Detecting Recombination in 4-Taxa DNA Sequence Alignments with Bayesian Hidden Markov Models and Markov Chain Monte Carlo

Dirk Husmeier and Gráinne McGuire†*

*Biomathematics and Statistics Scotland (BioSS), JCMB, King's Buildings, Edinburgh, United Kingdom; and †Department of Applied Statistics, University of Reading, Reading, United Kingdom

This article presents a statistical method for detecting recombination 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, 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. The proposed method is tested on various synthetic and real-world DNA sequence alignments, and we compare its performance with the established detection methods RECPARS, PLATO, and TOPAL, as well as with two alternative parameter estimation schemes.

Introduction

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

In the last few years, a plethora of methods for detecting recombination have been developed—following up on the seminal paper by Maynard 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 sequence alignment and computing a phylogenetic divergence score for each window position. Two well-established methods following this approach are PLATO and TOPAL.

PLATO (Grassly and Holmes, 1997) estimates a phylogenetic tree from the whole DNA sequence alignment, and then systematically looks for subsets with a low likelihood under this model by computing the statistic

$$Q = \frac{\frac{1}{W}\sum_{t=bW+1}^{(b+1)W} L_t}{\frac{1}{N-W}\left(\sum_{t=1}^{bW} L_t + \sum_{t=(b+1)W+1}^{N} L_t\right)} \qquad (1)$$

where L_t denotes the log likelihood of the tth column of the alignment, W is the size of the subset, and N is the length of the alignment (see figure 1, top). This measure is calculated for all possible positions b along the sequence alignment and for varying subset sizes, typically $5 \leqslant W \leqslant N/2$. Parametric bootstrapping is applied to generate the null distribution of the maximized Q value under the null hypothesis of no recombination. If the reference model, from which the log likelihoods are computed, were the true tree (meaning the tree one would get if no recombination event had happened), significantly large Q values would be a reliable indication for recombinant regions. However, the true tree is not known, and is approximated by a tree estimated from the whole sequence alignment. This alignment includes the recombinant regions, which perturb the parameter estimation for the reference tree (see fig. 1, bottom). Consequently, the method becomes increasingly unreliable as the recombinant regions grow in length.

TOPAL (McGuire, Wright, and Prentice 1997; McGuire and Wright 2000), illustrated in figure 2, replaces the global by a local reference tree. A window of typically 200–500 bases is slid along the DNA sequence alignment. The reference tree is estimated from the left half of the window, and used to computed a goodness-of-fit score for both parts of the window. The difference between these goodness-of-fit scores, the so-called DSS statistic, is likely to be small within a homogeneous part of the alignment, but large as the window is moved into a recombinant region. Parametric bootstrapping is applied to compute a distribution of DSS peaks under the null hypothesis of no recombination, and significantly large DSS peaks are indicators of putative recombinant breakpoints. While this method overcomes the principled shortcoming of PLATO, 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 follows up on earlier work 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 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, and (2)

Key words: phylogeny, DNA sequence alignment, recombination, hidden Markov models, Markov chain Monte Carlo.

E-mail: dirk@bioss.ac.uk.

Mol. Biol. Evol. 20(3):315–337. 2003
DOI: 10.1093/molbev/msg039

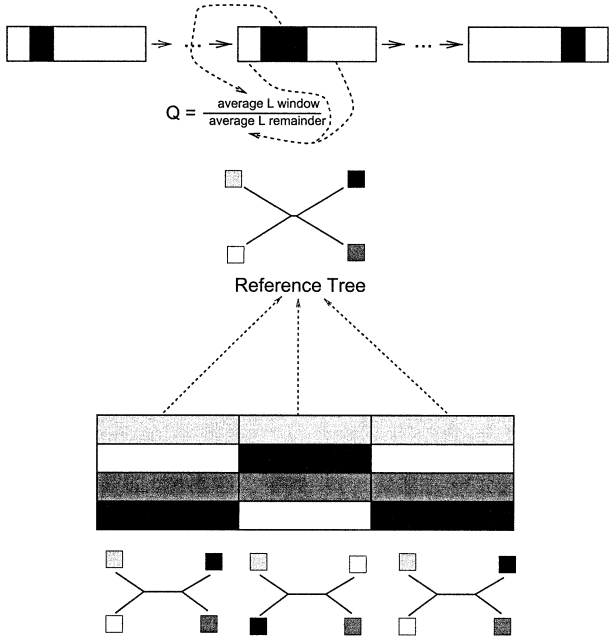

FIG. 1.—Illustration of PLATO. A window of varying size is moved along the DNA sequence alignment. The average log likelihood is computed for both the window and the remainder of the sequence, and the Q statistic is defined as the ratio of these values (*top*). If the reference model, from which the log likelihoods are computed, were the true tree (meaning the tree one would get if no recombination event had happened), large Q values would be a reliable indication for recombinant regions. However, the true tree is not known, and is approximated by a tree estimated from the whole sequence alignment. This includes the recombinant regions, which perturb the parameter estimation for the reference tree (*bottom*) and thus cause the test to lose power.

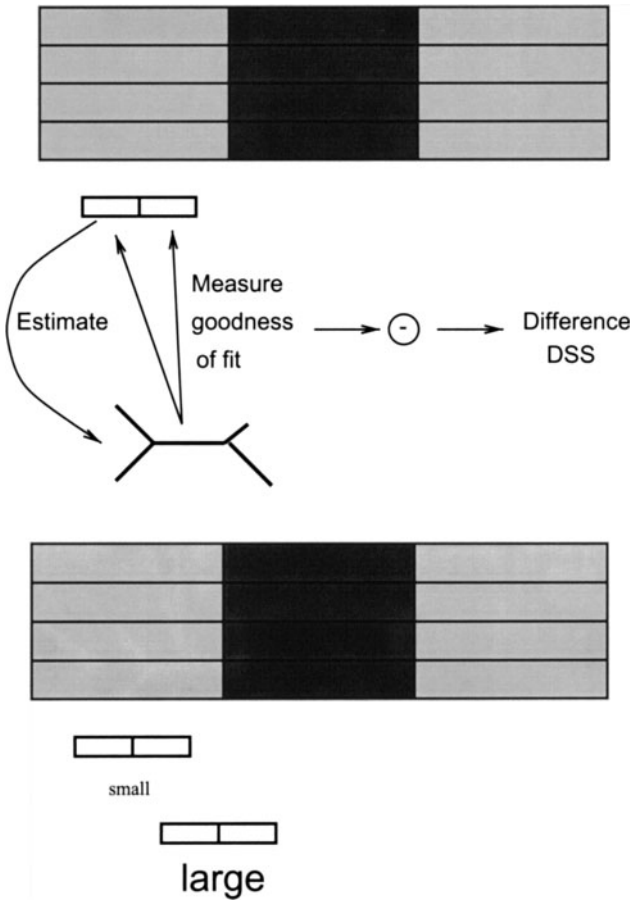

FIG. 2.—Illustration of TOPAL. A window is moved along the DNA sequence alignment. A tree is estimated from the left part of the window, and a goodness-of-fit score is computed for both parts of the window. The DSS statistic is defined as the difference between these scores. When the window is centred on or near the breakpoint of a recombinant region, the tree estimated from the left subwindow is not an adequate description for the data on the right, which leads to a large DSS value.

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-neighbor 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: the next section, *Method: Background and Earlier Approaches*, introduces the mathematical method and discusses the shortcomings of existing parameter estimation techniques. Then, under *Method: A Bayesian Approach*, we discuss how earlier approaches can be improved with a Bayesian approach using Markov chain Monte Carlo. In the section titled *Data* we describe various synthetic and real-world DNA sequence alignments on which the proposed scheme was tested. We next present the simulation study itself and discuss the results. The article ends with a conclusion and recommendations for future work.

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 $\mathbf{y}_t$, where the subscript t represents the site, $1 \leqslant t \leqslant N$. Hence $\mathbf{y}_t$ is an m-dimensional column vector containing the nucleotides at the tth site of the alignment, and $\mathcal{D} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. Attached to each site is a hidden state variable S_t, which represents the tree topology at site t. For m taxa, there are $K = (2m - 5)!!$ distinct unrooted topologies (where !! denotes double factorial), hence $S_t \in \{1, \ldots, K\}$. If a recombination 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

$$\mathbf{S} = (S_1, \ldots, S_N) \tag{2}$$

given the sequence alignment $\mathcal{D}$ and some optimality criterion to be discussed below.

Obviously, this optimization problem is, in general, intractable. First, the number of possible topologies at a given site, K, increases super-exponentially with the number of sequences m. Second, 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-neighbor interactions. This allows the application of a dynamic programing scheme which reduces the computational complexity to $\mathcal{O}(K^2N)$, that is, to an expression linear in N. To deal with the first source of complexity, the scheme has to be restricted to alignments with small numbers of sequences. In the current work, we restrict our approach to alignments with only $m = 4$ taxa. In the *Discussion* we describe how this restriction can be relaxed.

RECPARS

Hein (1993) defined optimality in a parsimony sense. His algorithm, RECPARS, searches for the most parsimonious state sequence $\mathbf{S}$, that is, the one that minimizes a given parsimony cost function $E(\mathbf{S})$. Interactions between sites are restricted to nearest-neighbor interactions, as discussed above, and the search is carried out with dynamic programing. Although RECPARS is faster than the methods to be discussed below, it suffers from the shortcomings inherent in parsimony, as discussed by Felsenstein (1988). Moreover, $E(\mathbf{S})$ depends only on the topology-defining sites; thus the algorithm discards a substantial proportion of sites in the alignment. The most serious disadvantage is that the cost function $E(\mathbf{S})$ depends on certain parameters—the mutation cost C_{mut}, and the recombination cost C_{rec}—which can *not* be optimized within the framework of this method. Consequently, these parameters have to be chosen by the user in advance, and the predictions depend on this rather arbitrary prior selection.

Detecting Recombination with Hidden Markov Models

Adopting a statistical approach to phylogenetics, illustrated in figure 3, 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 4, 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 represent columns in the DNA sequence alignment, $\mathbf{y}_t$. The joint probability of the DNA sequence alignment, $\mathcal{D}$, and the sequences of hidden states, $\mathbf{S}$, factorizes:

$$P(\mathcal{D}, \mathbf{S}) = P(\mathbf{y}_1, \ldots, \mathbf{y}_N, S_1, \ldots, S_N) = \prod_{t=1}^{N} P(\mathbf{y}_t \mid S_t) \prod_{t=2}^{N} P(S_t \mid S_{t-1}) P(S_1). \quad (3)$$

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

$$\hat{\mathbf{S}} = \operatorname{argmax}_{\mathbf{S}} P(\mathbf{S} \mid \mathcal{D}). \quad (4)$$

While, in general, this problem would be intractable because of the exponential increase in the number of state sequences (see above), the reduction to nearest-neighbor interactions between hidden states and the resulting factorization (3) allows the application of a dynamic programing technique, the so-called Viterbi algorithm (Rabiner 1989), to find the mode $\hat{\mathbf{S}}$ with computational complexity $\mathcal{O}(N)$. The factorization (3) contains three terms: $P(\mathbf{y}_t \mid S_t)$, $P(S_t \mid S_{t-1})$, and $P(S_1)$. The *transition probabilities* $P(S_t \mid S_{t-1})$ correspond to recombination events (if $S_t \neq S_{t-1}$). Let ν 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:

$$P(S_t \mid S_{t-1}, \nu) = \nu \delta(S_t, S_{t-1}) + \frac{1 - \nu}{K - 1} [1 - \delta(S_t, S_{t-1})] \quad (5)$$

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 $\sum_{S_t} P(S_t \mid S_{t-1}) = 1$. The *emission probabilities* $P(\mathbf{y}_t \mid S_t)$ can easily be computed with the pruning algorithm (Felsenstein 1981) if the branch lengths corresponding to the topology S_t, $\mathbf{w}_{S_t}$, and the parameters of the nucleotide substitution model, $\boldsymbol{\theta}_{S_t}$, are known. So, more precisely, we have $P(\mathbf{y}_t \mid S_t) = P(\mathbf{y}_t \mid S_t, \mathbf{w}_{S_t}, \boldsymbol{\theta}_{S_t})$. To simplify the notation, define the accumulated vectors $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ and define: $P(\mathbf{y}_t \mid S_t, \mathbf{w}_{S_t}, \boldsymbol{\theta}_{S_t}) = P(\mathbf{y}_t \mid S_t, \mathbf{w}, \boldsymbol{\theta})$. This means that S_t indicates which subvectors of $\mathbf{w}$ and $\boldsymbol{\theta}$ apply. We can depict the dependence of the probability distribution on the parameters $\mathbf{w}$ and ν in a probabilistic graphical model, shown in figure 4, left, with the emission and transition probabilities illustrated in figures 3 and 5, respectively. The prediction task is to find the most likely hidden

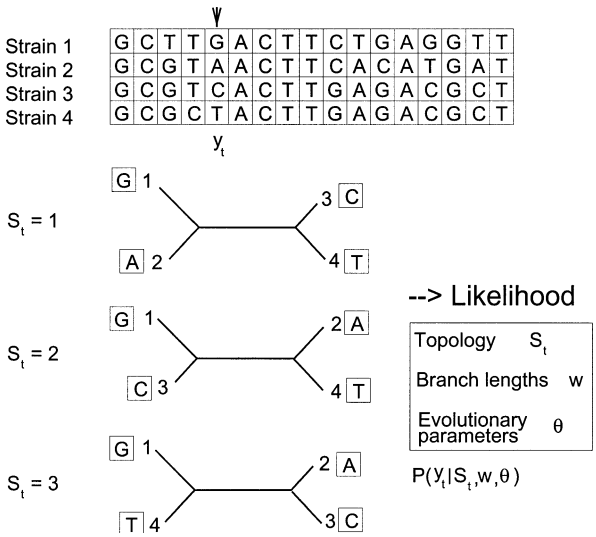

FIG. 3.—Statistical approach to phylogenetics and modeling recombination. For a given column $\mathbf{y}_t$ in the alignment, a probability $P(\mathbf{y}_t \mid S_t, \mathbf{w}, \boldsymbol{\theta})$ can be computed, which depends on the tree topology, S_t, the vector of branch lengths, $\mathbf{w}$, and the parameters of the nucleotide substitution model, $\boldsymbol{\theta}$. 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 $\mathbf{w}$ and $\boldsymbol{\theta}$ are accumulated vectors, as defined in the paragraph above equation (6).

