A Bayesian Model for Detecting Past Recombination Events in DNA Multiple Alignments.

Gráinne McGuire(1), Frank Wright(2) and Michael J. Prentice(3)

(1) Biomathematics and Statistics Scotland, JCMB, King's Buildings, Edinburgh, EH9 3JZ, Scotland

(2) Biomathematics and Statistics Scotland, Scottish Crop Research Institute, Dundee, DD2 5DA, Scotland

(3) Department of Maths and Statistics, JCMB, King's Buildings, University of Edinburgh, EH9 3JZ, Scotland

Abstract

Most phylogenetic tree estimation methods assume that there is a single set of hierarchical relationships among sequences in a data set for all sites along an alignment. Mosaic sequences produced by past recombination events will violate this assumption and may lead to misleading results from a phylogenetic analysis due to the imposition of a single tree along the entire alignment. Therefore, the detection of past recombination is an important first step in an analysis. A Bayesian model for the changes in topology caused by recombination events is described here. This model relaxes the assumption of one topology for all sites in an alignment and uses the theory of Hidden Markov models to facilitate calculations, the hidden states begin the underlying topologies at each site in the data set. Changes in topology along the multiple sequence alignment are estimated by means of the maximum *a posteriori* (MAP) estimate. The performance of the MAP estimate is assessed by application of the model to data sets of four sequences, both simulated and real.

Introduction

Phylogenetic analysis is concerned with the accurate reconstruction of the evolutionary history of a set of sequences. Given a multiple alignment of DNA sequences from different taxa, various methods exist which may be used to infer a phylogenetic tree (for an overview see Felsenstein, 1988 or Swofford et al., 1996). However, these procedures make the assumption that there is only one true tree or one set of hierarchical relationships for a given set of sequences. If this assumption is violated (e.g., through recombination) then the underlying model of inter-sequence relationships used by these methods will be invalid, and the resulting phylogenetic tree will often be misleading. Thus, the detection of recombination events is important if the evolutionary history of a set of DNA sequences is to be correctly inferred.

There have been various attempts in the literature to detect evidence of recombination events in alignments of DNA sequences. Stephens (1985) and Sawyer (1989) use the distribution of polymorphic sites along a sequence to detect recombination (but not recombination breakpoints), while Maynard Smith (1992) proposes the maximum chi-square test which finds the breakpoint maximising the chi-square statistic of sites before and after supporting different relationships among the sequences. Salminen et al. (1995) and Lawrence and Hartl (1992) use bootstrapping; the former detect recombination breakpoints by means of their graphical method, *bootscanning*, while the latter merely detect the presence of recombination within a data set. A graphical method was suggested by McGuire et al. (1997); this detects putative recombination breakpoints by finding changes in the local tree along a multiple alignment using a sliding window. PLATO (Grassly and Holmes, 1997) detects recombination by considering the likelihoods at each site in the alignment for the maximum likelihood tree. A set of sequential sites, with significantly lower site likelihoods may represent a recombination event. Recently, the homoplasy test (Maynard Smith and Smith, 1998) was suggested; this detects evidence of recombination but cannot find breakpoints, or estimate the number of recombination events that have occurred.

One possible way to model recombination within an alignment is to use split decomposition (Bandelt and Dress, 1992). This is a method which allows conflicting groupings or splits within a set of sequences and represents this information as a network diagram. A recombination event is suggested when sequences are linked by a network. However, Bandelt and Dress (1992) note that it is not obvious how to distinguish between random or systematic error within a data set, and recombination events, so this may not be the most powerful tool for detecting recombination.

