

Detection of Recombination in DNA Multiple Alignments with Hidden Markov Models

Dirk Husmeier and Frank Wright

Biomathematics and Statistics Scotland (BioSS)

SCRI, Invergowrie, Dundee DD2 5DA

United Kingdom

Phone: +44 1382 562731 X 2601

Fax: +44 1382 562426

Email: [dirk,frank]@bioss.sari.ac.uk

April 26, 2001

Keywords:

Phylogenetic trees, multiple alignments of DNA sequences, recombination, hidden Markov models, maximum likelihood and the expectation maximization (EM) algorithm.

Abstract

Conventional phylogenetic tree estimation methods assume that all sites in a DNA multiple alignment have the same evolutionary history. This assumption is violated in data sets from certain bacteria and viruses due to recombination, a process that leads to the creation of mosaic sequences from different strains and, if undetected, causes systematic errors in phylogenetic tree estimation. In the current work, a hidden Markov model (HMM) is employed to detect recombination events in multiple alignments of DNA sequences. The emission probabilities in a given state are determined by the branching order (topology) and the branch lengths of the respective phylogenetic tree, while the transition probabilities depend on the global recombination probability. The present study improves on an earlier heuristic parameter optimization scheme and shows how the branch lengths and the recombination probability can be optimized in a maximum likelihood sense by applying the expectation maximization (EM) algorithm. The novel algorithm is tested on a synthetic benchmark problem and is found to clearly outperform the earlier heuristic approach. The paper concludes with an application of this scheme to a DNA sequence alignment of the *argF* gene from four *Neisseria* strains, where a likely recombination event is clearly detected.

1 Introduction

Conventional phylogenetic tree estimation methods assume that all sites in a DNA multiple alignment have the same evolutionary history. This is a reasonable approach when applied to DNA sequences obtained from most species. However, this assumption is violated in certain bacteria and viruses due to sporadic *recombination*, which is a process that leads to the transfer of DNA subsequences between different strains. The resulting mixing of the genetic material by the formation of so-called *mosaic* sequences is likely to be an important source of genetic variation and is a process through which, for example, disease-causing bacteria may acquire resistance to antibiotics. Figure 1 shows an example in which the incorporation of the genetic material from another strain 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 mosaic sequences can lead to errors in phylogenetic tree estimation. Their detection, therefore, is a crucial prerequisite for inferring the evolutionary history of a set of DNA sequences.

Figure 1 approximately here

In the last few years, a plethora of methods for detecting interspecies recombination 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 et al., 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 et al., 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.

A different approach was taken in (Hein, 1993), where hidden states were introduced to represent different topologies, and a recombination event was interpreted as a transition between different states. Defining a cost for such a transition in addition to the substitution cost of weighted *parsimony* (Sankoff and Cedergren, 1983), a dynamic programming algorithm that finds the most parsimonious history of the sequences in terms of these two operations can be formulated. While this approach should, in principle, allow a more precise location of the breakpoints, it suffers from the shortcomings inherent to *parsimony*, as discussed in (Felsenstein, 1988).

The present article follows up on earlier work by (McGuire, 1998) and (McGuire et al., 2000), who translated the ideas of (Hein, 1993) into a *likelihood* framework and modelled changes in the topology due to recombination with a *hidden Markov model* (HMM). The idea is to use the *maximum likelihood* methodology, which on its own can lead to large fluctuations in the predictions due to statistical noise (see (McGuire, 1998) and Section 4), to form part of a Bayesian inference scheme, while the Markov chain is used to place a prior probability on the sequence of topologies along a multiple alignment.

While this approach led, overall, to encouraging results, it can be significantly improved in two important respects. First, the recombination probability, that is, the probability of a change in topology due to a recombination event, was not learned from the data but rather needed to be specified as a *constant* in advance. Given that recombination probabilities vary considerably among different species and data sets, an inference scheme that allows for an adjustment in light of the data would be more appropriate. Second,

the branch lengths of the phylogenetic trees can be estimated more accurately. In the previous method, they were optimized *separately*, for each of the possible topologies in turn. This approach is inaccurate. When the branch lengths of a recombinant tree (for instance, the tree in the middle of Figure 1) are estimated in this way, the valid signal from the recombinant region (the grey region in Figure 1) may be swamped by the conflicting signal coming from the nonrecombinant region (the white region in Figure 1) which has a different topology. This may adversely effect the estimation of the branch lengths and lead to suboptimal results, depending on the relative lengths of the recombinant and nonrecombinant regions. (McGuire, 1998) and (McGuire et al., 2000) tried a heuristic scheme and estimated the branch lengths from a subset of the DNA alignment, selected by a moving window of fixed size. Besides introducing a rather arbitrary cutoff between the window subset and the rest of the alignment, this *ad hoc* procedure is unsatisfactory in that the window size is not optimized on the basis of the data, and it is doubtful that enough prior knowledge is available on which a reasonable choice of this parameter can be based. (In (McGuire, 1998) the optimal subset size varied between 5 and 1000 base pairs, depending on the problem.) The approach taken in the present work is to consider all possible phylogenetic trees as part of a ‘super-model’, whose parameters are optimized simultaneously so as to maximize their joint likelihood. This is effected with the expectation maximization (EM) algorithm, which leads to a modified form of the Baum-Welch algorithm for standard HMMs. The resulting branch-length adaptation scheme turns out to be akin to the subset method mentioned above, but introduces *smooth* windows (without sharp cutoffs) whose effective widths are adapted, for each of the possible topologies separately, in light of the data.

The article is organized as follows. Section 2 briefly recapitulates the application of HMMs to the detection of recombination in multiple DNA sequence alignments. In Section 3, we apply the EM algorithm to maximizing the joint likelihood of all the HMM parameters, and discuss the improvements of the resulting training¹ scheme over

¹We use a terminology common in Neural Computation and Machine Learning, where the word

the earlier method of (McGuire, 1998) and (McGuire et al., 2000). The novel algorithm is tested in Section 4 on a synthetic benchmark problem, and the results are compared with those obtained with the earlier method of (McGuire, 1998) and (McGuire et al., 2000). Section 5 demonstrates an application to the detection of recombination among four *Neisseria* strains (*Neisseria gonorrhoeae*, *Neisseria meningitidis*, *Neisseria cinerea*, *Neisseria mucosa*). Section 6 gives the conclusions and an outlook on future work.

training is often used synonymously for *parameter estimation* or *adaptation*.

2 Detecting recombination with HMMs

Let $\mathbf{y}_t \in \{A, G, C, T\}^m$ denote the t th column in a multiple alignment of m DNA sequences of length N , $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$, and introduce the multinomial random variable $s_t \in \{1, \dots, K\}$ to indicate the tree topology that generated the nucleotide configuration at site t . Let \mathbf{w}_{s_t} denote the vector of all branch lengths in the tree corresponding to s_t . A phylogenetic tree is a generative probabilistic model, that is, given the topology s_t and the parameters \mathbf{w}_{s_t} , we can compute the probability for an observed column vector \mathbf{y}_t in the alignment²: $P(\mathbf{y}_t | s_t, \mathbf{w}_{s_t})$.

In the presence of recombination, the tree topology s_t becomes a site-dependent random variable, and our objective is to find the mode of

$$P(\mathbf{s} | \mathbf{Y}) = P(s_1, \dots, s_N | \mathbf{y}_1, \dots, \mathbf{y}_N) \quad (1)$$

that is, the most likely sequence of topologies, $\mathbf{s} = (s_1, \dots, s_N)$, given the data $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$. From Bayes rule we have:

$$P(\mathbf{s} | \mathbf{Y}) \propto P(\mathbf{Y} | \mathbf{s}) P(\mathbf{s}) \quad (2)$$

where the expression on the right needs to be normalized to turn the proportionality into an equality. Under the common assumption that mutations at different sites on the DNA strand are independent of each other (see, for instance, (Felsenstein, 1981)), the first term in the expression on the right factorizes:

$$P(\mathbf{Y} | \mathbf{s}) = \prod_{t=1}^N P(\mathbf{y}_t | s_t) \quad (3)$$

If we assume a uniform prior on the state sequences, $P(\mathbf{s}) = C$, where C denotes a

²More precisely, we should also have noted the dependence on the evolutionary model and its parameters (like the transition-transversion ratio), which we here assume to be known and fixed. See (Durbin et al., 1998), chapter 8 for further details.

constant, then inserting (3) into (2) and normalizing gives:

$$P(\mathbf{s}|\mathbf{Y}) = \prod_{t=1}^N P(s_t|\mathbf{y}_t), \quad P(s_t|\mathbf{y}_t) = \frac{P(\mathbf{y}_t|s_t)}{\sum_{s_t=1}^K P(\mathbf{y}_t|s_t)} \quad (4)$$

This approach, however, gives poor results, as found in (McGuire, 1998) and demonstrated again in Section 4. The problem is caused by the naive assumption of a uniform prior on \mathbf{s} . Recombination typically results in the exchange or transfer of regions of DNA consisting of many bases. This leads to strong correlations between the topologies at adjacent positions in the alignment, which is not captured by the uniform prior. In order to introduce spatial correlations at the lowest possible order, (McGuire, 1998) introduced a prior on the state sequences in form of a first-order Markov process:

$$P(\mathbf{s}) = P(s_1) \prod_{t=2}^N P(s_t|s_{t-1}) \quad (5)$$

Inserting (3) and (5) into (2) we obtain:

$$P(\mathbf{s}|\mathbf{Y}) \propto P(s_1) \prod_{t=2}^N P(s_t|s_{t-1}) \prod_{t=1}^N P(\mathbf{y}_t|s_t) \quad (6)$$

This expansion is of the form of a hidden Markov model (HMM), which is discussed at length in (Rabiner, 1989).

In a standard HMM, the probabilities on the right, that is, the *prior* probabilities $P(s_1)$, the *transition* probabilities $P(s_t|s_{t-1})$, and the *emission* probabilities $P(\mathbf{y}_t|s_t)$, are usually modelled as multinomial distributions. We will briefly recapitulate how this is to be modified for the current problem of detecting recombination.

Transition probabilities

In principle there are $K(K - 1)$ transition probabilities to be specified. Given that recombination is likely to be a rare event, it would hardly be possible to optimize these parameters in a maximum likelihood sense (overfitting), nor is it likely that detailed prior knowledge is available to decide on these parameters in advance. (McGuire, 1998) suggested considering only *one* parameter: the overall probability that no recombination occurs. This is similar to an approach taken in (Felsenstein and Churchill, 1996) for modelling rate variation among sites. In the present work we adopt a slightly modified parameterization in terms of the probability that no recombination is observed³, ν :

$$P(s_t|s_{t-1}) = \nu\delta(s_t, s_{t-1}) + \frac{1 - \nu}{K - 1}[1 - \delta(s_t, s_{t-1})] \quad (7)$$

where $\delta(s_t, s_{t-1})$ denotes the Kronecker delta function, which is 1 when $s_t = s_{t-1}$ and zero otherwise. It is easily checked that this satisfies the normalization constraint $\sum_{s_t} P(s_t|s_{t-1}) = 1$.

Emission probabilities

The emission probabilities $P(\mathbf{y}_t|s_t, \mathbf{w}_{s_t})$ are defined by the chosen evolution model (see, for instance, (Felsenstein, 1981), (Felsenstein and Churchill, 1996)) and depend on the topology of the phylogenetic tree, $s_t \in \{1, \dots, K\}$, and the respective vector of branch lengths, \mathbf{w}_{s_t} . To simplify our notation, we introduce the accumulated vector of all branch lengths in all possible topologies, $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_K)$ and define: $P(\mathbf{y}_t|s_t, \mathbf{w}_{s_t}) = P(\mathbf{y}_t|s_t, \mathbf{w})$. This means that s_t indicates which subvector of \mathbf{w} applies.

³If during recombination a DNA fragment is transferred between identical strains, this event cannot be detected.

Prior probabilities

In principle, one needs to adapt $K - 1$ prior probabilities $P(s_1)$. Due to the likely rarity of recombination events, however, a maximum likelihood approach would most probably lead to overfitting. Also, since DNA sequence alignments are usually sufficiently long, $N \gg K$, the influence of $P(s_1)$ on the mode of $P(s_1, \dots, s_N | \mathbf{Y})$ is negligible. We therefore decided to keep the prior probabilities constant: $P(s_1) = \frac{1}{K} \forall s_1 \in \{1, \dots, K\}$.

In what follows, we will use the symbol \mathbf{q} to denote the vector of all adaptable parameters, which are the branch lengths and the recombination probability: $\mathbf{q} = (\mathbf{w}, \nu)$.