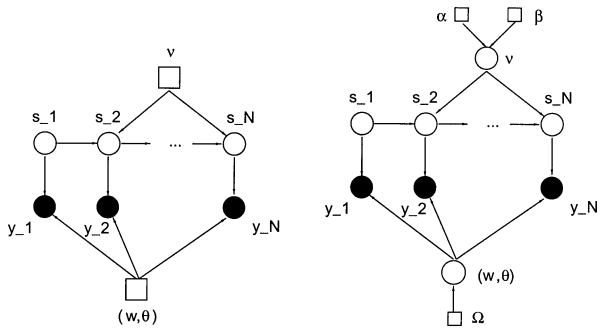

FIG. 4.—Modeling recombination with hidden Markov models. Positions in the model, labeled by the subscript t, correspond to sites in the DNA sequence alignment. Black nodes represent observed random variables; these are the columns in the DNA sequence alignment. White nodes represent hidden states; these are the different tree topologies, shown (for four sequences) in figure 3. Arcs represent conditional dependencies. Squares represent parameters of the model. The probability for observing a column vector $\mathbf{y}_t$ at position t in the DNA sequence alignment depends on the tree topology S_t, the vector of branch lengths $\mathbf{w}$, and the parameters of the nucleotide substitution model $\boldsymbol{\theta}$. 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 ν. *Left:* In the older approaches of McGuire, Wright, and Prentice (2000) and Husmeier and Wright (2001), ν, $\mathbf{w}$, and $\boldsymbol{\theta}$ are parameters that have to be estimated. *Right:* In the Bayesian approach, ν, $\mathbf{w}$, and $\boldsymbol{\theta}$ are random variables. The prior distribution for ν is a beta distribution with hyperparameters α and β. The prior distributions for the remaining parameters are discussed under *Method: A Bayesian Approach* and depend on some hyperparameters Ω. The parameters ν, $\mathbf{w}$, and $\boldsymbol{\theta}$ are sampled from the posterior distribution with Markov chain Monte Carlo.

state sequence conditional on the observations (that is, the DNA sequence alignment) and the parameters $\mathbf{w}$, $\boldsymbol{\theta}$, and ν:

$$\begin{aligned}&\operatorname{argmax}_{\mathbf{S}} P(\mathbf{S}\,|\,\mathcal{D}, \mathbf{w}, \boldsymbol{\theta}, \nu)\\ &\quad = \operatorname{argmax}_{S_1,\ldots,S_N} P(S_1, \ldots, S_N\,|\,\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{w}, \boldsymbol{\theta}, \nu) \qquad (6)\end{aligned}$$

The parameters $\mathbf{w}$, $\boldsymbol{\theta}$, and ν need to be estimated.

Heuristic Parameter Estimation: HMM-Heuristic

McGuire, Wright, and Prentice (2000) estimated the branch lengths $\mathbf{w}$ for each tree topology separately with maximum likelihood. This approach is suboptimal. For a proper estimation of the branch lengths of a recombinant tree, one would have to restrict the parameter estimation to 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), because the estimation includes regions of the alignment 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 ν.

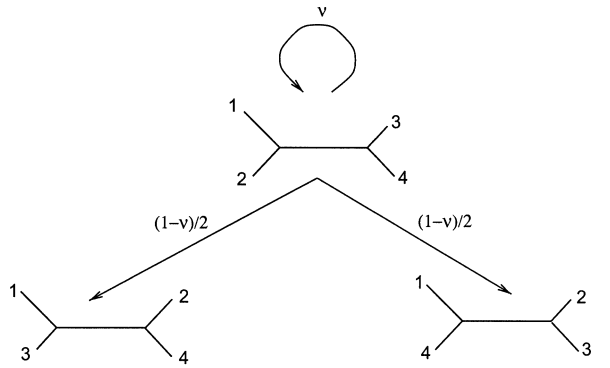

FIG. 5.—Transition probabilities. The hidden states of the HMM represent different tree topologies, and state transitions correspond to recombination events. The transition probability ν is the probability that on moving from a site in the DNA sequence alignment to an adjacent site, no topology change occurs. If a topology change does occur, we assume that, a priori, all transitions are equally likely.

Parameter Estimation with Maximum Likelihood: 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

$$L(\mathbf{w}, \nu) = \ln P(\mathcal{D}\,|\,\mathbf{w}, \nu) = \ln \sum_{\mathbf{S}} P(\mathcal{D}, \mathbf{S}\,|\,\mathbf{w}, \boldsymbol{\theta}, \nu) \qquad (7)$$

with respect to the vector of branch lengths $\mathbf{w}$, the parameters of the nucleotide substitution model $\boldsymbol{\theta}$, and the recombination parameter ν. This requires a summation over all state sequences $\mathbf{S} = (S_1, \ldots, S_N)$, 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 suffers from the shortcoming that the predicted state sequence does not only depend on the data, argmax $P(\mathbf{S}\,|\,\mathcal{D})$, but also on the parameters, argmax $P(\mathbf{S}\,|\,\mathcal{D}, \mathbf{w}, \boldsymbol{\theta}, \nu)$.The fact that these parameters are estimated from the data itself with maximum likelihood renders the approach susceptible to over-fitting. This calls for an independent hypothesis test with parametric bootstrapping, which, however, incurs prohibitively high computational costs, as demonstrated by Larget and Simon (1999).

To rephrase this problem, note that hidden Markov models and phylogenetic trees have many similarities with neural networks; in fact, all three models are instances of the more general class of graphical models (Heckermann 1999). Studies on neural networks and graphical models have shown that, for sparse data, maximum likelihood is susceptible to over-fitting, and that the generalization performance is significantly improved with the Bayesian approach. A detailed investigation of this approach can be found in Neal (1996). In a nutshell, maximum likelihood gives only a point estimate of the parameters, which

ignores the more detailed information contained in the curvature and (possibly) multimodality of the likelihood landscape. By sampling rather than optimizing parameters, the Bayesian approach captures more information about this landscape, and consequently gives improved and more reliable predictions.

Method: A Bayesian Approach

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

$$P(\mathbf{S} \mid \mathcal{D}) = \int P(\mathbf{S}, \mathbf{w}, \boldsymbol{\theta}, \nu \mid \mathcal{D}) d\mathbf{w} d\boldsymbol{\theta} d\nu. \qquad (8)$$

In principle this avoids the over-fitting scenario mentioned above and removes the need for a separate hypothesis test. The difficulty, however, is that the integral in (eq. 8) 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 discuss 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; (3) methods for accelerating the convergence of the Markov chain; (4) the prediction resulting from this scheme; and (5) a software implementation. We will then test this approach on various DNA sequence alignments.

Prior Probabilities

Inherent in the Bayesian framework is the choice of prior probabilities for all model parameters, as illustrated in figure 4, right. We make the usual assumption of parameter independence, $P(\nu, \mathbf{w}, \boldsymbol{\theta}) = P(\nu)P(\mathbf{w})P(\boldsymbol{\theta})$, 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 ν is a binomial random variable, for which the conjugate prior is a beta distribution,

$$P(\nu) = \mathcal{B}(\nu \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \nu^{\alpha-1}(1-\nu)^{\beta-1}, \qquad (9)$$

whose shape is determined by the hyperparameters α and β, as shown in figure 6.

The branch lengths $\mathbf{w}$ 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. Because for real DNA sequence alignments branch lengths are unlikely to approach 1, this restriction should not cause any difficulties.

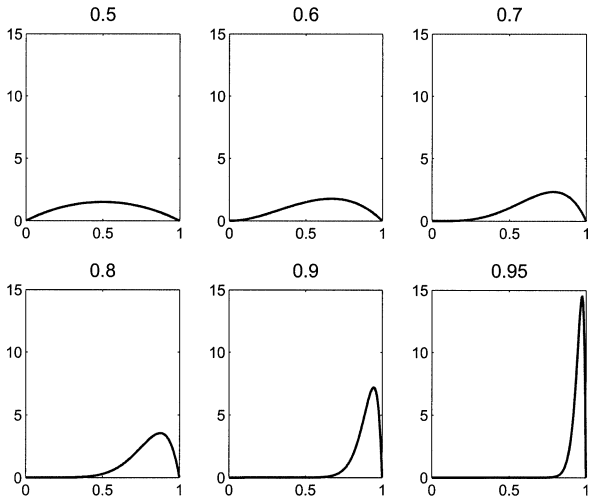

FIG. 6.—Prior distribution for the recombination parameter ν. The conjugate prior for ν is a beta distribution, which depends on two hyperparameters, α and β. The mean of the distribution is $\mu = \alpha/(\alpha + \beta)$. The subfigures show plots of the distribution for different values of μ, indicated at the top of each subfigure, when $\beta = 2$ is fixed.

The prior on $\boldsymbol{\theta}$ 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, π_A, π_C, π_G, and π_T, and the transition bias ρ. (Note that because of the normalization constraint $\pi_A + \pi_C + \pi_G + \pi_T = 1$, there are 4 rather than 5 free parameters.) In our approach, each tree is allowed to have a different value of ρ, whereas the nucleotide frequencies are assumed to be the same for all trees. This is for algorithmic efficiency: Allowing each tree to have a different set of frequencies means that some frequencies might be inferred from a small amount of data, leading to vague posterior distributions that slow down the convergence of the Markov chain. The total vector of nucleotide substitution parameters is thus of the form

$$\boldsymbol{\theta} = (\rho_1, \ldots, \rho_K, \pi_A, \pi_C, \pi_G, \pi_T). \qquad (10)$$

A natural prior for the nucleotide frequencies π_i is a Dirichlet distribution, which, as a multivariate generalization of the beta distribution (eq. 9), 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 ρ_k ($k = 1, \ldots, K$), we choose a uniform prior over the interval [0, 2]. Again, an upper bound is needed to prevent the prior from becoming improper. Allowing ρ_k to be as large as 2 will account for extreme cases of transition bias, which should not impose any serious restrictions in practice. Finally, we assume $P(S_1) = 1/K \; \forall \; S_1 \in \{1, \ldots, K\}$, that is, a uniform prior on the tree topologies.

The joint distribution of the DNA sequence alignment, the state sequences, and the model parameters is given by

$$P(\mathcal{D}, \mathbf{S}, \mathbf{w}, \boldsymbol{\theta}, \nu) = \prod_{t=1}^{N} P(\mathbf{y}_t \,|\, S_t, \mathbf{w}, \boldsymbol{\theta}) \times \prod_{t=2}^{N} P(S_t \,|\, S_{t-1}, \nu) P(S_1) P(\mathbf{w}) P(\boldsymbol{\theta}) P(\nu) \tag{11}$$

where $P(\mathbf{y}_t \,|\, S_t, \mathbf{w}, \boldsymbol{\theta})$ is the probability of the tth 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 (eq. 5), and $P(S_1)$, $P(\mathbf{w})$, $P(\boldsymbol{\theta})$, and $P(\nu)$ are the prior probabilities, discussed above.

Markov Chain Monte Carlo (MCMC) Sampling

Ultimately, we are interested in the marginal posterior probability of the state sequences, $P(\mathbf{S} \,|\, \mathcal{D})$, which requires a marginalization over the model parameters according to (eq. 8). The numerical approximation is to sample from the joint posterior distribution

$$P(\mathbf{S}, \mathbf{w}, \boldsymbol{\theta}, \nu \,|\, \mathcal{D}) \tag{12}$$

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

$$\begin{aligned}
\mathbf{S}^{(i+1)} &\sim P(\cdot \mid \mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}, \mathcal{D}) \\
\mathbf{w}^{(i+1)} &\sim P(\cdot \mid \mathbf{S}^{(i+1)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}, \mathcal{D}) \\
\boldsymbol{\theta}^{(i+1)} &\sim P(\cdot \mid \mathbf{S}^{(i+1)}, \mathbf{w}^{(i+1)}, \nu^{(i)}, \mathcal{D}) \\
\nu^{(i+1)} &\sim P(\cdot \mid \mathbf{S}^{(i+1)}, \mathbf{w}^{(i+1)}, \boldsymbol{\theta}^{(i+1)}, \mathcal{D}).
\end{aligned} \tag{13}$$

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 (eq. 5) and (eq. 9) it is seen that writing the joint probability (eq. 11) as a function of ν gives:

$$P(\mathcal{D}, \mathbf{S}, \mathbf{w}, \boldsymbol{\theta}, \nu) \propto \nu^{\Psi+\alpha-1} (1-\nu)^{N-\Psi+\beta-2}. \tag{14}$$

On normalization this gives

$$P(\nu \,|\, \mathcal{D}, \mathbf{S}, \mathbf{w}, \boldsymbol{\theta}) = \mathcal{B}(\nu \mid \Psi + \alpha,\ N - 1 - \Psi + \beta) \tag{15}$$

where $\mathcal{B}$ is the beta distribution (eq. 9), from which sampling is straightforward (Rubinstein 1981).

For sampling the state sequences $\mathbf{S}$, 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:

$$\begin{aligned}
S_1^{(i+1)} &\sim P(\cdot \mid S_2^{(i)}, S_3^{(i)}, \ldots, S_N^{(i)}, \mathcal{D}, \mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}) \\
S_2^{(i+1)} &\sim P(\cdot \mid S_1^{(i+1)}, S_3^{(i)}, \ldots, S_N^{(i)}, \mathcal{D}, \mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}) \\
&\vdots \\
S_N^{(i+1)} &\sim P(\cdot \mid S_1^{(i+1)}, S_2^{(i+1)}, \ldots, S_{N-1}^{(i+1)}, \mathcal{D}, \mathbf{w}^{(i)}, \boldsymbol{\theta}^{(i)}, \nu^{(i)}).
\end{aligned} \tag{16}$$

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

$$\begin{aligned}
&P(S_t \mid S_1, \ldots, S_{t-1}, S_{t+1}, \ldots, S_N, \mathcal{D}, \mathbf{w}, \boldsymbol{\theta}, \nu) \\
&\quad = P(S_t \mid S_{t-1}, S_{t+1}, \mathbf{y}_t, \mathbf{w}, \boldsymbol{\theta}, \nu) \\
&\quad \propto P(S_{t+1} \mid S_t, \nu) P(S_t \,|\, S_{t-1}, \nu) P(\mathbf{y}_t \,|\, S_t, \mathbf{w}, \boldsymbol{\theta})
\end{aligned} \tag{17}$$

where $P(S_t \,|\, S_{t-1}, \nu)$ and $P(S_{t+1} \,|\, S_t, \nu)$ are given by equation (5). Note that the expression after the α symbol is easily normalized to give a proper probability, from which sampling is straightforward (because $S_t \in \{1, \ldots, K\}$ is discrete).