Hein (1993) considers changes in the most parsimonious tree along an alignment. Given a particular topology (branching pattern), he describes the set of possible topologies that can arise from it assuming one or more recombination events have occurred. He then considers the problem in terms of a graph. Each node (i, T) consists of the data in the i^{th} column of the alignment and a given topology T. The node is assigned a weight, $w(i, T)$, the weight of position i given topology T. An edge connects two neighbouring nodes, i and $i-1$, and is assigned a weight $d(T, T')$, the recombinational distance between T (the topology at position i) and T' (the topology at position $i-1$). $W(i, T)$ is the weight of the most parsimonious history of the first i positions, given that the topology at position i is T. The most parsimonious history of the sequences is the path of lowest weight from node 1 to node N, where N is the total sequence length, the total weight being found by summing the weights of the nodes and the edges. This is given by $W(N, T)$ and is found by the following recursion:

$$\begin{aligned} W(1,T) &= w(1,T) && (i=1) \\ W(i,T) &= \min_{T'} \{W(i-1,T') + d(T,T') + w(i,T)\} && (i>1). \end{aligned} \tag{1}$$

Thus, the sequence will start in a particular topology and will only change topologies when it becomes worthwhile to do so. A sensible choice of $d(T, T')$ will prevent changes occurring too frequently.

While the graphical representation of this problem is intuitive, it uses the maximum parsimony principle, and is likely to suffer from the problems inherent with maximum parsimony (e.g., its inconsistency under certain circumstances; Felsenstein, 1978).

This paper describes a Bayesian model for changes in topology due to recombination events. As such, it uses maximum likelihood methodology to form part of the posterior distribution, while a Markov chain is used to place a prior probability on a particular sequence of topologies along a multiple alignment. To facilitate computations, the theory of Hidden Markov models is applied. Hidden Markov models have been previously used by Felsenstein and Churchill (1996) to incorporate unknown rate variation into likelihood calculations for phylogenetic trees and by Thorne et al. (1996) to predict secondary structure from protein alignments. Indeed, the potential use of Hidden Markov models to tackle the recombination problem has previously been discussed by Felsenstein and Churchill (1996) and N. Goldman (personal communication). The resulting model is conceptually very similar to Hein's (1993) parsimony approach. As before, the nodes will be the columns of the multiple alignment. $w(i, T)$ will correspond to the likelihood at each site given a particular topology while $d(T, T')$ will be replaced by the Markov chain probabilities.

Method

Consider an alignment of four DNA sequences, one of which has incorporated genetic material from another at some point in the past. Thus, this recombination event will result in a change of topology in the affected region of the alignment. If the likelihoods at each site for each possible topology are calculated and compared, one topology should correspond to the highest likelihoods in the non-recombinant region, while another should take over in the recombinant zone. Some noise would be expected (i.e., incorrect topologies corresponding to the largest site likelihoods for single or for short runs of sites) but prior beliefs about the relative infrequency of recombination events suggests that short runs of a particular topology should be ignored. Since both prior information and likelihoods have been discussed, this suggests a Bayesian approach to the problems of detecting evidence of recombination events.

Given an alignment of T DNA sequences, each N bp long, this may be considered as a $T \times N$ matrix $\mathbf{S}$, with each column of the matrix, $\mathbf{S}_t$, representing the nucleotides in each sequence at site t. Suppose that $\{C_t\}$, $t = 1, 2, \ldots, N$, is a process which allocates an (unobservable) topology c_t, $c_t = 1, 2, \ldots, m$, to $\mathbf{S}_t$, where m is the number of distinct topologies. Since recombination events may be detected by their effect on the topology along an alignment, it is clear that the problem of detecting evidence of recombination becomes that of estimating the most probable combination of topologies $c_1, \ldots, c_N$ given the data $\mathbf{S}$.

The process $\{C_t\}$ may be thought of as encoding prior information about recombination events, while the alignment, $\mathbf{S}$, may be used to calculate the likelihood given a particular sequence of topologies, $c_1, c_2, \ldots, c_N$ along an alignment. Using the fundamental result that the posterior distribution is proportional to the product of the prior distribution and the likelihood, the following model for topology change due to recombination events is obtained:

$$\begin{aligned}\mathrm{Prob}(C_1 = c_1, C_2 = c_2, \ldots, C_N = c_N|\mathbf{S}) \propto\ & \mathrm{Prob}(C_1 = c_1, \ldots, C_N = c_N) \\ & \times \mathrm{Prob}(\mathbf{S}|C_1 = c_1, \ldots, C_N = c_N).\end{aligned} \tag{2}$$

Therefore, to obtain the posterior distribution for a particular sequence of topologies, $c_1, \ldots, c_N$, along a multiple alignment, a prior distribution for sequences $c_1, \ldots, c_N$ and a likelihood for the data given a particular sequence of $c_1, \ldots, c_N$ must first be specified. For ease of notation, note that $C_1 = c_1, \ldots, C_N = c_N$ will generally be abbreviated to $c_1, \ldots, c_N$ below.