Comparison with Hein's parsimony-based algorithm

The most likely sequence of hidden states conditional on the DNA sequence alignment, $\max_{s_1, \dots, s_N} P(s_1, \dots, s_N | \mathbf{y}_1, \dots, \mathbf{y}_N) = \max_{s_1, \dots, s_N} \ln P(s_1, \dots, s_N, \mathbf{y}_1, \dots, \mathbf{y}_N)$, is computed with the Viterbi algorithm. This draws on the factorization (6), which leads to the recursion:

$$\begin{aligned}
 \gamma_n(s_n) &= \max_{s_1, \dots, s_{n-1}} \ln P(\mathbf{y}_1, \dots, \mathbf{y}_n, s_1, \dots, s_n) \\
 &= \max_{s_1, \dots, s_{n-1}} \sum_{t=1}^n \left[\ln P(\mathbf{y}_t | s_t) + \ln P(s_t | s_{t-1}) \right] + \ln P(s_1) \\
 &= \ln P(\mathbf{y}_n | s_n) + \max_{s_{n-1}} \left[\ln P(s_n | s_{n-1}) + \gamma_{n-1}(s_{n-1}) \right] \tag{8}
 \end{aligned}$$

Obviously, $\max_{s_1, \dots, s_N} \ln P(s_1, \dots, s_N, \mathbf{y}_1, \dots, \mathbf{y}_N) = \max_{s_N} \gamma_N(s_N)$. We compare this with the approach taken in (Hein, 1993), where recombination is detected by considering changes in the most parsimonious topology along an alignment. Let $e(\mathbf{y}_t | s_t)$ denote the substitutional cost of position t given topology s_t , as defined by a model of weighted parsimony (Sankoff and Cedergren, 1983). Let $\tau(s_{t+1}, s_t)$ denote the recombinational

distance between topologies s_t and s_{t+1} , and define $W_t(s_t)$ to be the cost of the most parsimonious history of the first t positions, given that the topology of position t is s_t . The objective of the approach in (Hein, 1993) is to find the most parsimonious history of the whole alignment, $\min_{s_N} W_N(s_N)$. This can be accomplished with a dynamic programming algorithm, drawing on the following recursion:

$$W_n(s_n) = e_n(s_n) + \min_{s_{n-1}} [W_{n-1}(s_{n-1}) + \tau(s_n, s_{n-1})] \quad (9)$$

which is initialized with $W_1(s_1) = e_1(s_1)$. Obviously, on defining

$$W_n(s_n) = -\gamma_n(s_n) \quad (10)$$

$$\tau(s_n, s_{n-1}) = -\ln P(s_n | s_{n-1}) \quad (11)$$

$$e_n(s_n) = -\ln P(\mathbf{y}_n | s_n) \quad (12)$$

equations (9) and (8) become formally identical, that is, the most parsimonious path obtained with the dynamic programming algorithm is equivalent to the Viterbi path of the HMM. The HMM approach discussed in the present paper is thus a consequent improvement on (Hein, 1993) in the sense that the maximum likelihood approach to phylogenetics overcomes well-known shortcomings of the parsimony method (Felsenstein, 1988). Moreover, the approach in (Hein, 1993) does not allow assessing the reliability of the prediction or estimating the parameters (recombination and substitution costs) from the data, that is, they have to be chosen as *a priori* known constants in advance. On the contrary, the HMM discussed in the present paper is a proper generative probabilistic model. This allows assessing the reliability of a prediction by, e.g., computing the probability of the Viterbi path. It also allows estimating the parameters from the data with the method of maximum likelihood, as will be shown in the next section.

3 Methods and Algorithms

In the method of (McGuire, 1998) and (McGuire et al., 2000), the recombination parameter ν was fixed (equivalent to assuming it is known *a priori*), whereas the branch lengths \mathbf{w} were optimized, for each of the possible topologies $k = 1, \dots, K$ separately, so as to maximize the constrained log likelihood:

$$L_c(\mathbf{w}, k) = \ln P(\mathbf{Y} | \mathbf{w}, s_t = k \forall t) \quad (13)$$

Since this keeps the topology fixed – even in regions where it does not apply – we will refer to this approach as the method of *constrained maximum likelihood* (CML). In the present article we will show how to adapt both the recombination parameter and the branch lengths in a joint maximum likelihood sense, that is, so as to maximize:

$$L(\mathbf{q}) = \ln P(\mathbf{Y} | \mathbf{q}) \quad (14)$$

In principle this could be achieved with a gradient ascent scheme, but this would involve a summation over all possible K^N combinations of hidden states: $P(\mathbf{Y} | \mathbf{q}) = \sum_{\mathbf{s}} P(\mathbf{Y}, \mathbf{s} | \mathbf{q})$. Obviously, such an approach becomes untractable for long sequences, $N \gg 1$. A viable alternative, however, is the expectation maximization (EM) algorithm (Dempster et al., 1977), which is based on the following decomposition of the log likelihood (Neal and Hinton, 1999):

$$L(\mathbf{q}) = U(\mathbf{q}) + KL(Q, P) \quad (15)$$

$$U(\mathbf{q}) = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln P(\mathbf{Y}, \mathbf{s} | \mathbf{q}) - \sum_{\mathbf{s}} Q(\mathbf{s}) \ln Q(\mathbf{s}) \quad (16)$$

$$KL(Q, P) = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \left(\frac{Q(\mathbf{s})}{P(\mathbf{s} | \mathbf{Y}, \mathbf{q})} \right) \quad (17)$$

Here, $Q(\mathbf{s})$ is an arbitrary probability distribution over the hidden states, and KL represents the Kullback-Leibler divergence between the distributions Q and $P(\mathbf{s} | \mathbf{Y}, \mathbf{q})$. Note

that $KL(Q, P)$ is always non-negative (and zero if and only if $Q = P$), which implies that U is a lower bound on L : $U(\mathbf{q}) \leq L(\mathbf{q})$. EM alternates between optimizing the distribution over hidden states $Q(\mathbf{s})$ (the E-step) and optimizing the parameters given $Q(\mathbf{s})$ (the M-step). The E-step holds the parameters fixed and sets Q to the posterior distribution over the hidden states given the parameters, $Q(\mathbf{s}) = P(\mathbf{s}|\mathbf{Y}, \mathbf{q})$. This sets $KL(Q, P) = 0$ and, consequently, $L(\mathbf{q}) = U(\mathbf{q})$. The M-step holds the distribution $Q(\mathbf{s})$ fixed and computes the parameters \mathbf{q} that maximize U . Since $L(\mathbf{q}) = U(\mathbf{q})$ at the beginning of the M-step, and since the E-step does not affect the model parameters, each EM cycle is guaranteed to increase the likelihood unless the system has already converged to a (local) maximum (or, less likely, a saddle point).

In the context of HMMs, the E-step is carried out with the *forward-backward* algorithm (Rabiner, 1989), which is a dynamic programming technique that reduces the order of complexity from $O(K^N)$ to $O(NK^2)$. Hence all that remains to be done is to derive update equations for the parameters in the M-step, that is, to maximize the function U defined in (16). We introduce the definition

$$\Psi = \sum_{\mathbf{s}} \sum_{t=2}^N Q(\mathbf{s}) \delta(s_t, s_{t-1}) = \sum_{t=2}^N \sum_{s_t=1}^K Q(s_t, s_{t-1} = s_t) \quad (18)$$

and note that

$$\sum_{\mathbf{s}} \sum_{t=2}^N Q(\mathbf{s}) [1 - \delta(s_t, s_{t-1})] = N - 1 - \Psi \quad (19)$$

Inserting (6), (7), and (18) into (16) gives:

$$\begin{aligned} U &= \sum_{\mathbf{s}} Q(\mathbf{s}) \sum_{t=1}^N \ln P(\mathbf{y}_t | s_t, \mathbf{w}) + \Psi \ln \nu \\ &\quad + (N - 1 - \Psi) \ln \left(\frac{1 - \nu}{K - 1} \right) + C \end{aligned} \quad (20)$$

Optimization of the recombination parameter

Setting the derivative of U with respect to ν to zero, $\frac{\partial U}{\partial \nu} = 0$, we obtain

$$\nu = \frac{\Psi}{N-1} \quad (21)$$

This optimization is straightforward since, as seen from (18), Ψ only depends on $Q(s_{t-1}, s_t)$, which is obtained by application of the forward-backward algorithm.

Optimization of the branch lengths

Only the first term on the left-hand side of (20) depends on the branch lengths \mathbf{w} . This requires a maximization of

$$\begin{aligned} U(\mathbf{w}) &= \sum_s Q(\mathbf{s}) \sum_{t=1}^N \ln P(\mathbf{y}_t | s_t, \mathbf{w}) \\ &= \sum_{t=1}^N \sum_{s_t=1}^K Q(s_t) \ln P(\mathbf{y}_t | s_t, \mathbf{w}) \end{aligned} \quad (22)$$

where $Q(s_t)$ is the t th marginal distribution obtained from $Q(\mathbf{s})$. The optimization can be achieved with standard phylogenetic programs like DNAML of the PHYLIP package⁴. The only modification required is the introduction of a weighting factor $Q(s_t)$ for each site, as illustrated in Figure 2.

Figure 2 approximately here

Note that the window method applied in (McGuire, 1998) and (McGuire et al., 2000)

⁴PHYLIP, developed by J. Felsenstein, is a package of programs for inferring phylogenies. It can be downloaded from <http://evolution.genetics.washington.edu/phylip.html>

can be interpreted as a special case of (22), where the weighting function

$$Q(s_t) = \begin{cases} 1 & \text{if } t \in [t' - \delta, t' + \delta] \\ 0 & \text{if } t \notin [t' - \delta, t' + \delta] \end{cases} \quad (23)$$

has a fixed ‘discontinuous’ functional form that is independent of the tree topology s_t . The window size δ has to be chosen according to some heuristics in advance. This is a rather unnatural choice, which suffers from the lack of adaptation to the given sequence alignment \mathbf{Y} . The weighting scheme proposed in this article, on the other hand, is naturally adapted in the E-step of the training scheme, that is, it is optimized in a maximum likelihood sense during the iterative parameter estimation process. Moreover, the explicit dependence on the tree topology s_t introduces extra flexibility.

Algorithm

The implementation of the parameter update scheme is straightforward and can be accomplished with the following algorithm:

1. Initialize the parameters \mathbf{w} and ν . This can be done as in (McGuire, 1998), that is, by choosing a plausible recombination rate and by estimating \mathbf{w} , for each of the topologies, with a phylogenetic program like DNAML on the whole data set.
2. Compute $Q(s_t)$ and $Q(s_{t-1}, s_t)$ with the forward-backward algorithm for HMMs.
3. Compute Ψ from (18) and adapt ν according to (21).
4. For $t = 1$ to N : weight the t th column in the multiple sequence alignment, \mathbf{y}_t , by $Q(s_t)$, and optimize the branch lengths \mathbf{w} so as to maximize $U(\mathbf{w})$ in (22). This can, in principle, be achieved with a standard phylogeny program, like DNAML of the PHYLIP package. The only change required is the introduction of a weighting

scheme for the sites in the alignment.

5. Test for convergence. If the algorithm has not yet converged, go back to step 2.

Note that this algorithm can be interpreted as a modified version of the Baum-Welch algorithm; see (Rabiner, 1989) for details.

Runtime and algorithmic complexity

In the simulations described later, we found that the EM algorithm usually converged within about 5-20 EM steps. The complexity of a single M step is identical to standard ML optimization (using, e.g., Phylip). The E step is identical to the forward-backward algorithm in standard HMMs, which is comparatively fast. This gives an effective run time that is increased by a factor of 5-20 over standard ML optimization methods (e.g. Phylip).

A committee of HMMs

The EM-algorithm is a greedy optimization scheme that finds the closest *local* maximum in the log likelihood landscape. If the latter is multimodal and the training simulation is repeated from different initializations, this usually results in a set of models $\{\mathcal{M}_i\}$ with different likelihood scores. Note that a model \mathcal{M} in this context refers to the whole HMM, which is defined by the recombination parameter ν and the branch lengths \mathbf{w} of all possible trees. A straightforward approach is to pick out the model $\hat{\mathcal{M}}$ with the highest likelihood score

$$L_i = \ln P(\mathbf{Y}|\mathcal{M}_i) \tag{24}$$

and to predict the recombination events from the mode of $P(\mathbf{s}|\mathbf{Y}, \hat{\mathcal{M}})$. However, from a Bayesian approach it would be more satisfactory to eliminate the explicit model dependence, which formally can be achieved by marginalisation:

$$P(\mathbf{s}|\mathbf{Y}) = \int P(\mathbf{s}, \mathcal{M}|\mathbf{Y})d\mathcal{M} = \int P(\mathbf{s}|\mathbf{Y}, \mathcal{M})P(\mathcal{M}|\mathbf{Y})d\mathcal{M} \quad (25)$$

This integral covers the space of all possible models, which is approximated by a sum over the set of selected models:

$$P(\mathbf{s}|\mathbf{Y}) = \sum_i P(\mathbf{s}|\mathbf{Y}, \mathcal{M}_i)P(\mathcal{M}_i|\mathbf{Y}) \quad (26)$$

Assuming a constant prior, $P(\mathcal{M}_i) = C \forall i$, and making use of Bayes' rule, the posterior probabilities $P(\mathcal{M}_i|\mathbf{Y})$ are given by the normalized likelihoods:

$$P(\mathcal{M}_i|\mathbf{Y}) = \frac{P(\mathbf{Y}|\mathcal{M}_i)P(\mathcal{M}_i)}{P(\mathbf{Y})} = \frac{P(\mathbf{Y}|\mathcal{M}_i)}{\sum_j P(\mathbf{Y}|\mathcal{M}_j)} \quad (27)$$

A direct computation of this expression is likely to run into numerical underflow problems. We therefore reformulate (27) in terms of the log likelihood (24):

$$P(\mathcal{M}_i|\mathbf{Y}) = \frac{1}{\sum_j \exp(L_j - L_i)} \quad (28)$$

Inserting (28) into (26) gives

$$P(\mathbf{s}|\mathbf{Y}) = \sum_i \left(\frac{P(\mathbf{s}|\mathbf{Y}, \mathcal{M}_i)}{\sum_j \exp(L_j - L_i)} \right) \quad (29)$$

Note that the combination of different models for improving the overall classification performance has been extensively studied in the Neural Computation literature; see, for instance, (Battiti and Colla, 1994) and (Husmeier, 1999). We here borrow a term frequently used in Neural Networks research and refer to (29) as a *committee of HMMs*.

4 Findings 1: A synthetic benchmark problem

In order to test if the proposed method can yield useful inferences about the presence of recombination in a phylogenetic data set, we carried out a simulation study similar to that described in (McGuire, 1998). A data set of four DNA sequences, $N = 1000$ nucleotides long, was simulated according to the tree on the left of Figure 3.

Figure 3 approximately here

The Kimura 2-Parameter model of evolution was employed (see, for instance, (Durbin et al., 1998)), with a fixed transition-transversion ratio of $\tau = 2$. We set the length of each of the exterior branches to $w = 0.1$ and the length of the interior branch to $w = 0.2$. The four sequences were evolved along the interior branch, and then along the outer branches until 75% of the mutations had occurred. At that point, we simulated two recombination events in the way depicted in Figure 1. In the first, strains 2 and 3 exchanged the subsequences between sites $t = 201$ and $t = 400$. As seen from Figure 3, this corresponds to a transition in the phylogenetic topology from state $s_t = 1$ to state $s_t = 2$. In the second recombination event, the subsequence between sites $t = 601 - 800$ was transferred from strain 2 to strain 4 (and vice versa). This changes the topology from state $s_t = 1$ to state $s_t = 3$. The sequences then continued to evolve along the exterior branches for the remaining 25% of their lengths. This simulates a realistic evolutionary scenario, in which a recombination occurred at some point in the past, followed by further mutations.