For sampling the remaining parameters, $\mathbf{w}$ and $\boldsymbol{\theta}$, we apply the Metropolis-Hastings algorithm (see Hastings [1970] and Chib and Greenberg [1995]). Let $\mathbf{z}^{(i)}$ denote the parameter configuration in the ith sampling step. A new parameter configuration $\tilde{\mathbf{z}}$ is sampled from a proposal distribution $Q(\tilde{\mathbf{z}} \,|\, \mathbf{z}^{(i)})$, and then accepted with probability

$$A(\tilde{\mathbf{z}}) = \min\left\{ \frac{P(\tilde{\mathbf{z}}) Q(\mathbf{z}^{(i)} \,|\, \tilde{\mathbf{z}})}{P(\mathbf{z}^{(i)}) Q(\tilde{\mathbf{z}} \mid \mathbf{z}^{(i)})}, 1 \right\} \tag{18}$$

in which case $\mathbf{z}^{(i+1)} = \tilde{\mathbf{z}}$. Otherwise, $\mathbf{z}^{(i+1)} = \mathbf{z}^{(i)}$. The distribution P is given by equation (11).

Improving the Convergence of the Markov Chain

In theory the algorithm converges to the posterior distribution (eq. 12) 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 $\mathbf{w}$ and for the transition biases ρ_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 equation (18). For the nucleotide frequencies π_A, π_C, π_G, π_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 $Q(. \,|\, .)$ terms must be calculated in equation (18).

If too few proposed values are accepted, the corresponding proposal distributions $Q(. \,|\, .)$ may be tuned to make acceptance more likely and thereby to accelerate convergence. For the branch lengths w_l and the transition biases ρ_k, this is done by decreasing the width of the uniform interval from which the new value is sampled. For the nucleotide frequencies π_A, π_C, π_G, π_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 first initializing the chain. The sequence of topologies $\mathbf{S}$ is chosen randomly or from

some initial estimation, e.g., using RECPARS. The branch lengths are set to some plausible value, e.g., the average branch length of the global maximum likelihood tree obtained with DNAML of the PHYLIP package (available from http://evolution.genetics.washington.edu/phylip.html). Initial values for the transition biases ρ_k and the nucleotide frequencies π_A, π_C, π_G, π_T can be estimated from the data, as described later under *Simulations*. The parameter groups are then updated in order according to (eq. 13) 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 $Q(.\,|\,.)$ are tuned as described above. This burn-in is followed by the sampling phase of the simulation, in which the state sequences $\mathbf{S}$ (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 (eq. 12).

In principle, the initialization of the hidden states is unimportant because the Markov chain will forget its initial configuration and converge toward the equilibrium distribution irrespective of its starting point. In practice, however, extreme starting values can 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 toward the main support of the posterior distribution in the available simulation time. To address this problem, we combined simulations from different initializations and explored a method akin to simulated annealing (Kirkpatrick, Gelatt, and Vecchi 1983). Note that the bottleneck of the presented Markov chain Monte Carlo scheme is the sampling of the topology sequences $\mathbf{S}$. Because recombination events are quite rare, the number of topology changes along the DNA sequence alignment is usually small and, consequently, the posterior distribution of ν concentrated on values close to 1. This discourages state transitions and may slow down the mixing of the Markov chain. To increase the transition rate during equilibration, we therefore modified the transition probabilities as follows: Let T denote the total length of the equilibration phase, and let $i \in \{1, \ldots, T\}$ denote the ith sample of the Markov chain during equilibration. Then, during equilibration, equation (15) is replaced by

$$P^{(i)}(\nu \mid \mathcal{D}, \mathbf{S}, \mathbf{w}, \boldsymbol{\theta}) = \mathcal{B}\left(\nu \mid \frac{i}{T}\Psi + \alpha, \frac{i}{T}(N - 1 - \Psi) + \beta\right). \tag{19}$$

For $i = 0$, this distribution is identical to the prior distribution (eq. 9), whereas for $i = T$ it is identical to the posterior distribution (eq. 15). For intermediate values, $0 < i < T$, the hyperparameters of equation (19) are a mixture of the prior and posterior hyperparameters, and the distribution (eq. 19) 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 ν will be sampled with a higher probability than with the standard (unannealed) scheme (assuming the prior has been chosen sufficiently vague). This facilitates transitions between state sequences and can be expected to improve the mixing of the Markov chain.

Prediction

Recall that the proposed Bayesian method samples topology sequences from the joint posterior probability $P(\mathbf{S} \,|\, \mathcal{D}) = P(S_1, \ldots, S_N \,|\, \mathcal{D})$, where S_t ($t = 1, \ldots, N$) represents the topology at site t. To display the results graphically, we marginalize, for each site in turn, over all the remaining sites so as to return the marginal posterior probabilities $P(S_t \,|\, \mathcal{D})$. These are then plotted, for each topology $P(S_t = 1 \,|\, \mathcal{D})$, $P(S_t = 2 \,|\, \mathcal{D})$, and $P(S_t = 3 \,|\, D)$, along the sequence alignment. Assigning each site t to the mode of the posterior probability $P(S_t \,|\, \mathcal{D})$ gives a list of putative recombinant regions, identical to the output of RECPARS. This is a useful reduction of information that allows the comparison of classification scores, as shown later (see figure 12). Note, however, that the posterior probabilities $P(S_t \,|\, \mathcal{D})$ contain further, additional information, as they also indicate the uncertainty of the prediction.

Implementation

The method discussed above has been implemented in the C++ program package BARCE, which is freely available from http://www.bioss.sari.ac.uk/~dirk/My_software.html.

Data

We have tested the viability of the proposed method on various DNA sequence alignments, including a simulated recombination and the sequences of maize, hepatitis B virus, and *Neisseria*.

Simulated Recombination

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 events were simulated, as shown in figure 7. Topology 1 is the "true" topology, which applies to those parts of the alignment that are not affected by recombination. The four sequences are evolved along the interior branch and the first quarter of the exterior branches of a phylogenetic tree (top left). At this point, the subsequence between sites 201 and 400 in Strain 3 is replaced by the corresponding subsequence in Strain 1 (top right). The sequences then continue to evolve along the exterior branches until the branch length is 0.75 times the final exterior branch length (middle, left). This is followed by a second recombination event, where the subsequence between sites 601 and 800 in Strain 2 replaces the corresponding subsequence in Strain 3 (middle right). The sequences then continue to evolve along the exterior branches for the remaining length (bottom left). The resulting mosaic structure of the alignment is shown in figure 7, bottom right. In the main

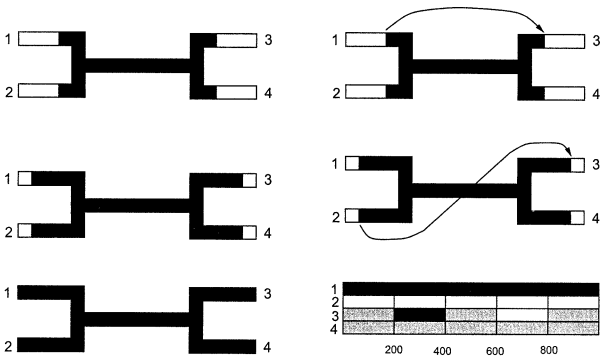

FIG. 7.—Simulation of recombination. Four sequences are evolved along the interior branch and the first quarter of the exterior branches of a phylogenetic tree (*top left*). At this point, the subsequence between sites 201 and 400 in Strain 3 is replaced by the corresponding subsequence in Strain 1 (*top right*). The sequences then continue to evolve along the exterior branches until the branch length is 0.75 times the final exterior branch length (*middle, left*). This process is followed by a second recombination event, where the subsequence between sites 601 and 800 in Strain 2 replaces the corresponding subsequence in Strain 3 (*middle right*). The sequences then continue to evolve along the exterior branches for the remaining length (*bottom left*). The resulting mosaic structure is shown in the *bottom right*.

part of the alignment, Strain 3 is most closely related to Strain 4. However, in the region between sites 201 and 400, it is most closely related to Strain 1, and in the region between 601 and 800, it is most closely related to Strain 2 (bottom right). Thus, the first, more ancient, recombination event corresponds to a transition from Topology 1 into Topology 2. The second, more recent, recombination event corresponds to a transition from Topology 1 into Topology 3. This model simulates a realistic scenario where an ancestor of Strain 3 incorporates genetic material from ancestors of other extant strains, which in each case is followed by subsequent evolution. The simulations were repeated for a different mosaic structure, where the first recombinant region was extended by 100 nucleotides (region 201–500), and the second region was shortened by 100 nucleotides (region 701–800). The mosaic structures are shown in the top-left subfigure of figures 8–11. For each mosaic structure, we repeated the simulation with three different tree heights, where the tree height is defined as half the sum of all the branch lengths between the two strains that are farthest apart. Note that as the tree height becomes smaller, the numbers of polymorphic and topology-defining sites decrease. This reduces the information content in the alignment and makes the detection of recombinant regions more difficult.

The software used for this simulation can be downloaded from http://www.bioss.sari.ac.uk/~dirk/software/Sirens/INFO.html.

Maize

Gene conversion is a process equivalent to recombination, which occurs in multigene families, where a DNA subsequence of one gene can be replaced by the DNA subsequence from another. 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 ClustalW (Thompson, Higgins, and Gibson 1994), using the default parameter settings. We define the states of the HMM as follows: Topology 1: [(Maz56,Maz63),(Maz89,Maz95)]; topology 2: [(Maz56,Maz89),(Maz63,Maz95)]; topology 3: [(Maz56,Maz95),(Maz63,Maz89)]. (See figure 15, top, for an illustration.)

Hepatitis B

A DNA virus with a short genome of only 3200 bases causes hepatitis B infection. Evidence for recombination was first found by Bollyky et al. (1996), and in this study we investigated a subset of four strains with the following GenBank identifiers (accession numbers in square brackets): (1) HPBADW1 [D00329], (2) HPBADW2 [D00330], (3) HPBADWZCG [M57663], (4) HPBADRC [D00630]. The sequences were aligned with ClustalW, using the default parameters. Columns with gaps were discarded, giving a total alignment length of 3049 bases. Bollyky and associates (1996) found a recombinant region of 189 bases in HPBADWZCG between $t = 1865$ and $t = 2054$ (when not removing gaps: $t = 2014$–2203), corresponding to a transition from topology $S_t = 1$ (HPBADW1 grouped with HPBADW2) into topology $S_t = 2$ (HPBADW1 grouped with HPBADWZCG).

Neisseria

One of the first indications for sporadic recombination was found in the bacterial genus *Neisseria* (Maynard 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 labeling 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 topology $S_t = 1$ in our HMM), whereas between $t = 1$ and $t = 202$, they found that *N. meningitidis* is grouped with *N. cinera* (defined as state $S_t = 3$). Zhou and Spratt (1992) suggested that the region $t = 507$–538 might be the result of rate variation. The situation is illustrated in figure 16.

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

Simulations

We applied RECPARS with different ratios of the

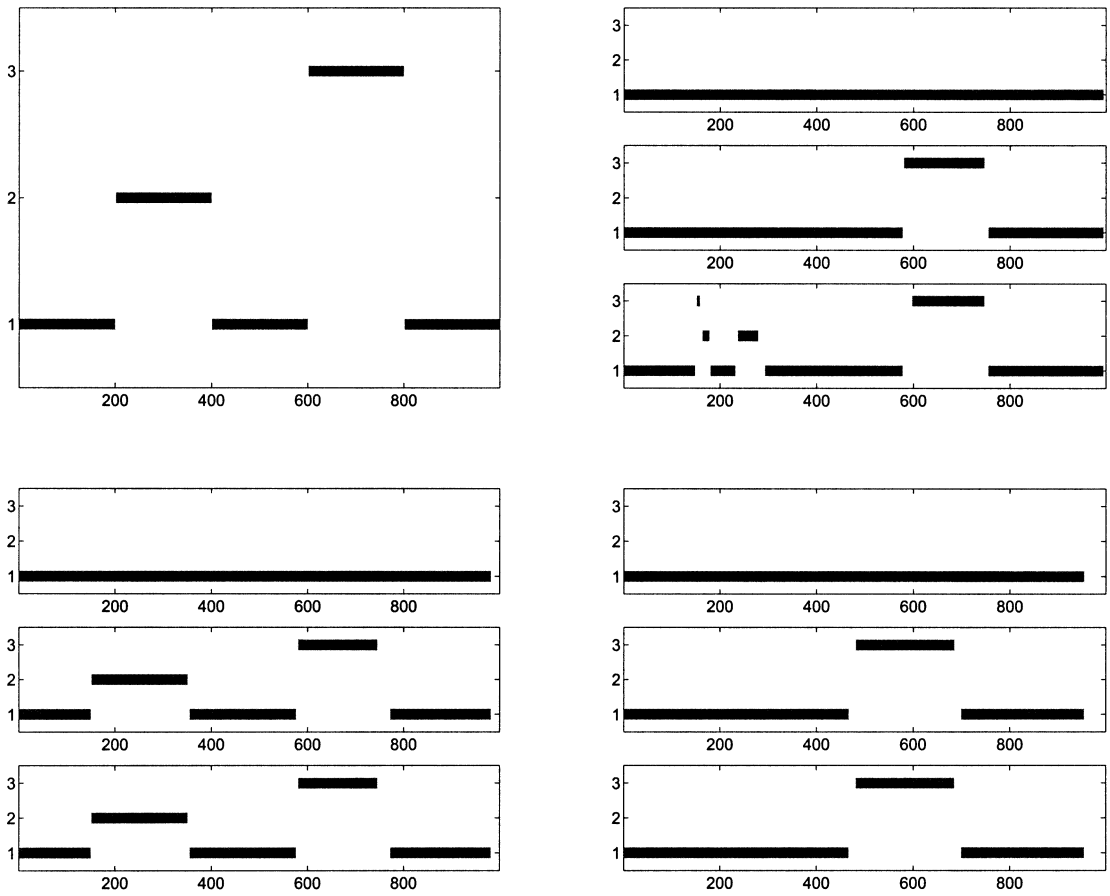

FIG. 8.—Prediction with RECPARS for mosaic structure A. The true mosaic structure is shown in the figure on the *top, left*, where the horizontal axis shows the position in the DNA sequence alignment, and the vertical axis represents the three possible tree topologies. The other figures show predictions with RECPARS for different tree heights. *Top, right*: Tree height = 0.3. *Bottom, left*: Tree height = 0.2. *Bottom, right*: Tree height = 0.1. Each figure contains three subfigures, which show the results for different recombination-mutation cost ratios, C_{rec}/C_{mut}. *Top subfigure*: C_{rec}/C_{mut} = 10.0; *middle subfigure*: C_{rec}/C_{mut} = 3.0; *bottom subfigure*: C_{rec}/C_{mut} = 1.5. The horizontal axis in each subfigure represents sites in the alignment, and the vertical axis represents the three possible tree topologies.

mutation and recombination costs, C_{rec}/C_{mut}, using the program written by Kim Fisker (ftp://ftp.daimi.aau.dk/pub/empl/kfisker/programs/RecPars/).

