz de Sa and Drouin (1996). A clear improvement is obtained with HMM-Bayes, see Figure 7, bottom right, which predicts a sharp transition from state \( S_t = 1 \) to state \( S_t = 3 \) at the location \( t \) predicted by Moniz de Sa and Drouin (1996).

**Comparison with HMM-Heu: Neisseria**

Figure 8 shows the prediction of \( P(S_t|D) \) for the *Neisseria* sequence alignment. The subfigure in the top left shows the prediction obtained with HMM-Heu, setting \( \nu = 0.9 \). The signal is very noisy and only gives a vague indication of a topology change at the beginning of the alignment. The subfigure in the top right shows the prediction with HMM-Heu when using the Bayesian posterior mean of \( \nu \) obtained with HMM-Bayes, \( \nu = 0.988 \). This leads to a considerable reduction of the noise and a qualitatively correct prediction of a topology change from state \( S_t = 3 \) to \( S_t = 1 \) at the beginning of the alignment, in accord with the findings by Zhou and Spratt (1992). However, the breakpoint at site \( t = 202 \), found by Zhou and Spratt (1992), is not clearly predicted. Also, a transition into state \( S_t = 2 \) is indicated, which was not found by Zhou and Spratt (1992). The prediction with HMM-Bayes, shown in Figure 8, bottom right, is in better agreement with the results of Zhou and Spratt (1992) in that it predicts a clear breakpoint at site \( t = 202 \). The other breakpoints predicted with HMM-Bayes are discussed below.
Comparison with RecPars

The results obtained with RecPars were found to depend sensitively on the recombination cost parameter. With a value of $C_{\text{recomb}} = 100$, as used by Hein (1993), none of the recombinant regions was detected. Reducing the recombination cost and selecting the best value, RecPars detected the recombinant regions in the synthetic DNA alignment with a precision of 5-10 bases. In the maize actin gene alignment, it detected the gene conversion event, but the breakpoint deviated by 24 bases from the prediction by Moniz de Sa and Drouin (1996). On the Neisseria data, the first 42 bases were detected as recombinant, whereas Zhou and Spratt (1992) predicted a region of 202 bases.

This suggests that HMM-Bayes locates the breakpoints of the recombinant regions more accurately than RecPars, partly due to the fact that RecPars uses only topology-defining sites, which inherently limits its resolution. A more substantial shortcoming of RecPars is the need to select the recombination cost $C_{\text{recomb}}$ – the parsimony equivalent to the recombination parameter $\nu$ – in advance. The results described above were obtained by selecting the best value of $C_{\text{recomb}}$, which is not possible in real application where the locations of the (putative) recombinant regions are not known beforehand. HMM-Bayes achieves a significant improvement in this respect by sampling the parameters from the posterior distribution (11) and thus does away with any heuristic tuning of parameters by hand.

Comparison with HMM-ML

On the synthetic and the maize sequence alignments, the predictions with HMM-ML (graphs not included in this paper) and HMM-Bayes were practically indistinguishable. The difference between the two approaches is in the confidence that we have in the
prediction. The prediction with HMM-ML, \( P(S|D, w, \theta, \nu) \), is dependent on the model parameters \( w, \theta, \nu \), which were fitted with maximum likelihood and could therefore be subject to over-fitting. This calls for an independent statistical significance test, using, e.g., parametric bootstrapping, as in Husmeier and Wright (2001). The prediction with HMM-Bayes, on the other hand, is only dependent on the data, \( P(S|D) \), since the model parameters have been integrated out. This means that the prediction is consistent within the Bayesian framework and does not require an independent significance test.

Figure 8 shows the prediction of \( P(S_t|D) \) on the *Neisseria* alignment, where the subfigure in the bottom left was obtained with HMM-ML, and the subfigure in the bottom right with HMM-Bayes. Both methods agree in predicting a sharp transition from topology \( S_t = 3 \) to \( S_t = 1 \) at breakpoint \( t = 202 \), which is in agreement with the findings by Zhou and Spratt (1992). Both methods also agree in predicting a short recombinant region of the same topology change at the end of the alignment. However, while the prediction with HMM-ML could have been the result of over-fitting, HMM-Bayes, by integrating out the model parameters, is not susceptible to this fallacy. This corroborates the prediction with HMM-ML, in the same way as a frequentist hypothesis test, and thus suggests that we have discovered a new recombinant region undetected by Zhou and Spratt (1992).