Detection of recombination without HMMs

In an initial study, we tested the performance of a naive classifier that assumes a uniform prior on the sequence of topologies: $P(\mathbf{s}) = C$. The branch lengths of the three possible

trees (see Figure 3) were optimized with maximum likelihood on the whole data set, and we computed the likelihoods $P(\mathbf{y}_t|s_t)$ for all sites $t = 1, \dots, N$ and all topologies $s_t \in \{1, 2, 3\}$. We then calculated the posterior probabilities for the different topologies according to (4). Each site t was assigned to the topology s_t at the mode, that is, the topology that maximizes $P(s_t|\mathbf{y}_t)$.

Figure 4 approximately here

Figure 4 shows a (smoothed) plot of the posterior probability $P(s_t|\mathbf{y}_t)$ against the site t in the multiple alignment. The signal is obviously very noisy, and the resulting classification performance, depicted at the bottom of Figure 4, is rather poor. This is a result of the poor prior, $P(\mathbf{s}) = C$, which does not take correlations between adjacent sites into account. The application of an HMM redeems this deficiency.

HMM, parameter optimization with CML

In the second part of the study, we applied an HMM to the detection of recombination, as discussed in Section 2. The parameters of the HMM were estimated as in (McGuire, 1998), that is, the recombination parameter was kept fixed, while the branch lengths were optimized, for each of the three topologies in turn, by maximizing the constrained log likelihood L_c [the method of *constrained maximum likelihood* (CML), see (13) on page 11]. Note that (McGuire, 1998) pointed out the sub-optimality of this scheme and tried maximizing the likelihood on a subset of the alignment (the heuristic *window* method, described in Section 1). However, in her simulation experiments she found ((McGuire, 1998), p.102) that for the synthetic data this did not give any improvement and that it seemed reasonable to use the whole data set to estimate the branch lengths. Since the simulation experiment of our study is similar to that performed in (McGuire, 1998), we also optimized the branch lengths on the basis of the entire data set. The

recombination parameter ν was kept fixed, as in (McGuire, 1998), with a choice of two different values: $\nu = 0.6$ and $\nu = 0.8$ (only the better results are reported).

Figure 5 approximately here

Figure 5, top, shows a plot of the posterior probability $P(s_t|\mathbf{Y})$ against the sites in the multiple sequence alignment, t , where the recombination zones (between nucleotides $t = 201 - 400$ and $t = 601 - 800$) are framed by vertical lines. Comparing this with Figure 4, we see that the application of an HMM leads to a graph that makes transitions into and out of the recombination zones much more noticeable. The bottom row in Figure 5 shows the classification scores obtained from the Viterbi path, that is, the joint mode of $P(s_1, \dots, s_N|\mathbf{Y})$. Again, a considerable improvement is found over the method of the previous section.

HMM, novel parameter optimization scheme

Finally we tested the novel training algorithm of Section 3, by which all parameters are optimized in a proper maximum likelihood sense. Since this is effected with the EM algorithm, we will henceforth refer to this scheme as the *EM method* (as opposed to the *CML method* tested in the previous section).

Figure 6 approximately here

Figure 6 shows, again, a plot of the posterior probability $P(s_t|\mathbf{Y})$ against the site in the multiple sequence alignment (top), and three histograms for the classification scores in the different regions (no recombination, first recombination event, and second recombination event). A comparison with Figure 5 suggests that the novel training scheme does not only achieve a slight improvement in the classification performance, but that it

also leads to transitions in the posterior probability $P(s_t|\mathbf{Y})$ that are considerably more pronounced. This is reflected by the relative classification entropy,

$$H = -\frac{1}{N \ln K} \sum_{t=1}^N \sum_{s_t=1}^K P(s_t|\mathbf{Y}) \ln P(s_t|\mathbf{Y}) \quad (30)$$

which measures the uncertainty of the classifier and ranges from 0 (perfect prediction) to 1 (random classification). Applying this measure to the graphs of $P(s_t|\mathbf{Y})$ in Figures 5 and 6, we find that the earlier approach (CML) leads to a comparatively high value of $H = 0.82$, while the novel training method (EM) reduces this classification uncertainty down to $H = 0.06$.

Figures 7 approximately here

We are finally interested in how well the true phylogenetic trees are approximated with the two training methods. The top row in Figure 7 shows the true trees, which correspond to the three states $s_t = 1$ (predominant topology), $s_t = 2$ (first recombination), and $s_t = 3$ (second recombination). The second row shows the predictions obtained with CML (Section 4). Obviously, the exterior branches are too long, while, for the recombinant topologies, the interior branch suffers from a considerable contraction. This deficiency is redeemed when applying the novel training scheme EM (bottom row of Figures 7), which leads to a much closer agreement with the true trees. This improvement in the tree estimation of EM over CML is confirmed in Table 1, which shows the root-mean-square deviation between the branch lengths of the correct and the estimated tree:

$$\delta w = \sqrt{\frac{1}{M}(\hat{\mathbf{w}} - \mathbf{w}_o) \cdot (\hat{\mathbf{w}} - \mathbf{w}_o)} \quad (31)$$

where \mathbf{w}_o denotes the correct vector of branch lengths, $\hat{\mathbf{w}}$ is the vector resulting from the parameter estimation scheme, and $M = \dim(\mathbf{w}_o)$.

Table 1 approximately here

Significance of the results

In order to test whether the results of the previous section are significant, we simulated four different recombination scenarios at two different times on two phylogenetic trees. This gave a set of $4 \times 2 \times 2 = 16$ synthetic DNA sequence alignments. The phylogenetic trees are shown in Figure 8, and the recombination scenarios are described in the caption of Table 2.

Figure 8 approximately here

Table 2 approximately here

On each DNA sequence alignment, we compared the earlier parameter adaptation scheme (CML) with the method proposed in the present paper (EM). The training simulations for CML were repeated with three different recombination parameters, $\nu = 0.7, 0.8, 0.9$, and the best results were recorded for the comparison.

We estimated the prediction performance with two different scores. The *sensitivity* is the probability for correctly identifying a recombination site. The *specificity* is the probability for correctly identifying a non-recombinant site.

Figure 9 approximately here

Figure 9 shows scatter plots of the two classification scores (top: sensitivity, bottom: specificity) obtained with CML against those obtained with EM. This suggests that EM gives a better fit to the data than CML, which was confirmed with a matched *t-test* (for which the null hypothesis of equal performance was rejected at a 95% significance level in both cases).

Differentiation between rate variation and recombination

To test whether our approach is able to suppress false positives and to distinguish between recombination and rate heterogeneity, we applied it to another series of synthetic sequence alignments, whose characteristics are depicted in the top of Figure 10. Otherwise, the data were generated in the way described at the beginning of Section 4. The total length of the alignment is 1200 bases, and the dominant topology is that of State 1 in Figure 3, $s_t = 1$.

Figure 10 approximately here

The *first* sequence alignment contains one recombinant region, corresponding to a transition from state $s_t = 1$ into state $s_t = 2$, while *no* part of the alignment was generated from a tree topology in state $s_t = 3$. A plot of the marginal posterior probabilities $P(s_t|\mathbf{Y})$, shown in the three subgraphs in the middle left of Figure 10, confirms that $P(s_t = 3|\mathbf{Y}) \approx 0$; thus false positives are successfully suppressed. The *second* sequence alignment is similar to the first but contains a differently diverged region, in which all the branch lengths of the underlying phylogenetic tree (in state $s_t = 1$) have been doubled. The three subgraphs in the middle right of Figure 10 show a plot of the marginal posterior probabilities $P(s_t|\mathbf{Y})$. They are similar to those of the first sequence alignment, with $P(s_t = 3|\mathbf{Y}) \approx 0$. This suggests that the model successfully differentiates between *recombination* and *rate variation*. For the *third* sequence alignment, the effect of rate variation is increased by multiplying the branch lengths of the phylogenetic tree by a factor of *three* rather than *two*. This now leads to an erroneous classification of this region as being in state $s_t = 3$, as seen from the three subgraphs in the bottom left of Figure 10, and the model misclassifies the *differently diverged* region as *recombinant*. An inspection of the tree corresponding to state $s_t = 3$ reveals a strong distortion, with long external branches and a contracted internal branch of length zero. This result is not surprising. Since the distribution of nucleotide column vectors in the differently di-

verged region deviates strongly from that of the rest of the alignment, employing a new state of the HMM gives an increase in the likelihood even though the new state itself is ill-matched to the data. Finally, the *fourth* sequence alignment differs from the third in that it contains *two* recombinant regions, corresponding to different states, $s_t = 2$ and $s_t = 3$. The presence of a region whose true nucleotide distribution is in state $s_t = 3$ should make it easier for the model to avoid a misclassification of the differently diverged region. This is in fact borne out in the simulation, as seen from the three graphs in the bottom right of Figure 10.

The overall conclusion of this study is that the model can deal with moderate rate variation, but fails to distinguish between rate variation and recombination if the changes in the branch lengths become too pronounced. This will be discussed again in Section 6.

5 Findings 2: Recombination in *Neisseria*

In the second simulation experiment, we applied our model to a real DNA alignment with evidence of a likely recombination event. The data used was a subset of the *Neisseria argF* DNA multiple alignment, studied by (Zhou and Spratt, 1992). We selected four strains of *Neisseria*: (1) *N. gonorrhoeae*, (2) *N. meningitidis*, (3) *N. cinerea*, and (4) *N. mucosa*.⁵ The alignment of these sequences was carried out using CLUSTAL W (Thompson et al., 1994) with the default parameter settings. Discarding columns with gaps, this leads to an alignment of $N = 787$ base pairs. We used the Kimura 2-parameter model for simplicity, as the base composition of this data set was not highly skewed from uniform usage. The transition-transversion ratio was set to $\tau = 2.3$, as estimated in (McGuire, 1998).

(Zhou and Spratt, 1992) found two anomalous regions in the *Neisseria argF* DNA sequence alignment. The two regimes occur at positions $t = 1 - 202$ (region *R1*) and $t = 507 - 538$ (region *R2*). In the rest of the sequence (region *R0*), *N.gonorrhoeae* clusters with *N.meningitidis* ($s = 1$) while in *R1* they found that it is grouped with *N.mucosa* ($s = 3$). The authors were unable to determine the cause of region *R2* but suggested that this segment might have evolved at a different rate (rate heterogeneity).

We performed 12 simulations with both the CML and the EM methods, starting from 4 different initial branch-length vectors ($\mathbf{w} = [0.1 \ 0.1 \ 0.1 \ 0.1 \ 0.1]$, $[0.1 \ 0.2 \ 0.1 \ 0.2 \ 0.1]$, $[0.2 \ 0.1 \ 0.2 \ 0.1 \ 0.2]$, $[0.2 \ 0.2 \ 0.2 \ 0.2 \ 0.2]$) and 3 initial recombination parameters $\nu = 0.6, 0.75, 0.9$. Note that CML does not allow an optimization of ν , so we reported the best of the three results obtained.

⁵The strains have the GenBank/EMBL accession numbers X64860, X64866, X64869, X64873, respectively. Note that (Zhou and Spratt, 1992) used a different labelling scheme, with the first nucleotide at $t = 296$, and the last one at $t = 1082$.

Evaluation of the EM training scheme

To compare the performance of the two algorithms, Figure 11 shows a scatter plot of the normalized log likelihood scores obtained with the two methods, where the horizontal line represents CML and the vertical line EM. The dashed diagonal line indicates where the two approaches are equal. However, all simulations lead to entries that are located far above the diagonal line, which clearly demonstrates that EM outperforms CML.

Figure 11 approximately here

Figure 12 approximately here

To estimate the sensitivity of the parameter estimates on the initialization, Figure 12 depicts a spectrum of the recombination parameters ν obtained from the various training simulations. The different symbols refer to different initial values ν_0 , and the graph suggests that this value has a negligible influence on the final results. Moreover, this spectrum shows that the conclusion of (McGuire, 1998) and (McGuire et al., 2000) that a maximum likelihood approach to optimizing ν leads to a recombination parameter of $\nu = 1$ was incorrect. This value would make state changes impossible and therefore would prevent the detection of any recombination event. The simulations performed in the present study demonstrate that although ν reaches large values of typically $\nu > 0.98$, it always steers clear of 1: $\nu \neq 1$. The reason for this deviation is the fact that our approach makes use of a continuous parameter update scheme – see (21) – whereas (McGuire, 1998) and (McGuire et al., 2000) employed a discrete grid method whose resolution ε apparently was too low to find $1 - \varepsilon < \nu < 1$.

Maximum likelihood prediction

In order to compare with the results of (Zhou and Spratt, 1992), we selected the HMM with the highest likelihood score and computed the Viterbi path, that is, the mode of $P(s_1, \dots, s_N | \mathbf{Y})$. From this we determined the classification scores in the three regions $R0$, $R1$, $R2$, which are shown as histograms in the bottom row of Figure 13.

Figure 13 approximately here

It is clearly seen that most of the sites in $R0$ are classified as State 1, $s_t = 1$, while most of the sites in $R1$ are classified as State 3, $s_t = 3$. This supports the earlier findings by (Zhou and Spratt, 1992), mentioned above. In contrast to this earlier study, however, the HMM classifies most of the sites in $R2$ as being in State 2, $s_t = 2$, whereas (Zhou and Spratt, 1992) only found that this region was ‘irregular’. An inspection of the posterior probabilities $P(s_t | \mathbf{Y})$, plotted in the top row of Figure 13 (left: $P(s_t = 1 | \mathbf{Y})$, middle: $P(s_t = 2 | \mathbf{Y})$, right: $P(s_t = 3 | \mathbf{Y})$), shows that the classification is very crisp and the transitions between the regions are very distinct. The HMM thus successfully identifies the putative mosaic structure found by (Zhou and Spratt, 1992). There is a transition from $s_t = 1$ to $s_t = 3$ at the end of the alignment, which was not found by (Zhou and Spratt, 1992). The affected region, however, is very short, and is not picked up by the Viterbi path⁶.