TOPAL was applied with two different window lengths: $W = 100$ and $W = 200$. We used Version 2 of the program (McGuire and Wright 2000) with the default options except for the nucleotide substitution model, where we replaced the (default) Jukes-Cantor model with the Kimura model. The transition-transversion ratio was estimated with Puzzle (Strimmer and von Haeseler 1996).

For PLATO, we used Version 2.11 of the program developed by Grassly and Holmes (1997), available from http://evolve.zoo.ox.ac.uk/software/Plato/Plato2.html, and we varied the window length between five bases and half the sequence length. The reference tree was obtained with maximum likelihood from the whole DNA sequence alignment, using DNAML of the PHYLIP package (available from http://evolution.genetics.washington.edu/phylip.html) or Puzzle. On the synthetic data, we used a uniform substitution rate, and set the transition-transversion ratio to the known true value. On the real data, we used two models of rate heterogeneity: (1) a uniform rate and (2) gamma distributed rates with five rate categories. The respective PLATO commands are these: (1) plato -mHKY -tTAU and (2) plato -g5 -aALPHA -mHKY -tTAU, where TAU and ALPHA are the transition-transversion ratio and the alpha parameter of the discrete gamma distribution, respectively, both estimated with Puzzle.

The application of HMM-heuristic 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 Puzzle), and estimating the nucleotide frequencies π_A, π_C, π_G, π_T from the data according to $\pi_X = N_X/\sum_{X'} N_{X'}$, 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

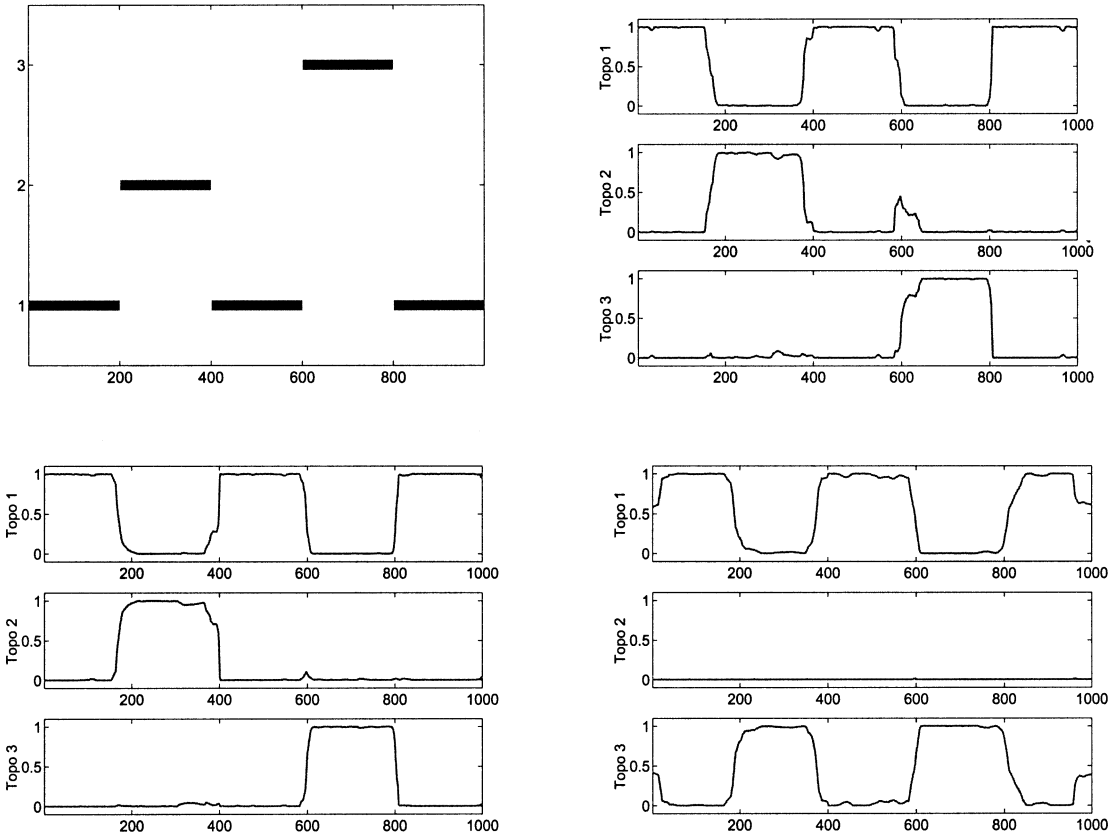

FIG. 9.—Prediction with HMM-Bayes for mosaic structure A. The true mosaic structure is shown in the figure on the *top, left,* where the horizontal axis shows the position in the DNA sequence alignment, and the vertical axis represents the three possible tree topologies. The other figures show the probabilities $P(S_t \mid \mathcal{D})$ predicted with HMM-Bayes for different tree heights. *Top, right*: Tree height = 0.3. *Bottom, left*: Tree height = 0.2. *Bottom, right*: Tree height = 0.1. Each figure contains three subfigures, which show the posterior probabilities for the three topologies, $P(S_t = 1 \mid \mathcal{D})$ (*top*), $P(S_t = 2 \mid \mathcal{D})$ (*middle*), $P(S_t = 3 \mid \mathcal{D})$ (*bottom*), plotted against the site t in the DNA sequence alignment.

and Wright (2001) and optimized the recombination parameter ν and all the branch lengths simultaneously in a maximum likelihood sense with the EM algorithm, using the MATLAB programs written by the authors (available from http://www.bioss.sari.ac.uk/~dirk/My_software.html).

Finally, the proposed Bayesian MCMC scheme, HMM-Bayes, was applied as follows. We used the Felsenstein 84 model of nucleotide substitution, with a prior on the parameters as described earlier under *Method: A Bayesian Approach*. For the prior on $\nu \in [0,1]$, we chose the three distributions shown at the bottom of figure 6, which incorporate our knowledge that $P(\nu)$ must be skewed toward the right of the interval [0,1]. This is so because $\nu = 0.5$ means that, on average, every second site is subject to recombination, and we know that recombination events are much rarer, that is, that $\nu \gg 0.5$. We found that for the chosen priors, the simulations gave very similar results. Recall that the posterior probability is the product of the prior and the likelihood. Whereas the first term is constant, the second term scales like N, the number of sites in the DNA sequence alignment. Consequently, for a sufficiently long alignment and a reasonable prior, the weight of the likelihood is considerably higher than that of the prior, and variations of the latter have therefore only a marginal influence on the prediction. This was borne out in the simulations, as shown in figure 21 and discussed in the Appendix.

The initial nucleotide frequencies and the initial transition-transversion ratio were estimated from the data, as described above. Equilibration was carried out over $10^5 - 10^6$ MCMC steps. This was followed by a sampling phase of the same length, during which the parameters ν, $\mathbf{w}$, $\boldsymbol{\theta}$, and topology sequences $\mathbf{S}$ were sampled in intervals of 1000 MCMC steps. From the recorded topology sequences $\mathbf{S}$, we computed the marginal posterior probabilities $P(S_t = 1 \mid \mathcal{D})$, $P(S_t = 2 \mid \mathcal{D})$, and $P(S_t = 3 \mid \mathcal{D})$ for all sites in the DNA sequence alignment, $1 \leqslant t \leqslant N$.

Results

Comparison with RECPARS

We applied both HMM-Bayes and RECPARS to the synthetic DNA sequence alignments described earlier under *Data*. The objective of this simulation study was to test the performance of both methods on different (a priori known) mosaic structures and for varying levels of difficulty of the detection problem (which is related to the tree height, also discussed under *Data*). The results are shown in figures 8–11. When the tree height is sufficiently large (0.3, 0.2), both HMM-Bayes and RECPARS predict

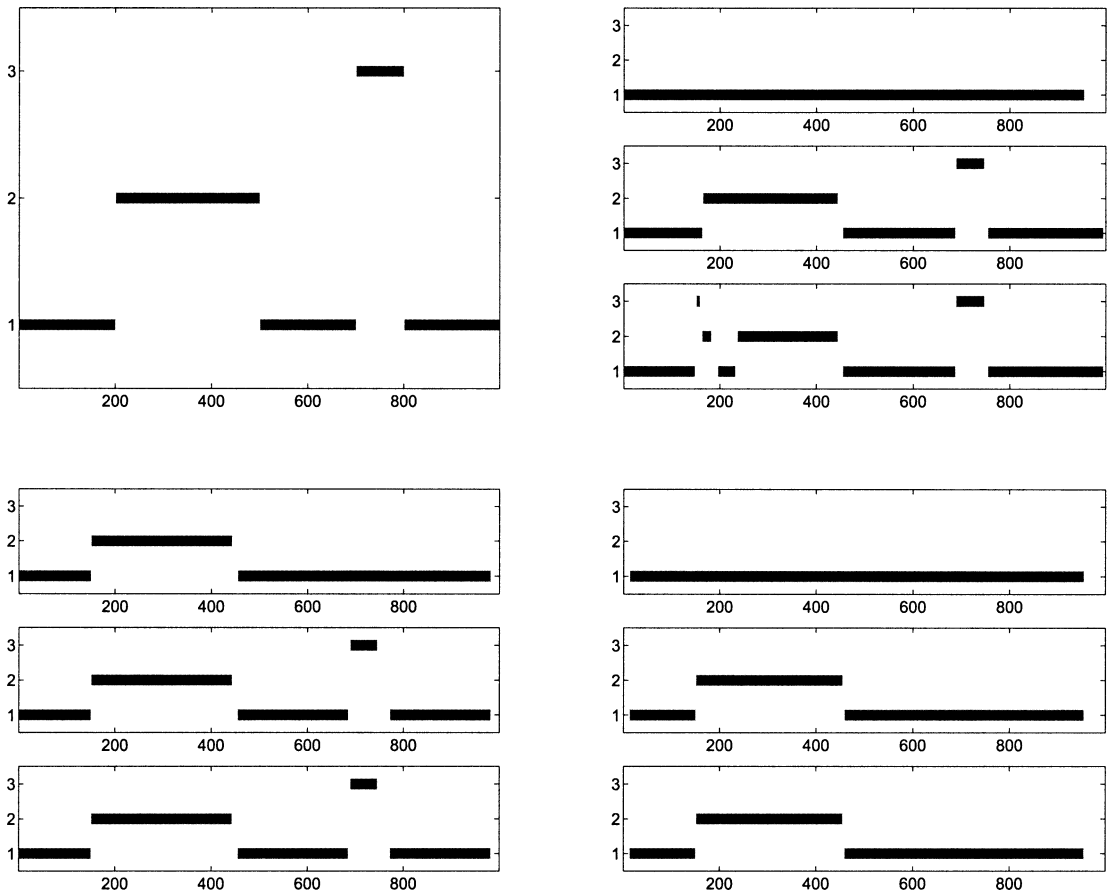

FIG. 10.—Prediction with RECPARS for mosaic structure B. The figure is explained in the legend for figure 8.

the true mosaic structure, but with two important differences. First, RECPARS gives only an accurate prediction if the recombination and substitution costs, C_{rec} and C_{mut}, have been set "appropriately." Note that these parameters can *not* be inferred from the data, but rather have to be chosen in advance. It was suggested by Wiuf, Christensen, and Hein (2001) that a ratio of the recombination and substitution costs of $C_{rec}/C_{mut} = 1.5$ works fine quite generally. However, this was not confirmed in our simulations, where for the largest tree height of 0.3 the predictions with this ratio were wrong, leading to a mosaic structure that is over-tessellated (fig. 8, top right; fig. 10, top right). Because HMM-Bayes infers all the parameters from the data, it does not suffer from this shortcoming (fig. 9, top right; fig. 11, top right). Second, even when RECPARS predicts the nature of the mosaic structure correctly, it is less accurate than HMM-Bayes in locating the breakpoints: From figures 8 and 10 it can be seen that the breakpoints predicted with RECPARS are typically misplaced by 20–30 nucleotides. This is a consequence of the fact that RECPARS uses only the topology-defining sites, and thus discards a considerable proportion of sites in the DNA sequence alignment.

When the tree height is decreased to 0.1, neither RECPARS nor HMM-Bayes predicts the mosaic structure of the alignment correctly. RECPARS finds only one recombinant region, which for the first alignment is even badly misplaced (fig. 8, bottom right). HMM-Bayes detects both recombinant regions and even locates them rather accurately, but it misclassifies the topology change for one of these regions (fig. 9, bottom right; fig. 11, bottom right). This is most likely a consequence of the fact that for small tree heights, the number of mutations and, consequently, the number of polymorphic sites is small. Thus, there is less information in the data, and any inference is inevitably less accurate.

For a more quantitative comparison between RECPARS and HMM-Bayes, recall that the detection of recombination is basically a classification problem: Each site in the sequence alignment is assigned to one of the three possible tree topologies. For RECPARS, this is done directly. For HMM-Bayes, it is done by assigning each site to the mode of the posterior probability. We use two criteria to rate the performance of the methods: The *sensitivity*, which is the percentage of correctly classified recombinant sites, and the *specificity*, which measures the percentage of correctly classified non-recombinant sites. Comparing the performance of RECPARS and HMM-Bayes across all simulations, shown in figure 12, we found that HMM-Bayes gives a consistent and significant improvement on RECPARS in the accuracy of locating and classifying the recombinant regions, as indicated by a systematically increased sensitivity score.