Differences between the predictions of HMM-ML and HMM-Bayes are found in the middle of the alignment, where two further breakpoints occur at sites \( t = 506 \) and \( t = 537 \). This is in agreement with Zhou and Spratt (1992). However, while Zhou and Spratt (1992) suggested that the region between \( t = 506 \) and \( t = 537 \) was differently diverged due to rate heterogeneity, HMM-ML predicts a recombination event with a transition from topology \( S_t = 1 \) into \( S_t = 2 \). The reason for this difference was discussed by Husmeier and Wright (2001): Since the distribution of the nucleotide column vectors \( y_t \) in the indicated region is significantly different from the rest of the alignment, mod-
elling this region with a different hidden state can increase the likelihood although the hidden state itself (topology $S_t = 2$) might be ill-matched to the data. Consequently, over-fitting arises as a consequence of model misspecification (discussed further below). Interestingly, this deficiency is partially redeemed when using HMM-Bayes, whose prediction is shown in Figure 8, bottom right. The critical region between sites $t = 506$ and $t = 537$ is again identified, indicated by a strong drop in the posterior probability for the dominant topology, $P(S_t = 1|\mathcal{D})$. Again, the model suffers from misspecification in that the hidden states available only represent different tree topologies, but do not allow the modelling of different evolutionary rates. However, this misspecification is indicated by a distributed representation, where both alternative hidden states, $S_t = 2$ and $S_t = 3$, are assigned a significant probability mass. With the prediction of this uncertainty, HMM-Bayes indicates the model misspecification inherent to the current scheme, and thus avoids the over-fitting incurred when applying HMM-ML.

7 Conclusion

In this article, we have proposed a Bayesian MCMC method (HMM-Bayes) for detecting recombination and gene conversion with HMMs. This follows up on earlier work by McGuire, Wright, and Prentice (2000) and Husmeier and Wright (2001), where the parameters were estimated heuristically (HMM-Heu) or with maximum likelihood (HMM-ML). We have tested the methods on three DNA sequence alignments and found that HMM-Bayes leads to a considerable improvement in the detection of recombination and gene conversion over RecPars and HMM-Heu. On the synthetic and maize DNA sequence alignments, the predictions with HMM-ML and HMM-Bayes were similar, with the difference that HMM-Bayes includes an automatic hypothesis test (within the Bayesian framework and subject to the HMM structure being appropriate), while the
predictions with HMM-ML have to be tested separately. Thus, the Bayesian method proposed in this paper offers an alternative scheme for validating the predictions with HMM-ML, which is mathematically more sound (avoiding the multiple testing problem) and computationally more efficient (as discussed by Larget and Simon (1999)) than the parametric bootstrapping approach used by Husmeier and Wright (2001). On the *Neisseria* DNA sequence alignment, HMM-Bayes gives a better prediction than HMM-ML in that it indicates uncertainty in a differently diverged region, where HMM-ML leads to over-fitting as a result of model misspecification. This points to an inherent limitation of the current approach: The hidden states represent different tree topologies, but do not allow for different rates of evolution. A way to redeem this deficiency is to employ a factorial hidden Markov model (FHMM), as discussed by Ghahramani and Jordan (1997), and to introduce two different types of hidden states: one representing different topologies, the other representing different evolutionary rates. This effectively combines the method of the present paper with the approach of Felsenstein and Churchill (1996). While parameter estimation with maximum likelihood would lead to a considerable increase of the computational complexity (Ghahramani and Jordan, 1997), it seems that this increase will be less dramatic for the MCMC method, since the Gibbs sampling scheme of (16) can be applied to both types of hidden states separately. A detailed investigation of this approach is the subject of future research.

As mentioned in Section 2, the method presented here is restricted to DNA sequence alignments with small numbers of sequences. This is because each possible tree topology constitutes a separate state of the HMM. In practical applications, our method is therefore at best combined with a fast low-resolution preprocessing step that can analyze more sequences simultaneously. For example, one can conduct the initial search for recombination with the methods of Bandelt and Dress (1992) or Strimmer and Moulton (2000), which represent evolutionary relationships between sequences by a network if there are conflicting phylogenetic signals in the data. While this does not allow indi-
vidual recombination events to be identified, it is a useful preprocessing step in that a network that strongly deviates from a bifurcating tree is suggestive of recombination and gives hinds as to which sequences might belong to candidate recombinant strains. After identifying a small set of putative recombinant sequences, the exact nature of the recombination processes and the location of the breakpoints can be further investigated with the high-resolution method discussed in the present paper.

Acknowledgements

Both authors were supported by the Biotechnology and Biological Sciences Research Council (BBSRC) Bioinformatics Initiative. Gráinne McGuire was also partly supported by the Isaac Newton Institute, Cambridge, UK, where she carried out a substantial part of the work leading to the results presented here.