Likelihood

Likelihood methods for constructing phylogenetic trees generally make the assumption that each site evolves independently of each other. Hence, the likelihood calculation simplifies to

$$\mathrm{Prob}(\mathbf{S}|c_1, \ldots, c_N) = \prod_{j=1}^{N} \mathrm{Prob}(\mathbf{S}_j|c_j). \tag{3}$$

At first glance, this calculation appears simple: for each site, $\mathbf{S}_t$, in the multiple alignment, the likelihoods for all possible topologies are calculated using, for example, the pruning algorithm outlined by Felsenstein (1981) and implemented in the program DNAML from his PHYLIP package (Felsenstein, 1993). However, to calculate the likelihood values, appropriate values for the branch lengths must be chosen. It is not immediately obvious how these should be estimated.

One suggestion would be to use the entire data set to estimate the branch lengths for each possible topology since the more data used, the smaller the variance of the branch lengths estimates should be.

Suppose, however, that a short recombination event relative to the entire sequence has occurred. The correct signal for the branch lengths for the recombinant topology may be swamped by the incorrect signal from the sites in the non-recombinant region, leading to incorrect estimates. It is possible that the resulting site likelihoods in the recombinant region will be reduced as a result. Thus, the method will lose power.

This problem could be avoided by calculating the site likelihoods using subsets of the data, e.g., subalignments of 50 bp, or less. This should minimise conflicting signal from the data but will increase the variance of the branch lengths. It is not immediately clear whether a trade-off is necessary between these two effects or whether one will dominate the other. Studies involving four-sequence simulated data sets and some real data sets containing known recombination events (data not shown) suggest that for homogeneous data sets (those obeying a single model of nucleotide substitution), reducing the variance of the branch lengths is the more important. Therefore, for such data sets (used in a simulation study below), the entire sequence length will be used to estimate the branch lengths. It must be noted, however, that the results using small subsets of the data for branch length estimation and likelihood did not greatly decrease the efficiency of the model (data not shown).

For real data sets, containing various heterogeneities in the model of nucleotide substitution, conflicting phylogenetic signal appeared to be the more important effect. Hence, small subsets of the data (5–50 bp) were a sensible choice for likelihood calculations. This suggests a rule of thumb: unless it can be stated with a reasonable degree of certainty that the data set is homogeneous apart from putative recombination events, small subsets should be used in the calculation of the site likelihoods.

Prior

The prior distribution for the sequence of topologies for a data set of length N bp, should specify a probability for every possible sequence of N numbers, with the number at each position taking a value in $\{1, 2, \ldots, m\}$, where m is the total number of possible topologies. It should incorporate prior knowledge about recombination events, e.g., that recombination is a relatively infrequent process. In mathematical terms, this translates roughly into dependence of the value at a point t on the values in its neighbourhood.

One way of incorporating a limited form of dependence into the prior distribution is to use a discrete, first-order Markov model. This is a model having the property that

$$\text{Prob}[N(t+1)|N(t), N(t-1), \ldots] = \text{Prob}[N(t+1)|N(t)],$$

where $t \in \mathbb{N}_0$. In other words, the state of the process at time $t+1$ depends only on the current state of the process, $N(t)$.

To define a first-order Markov chain for the sequence of topologies, C_t, $1 \leq t \leq N$, the transition probabilities, p_{ij}, that a chain in state i is in state j after one time step may be specified. This has been done as follows:

$$p_{ij} = \lambda\delta_{ij} + (1-\lambda)f_j \tag{4}$$

where f_j is the stationary frequency of topology j, $j = 1, 2, \ldots, m$;

δ_{ij} is the Kronecker delta function (1 if $i = j$; 0 otherwise).