Figure 13 compares the predictions of RECPARS and

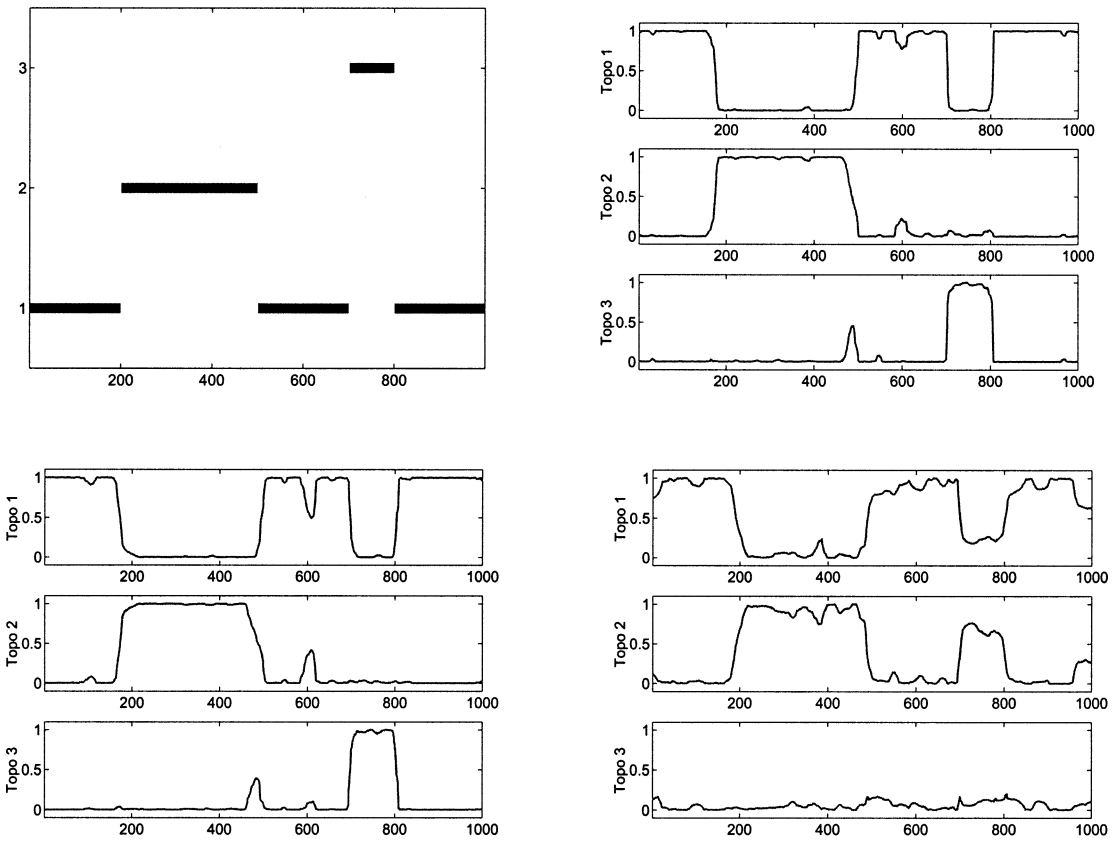

FIG. 11.—Prediction with HMM-Bayes for mosaic structure B. The figure is explained in the legend for figure 9.

HMM-Bayes on the real-world DNA sequence alignments. First, it can be seen that the predictions with RECPARS depend sensitively on the recombination-mutation cost ratio C_{rec}/C_{mut}. The best predictions show a qualitative agreement with the predictions from the literature, but note that selecting the best value of C_{rec}/C_{mut} in this way, with the benefit of hindsight, is not possible in real applications where the location of the recombinant regions is not known beforehand. Also note that setting $C_{rec}/C_{mut} = 1.5$, as suggested by Wiuf Christensen, and Hein (2001), is not guaranteed to give reliable results, as can be seen for the hepatitis B sequence alignment (fig. 13, middle left), where this parameter setting leads to an over-tessellated mosaic structure with false positive predictions of spurious recombinant regions.

Second, even when selecting the best C_{rec}/C_{mut} ratio, the agreement between the breakpoints predicted with RECPARS and those from the literature is good only for the hepatitis B sequence alignment (fig. 13, middle left). For the maize alignment (fig. 13, top left), RECPARS predicts a large uncertainty in the location of the breakpoint. This is a consequence of the lack of topology-defining sites in this part of the alignment (recall that RECPARS uses only these sites): between $t = 706$ and $t = 968$, the alignment contains only a single topology-defining site. For the *Neisseria* sequence alignment (fig. 13, bottom left), the recombinant region predicted with RECPARS is far too short.

The subfigures on the right of figure 13 show the predictions with HMM-Bayes. Note that these predictions do not depend on any heuristic tuning of parameters, since all the parameters are properly inferred from the data by sampling them from the posterior distribution (12) with MCMC.

The mosaic structures and the breakpoints predicted with HMM-Bayes for the maize and the hepatitisB sequence alignments are in agreement with the results from the literature (fig. 13, top and middle). Also, the location of the predicted recombinant region in *Neisseria* accords with the prediction by Zhou and Spratt (1992). Their second anomalous region, shown by the gap in figure 13, bottom left, is modeled by a distributed representation with HMM-Bayes; this reflects the uncertainty about the nature of this region. The only difference between the literature and the prediction with HMM-Bayes is the presence of a further peak of $P(S_t = 3 \mid \mathcal{D})$ at the end of the alignment (see fig. 13, bottom right). Because this region is very short (less than 20 bases long) and therefore difficult to detect with other methods, we assume that we have found a true recombinant region that has not been discovered before.

Comparison with the Heuristic HMM

Figure 14 shows the results obtained on one of the synthetic DNA sequence alignments (mosaic structure A, tree height 0.2). The subfigure on the left shows the prediction of $P(S_t \mid \mathcal{D})$ with HMM-heuristic. For this

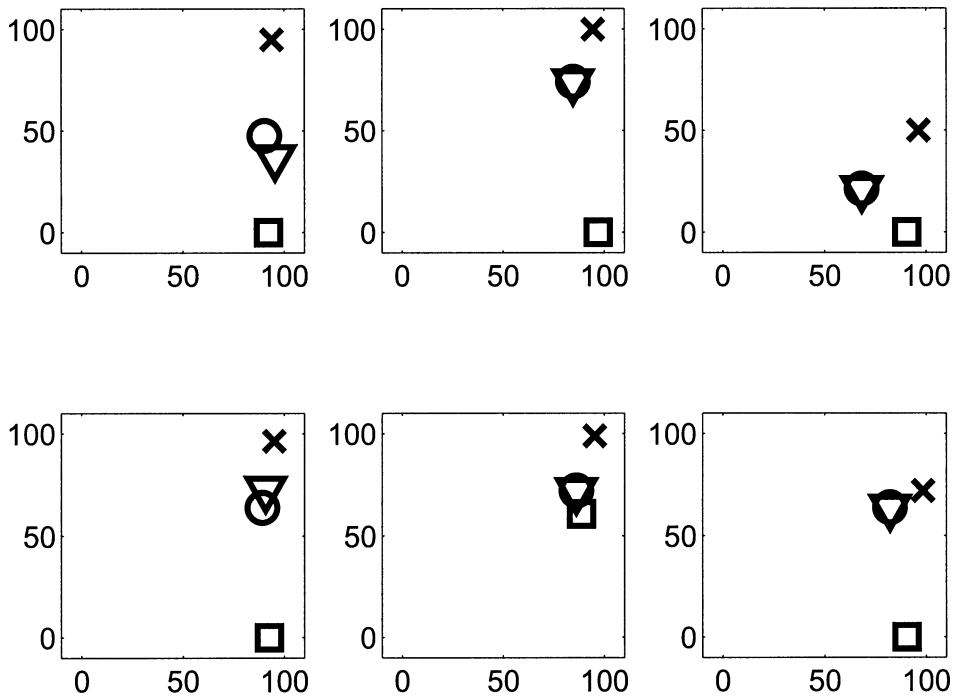

FIG. 12.—Comparison between RECPARS and HMM-Bayes on synthetic sequence alignments. Each subfigure shows a plot of the sensitivity/specificity classification scores, where the vertical axis represents the *sensitivity*, and the horizontal axis represents the *specificity*. Note that an optimal method that classifies all sites correctly has a score of 100/100, whereas a method that does not predict any recombination event has a score of 0/100. The symbols in the figure represent the different methods compared: Crosses: HMM-Bayes; squares: RECPARS, $C_{rec}/C_{mut} = 10.0$; circles: RECPARS, $C_{rec}/C_{mut} = 3.0$; triangles: RECPARS, $C_{rec}/C_{mut} = 1.5$. The subfigures show the results for the different sequence alignments. *Top row*: mosaic structure A (see figs. 8 and 9); *bottom row*: mosaic structure B (see figs. 10 and 11). *Left column*: tree height = 0.3; *middle column*: tree height = 0.2; *right column*: tree height = 0.1.

method, the recombination parameter ν has to be specified in advance, and we have set it to the mean of the prior distribution: $\nu = 0.8$. It can be seen that the overall pattern of the posterior probabilities is correct, showing an increase for topology $S_t = 2$ in the region $200 < t < 400$, and an increase for topology $S_t = 3$ 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 14, overcomes this shortcoming. The predicted state transitions coincide with the true breakpoints. The posterior probabilities for the topologies, $P(S_t \mid \mathcal{D})$, are close to zero or one, which significantly reduces the noise. By assigning each site in the alignment to the mode of $P(S_t \mid \mathcal{D})$, the tree topologies and the mosaic structure of the alignment are predicted correctly. The mean and the standard deviation of the posterior distribution $P(\nu \mid \mathcal{D})$ are $\langle \nu \rangle_{posterior} = 0.992$ and $\sigma_{posterior} = 0.004$. With four breakpoints in an alignment of length 1000 bases, the true value for the recombination parameter is $\nu = 0.996$, which deviates from the prediction by only 0.4%.

Figure 15 shows the prediction of $P(S_t \mid \mathcal{D})$ for the maize sequence alignment. The subfigures in the middle row show predictions obtained with HMM-heuristic, 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 topology $S_t = 3$. 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 15 shows the prediction with HMM-heuristic when setting ν 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 Moinz de Sa and Drouin (1996). A clear improvement is obtained with HMM-Bayes (figure 15, bottom right), which predicts a sharp transition from topology $S_t = 1$ into topology $S_t = 3$ at the location t predicted by Moinz de Sa and Drouin (1996).

Comparison with Maximum Likelihood

On the maize and hepatitis B sequence alignments, the predictions with HMM-ML and HMM-Bayes were practically indistinguishable (graphs not included in this article). The difference between the two approaches is in the confidence that we have in the prediction. The prediction with HMM-ML, $P(S_t \mid \mathcal{D}, \mathbf{w}, \boldsymbol{\theta}, \nu)$, is dependent on the model parameters $\mathbf{w}$, $\boldsymbol{\theta}$, ν, which were fitted with maximum likelihood and might therefore be subject to over-fitting. This calls for an independent statistical significance test, using, e.g., parametric bootstrapping. This approach is extremely computationally expensive, as discussed by Larget and Simon (1999). The prediction

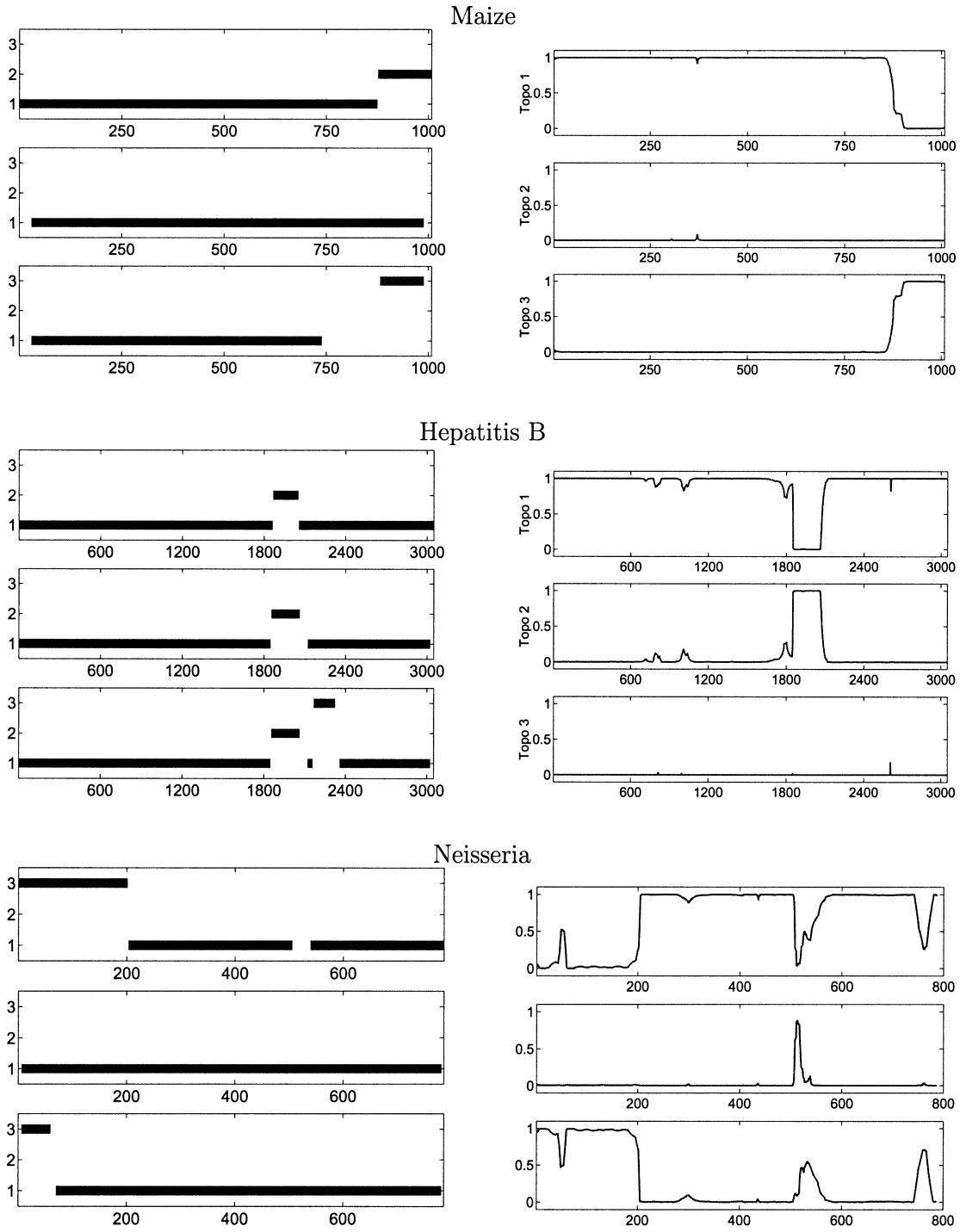

FIG. 13.—Comparison between RECPARS and HMM-Bayes on real-world sequence alignments. The figure contains six subfigures, where each subfigure comprises three graphs. The rows show the results on different sequence alignments. *Top*: Maize. *Middle*: Hepatitis B. *Bottom*: Neisseria. *Left column*: Comparison between the predictions from the literature and those obtained with RECPARS. The horizontal axis represents sites in the DNA sequence alignment; the vertical axis represents the three possible tree topologies. The *top graph* in each subfigure shows the prediction from the literature. The *middle graph* shows the prediction with RECPARS for $C_{rec}/C_{mut} = 10$. The *bottom graph* shows the prediction with RECPARS for $C_{rec}/C_{mut} = 1.5$. *Right column*: Predictions with HMM-Bayes. Each subfigure is composed of three graphs. These graphs show the posterior probabilities for the three topologies, $P(S_t = 1 \mid \mathcal{D})$ (*top*), $P(S_t = 2 \mid \mathcal{D})$ (*middle*), $P(S_t = 3 \mid \mathcal{D})$ (*bottom*), plotted along the DNA sequence alignment (the subscript t denotes the position in the alignment).