CML prediction

Figure 14 shows that the results obtained with CML are considerably worse: the noise in the ‘signal’ $P(s_t | \mathbf{Y})$ has increased, and the agreement with (Zhou and Spratt, 1992)

⁶In this respect we found a small deviation between predictions based on the Viterbi path $P(s_1, \dots, s_N | \mathbf{Y})$ and those based on the single-site mode $P(s_t | \mathbf{Y})$.

is rather poor.

Figure 14 approximately here

(McGuire, 1998) and (McGuire et al., 2000) found that their *window method*, mentioned in Section 1, led to considerably better results, and they report an optimal window size of $\Delta t = 5$ for this data set. However, this optimal value did not result from any parameter optimization scheme based on the data \mathbf{Y} , but was chosen, with the benefit of hindsight, *after* inspecting the results. This procedure is methodologically dubious since the ‘optimal’ value for Δt could not have been found if the putative mosaic structure of the sequence alignment had not already been known beforehand.

Prediction with a committee of HMMs

As an alternative to the prediction with the maximum likelihood model, we studied the prediction with a committee of HMMs, where we applied the Bayesian weighting scheme (29), described in Section 3. The results are plotted in Figure 15.

Figure 15 approximately here

A comparison with the maximum likelihood prediction, Figure 13, reveals the following differences:

1. Region *R2* is still clearly classified as $s_t \neq 1$, that is, a recombination event is predicted. However, the committee of models is less certain about whether this topology corresponds to $s_t = 2$ or to $s_t = 3$.
2. Only the first part of region *R1*, $t = 1 - 77$, shows a *distinct* classification as being in state $s_t = 3$. Between $t = 78 - 202$, the probability $P(s_t = 3|\mathbf{Y})$ decays to smaller

values $P(s_t = 3|\mathbf{Y}) \approx 0.5$. Note, however, that both the maximum likelihood and the committee approach agree in predicting a sharp transition between sites $t = 202$ and $t = 203$, thereby clearly marking the recombinant zone.

3. The committee is less confident in its classification of the short region at the end of the alignment.

Figure 16 approximately here

Figure 16 shows a plot of the local classification entropy H_t ,

$$H_t = -\frac{1}{\ln K} \sum_{s_t=1}^K P(s_t|\mathbf{Y}) \ln P(s_t|\mathbf{Y}) \quad (32)$$

along the sequence alignment. This demonstrates that those regions for which the maximum likelihood and the committee predictions disagree are characterized by a high degree of classification uncertainty (expressed by a value of H_t close to 1).

Phylogenetic informativeness

We compare this classification uncertainty with a measure of variability and phylogenetic informativeness in the data, for which the number of topology-defining sites, N_{tds} , is a good candidate – that is, the number of those sites that are informative under a parsimony model. For instance, the nucleotide configurations AACC and AACG are topology-defining, supporting the topology $((1,2),(3,4))$, while AAAA, AAAG, and ACGT are not. More formally, let $n_i(t)$ denote the number of times nucleotide $i \in \{A, C, G, T\}$ is found at the t th site of the alignment, then N_{tds} is the number of sites that satisfy the condition $\max_i n_i(t) = 2$.

Table 3 approximately here

Table 3 shows the number of topology-defining sites in different regions of the alignment. It is seen that the high-entropy region between $t = 78$ and $t = 202$ (region *R1b*) is indeed characterized by a significantly low N_{tds} score; thus the high degree of uncertainty results from a lack of phylogenetic information in this part of the alignment. The high-entropy region *R2* (between $t = 507$ and $t = 538$), however, has an over-the-average N_{tds} score. This suggests that model-misspecification might be a more likely cause of uncertainty than low data informativeness. If region *R2* has diverged at a different rate, as suggested by (Zhou and Spratt, 1992), then our model might be misspecified, as discussed in the last subsection of Section 4. This would explain the pattern found in Figure 15, where region *R2* is recognized as being different ($P(s_t = 1|\mathbf{Y})$ is small), but the model cannot decide whether it is in state $s_t = 2$ or in state $s_t = 3$ (because none of these states offers a good model for a differently diverged region).

The committee prediction thus captures the inherent uncertainty, whereas the single maximum likelihood model tends to be over-confident.

Discussion: Significance of the results

An HMM is a probabilistic generative model. Rather than just predicting the location of the recombinant regions with the Viterbi path, it also predicts the significance of its prediction via the posterior probabilities for the states, $P(s_t|\mathbf{Y})$. Note that this significance estimation is not available in Hein’s parsimony algorithm (Hein, 1993), discussed in Section 2.

However, the previous subsections have indicated that the significance estimation of the maximum likelihood model is over-confident. This is because the probability for the state is not only conditional on the data, but also on the model parameters, $P(s_t|\mathbf{Y}, \mathbf{w}, \nu)$. The latter have been optimized with maximum likelihood and, consequently, the significance

estimation itself can be biased (overfitting).

A Bayesian way to overcome this deficiency is to marginalise over the model parameters; see (25). The required integral is usually analytically intractable and one has to resort to numerical methods, like Markov chain Monte Carlo (MCMC). In the present paper we have applied the committee approach as a simpler approximation to this integral. The results suggest that this gives a significant improvement over the maximum likelihood model: a large classification uncertainty is predicted for regions where the data informativeness is low or where the model is misspecified.

A remaining problem is that of model selection, that is, whether the employed 3-state HMM gives a significant improvement over a single phylogenetic tree. A Bayesian approach would have to integrate over different model orders with reversible jump MCMC, as in (Robert et al., 2000), but we have not implemented this approach. Instead, we follow a frequentist approach and use parametric bootstrapping.

We first estimate the maximum likelihood scores for the competing hypotheses: $L(H_0)$ and $L(H_1)$, where H_0 – the null hypothesis – is the assumption that the data can adequately be described with a single phylogenetic tree. We then generate B bootstrap replicas under the null hypothesis H_0 . For each of these replicas, $1 \leq i \leq B$, we estimate the maximum likelihood score, $L_i(H_0)$, and compute the p-value according to

$$p = \frac{1}{B} \left| \left\{ i \mid L_i(H_0) > L(H_1), 1 \leq i \leq B \right\} \right| \quad (33)$$

where $|\cdot|$ denotes cardinality. The values thus obtained are shown in Table 4.

Taking a standard critical region of $P < 0.05$ ⁷, the null hypothesis of a single tree has to be rejected in favour of the 3-state HMM. Note that for both 2-state HMMs, the

⁷Strictly speaking this value has to be reduced because of multiple testing, but in the present application this does not change the result.

improvement over the single tree is *not* significant.

Table 4 approximately here

In a further test, we investigated the various segments $\delta\mathbf{Y}$ of the alignment (listed in Table 5) separately and applied the method of (Shimodaira and Hasegawa, 1999) to test whether the data in these segments allow a significant discrimination between the tree topologies.

In detail, we determine the maximum-likelihood estimate of the branch lengths – for each of the *a priori* chosen regions $\delta\mathbf{Y}$ and each of the three tree topologies in turn – and compute the normalized log likelihoods $L_k^\circ = \frac{1}{|\delta\mathbf{Y}|} \log P(\delta\mathbf{Y}|\hat{\mathbf{w}}_k, s = k)$, $k = 1, 2, 3$, where $|\delta\mathbf{Y}|$ denotes the length of the fragment $\delta\mathbf{Y}$. We then repeat the simulations on B (non-parametric) bootstrap replicas ($B = 100$ in our study) of the respective sequences and test, for each of the subsets $\delta\mathbf{Y}$ and each of the topologies $k = 1, 2, 3$, the *null hypothesis* that $\Delta L_k^\circ = \max_i\{L_i^\circ\} - L_k^\circ = 0$. Only if the null hypothesis can be rejected for all but one topology does the subset $\delta\mathbf{Y}$ allow the identification of the ‘*correct*’ tree in the respective region. Otherwise, the information in $\delta\mathbf{Y}$ is not sufficient to discriminate between the different topologies.

Table 5 approximately here

A more detailed exposition of this test can be found in (Shimodaira and Hasegawa, 1999). The results are shown in Table 5 and are in accord with our earlier findings. Regions *R1b* ($t = 78 - 202$) and *R2* ($t = 507 - 538$), for which the HMM committee predicts a high degree of uncertainty, do not allow a discrimination between different topologies. Note that this does not imply that segment *R2* in itself is not significant, but rather that rate variation, as suggested by (Zhou and Spratt, 1992), is a more likely cause for the mosaic structure than recombination. The remaining segments allow a

significant classification of the topologies as either $s_t = 1$ or $s_t = 3$. This supports the prediction of the HMM committee and the earlier findings by (Zhou and Spratt, 1992). We note, however, that the test described here suffers from a certain selection bias in that the investigated segmentation is not chosen independently *a priori* but rather based on inspecting the HMM results.

6 Conclusions

The present article follows up on earlier work by (McGuire, 1998) and (McGuire et al., 2000), where a hidden Markov model was applied to detect recombinations in DNA multiple sequence alignments. The parameter optimization method of this earlier approach, however, left scope for significant improvements as (1) the recombination probability was not optimized, and (2) the branch lengths were estimated, for each possible phylogenetic tree separately, with *constrained maximum likelihood* – this keeps the tree topology fixed even in regions where it is incorrect and leads to systematic errors in the branch-length estimations. A heuristic remedy (the *window method*) was discussed in the earlier work, but also suffered from the absence of an optimization scheme for the newly introduced parameter (the window size). In the present study, by applying the EM algorithm, we could show how all parameters – the recombination probability and the branch lengths of all possible phylogenetic trees – can be optimized simultaneously in an *unconstrained maximum likelihood* sense. This (1) overcomes the systematic errors in the branch-length estimation and (2) leads to a straightforward update algorithm for the recombination probability. Unlike (McGuire, 1998) and (McGuire et al., 2000) we were able to estimate the recombination probability even when the true value was close to one. The novel training scheme is easy to implement as it incorporates well-established algorithms for standard HMMs. We tested the method on a synthetic benchmark problem, where we found (1) improved classification scores for recombinant and nonrecombinant regions, (2) a reduced uncertainty in the prediction, as indicated by a reduced classification entropy, and (3) more accurate branch-length estimations. On a real data set, where the location of the recombinant regions is unknown, we obtained an improved likelihood score.

Since the novel approach allows the computation of the likelihood of the total HMM, it also opens the way for a Bayesian model combination. While the maximum likeli-

hood model made accurate predictions about regions of known phylogenetic topology, it seemed to be over-confident in its classification of short stretches in the alignment whose topology, in fact, had not been identified. A Bayesian committee was more successful in capturing this uncertainty, in consistency with the Shimodaira-Hasegawa significance test and the estimated informativeness of the data.

The consequent extension of our work is to improve the Bayesian approach by development of a Markov chain Monte Carlo (MCMC) sampler. This would overcome the selection bias inherent in the committee approach and would ensure, at least in principle, the sampling from the correct posterior distribution. The present work has extended the probabilistic generativeness from the single-tree level to the combination of all trees in the HMM: given the recombination probability and the branch lengths in all possible trees, we can compute the probability of any column vector in the DNA sequence alignment. This enables the computation of the Metropolis-Hastings ratio and thus forms the basis for the construction of an MCMC sampler. Further elements required are sampling schemes for the parameters of the emission probabilities (branch lengths) and the transition probabilities (recombination parameter), which, however, have recently been developed – (Larget and Simon, 1999) and (Robert et al., 2000) – and should therefore be straightforward to implement.

Two limitations of the earlier work still apply to ours. Since each possible topology constitutes a separate state of the HMM, the feasibility of an exhaustive search in the tree space is an indispensable prerequisite for the applicability of the current scheme. This implies a limitation of our algorithm to alignments of small numbers of species. In practical applications, the HMM method presented here is therefore at best combined with a fast low-resolution preprocessing step that can analyze more taxa simultaneously. For example, one can proceed as in (Holmes et al., 1999) and conduct the initial search for recombination with split decomposition (Bandelt and Dress, 1992), a method that represents evolutionary relationships between sequences by a network if there are conflicting

phylogenetic signals in the data. Split decomposition itself does not allow individual recombination events to be identified nor the statistical support for them to be assessed. It is, however, a useful preprocessing step in that a network that strongly deviates from a bifurcating tree is suggestive of recombination and gives hints as to which sequences might belong to candidate recombinant strains. This can then be further investigated with the high-resolution method discussed in the present paper.

The second limitation is that the hidden states represent different tree topologies, but do not allow for different rates of evolution. The simulations in the last subsection of Section 4 suggest that this might not be a problem if the variations in the branch lengths are moderate. However, if a region has evolved at a drastically different rate, employing a new state for modelling this region might increase the likelihood even though the new state itself – representing a different (wrong) topology – is ill-matched to the data. Consequently, a *differently diverged* region might be erroneously classified as *recombinant*. A way to redeem this deficiency is to employ a *factorial hidden Markov model* (FHMM), as proposed in (Ghahramani and Jordan, 1997), and to introduce two separate 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). A detailed investigation of this idea is the subject of future research.

Acknowledgements

This work was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) Bioinformatics Initiative and the Scottish Executive Rural Affairs Department (SERAD). The simulations were carried out with programs written in MATLAB, which can be downloaded from http://www.bioss.sari.ac.uk/~dirk/My_software.html. We would like to thank Gráinne McGuire for stimulating discussions, and Jim McNicol and Rob Kempton for commenting on the manuscript.

References

- Bandelt, H. and Dress, A. W. M. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, 1:242–252.
- Battiti, R. and Colla, A. M. (1994). Democracy in Neural Nets: Voting Schemes for Classification. *Neural Networks*, 7:691–707.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, B39(1):1–38.
- Durbin, R., Eddy, S. R., Krogh, A., and Grisham, M. (1998). *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.
- Felsenstein, J. (1981). Evolution trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 17, 368-376, 17:368–376.
- Felsenstein, J. (1988). Phylogenies from Molecular Sequences: Inference and Reliability. *Annual Review of Genetics*, 22:521–565.
- Felsenstein, J. and Churchill, G. A. (1996). A Hidden Markov Model Approach to Variation Among Sites in Rate of Evolution. *Molecular Biology and Evolution*, 13(1):93–104.
- Ghahramani, Z. and Jordan, M. I. (1997). Factorial Hidden Markov Models. *Machine Learning*, 29:245–273.
- Grassly, N. C. and Holmes, E. C. (1997). A Likelihood Method for the Detection of Selection and Recombination Using Nucleotide Sequences. *Molecular Biology and Evolution*, 14(3):239–247.