References


Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977, Maximum Likelihood from


Figure 1: Statistical approach to phylogenetics and modelling recombination. For a given column $y_t$ in the alignment, a probability $P(y_t|S_t, w, \theta)$ can be computed, which depends on the tree topology, $S_t$, the vector of branch lengths, $w$, and the parameters of the nucleotide substitution model, $\theta$. In the presence of recombination, the tree topology can change and thus becomes a random variable that depends on the site label $t$. For four taxa, there are three different tree topologies. The vectors $w$ and $\theta$ are accumulated vectors, as defined in the paragraph above equation (5).
Figure 2: **Modelling recombination with hidden Markov models.** Positions in the model, labelled by the subscript $t$, correspond to sites in the DNA sequence alignment. Black nodes represent observed random variables; these are the columns in the DNA sequence alignment. White nodes represent hidden states; these are the different tree topologies, shown (for four sequences) in Figure 1. Arcs represent conditional dependencies. Squares represent parameters of the model. The probability for observing a column vector $y_t$ at position $t$ in the DNA sequence alignment depends on the tree topology $S_t$, the vector of branch lengths $w$, and the parameters of the nucleotide substitution model $\theta$. The tree topology at position $t$ depends on the topologies at the adjacent sites, $S_{t-1}$ and $S_{t+1}$, and the recombination parameter $\nu$. **Left:** In the older approaches of McGuire, Wright, and Prentice (2000) and Husmeier and Wright (2001), $\nu$, $w$, and $\theta$ are parameters that have to be estimated. **Right:** In the Bayesian approach, $\nu$, $w$, and $\theta$ are random variables. The prior distribution for $\nu$ is a beta distribution with hyperparameters $\alpha$ and $\beta$. The prior distributions for the remaining parameters are discussed in Section 3 and depend on some hyperparameters $\Omega$. The parameters $\nu$, $w$, and $\theta$ are sampled from the posterior distribution with Markov chain Monte Carlo.
Figure 3: Prior distribution for the recombination parameter $\nu$. The conjugate prior for $\nu$ is a beta distribution, which depends on two hyperparameters, $\alpha$ and $\beta$. The mean of the distribution is $\mu = \frac{\alpha}{\alpha + \beta}$. The subfigures show plots of the distribution for different values of $\mu$, indicated at the top of each subfigure, when $\beta = 2$ is fixed.
Figure 4: **Synthetic DNA sequence alignment.** Two recombination or gene conversion events are simulated by exchanging the indicated lineages. Defining the predominant tree topology as state 1, the first recombination event corresponds to a transition into state 2, while the second event corresponds to a transition into state 3.
Figure 5: **Recombination in Neisseria.** According to Zhou and Spratt (1992), a recombination event corresponding to a transition from state 1 to state 3 has affected the first 202 nucleotides of the DNA sequence alignment. A second more diverged region seems to be the result of rate variation.
Figure 6: Detection of recombination in the synthetic sequence alignment. The figure contains two subfigures, where each subfigure is composed of three graphs. These graphs show the posterior probabilities for the three topologies, $P(S_t = 1|\mathcal{D})$ (top), $P(S_t = 2|\mathcal{D})$ (middle), $P(S_t = 3|\mathcal{D})$ (bottom), plotted along the DNA sequence alignment (the subscript $t$ denotes the position in the alignment). Left: Prediction with HMM-Heuristic, $\nu = 0.8$. Right: Prediction with HMM-Bayes.
Figure 7: Detection of gene conversion between two maize actin genes. The figure contains four subfigures, where each subfigure is composed of three graphs, as explained in the caption of Figure 6. **Top left:** HMM-Heuristic, $\nu = 0.8$. **Top right:** HMM-Heuristic, $\nu = 0.95$. **Bottom left:** HMM-Heuristic, $\nu = 0.997$. **Bottom right:** HMM-Bayes.
Figure 8: Detection of recombination in the Neisseria DNA sequence alignment. The figure contains four subfigures, where each subfigure is composed of three graphs, as explained in the caption of Figure 6. **Top left:** HMM-Heuristic, $\nu = 0.9$.  **Top right:** HMM-Heuristic, $\nu = 0.988$.  **Bottom left:** HMM-ML.  **Bottom right:** HMM-Bayes.
Figure 9: Simulated annealing applied to the Neisseria DNA sequence alignment. **Top left:** Evolution of the unnormalized posterior probability (ordinate) without annealing (thin solid line) and with the simulated annealing scheme described in the text (thick line). The abscissa shows the number of MCMC steps. **Top right:** Values of $\nu$ (ordinate) sampled during the MCMC simulation with (circles) and without (crosses) simulated annealing. The abscissa shows the number of MCMC steps. **Bottom left:** HMM-Bayes, equilibrated *without* simulated annealing, using the prediction by RecPars for initialization. The three subgraphs are explained in the caption of Figure 6. **Bottom right:** HMM-Bayes, equilibrated *with* simulated annealing, using the prediction by RecPars for initialization. The three subgraphs are explained in the caption of Figure 6.