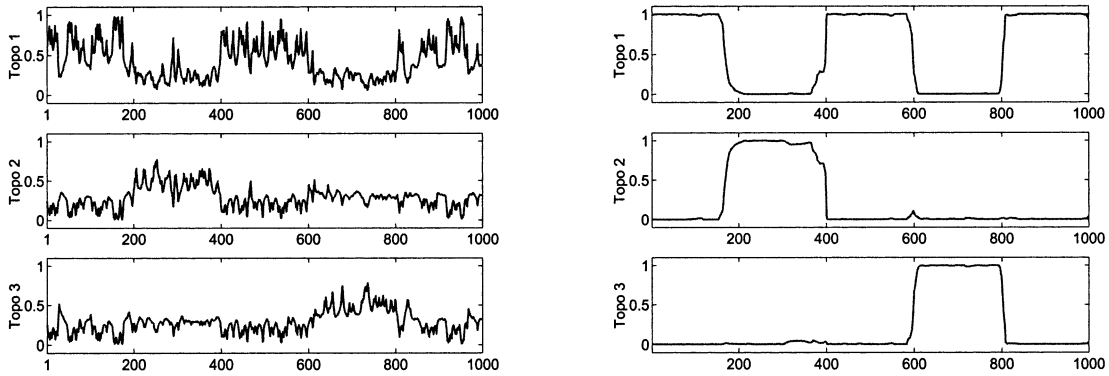

FIG. 14.—Comparison between HMM-heuristic and HMM-Bayes on the synthetic sequence alignment. The figure contains two subfigures, each of which is composed of three graphs. These graphs show the posterior probabilities for the three topologies, $P(S_t = 1 \mid \mathcal{D})$ (*top*), $P(S_t = 2 \mid \mathcal{D})$ (*middle*), $P(S_t = 3 \mid \mathcal{D})$ (*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.

with HMM-Bayes, on the other hand, is only dependent on the data, $P(S_t \mid \mathcal{D})$, because 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 (assuming sufficient convergence of the MCMC simulation).

Figure 16 shows the prediction of $P(S_t \mid \mathcal{D})$ on the *Neisseria* sequence 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 $t = 537$ might be the result of rate heterogeneity, HMM-ML predicts a recombination event with a clear transition from topology $S_t = 1$ into $S_t = 2$. This seems to be the result of over-fitting: Because the distribution of the nucleotide column vectors $\mathbf{y}_t$ in the indicated region is significantly different from the rest of the alignment, modeling 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. This deficiency is redeemed with HMM-Bayes, whose prediction is shown in figure 16, 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, $P(S_t = 1 \mid \mathcal{D})$. However, the uncertainty in the nature of this region 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 also indicates a certain model misspecification inherent in the current scheme—the absence of hidden states for representing different evolutionary rates—and thus avoids the over-fitting incurred when applying HMM-ML.

To test the conjecture that the Bayesian approach is more robust to over-fitting than maximum likelihood, we carried out two further simulation studies. In the first study, we simulated the effect of rate heterogeneity (fig. 17, top). We simulated the evolution process along the branches of a four-species phylogenetic tree with the Kimura model, as described earlier under *Data*, but reduced the rate of nucleotide substitution by a factor of 1/5 in the center region, between sites $t = 301$ and $t = 600$. The results are shown in the bottom of figure 17. The maximum likelihood approach clearly over-fits and erroneously predicts a recombinant region. The Bayesian MCMC approach is only slightly affected in its prediction of the posterior probabilities: the graph of $P(S_t = 1 \mid \mathcal{D})$ shows a small dent. This does not lead to classifying any site in the alignment as a topology different from $S_t = 1$, though, hence no recombination is predicted, and over-fitting is avoided.

The objective of the second simulation study was to test how reliable the prediction of the prediction *uncertainty* is. This is important for medical applications: When predicting that a certain HIV strain, for instance, is a mosaic of well-established strains, a pharmaceutical company would like to know how reliable this prediction is before launching an expensive drug or vaccine development project.

To this end, we first simulated a recombination process according to the method described in the *Data* section, and then simulated observational noise, corresponding to wrong base calls or typing errors in DNA sequencing, by randomly replacing 20% of the columns in the alignment by those of a second alignment that was unaffected by recombination. The process is illustrated in the top of figure 18. The consequence is that the uncertainty in the prediction of the recombinant region should increase, and the recombination event should be predicted with a probability less than 1. Figure 18 (bottom

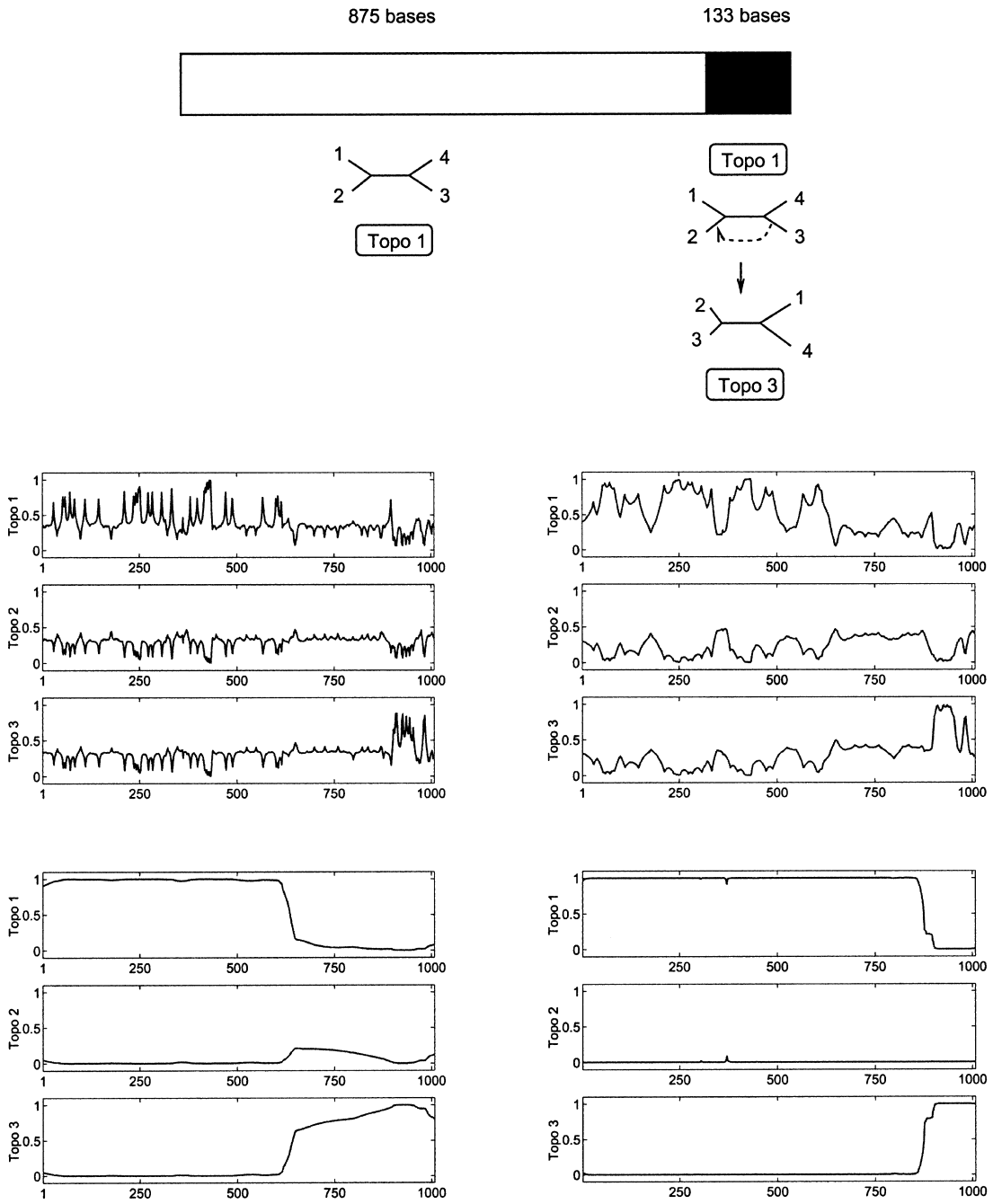

FIG. 15.—Gene conversion between two maize actin genes. *Top*: Indication of gene conversion between a pair of maize actin genes, corresponding to a transition from topology 3 into topology 1 in the first 875 nucleotides of their coding regions, has been reported by Moniz de Sa and Drouin (1996). *Bottom*: The figure contains four subfigures, each composed of three graphs, as explained in the legend for figure 14. *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.

left) shows the prediction with maximum likelihood. The location of the breakpoint (in the middle of the alignment at site 200) is fairly accurate, and the nature of the mosaic structure has been identified correctly in that a recombination event corresponding to a change from topology 1 (left) into topology 3 (right) is predicted. However, this prediction is overconfident in that the topology change is predicted with probability 1, which ignores the effect of typing errors. When applying the Bayesian approach, we found that the MCMC trajectories converged to two different semiconverged states, one corresponding to a clear recombination event and the other to the absence of any recombination. While this bimodality indicates insufficient mixing of the Markov chain, it also points to the intrinsic uncertainty in the prediction problem, caused by the introduction of typing errors (we did not observe any bimodality in the absence of typing errors). When combining several MCMC trajectories started from different initializations, we obtained the prediction shown in figure 18, bottom right, which, as opposed to the

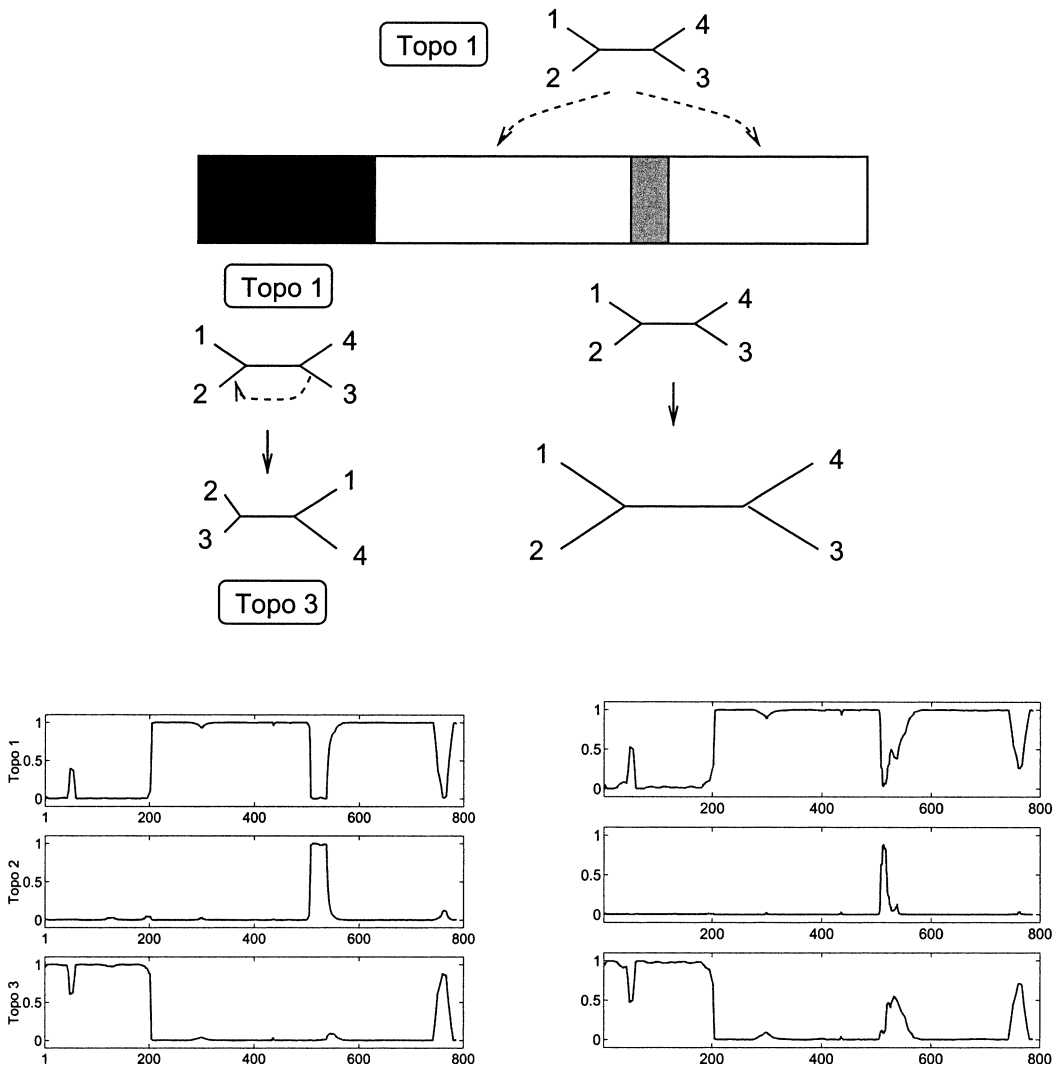

FIG. 16.—Recombination in Neisseria. *Top*: According to Zhou and Spratt (1992), a recombination event corresponding to a transition from topology 1 into topology 3 has affected the first 202 nucleotides of the DNA sequence alignment. A second more diverged region might be the result of rate variation. *Bottom*: The figure contains two subfigures, each composed of three graphs, which are explained in the legend for figure 14. *Left:* Maximum likelihood (HMM-ML). *Right*: Bayesian MCMC (HMM-Bayes).

maximum likelihood approach, indicates the intrinsic uncertainty by predicting a recombination event with probability less than one.

Comparison with the Window Methods PLATO and TOPAL

Finally, we have compared the performance of HMM-Bayes with the window methods PLATO and TOPAL. The results are shown in figure 19. For PLATO, the agreement with the "true" (meaning *true* for the synthetic data and *predicted in the literature* for the real data) locations is poor. This is most likely a consequence of the principled shortcoming of PLATO: Because the recombinant regions are rather long, they have a substantial impact on the estimation of the reference tree, causing the Q-statistic of equation (1) to lose power and rendering the detection method unreliable.

The last two rows of figure 19 show the DSS statistic of TOPAL for the two window sizes $W = 100$ (third row) and $W = 200$ (bottom row). It can be seen that the results depend critically on this parameter, which has to be chosen sufficiently large. For $W = 100$, the agreement between the predicted and the true locations of the breakpoints is poor. Doubling the window length, $W = 200$, gives a qualitatively correct prediction of the breakpoints, except for a spurious peak at the beginning of the hepatitis B alignment. Notice that a short window of $W = 100$ is not recommended. However, increasing the window size to $W = 200$ degrades the spatial resolution and leads to a larger uncertainty in locating the breakpoints.

On comparing these methods with HMM-Bayes (figs. 13 and 14), we find that the latter gives more accurate predictions than PLATO and more precise locations of the breakpoints than TOPAL, with the further advantage that all parameters are estimated from the data, and no arbitrary window parameter has to be chosen in advance.