- Hein, J. (1993). A Heuristic Method to Reconstruct the History of Sequences Subject to Recombination. *Journal of Molecular Evolution*, 36:396–405.
- Holmes, E. C., Worobey, M., and Rambaut, A. (1999). Phylogenetic Evidence for Recombination in Dengue Virus. *Molecular Biology and Evolution*, 16(3):405–409.
- Husmeier, D. (1999). *Neural Networks for Conditional Probability Estimation: Forecasting Beyond Point Predictions*. Perspectives in Neural Computing. Springer, London. ISBN 1-85233-095-3.
- Larget, B. and Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16(6):750–759.
- McGuire, G. (1998). *Statistical Methods for DNA Sequences: Detection of Recombination and Distance Estimation*. PhD thesis, University of Edinburgh.
- McGuire, G., Wright, F., and Prentice, M. (1997). A Graphical Method for Detecting Recombination in Phylogenetic Data Sets. *Molecular Biology and Evolution*, 14(11):1125–1131.
- McGuire, G., Wright, F., and Prentice, M. (2000). A Bayesian Method for Detecting Recombination in DNA Multiple Alignments. *Journal of Computational Biology*, 7(1/2):159–170.
- Neal, R. M. and Hinton, G. E. (1999). A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 355–368, Cambridge, MA. MIT Press.
- Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.
- Robert, C. P., Ryden, T., and Titterton, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B*, 62(1):57–75.

- Salminen, M. O., Carr, J. K., Burke, D. S., and McCutchan, F. E. (1995). Identification of Breakpoints in Intergenotypic Recombinants of HIV Type 1 by Bootscanning. *Aids Research and Human Retroviruses*, 11(11):1423–1425.
- Sankoff, D. and Cedergren, R. J. (1983). Simultaneous comparison of three or more sequences related by a tree. In Sankoff, D. and Kruskal, J. B., editors, *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*, pages 253–264. Addison-Wesley.
- Shimodaira, H. and Hasegawa, M. (1999). Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular Biology and Evolution*, 16(8):1114–1116.
- Smith, J. M. (1992). Analyzing the Mosaic Structure of Genes. *Journal of Molecular Evolution*, 34:126–129.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680.
- Zhou, J. and Spratt, B. G. (1992). Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Molecular Microbiology*, 6:2135–2146.

Training method	State 1	State 2	State 3
CML	0.13	0.23	0.22
EM	0.04	0.06	0.04

Table 1: Root-mean-square deviation δw between the branch lengths of the true and the estimated phylogenetic tree.

Recombination Scenario	Rec-A	Rec-B	Rec-C	Rec-D
1st recombinant region:				
Beginning	201	201	201	201
Length	200	100	100	50
2nd recombinant region:				
Beginning	601	601	801	651
Length	200	300	100	250

Table 2: Overview of the recombination scenarios for the synthetic DNA sequence alignments studied in the text. We generated an alignment of 1000 base pairs from the phylogenetic trees depicted in Figure 8 and simulated two recombination events. In the first event, strains 2 and 3 exchanged a DNA subsequence of the indicated length, which corresponds to a state transition $s = 1 \rightarrow 2$. In the second event, strains 2 and 4 exchanged genetic material, corresponding to a transition $s = 1 \rightarrow 3$. After the recombination events we simulated the evolution process along the external branches for further $(100 - \rho)\%$ of the branch lengths. The simulations were repeated twice with different random number generator seeds and different values for ρ : $\rho = 70, 80$.

Region	1-787	1-77	78-202	507-538
N_{tds}	9.8	14.3	1.6	18.8

Table 3: Number of topology-defining sites N_{tds} (in percent) for different sub-regions of the *Neisseria* DNA sequence alignment.

Null hypothesis H_0	1 state: $s_t = 1$	1 state: $s_t = 1$	1 state: $s_t = 1$
Alternative hypothesis H_1	2 states: $s_t = 1 \vee 2$	2 states: $s_t = 1 \vee 3$	3 states
p-value	0.12	0.08	0.01

Table 4: Hypothesis testing with parametric bootstrapping. The table shows the p-values obtained from 100 bootstrap replicas.

Region	Base pairs	L_1	L_2	L_3	$p(s_1)$	$p(s_2)$	$p(s_3)$	H_1	H_2	H_3
R0	203-506	-2.81	-3.42	-3.53	0.82	0.03	0.02	+	-	-
	539-787									
R1	1-202	-2.95	-2.97	-2.91	0.05	0.01	0.59	-	-	+
R2	507-538	-3.49	-3.53	-3.49	0.72	0.42	0.55	+	+	+
R1a	1-77	-3.22	-3.25	-3.10	0.03	0.00	0.59	-	-	+
R1b	78-202	-2.79	-2.79	-2.80	0.74	0.56	0.26	+	+	+

Table 5: Shimodaira-Hasegawa test applied to the *Neisseria* data. Define $T_k = \max_i \{L_i^\circ - L_k^\circ | i = 1, 2, 3\}$, where L_k° , shown in Columns 3-5, represents the normalized log likelihood for topology $s = k$. Let H_k be the null hypothesis that $E(T_k) = 0$, i.e., $s = k$ is the best model. The distribution of T_k under H_k is simulated with non-parametric bootstrapping. Columns 6-8 show the p -values of the tests of H_k , Columns 9-11 indicate which hypotheses are to be rejected (at a 95 % significance level, where '-' means that H_k is to be rejected and, conversely, '+' means that H_k is to be accepted). Note that for two regions (R1b and R2) none of the hypotheses is rejected, indicating that there is not enough information in the data to discriminate between the models.

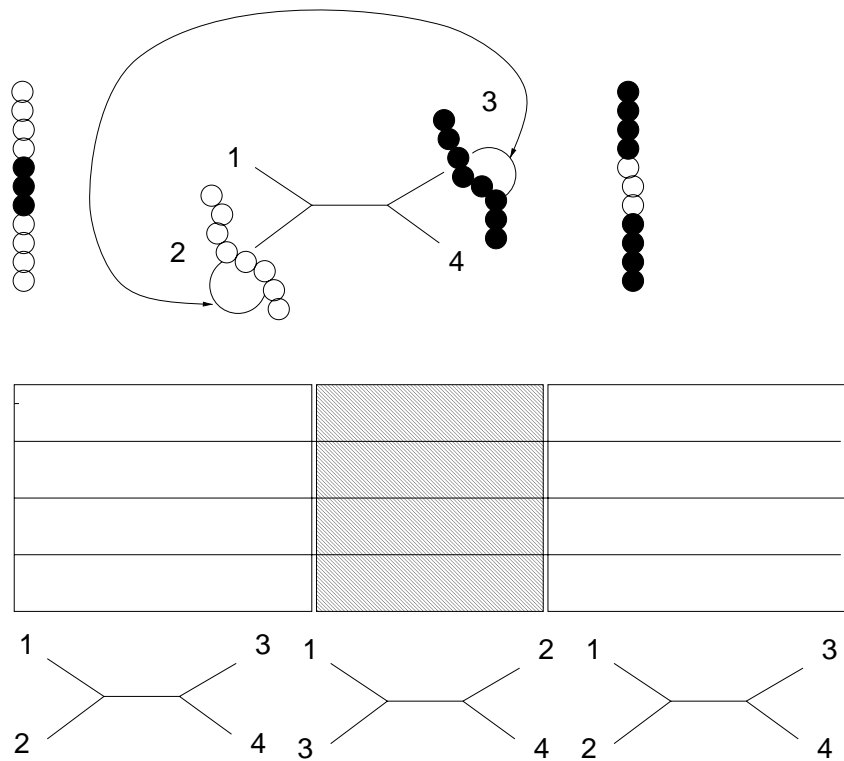


Figure 1: Influence of recombination on phylogenetic inference. The figure shows a hypothetical phylogenetic tree of four strains. Recombination is the exchange of DNA subsequences between different strains (top diagram, middle), which results in two so-called mosaic sequences (top diagram, margins). The affected region in the multiple DNA sequence alignment (shown by the shaded area in the middle diagram) seems to originate from a different phylogenetic topology, in which two branches of the phylogenetic tree have been swapped (bottom diagram, where the numbers at the leaves represent the four strains).

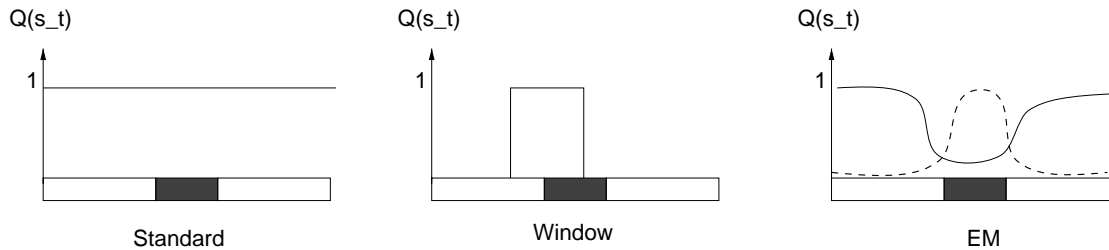


Figure 2: Three nucleotide weighting schemes for adapting the branch lengths. The bottom of each figure represents a multiple DNA sequence alignment with a recombinant zone, printed in grey, in the middle. *Left*: Conventional constrained maximum likelihood, where the tree parameters are optimized on the whole data set. This corresponds to constant weights $Q(s_t) = 1 \forall t$. *Middle*: Heuristic window method, suggested in (McGuire, 1998). *Right*: Unconstrained maximum likelihood with the EM algorithm. The dashed line shows the site-dependent weights $Q(s_t = T_R)$ for the recombinant topology T_R , the solid line represents the weights for the non-recombinant topology T_0 : $Q(s_t = T_0)$. Note that in this scheme the weights $Q(s_t)$ are updated automatically in every iteration of the algorithm as a natural consequence of the optimization procedure.

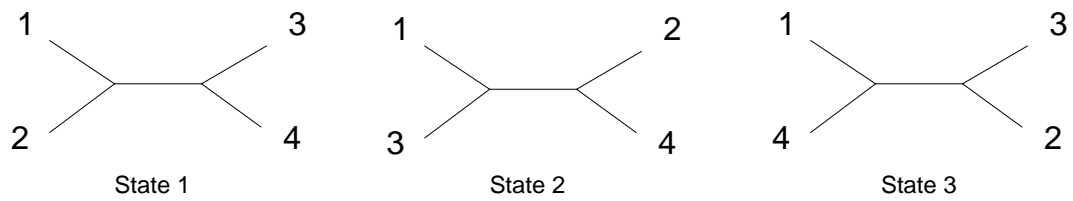


Figure 3: Possible unrooted tree topologies for a set of four DNA sequences. The numbers at the leaf nodes represent different strains. The numbers at the bottom indicate the corresponding state in the HMM.

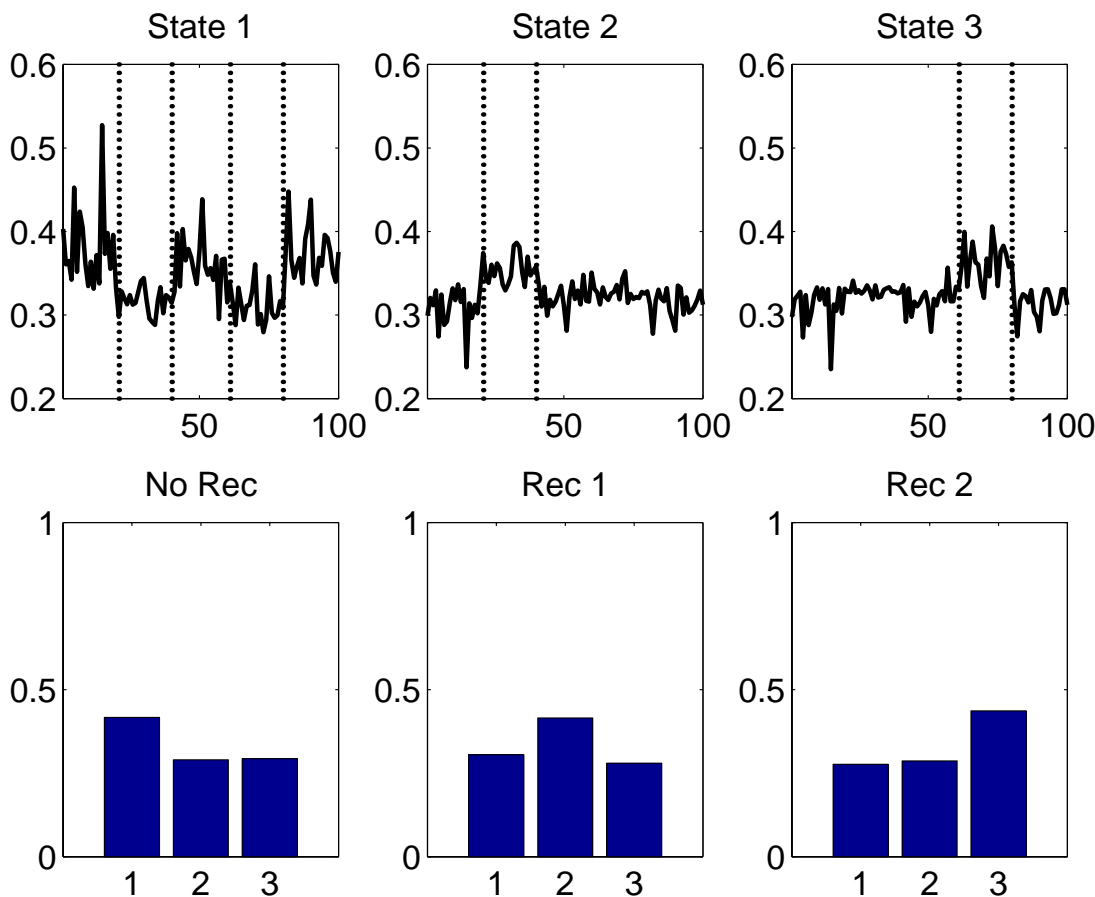


Figure 4: Predicting recombinations *without* HMMs. The figures in the top row plot the posterior probabilities $P(s_t|\mathbf{y}_t)$ for the three states $s_t \in \{1, 2, 3\}$ against the position t in the multiple alignment. States $s_t = 2$ and $s_t = 3$ represent the recombination events, whose locations are marked by the vertical lines. The signal was smoothed by taking the average over a fixed-size window of length 10, so each position on the horizontal axis represents 10 adjacent sites in the multiple alignment. The histograms in the bottom row show the classification scores for the three regions *no recombination* ('No Rec', $s_t = 1$), *1st recombination* ('Rec 1', $s_t = 2$), and *2nd recombination* ('Rec 2', $s_t = 3$), where the classification scheme assigned each site t to the state that maximizes $P(s_t|\mathbf{y}_t)$.