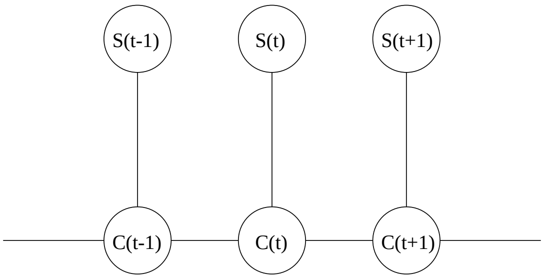

Figure 1: Conditional independence graph of the recombination model.

λ is a value between 0 and 1, representing the difficulty of changing topology, with a value of 0 representing no dependence on the topology at the previous site (an easy change of topology), while a value of 1 makes it impossible to switch topologies. Using (4), the prior probability of a particular sequence $c_1, \ldots, c_N$ is

$$\text{Prob}(C_1 = c_1, \ldots, C_N = c_N) = f_{c_1} p_{c_1 c_2} p_{c_2 c_3} \ldots p_{c_{N-1} c_N}. \tag{5}$$

It is difficult to select a vague prior. While the stationary frequencies of all the possible topologies may be allowed to be equal, a value must be selected for λ and this may introduce a degree of subjectivity. Therefore, it will be necessary to investigate the sensitivity of results to the choice of prior.

Posterior

It is now possible to obtain an expression for the posterior distribution. Substituting (5) and (3) into (2) yields the following:

$$\text{Prob}(c_1, \ldots, c_N | \mathbf{S}_1, \ldots, \mathbf{S}_N) \propto f_{c_1} p_{c_1 c_2} \ldots p_{c_{N-1} c_N} \prod_{j=1}^{N} \text{Prob}(\mathbf{S}_j | c_j). \tag{6}$$

To be of any use, it must be possible to use the posterior distribution to obtain inferences about the presence of recombinants in the DNA multiple alignment. It might be thought that to find, for example, the maximum *a posteriori* (MAP) estimate (i.e., that sequence of topologies which maximises the posterior probability), it would be necessary to evaluate the probability of every possible combination and select the largest one, a calculation which will not usually be feasible in practice. Fortunately the posterior distribution has some elegant mathematical properties.

The Bayesian model may be represented by the conditional independence graph shown in Figure 1. In such a graph, the absence of an edge between two vertices indicates that the two variables in question are conditionally independent given the other variables. The graph displays the independence of the observations $\{\mathbf{S}_t\}$ given the states of the Markov process $\{C_t\}$; it also shows the conditional independence of C_{t+1} and C_{t-1} given C_t (the Markov property). A model with this dependency structure is known as a Hidden Markov model, a very useful model with a wide range of applications, e.g., as a tool for discrete time series analysis (MacDonald and Zucchini, 1997) or in the speech processing literature (Juang and Rabiner, 1991). The advantages of this model lie in the fact that algorithms have been developed that make certain calculations (e.g., finding the MAP estimate , or calculating the renormalisation constant for the posterior distribution in eq. 6 above) feasible. These algorithms are described below.

A sensible way to infer the presence of recombination events in a data set is to use the MAP estimate, i.e., to find $\widehat{c}_1, \dots, \widehat{c}_N$ which maximise

$$\text{Prob}(C_1 = c_1, \dots, C_N = c_N | \mathbf{S}_1, \dots, \mathbf{S}_N). \tag{7}$$

This is referred to as the *global decoding* problem in the speech processing literature, and may be solved using the Viterbi algorithm (see Juang and Rabiner, 1991; MacDonald and Zucchini, 1997, p. 65). An outline of this method is given below.