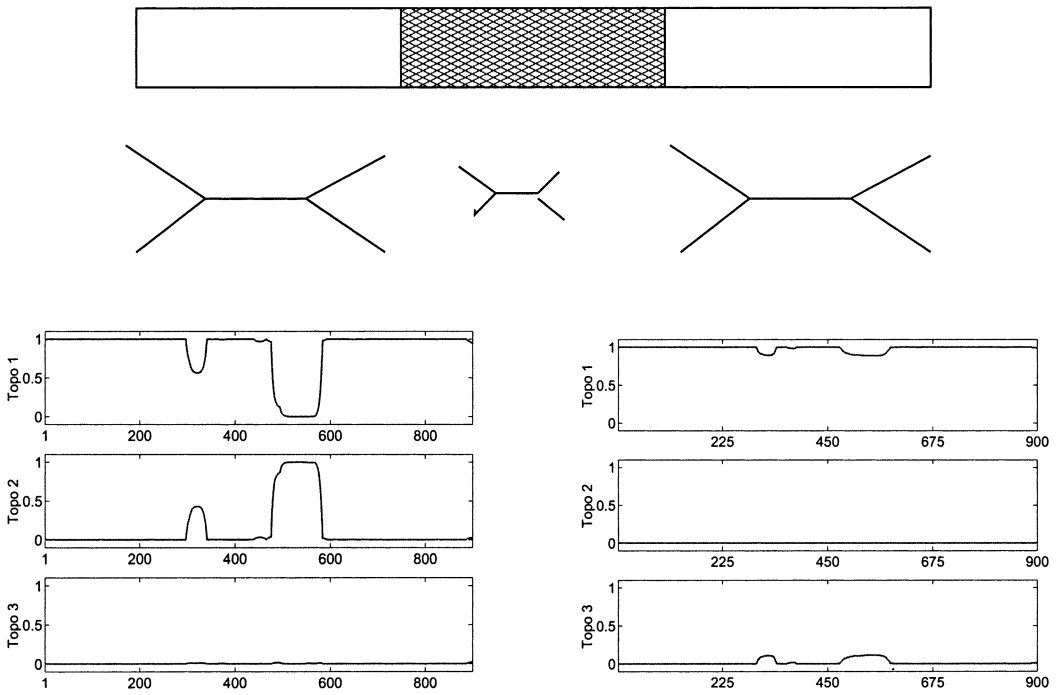

FIG. 17.—Comparison between HMM-ML and HMM-Bayes on sequence alignments subject to rate variation. *Top*: Simulation experiment. An alignment of 900 nucleotides was generated by simulating the evolution process along the branches of a four-species phylogenetic tree with the Kimura model. Rate heterogeneity was simulated by reducing the nucleotide substitution rate by a factor of 1/5 in the center region, that is, between sites $t = 301$ and $t = 600$. *Bottom*: Results. The figure contains two subfigures, each composed of three graphs. These graphs show the posterior probabilities for the three topologies, $P(S_t = 1 \mid \mathcal{D})$ (*top*), $P(S_t = 2 \mid \mathcal{D})$ (*middle*), $P(S_t = 3 \mid \mathcal{D})$ (*bottom*), plotted against the sites t of the DNA sequence alignment. *Left*: Prediction with HMM-ML. *Right*: Prediction with HMM-Bayes.

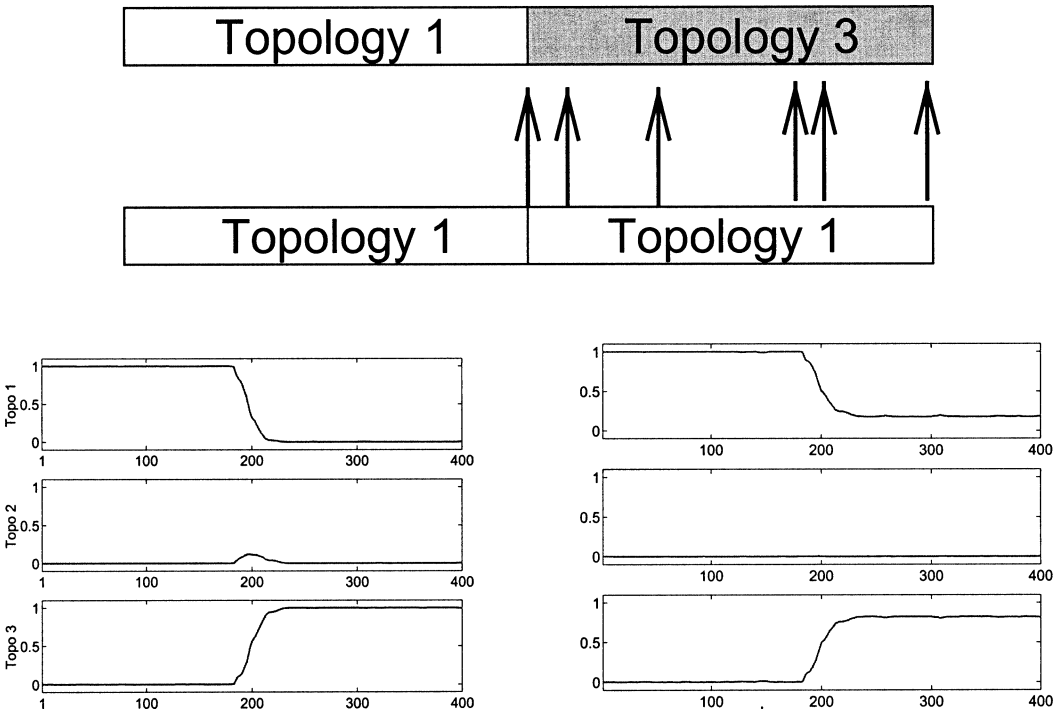

FIG. 18.—Comparison between HMM-ML and HMM-Bayes on noisy mosaic structures. *Top*: Simulation experiment. Recombination in a population of 4 taxa was simulated according to figure 7. To simulate noise, we took a second alignment that was unaffected by recombination, and randomly replaced 20% of the columns of the first alignment by those of the second. *Bottom*: Results. The figure contains two subfigures, each composed of three graphs. These graphs show the posterior probabilities for the three topologies, $P(S_t = 1 \mid \mathcal{D})$ (*top*), $P(S_t = 2 \mid \mathcal{D})$ (*middle*), $P(S_t = 3 \mid \mathcal{D})$ (*bottom*), plotted against the sites t of the DNA sequence alignment. *Left:* Prediction with HMM-ML. *Right:* Prediction with HMM-Bayes.

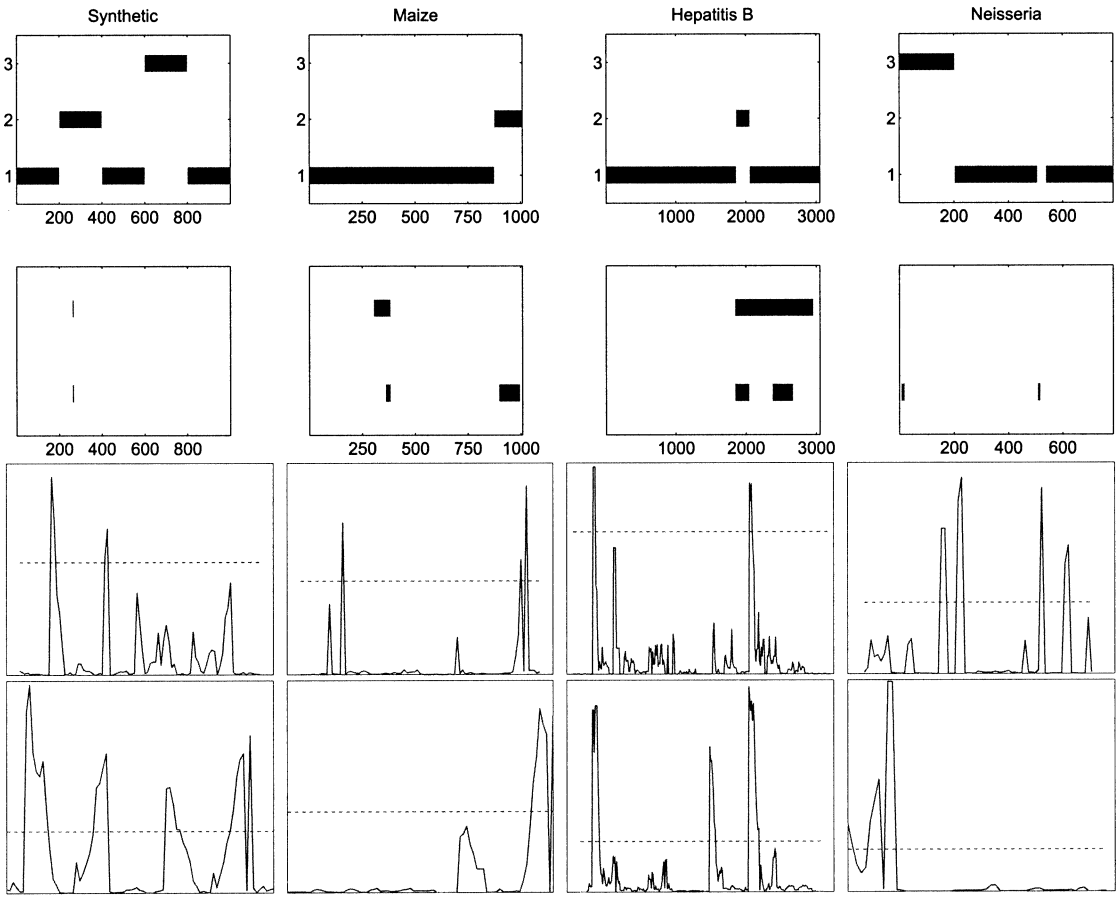

FIG. 19.—Mosaic structures predicted with PLATO and TOPAL. The *top row* shows the "true" mosaic structures of the DNA sequence alignments. For the synthetic DNA sequence alignment, this is the correct mosaic structure, which is known. For the real-world DNA sequence alignments, these are the mosaic structures predicted in the cited literature. In each subfigure, the horizontal axis represents sites in the sequence alignment, and the vertical axis shows the three possible tree topologies. The *second row* shows the recombinant regions predicted with PLATO, where the Q-statistic of (eq. 1) is significantly larger than under the null hypothesis of no recombination. Each figure shows two different predictions for different models of rate heterogeneity. *Bottom:* Uniform rate. *Top:* Gamma distributed rates. Again, the horizontal axis represents sites in the sequence alignment. The *last two rows* show the predictions with TOPAL for a window size of 100 bases (*third row*) and a window size of 200 bases (*fourth row*). In each subfigure, the horizontal axis represents the site t in the alignment, the vertical axis represents the DSS statistic, and the horizontal (dashed) line shows the 95 percentile under the null hypothesis of no recombination. The *columns* represent the different alignments. From left to right: Synthetic, maize, hepatitis B virus, and *Neisseria*.

Discussion

In this article, we have proposed a Bayesian MCMC method (HMM-Bayes) for detecting recombination with HMMs. This follows up on earlier work by Hein (1993); McGuire, Wright, and Prentice (2000); and Husmeier and Wright (2001), where the parameters were not estimated (RECPARS), were estimated heuristically (HMM-heuristic), or were estimated with maximum likelihood (HMM-ML). We have compared the methods on various synthetic and real-world DNA sequence alignments and found that HMM-Bayes leads to a considerable improvement on RECPARS and HMM-heuristic in predicting the nature and breakpoints of recombinant regions. All model parameters are properly inferred from the data, removing the need for arbitrarily setting these parameters by hand in advance. In comparison with the older maximum likelihood scheme (HMM-ML), the Bayesian approach has been found to be more robust against over-fitting, and to give a more reliable estimation of the uncertainty of the prediction.

Note that our approach focuses on topology changes rather than recombination in general. If a recombination event only changes the branch lengths of a tree, without affecting its topology, it will not be detected. However, the main motivation for our method is a prescreening of an alignment for topology changes as a crucial prerequisite for a consistent phylogenetic analysis. Most standard phylogenetic methods are based on the implicit assumption that the given alignment results from a single phylogenetic tree. In the presence of recombination they would therefore infer some "average" tree. If the recombination event leads to different trees with the same topology but different branch lengths, then the wrong assumption of having only one tree is not too dramatic: The topology will still be correct (as it has not been changed by the recombination event), and the branch lengths will show some average value. This is still

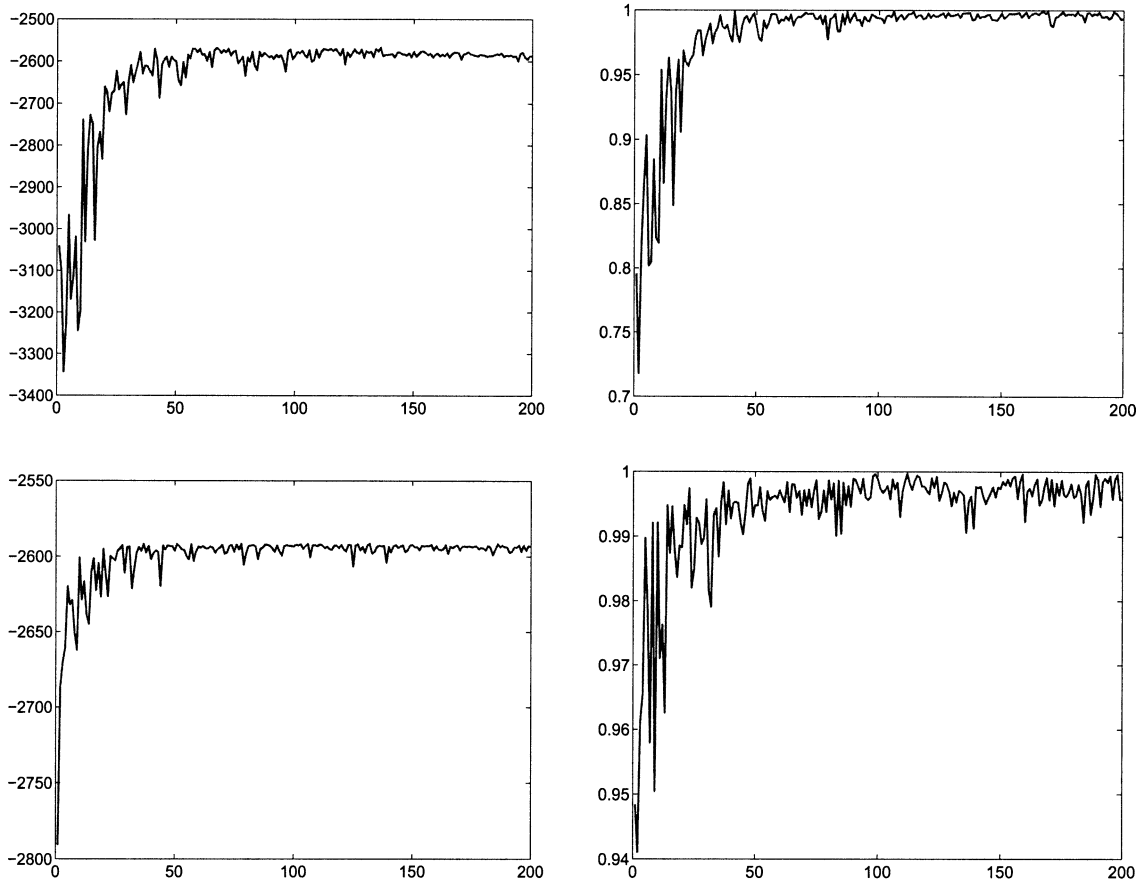

FIG. 20.—Monitoring convergence of the Markov chain. The figure shows various trace plots obtained from MCMC simulations on the maize sequence alignment. The subfigures in the *left column* show trace plots of the log unnormalized posterior, that is, the log likelihood plus the log prior. The subfigures in the *right column* show trace plots of the recombination parameter ν. In each case, the horizontal axis represents MCMC steps divided by 1000, that is, the total simulation length was 200,000 MCMC steps. The rows correspond to different simulations. *Top row*: Prior on ν with mean $\mu = 0.8$. *Bottom row*: Prior on ν with mean $\mu = 0.95$ (see fig. 6). In both cases, ν was initially set to its prior mean. Note the different scales on the vertical axes.

reasonable, because branch lengths are continuous numbers, and the mean of continuous numbers is well defined. If the recombination event, however, leads to a change of the topology, inferring an average tree is no longer reasonable: Tree topologies are cardinal entities, for which an average value is *not* defined. In fact, it is well known that in this case the resulting average tree will be in a distorted "limbo" state between the dominant and recombinant trees, which renders the whole inference scheme unreliable. Thus, by focusing on recombination-induced topology changes we focus on those events that cause the main problems with standard phylogenetic analysis methods.