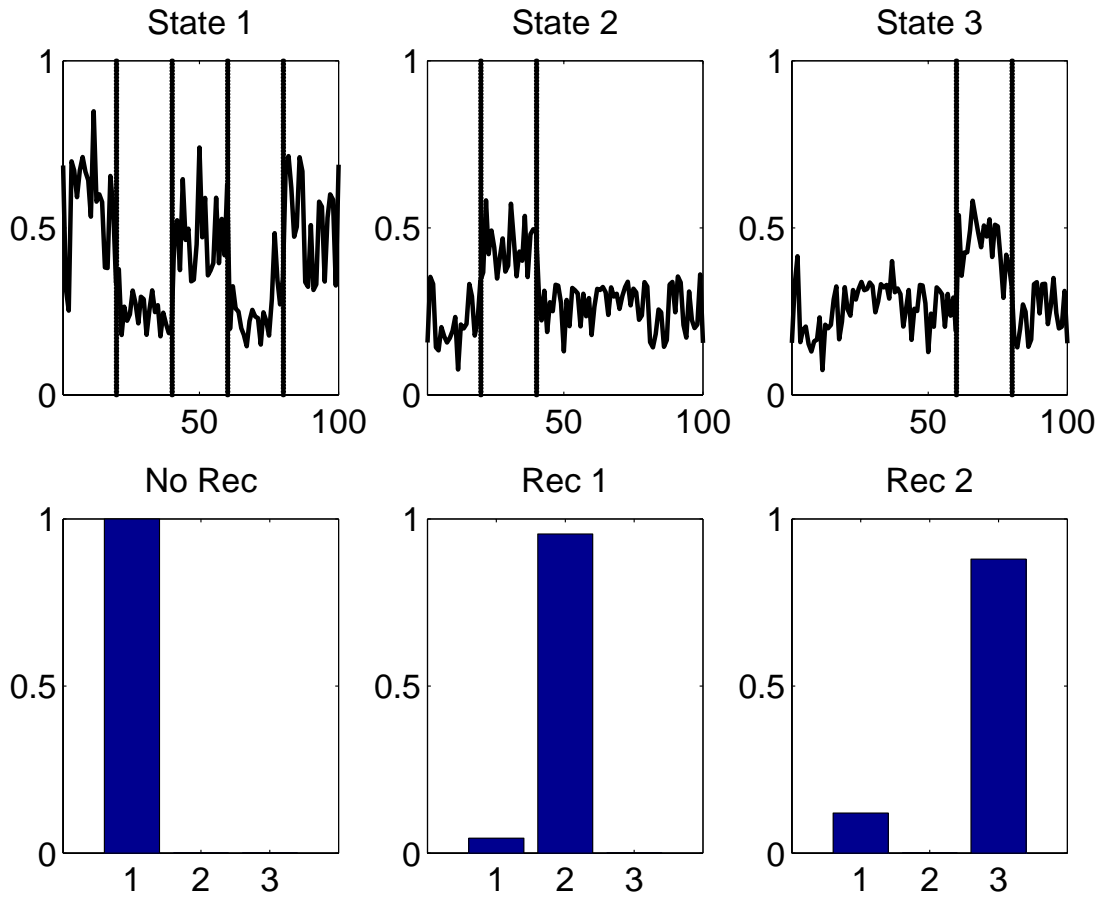


Figure 5: Predicting recombinations with an HMM – parameter optimization with CML. The branch lengths of the phylogenetic trees were optimized with constrained maximum likelihood (see text). The recombination parameter was kept at a fixed value of $\nu = 0.8$. An explanation of the curves is given in the caption of Figure 4. The histograms at the bottom show the classification scores obtained from the Viterbi path, that is, the mode of $P(s_1, \dots, s_N | \mathbf{Y})$.

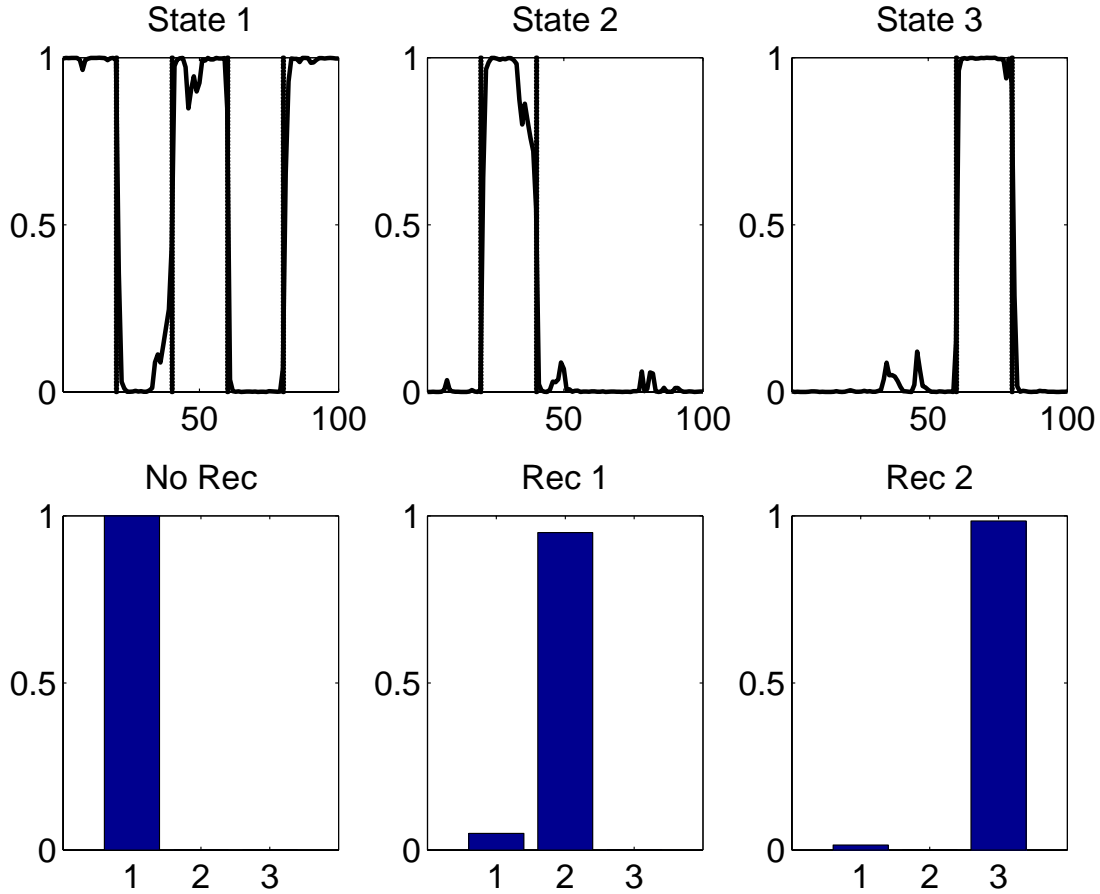


Figure 6: Predicting recombinations with an HMM – parameter optimization with EM. The branch lengths and the recombination parameter were optimized with the novel training algorithm described in Section 3. An explanation of the curves is given in the caption of Figure 4. The histograms at the bottom show the classification scores obtained from the Viterbi path, that is, the mode of $P(s_1, \dots, s_N | \mathbf{Y})$.

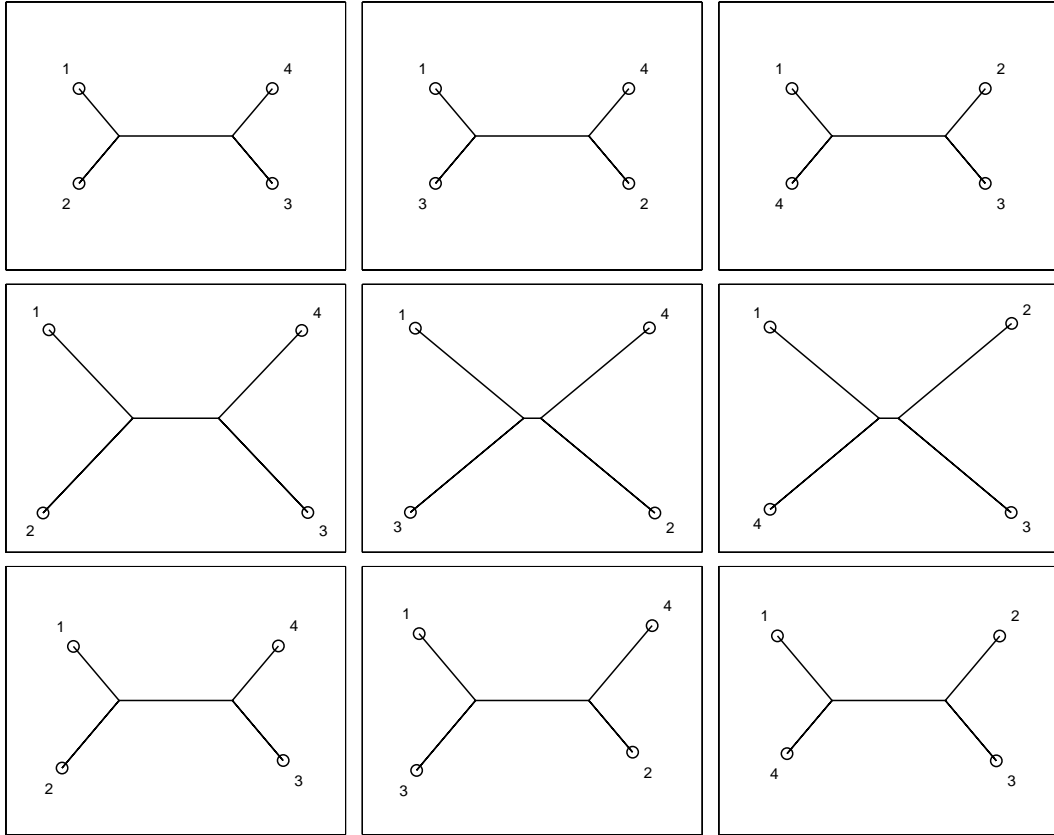


Figure 7: Comparison of two tree estimation methods. *Top row:* True phylogenetic trees. *Second row:* Trees obtained with CML (Section 4). *Third row:* Trees obtained with EM (Section 4). The columns represent the different topologies. *Left:* Predominant topology, $s_t = 1$. *Middle:* Topology of the first recombination, $s_t = 2$. *Right:* Topology of the second recombination, $s_t = 3$.

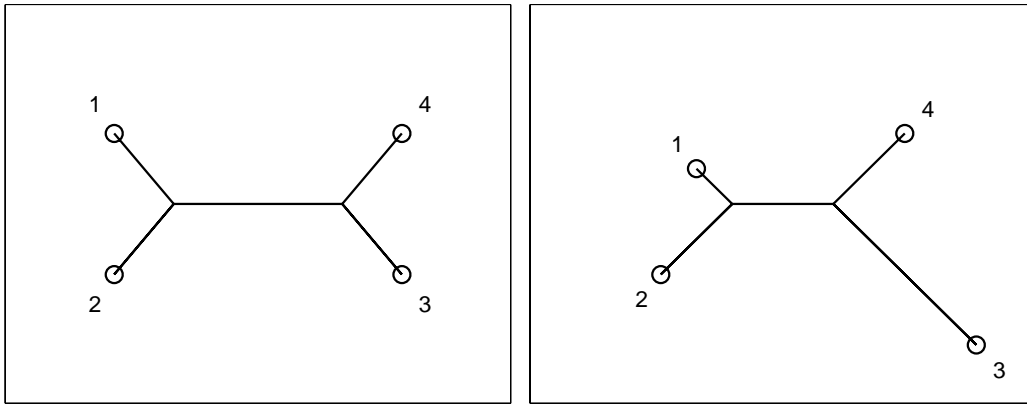


Figure 8: Phylogenetic trees used for generating the synthetic data. The vectors of branch lengths are: *left*: $\mathbf{w}_1 = (0.1 \ 0.1 \ 0.1 \ 0.1 \ 0.2)$, *right*: $\mathbf{w}_3 = (0.05 \ 0.1 \ 0.2 \ 0.1 \ 0.1)$. The first 4 elements of the vectors refer to the exterior branches (anti-clockwise, starting from the top-left node), while the last element represents the interior branch.