Finding the MAP estimate is equivalent to maximising the joint probability $\text{Prob}(C_1 = c_1, \dots, C_N = c_N, \mathbf{S}_1, \dots, \mathbf{S}_N)$ which is also equivalent to maximising

$$f_{c_1} p_{c_1 c_2} \dots p_{c_{N-1} c_N} \prod_{j=1}^{N} \text{Prob}(\mathbf{S}_j | C_j = c_j). \tag{8}$$

Define the quantities

$$R^{(N)}_{c_N} = \text{Prob}(\mathbf{S}_N | C_N = c_N) \tag{9}$$

and

$$\begin{aligned} R^{(t)}_{c_t} &= \max_{c_{t+1}, \cdots, c_N} \text{Prob}(\mathbf{S}_t = \mathbf{s}_t, \dots, \mathbf{S}_N = \mathbf{s}_N, C_{t+1} = c_{t+1}, \dots, C_N = c_N | C_t = c_t) \\ &= \max_{c_{t+1}, \cdots, c_N} \text{Prob}(c_{t+1}, \dots, c_N | c_t) \text{Prob}(\mathbf{S}_t = \mathbf{s}_t, \dots, \mathbf{S}_N = \mathbf{s}_N | c_t, \dots, c_N) \end{aligned} \tag{10}$$

where, for ease of notation, the event that $C_t = c_t$ may also be represented simply as c_t. Note that $R^{(t)}_{c_t}$ gives the maximisation over $c_{t+1}, \dots, c_N$ of the probability of observing a particular sequence of topologies from site t to the end of the sequences given that site t has topology c_t.

The maximisation of (8) may be simplified by noting that the following recursion exists between $R^{(t)}_{c_t}$ and $R^{(t+1)}_{c_{t+1}}$:

$$R^{(t)}_{c_t} = \text{Prob}(\mathbf{S}_t = \mathbf{s}_t | c_t) \max_{c_{t+1}} \left[p_{c_t c_{t+1}} R^{(t+1)}_{c_{t+1}} \right], \tag{11}$$

with starting point $R^{(N)}_{c_N}$. By applying the recursion repeatedly from $t = N-1, N-2, \dots, 2, 1$, the quantity $R^{(1)}_{c_1}$ is obtained for all possible values of c_1. Selecting the largest of the quantities $f_{c_1} R^{(1)}_{c_1}$ gives the relative size of the maximum probability specified in (8).

To find the MAP estimate (the global decoding step), note that from (11), for each topology c_t at site t, the topology c^*_{t+1} at the next site, $t+1$, which maximises the contribution to the posterior probability is known. Once the topology $\widehat{c}_1$ at the first site which maximises the posterior probability is known, (11) gives $\widehat{c}_2$, and then $\widehat{c}_3$ and so on. Hence, the algorithm requires another pass through the data, this time from positions 1 to N.

This algorithm only yields the relative size of the maximum probability since the posterior distribution has not been renormalised. The renormalisation constant, K, may be found by summing the contributions of possible terms, or more easily by noting that

$$K = \sum_{i=1}^{m} \alpha_1(i) \beta_1(i) \tag{12}$$

where

$$\begin{aligned}\alpha_1(i) &= \mathrm{Prob}(\mathbf{S}_1 = \mathbf{s}_1, C_1 = i)\\ &= \mathrm{Prob}(C_1 = i)\mathrm{Prob}(\mathbf{S}_1 = \mathbf{s}_1 | C_1 = i)\\ \beta_t(i) &= \mathrm{Prob}(\mathbf{S}_{t+1} = \mathbf{s}_{t+1}, \ldots, \mathbf{S}_N = \mathbf{s}_N | C_t = i)\end{aligned}$$

implying that

$$\beta_1(i) = \mathrm{Prob}(\mathbf{S}_2 = \mathbf{s}_2, \ldots, \mathbf{S}_N = \mathbf{s}_N | C_1 = i).$$

$\beta_1(t)$ may be found using the following recursion

$$\beta_t(i) = \sum_{j=1}^{m} \mathrm{Prob}(\mathbf{S}_{t+1} = \mathbf{s}_{t+1} | C_{t+1} = j)\beta_{t+1}(j)p_{ij},$$

with starting point $\beta_N(i)$, assumed by convention to be equal to one. Note that this algorithm is one possible way to calculate the renormalisation constant using the forward $[\alpha_t(i)]$ and backward $[\beta_t(i)]$ probabilities. More details are in MacDonald and Zucchini (1997, p. 59).