We should further point out that our approach is not about estimating recombination rates, but rather about inferring the mosaic structure of a particular sequence alignment. The proposed method has a parameter that might be confused with a recombination rate: the parameter ν, which should actually be referred to as a recombination probability (the probability that on moving from the nth site to the $(n + 1)$th site in the sequence alignment no topology change occurs). This parameter corresponds to the recombination cost in RECPARS; it is *not* a recombination rate in population genetics terms. By estimating this parameter from the data—that is, the sequence alignment—the prediction performance of the algorithm improves considerably, thereby overcoming the main shortcomings of RECPARS and HMM-heuristic, where this parameter has to be chosen arbitrarily in advance. We do not claim, however, that ν has any meaning in itself: it is a parameter whose proper estimation from the sequence alignment improves the detection of phylogenetic topology changes, and this is what we are interested in.

A limitation of the proposed method is that the states of the HMM represent only 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

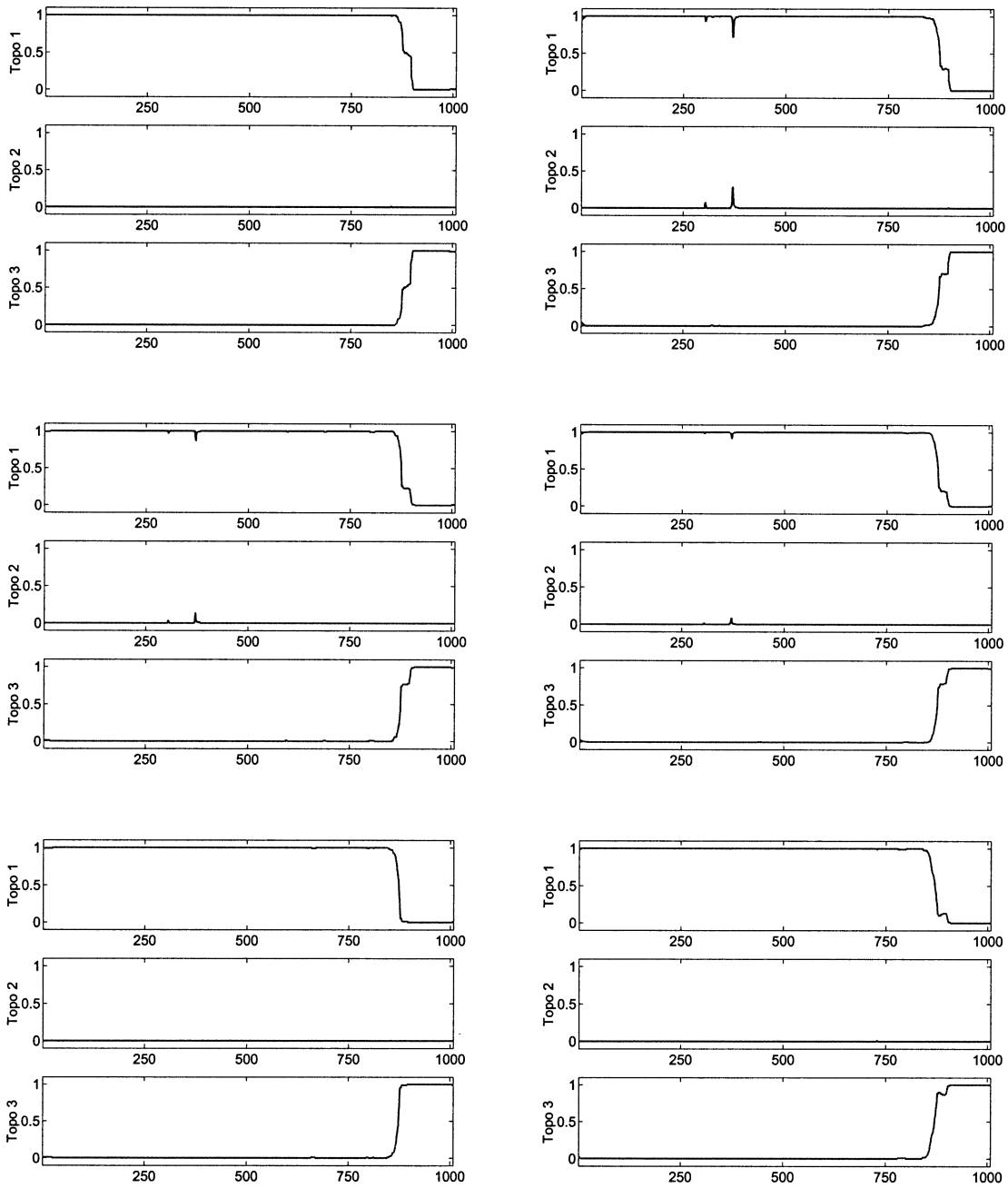

FIG. 21.—Dependence of the prediction on the prior and the initialization. The figure compares the results on the maize alignment obtained from different priors and different initializations. The figure contains six subfigures, each composed of three graphs. These graphs show the predicted posterior probabilities for the three topologies, $P(S_t = 1 \mid \mathcal{D})$ (*top*), $P(S_t = 2 \mid \mathcal{D})$ (*middle*), $P(S_t = 3 \mid \mathcal{D})$ (*bottom*), plotted along the DNA sequence alignment (the subscript t denotes the position in the alignment). The six subfigures are arranged in three rows and two columns, where the *rows* represent different priors $P(\nu)$ (shown in fig. 6), while the columns represent different initializations of the topology sequences $\mathbf{S} = (S_1, \ldots, S_N)$. ν was initially set to its prior mean. *Top row*: Prior with $\mu = 0.8$. *Middle row*: Prior with $\mu = 0.9$. *Bottom row*: Prior with $\mu = 0.95$. *Left column*: Initialization of $\mathbf{S}$ from RECPARS with $C_{rec}/C_{mut} = 10.0$. *Right column*: Initialization of $\mathbf{S}$ from RECPARS with $C_{rec}/C_{mut} = 1.5$.

complexity (Ghahramani and Jordan, 1997), it seems that this increase will be less dramatic for the MCMC method, because the Gibbs sampling scheme of (eq. 17) can be applied to both types of hidden states separately. A detailed investigation of this approach is the subject of future research.

As mentioned earlier under *Method: Background and Earlier Approches*, the method presented here is restricted to DNA sequence alignments with small numbers of taxa, and our current software implementation can only deal with alignments of four sequences. This is because each possible tree topology constitutes a separate state of the HMM. In practical applications, our method has therefore to be combined with a fast low-resolution preprocessing method, like RECPARS or TOPAL: In a first, preliminary analysis, apply RECPARS or TOPAL to identify putative

recombinant sequences and their approximate mosaic structures. In a second, subsequent step, apply HMM-Bayes to the tentative sets of four sequences that result from the previous step. This will allow a more accurate analysis of the mosaic structure than can be obtained from a window method like TOPAL with its inherently low resolution, and it will resolve contradictions that are likely to arise from different (arbitrary) settings of the parsimony cost parameters of RECPARS. In general, after identifying a small set of putative recombinant sequences with any fast low-resolution method, the exact nature of the recombination processes and the location of the breakpoints can be further investigated with the high-resolution method proposed in this article.

Note that, in principle, our method can deal with more than four sequences. All that is required is that the set of different candidate topologies, which constitute the states of the HMM, is sufficiently small (fewer than 10, say). Such a sparse set of candidate topologies can be obtained from the preliminary analysis with one of the lower-resolution methods. The proposed Bayesian HMM method can then be applied in a subsequent step, with the hidden states set to the topologies obtained from the preliminary analysis. This will, obviously, not be able to detect any topology changes not detected before, but it would still be likely to give an improvement on the low-resolution method of the previous step in that recombination breakpoints and the nature of the mosaic structure would be predicted more accurately.

A more ambitious goal would be to improve the methodology itself by introducing a more informative form of the transition probabilities between the topologies. In the current version, all changes into other topologies are equally likely, as expressed by (eq. 5). For four sequences with only three possible tree topologies, this seems to be a valid assumption. However, on increasing the number of sequences, this uniform prior on the topologies leads to a superexponential explosion of the parameter space. Choosing a more informative prior that, given a topology, is only nonzero for closely related topologies—as suggested by Hein (1993) and implemented in RECPARS—might offer a way to apply the proposed method to alignments of more than four sequences. This approach, however, has not yet been explored and will certainly offer an interesting topic for future research.

Appendix

To monitor the convergence of the MCMC simulations, we inspected trace plots, like those in figure 20, to check whether the MCMC trajectories had reached stationarity. We then repeated the simulations from different initializations, as shown in figure 21, and tested for consistency of the results.

Figure 21 shows the dependence of the predictions on both the prior and the initialization. All predicted posterior probabilities $P(S_t \mid \mathcal{D})$ are similar in that they show a transition from topology 1 into topology 3 around position $t = 875$. The graphs differ slightly in the exact form of this transition. This implies that when deriving from the posterior probabilities a crisp classifier that flags a recombinant region when $P(S_t = 3 \mid \mathcal{D}) > 0.5$, the predicted breakpoints may vary slightly between the simulations. In fact, we found that the predictions for the breakpoints varied between $t = 771$ and $t = 781$, whereas Moniz de Sa and Drouin (1996) predicted a breakpoint at $t = 775$. This is not surprising. All simulations predict uncertainty in the immediate neighborhood of $t = 775$. Consequently, when estimating the breakpoint from the posterior probabilities, this estimate itself is subject to this uncertainty. Note, however, that the uncertainty inherent in our estimation is of the order of ± 5 nucleotides, which is considerably more precise than that of the alternative detection methods discussed in the article.

Acknowledgments

This work was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) Bioinformatics Initiative and the Scottish Executive Environmental and Rural Affairs Department (SEERAD). The research leading to this paper was carried out in part at the Isaac Newton Institute, Cambridge, United Kingdom. We would like to thank Frank Wright and Karen Ayres for helpful discussions. We are also grateful to the associated editor and to four anonymous referees for their detailed comments on earlier versions of this manuscript.

Literature Cited

Bollyky, P. L., A. Rambaut, P. H. Harvey, and E. C. Holmes. 1996. Recombination between sequences of hepatitis B virus from different genotypes, J. Mol. Evol. **42**:97–102.

Casella, G., and E. I. George. 1992. Explaining the Gibbs sampler. Am. Statist. **46**:167–174.

Chib, S., and E. Greenberg. 1995. Understanding the Metropolis-Hastings Algorithm. Am. Statist. **49**:327–335.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Statist. Soc. **B39**:1–38.

Felsenstein, J. 1981. Evolution trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1988. Phylogenies from molecular sequences: inference and reliability, Annu. Rev. Genet. **22**:521–565.

Felsenstein, J., and G. A. Churchill. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13**:93–104.

Ghahramani, Z., and M. I. Jordan. 1997. Factorial hidden Markov models, Machine Learn. **29**:245–273.

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

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

Heckermann, D. 1999. A tutorial on learning with Bayesian networks. Pp. 301–354 *in* M. I. Jordan, ed. *Learning in Graphical Models*, Adaptive computation and machine learning. MIT Press, Cambridge, Massas.

Hein, J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. **36**:396–405.

Husmeier, D., and F. Wright. 2001. Detection of recombination in DNA multiple alignments with hidden Markov models. J. Comput. Biol. **8**:401–427.

Kimura, M. 1980. A simple method for estimating evolutionary

rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. Science **220**:671–680.

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

Mau, B., M. A. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55**:1–12.

Maynard Smith, J. 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. **34**:126–129.

McGuire, G., and F. Wright. 2000. TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. Bioinformatics **16**:130–134.

McGuire, G., F. Wright, and M. J. Prentice. 1997. A graphical method for detecting recombination in phylogenetic data sets. Mol. Biol. Evol. **14**:1125–1131.

———. 2000. A Bayesian method for detecting recombination in DNA multiple alignments. J. Comput. Biol. **7**:159–170.

Moniz de Sa, M., and G. Drouin. 1996. Phylogeny and substitution rates of angiosperm actin genes. Mol. Biol. Evol. **13**:1198–1212.

Neal, R. M. 1996. Bayesian learning for neural networks, vol. 118, Lecture notes in statistics. Springer-Verlag, New York.

Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**: 257–286.

Robert, C. P., G. Celeux, and J. Diebolt. 1993. Bayesian estimation of hidden Markov chains: A stochastic implementation. Statist. Prob. Lett. **16**:77–83.

Rubinstein, R. Y. 1981. Simulation and the Monte Carlo method. John Wiley & Sons, New York.

Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Wiuf, C., T. Christensen, and J. J. Hein. 2001. A simulation study of the reliability of recombination detection methods. Mol. Biol. Evol. **18**:1929–1939.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method, Mol. Biol. Evol. **14**:717–724.

Zhou, J., and B. G. Spratt. 1992. Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene, Mol. Microbiol. **6**:2135–2146.

Naruya Saitou, Associate Editor

Accepted October 11, 2002