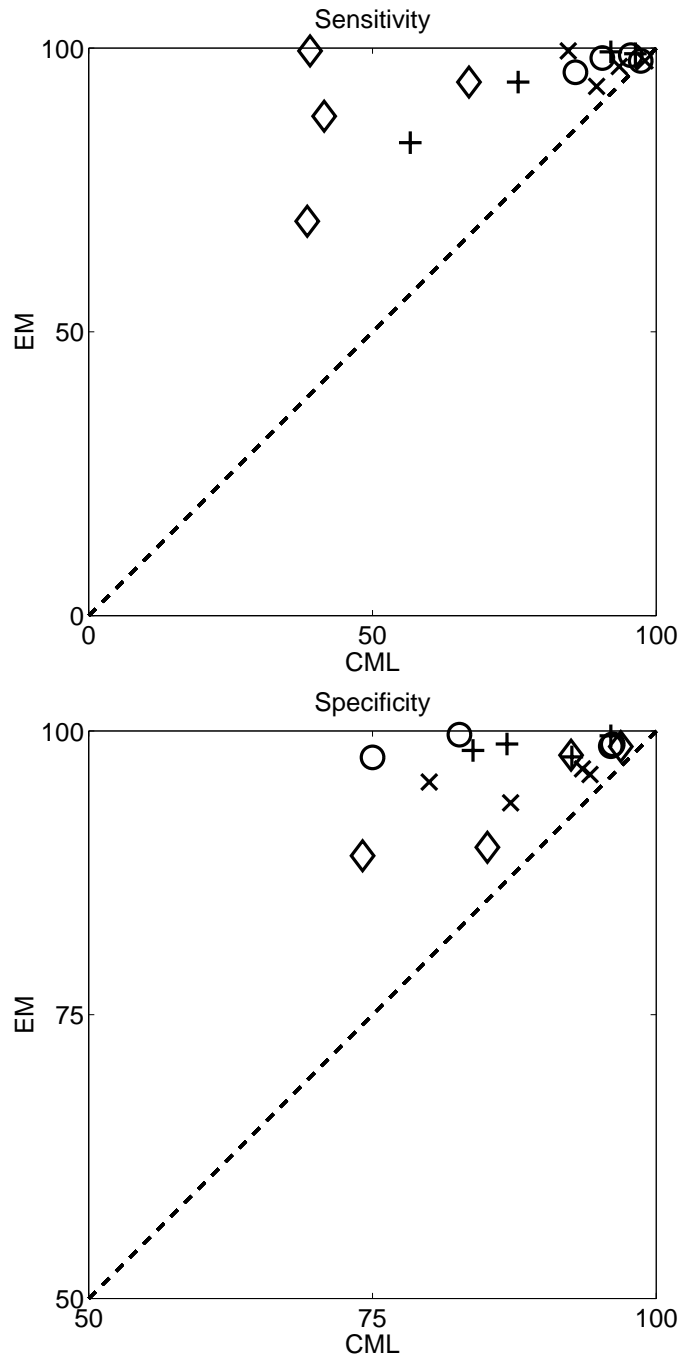


Figure 9: Scatter plots of the classification scores obtained with CML (horizontal axis) and EM (vertical axis). *Top:* Sensivities. *Bottom:* Specificities. The diagonal line indicates an equal performance, for symbols above this line EM outperforms CML. The symbols represent different recombination events, described in the caption of Table 2. x: Rec-A, o: Rec-B, ◇: Rec-C, +: Rec-D.

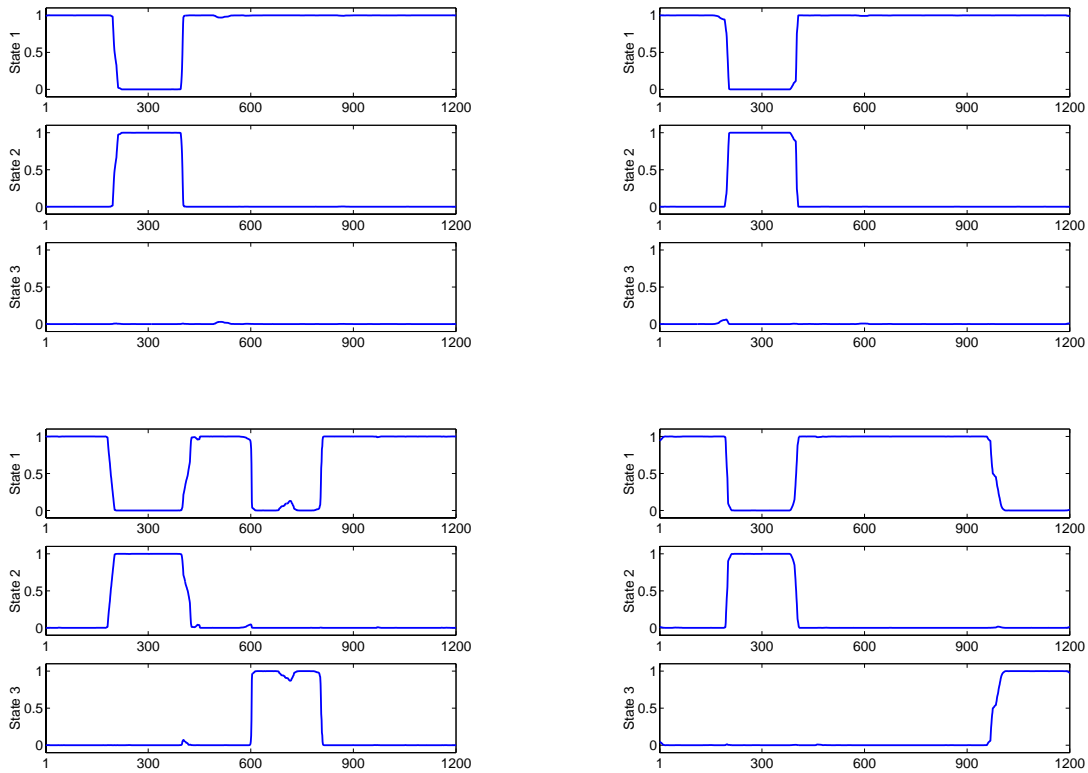
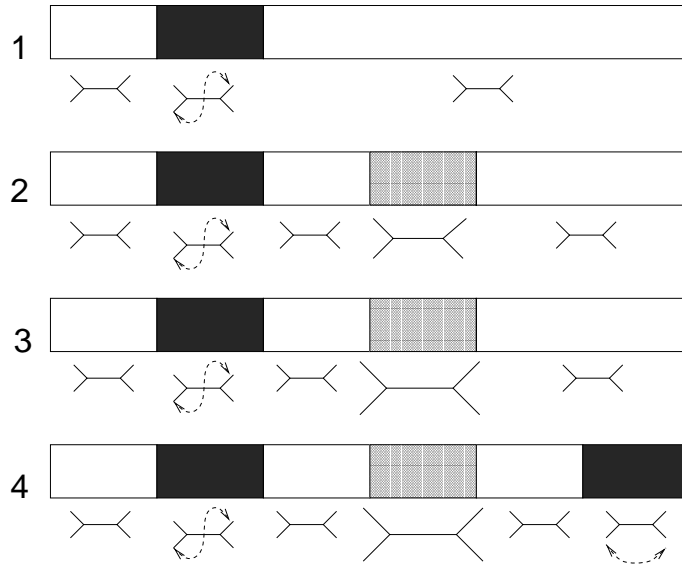


Figure 10: Differentiating between recombination and rate heterogeneity. The figure in the *top* shows four synthetic sequence alignments, for which the marginal posterior state probabilities $P(s_t|\mathbf{Y})$ are shown in the *middle* and *bottom* figures. Each of these figures contains three graphs: $P(s_t = 1|\mathbf{Y})$ (top), $P(s_t = 2|\mathbf{Y})$ (middle), and $P(s_t = 3|\mathbf{Y})$ (bottom). *Sequence 1, middle left*: One recombination event. *Sequence 2, middle right*: One recombinant region and a differently diverged region, factor 2. *Sequence 3, bottom left*: One recombinant region and a strongly differently diverged region, factor 3. *Sequence 4, bottom right*: Two recombinant regions and a strongly differently diverged region, factor 3. The total length of the alignment is 1200 bases.

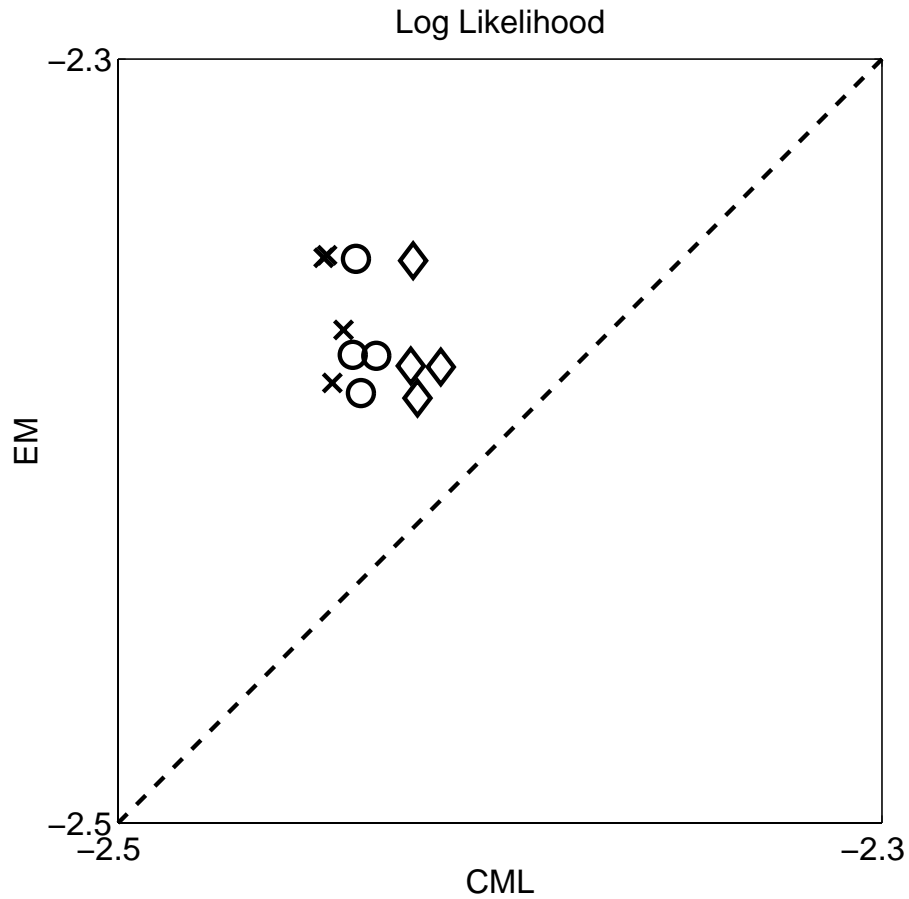


Figure 11: Scatter plot of the normalized log likelihood L° obtained on the *Neisseria* data. *Horizontal axis*: CML. *Vertical axis*: EM. The diagonal line indicates an equal performance of the two methods, for entries above this line EM is superior. The symbols indicate different initial values for the recombination parameter ν ; x: $\nu_0 = 0.6$; o : $\nu_0 = 0.75$, \diamond : $\nu_0 = 0.9$. Note the strong dependence of CML on ν_0 , which follows from the fact that this parameter is not adapted by the training algorithm.

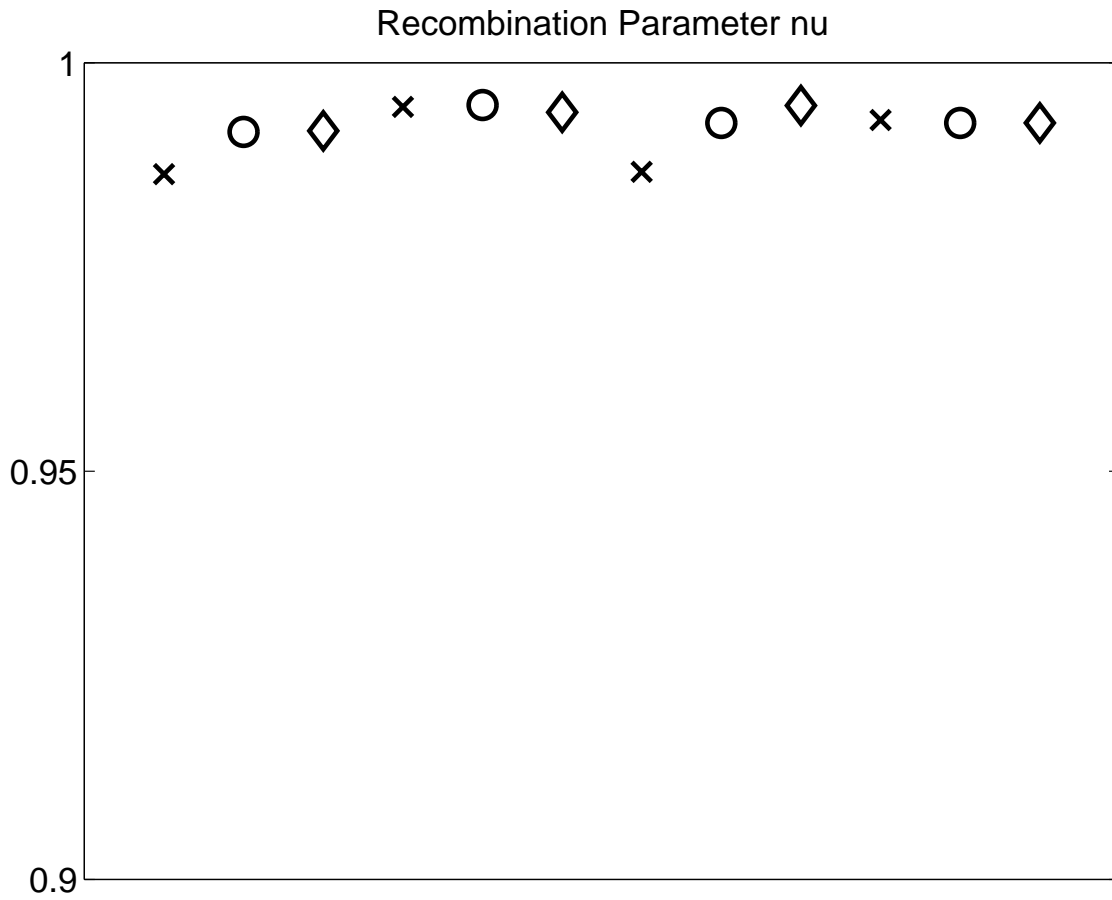


Figure 12: Spectrum of recombination parameters ν obtained from different training simulations. The symbols indicate the initial value for ν . x : $\nu_0 = 0.6$; o : $\nu_0 = 0.75$, \diamond : $\nu_0 = 0.9$.

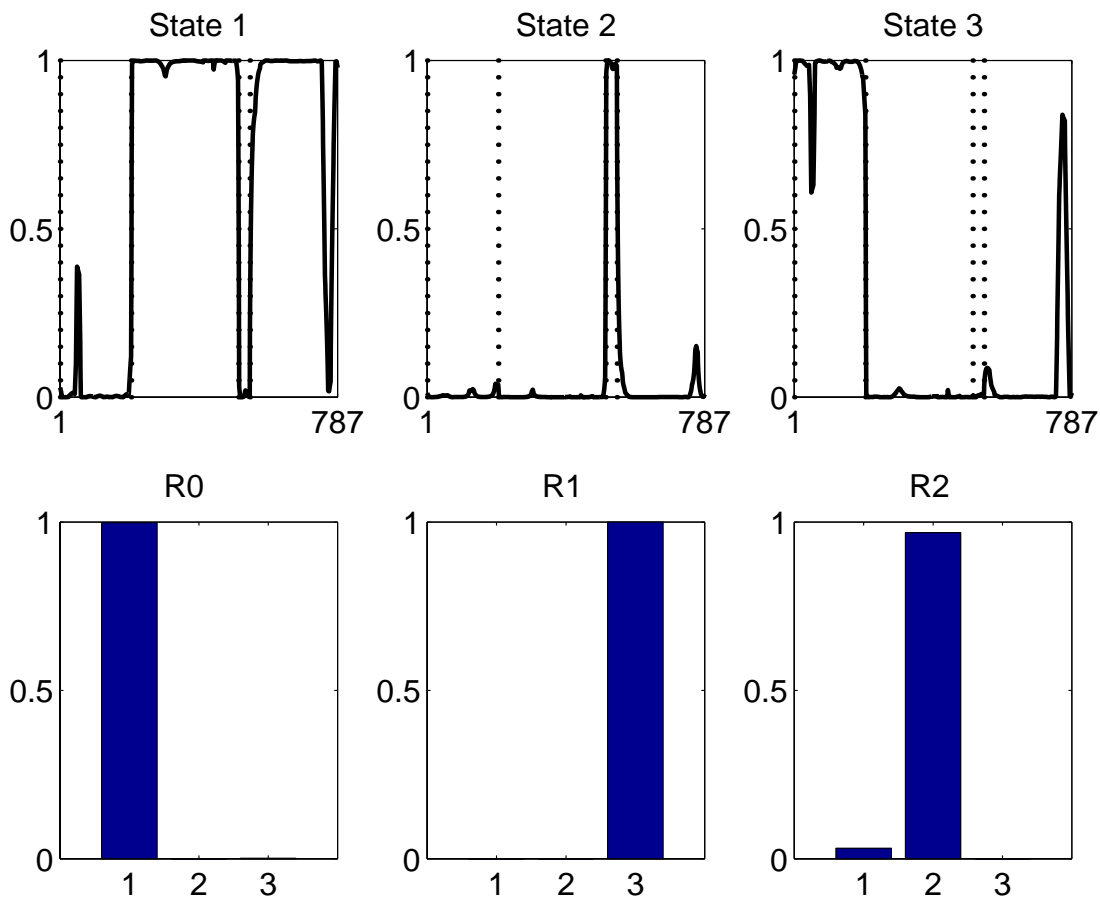


Figure 13: Prediction of recombinant regions in the *Neisseria* data with a single HMM. The *top row* shows a plot of the posterior probabilities $P(s_t|\mathbf{Y})$ along the sequence alignment, where s_t represents one of the three tree topologies $s_t = 1$ (*left graph*), $s_t = 2$ (*middle graph*), and $s_t = 3$ (*right graph*). The vertical lines indicate candidate regions for recombinations. The histograms in the *bottom row* show the classification scores in the following regions. *Left*: R0 ($t = 203 - 506, 539 - 787$), classified as $s_t = 1$ in earlier work. *Middle*: R1 ($t = 1 - 202$), believed to result from a recombination equivalent to a transition into $s_t = 3$. *Right*: R2 ($t = 507 - 538$), an irregular region not classified before.

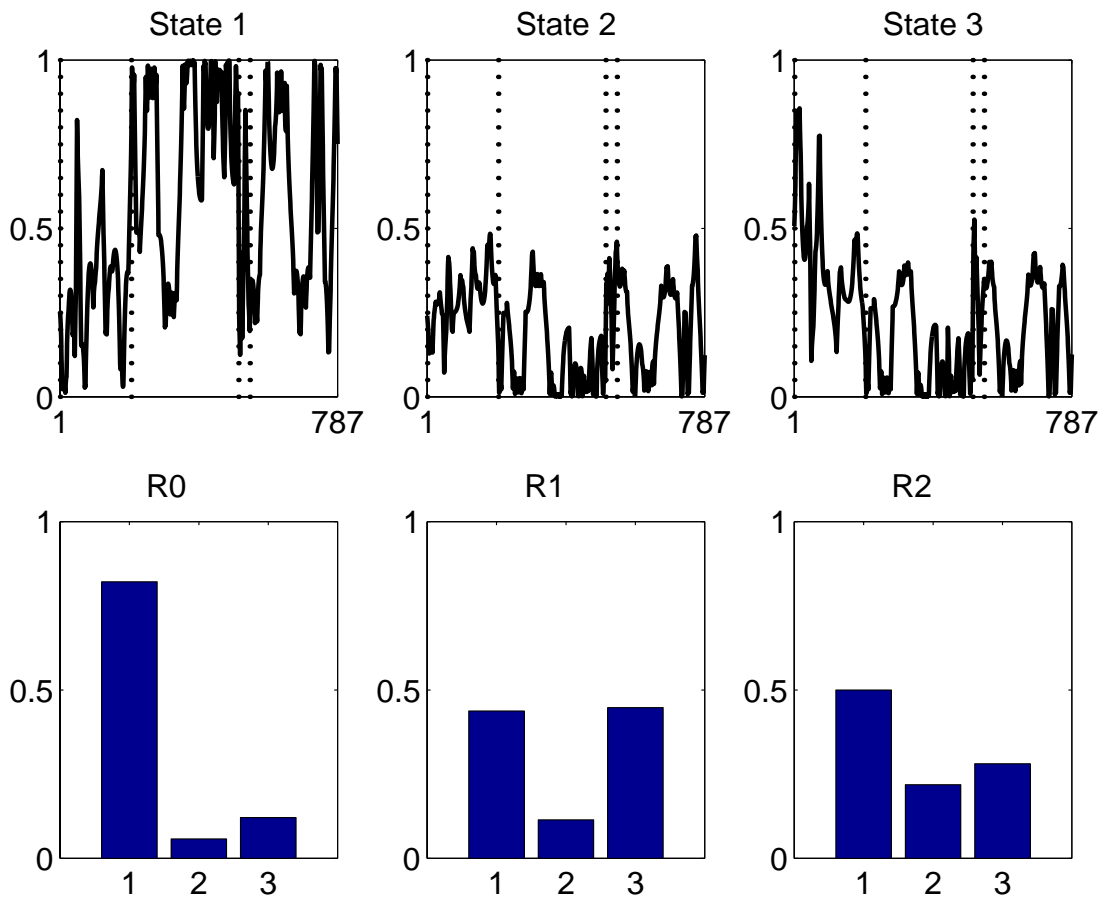


Figure 14: Prediction of recombinant regions in the *Neisseria* data with an HMM trained with CML. Note the increase in the noise of the ‘signal’ $P(s_t|\mathbf{Y})$. For details, see the caption of Figure 13.

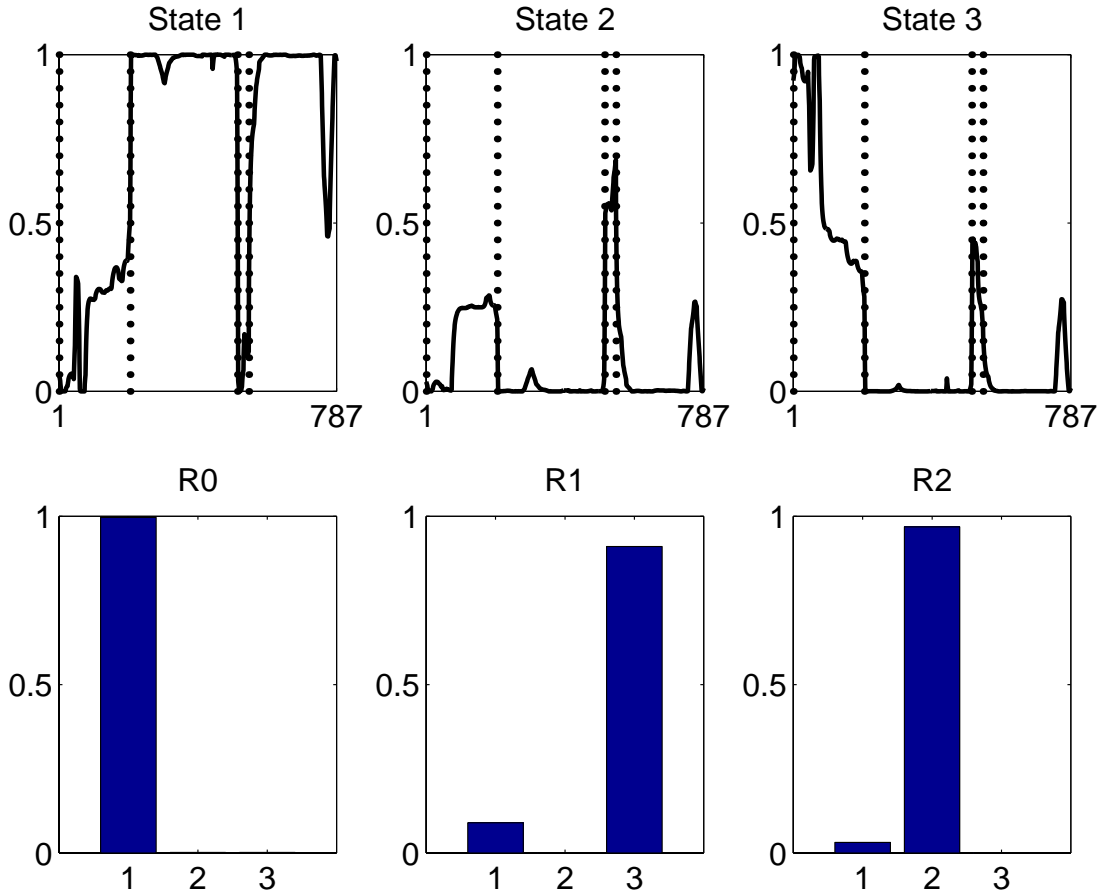


Figure 15: Prediction of recombinant regions in the *Neisseria* data with a committee of HMMs. The graphs are similar to those of Figure 13, except that the predictions are based on a committee of HMMs. Note the decrease of $P(s_t = 3|\mathbf{Y})$ in the right part of region R1 ($t = 1-202$), which is consistent with the significance test reported in the text. Also, note the increased uncertainty in the classification of region R2 ($t = 507 - 538$). For details, see the caption of Figure 13.

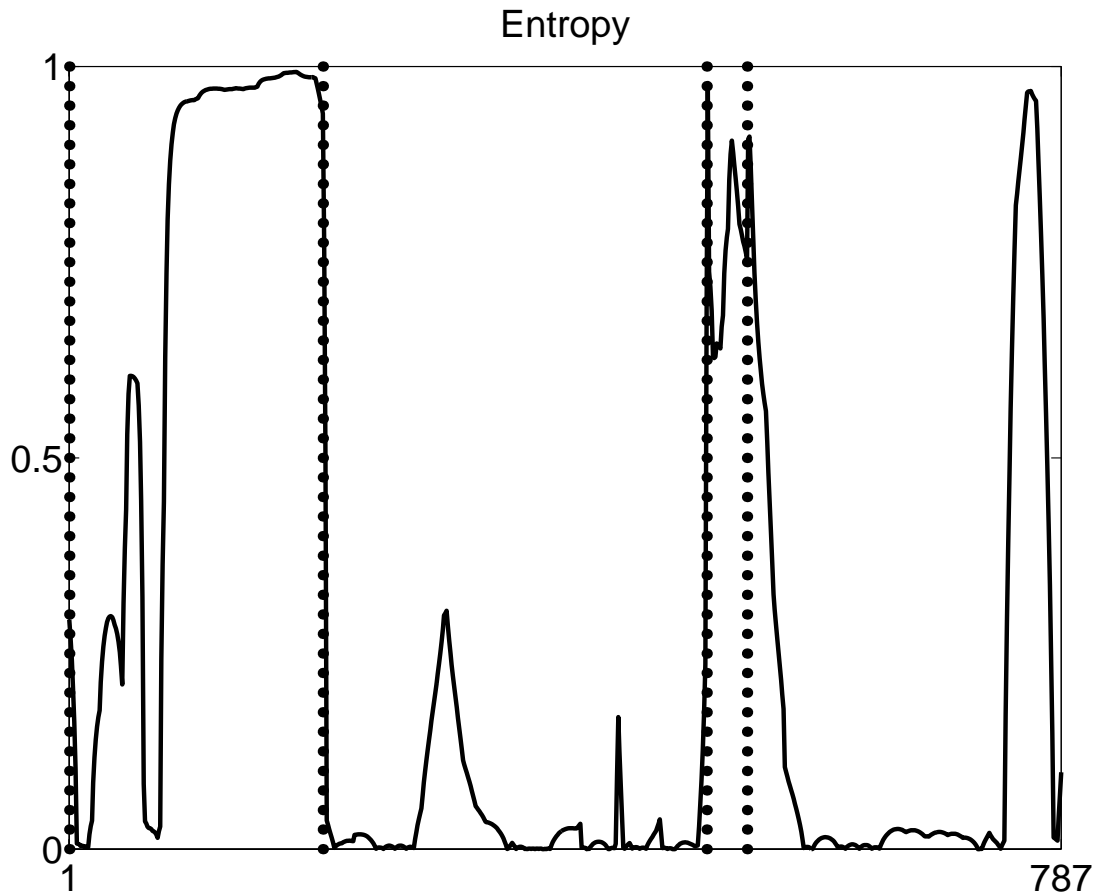


Figure 16: Classification entropy H_t , defined in (32), plotted along the DNA sequence alignment. The uncertainty of the classifier increases with increasing values of H_t . The vertical dotted lines mark candidate regions for recombination (R1: $t = 1 - 202$, R2: $t = 507 - 538$). Note the increase in uncertainty at the end of zone R1 and in zone R2.