The forward and backward probabilities may also be used to calculate the probability that the sequence of topologies changes state at a particular point. The probability that site $t-1$ has topology i while site t has topology j given the data, $\mathrm{Prob}(C_{t-1} = i, C_t = j | \mathbf{S}_1, \ldots, \mathbf{S}_N)$, is proportional to the joint probability $\mathrm{Prob}(C_{t-1} = i, C_t = j, \mathbf{S}_1, \ldots, \mathbf{S}_N)$ which can be written as

$$\begin{aligned}\mathrm{Prob}(\mathbf{S}_1, \ldots, \mathbf{S}_{t-1}, C_{t-1} = i)p_{ij}\mathrm{Prob}(\mathbf{S}_t | C_t)\mathrm{Prob}(\mathbf{S}_{t+1}, \ldots, \mathbf{S}_N | C_t = j)\\ = \alpha_{t-1}(i)p_{ij}\mathrm{Prob}(\mathbf{S}_t | C_t)\beta_t(j).\end{aligned} \tag{13}$$

Since the forward and backward probabilities may be found using the recursions described above, this quantity may be easily calculated, assuming that the posterior distribution has been renormalised. This probability could be used to find a range of sites for a putative recombination breakpoint.

Programs to implement these calculations for DNA multiple alignments for four sequences have been written in C and in unix Bourne shell scripts. The likelihood calculations are carried out using DNAML from the PHYLIP package. Further details or information on how to obtain these programs may be found at http://www.bioss.sari.ac.uk/~frank/Genetics.

Findings

To investigate the effectiveness of this model and to determine the sensitivity of the results to the choice of the prior distribution (i.e., the choice of λ), a small simulation study was carried out. Data sets of four sequences were used, there being only three possible unrooted topologies to consider in this case. Trees were simulated according to the phylogeny shown in Figure 2. The value of x was chosen to be 0.05 (a realistic branch length for many potentially recombinant sequences), while sequences were 1000 bp long. The data were simulated using a Kimura two parameter model, with a transition-transversion ratio of 2. The sequences were evolved along the internal branch, and then

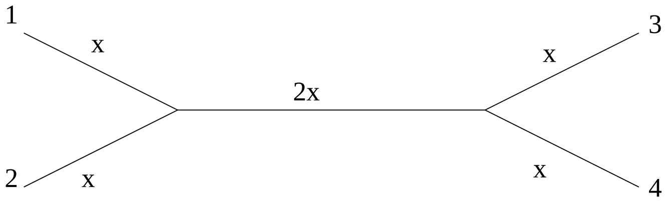

Figure 2: The tree used to simulate recombinant data sets. The length of each of the exterior branches is x while the length of the interior branch is $2x$.

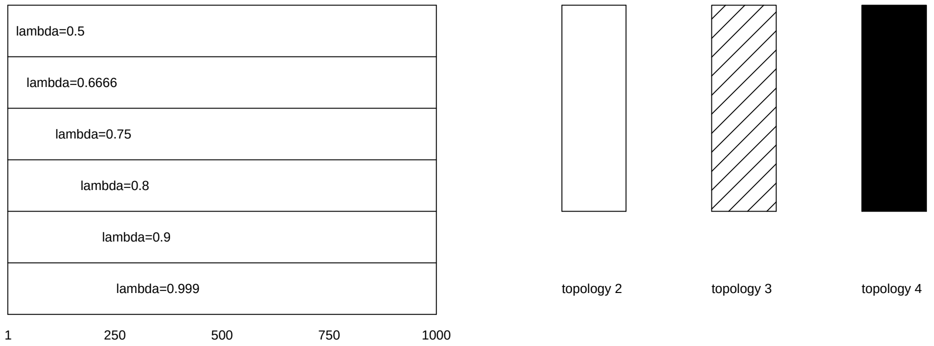

Figure 3: Key to the graphs in Figure 4. The graph on the left shows the horizontal axis, depicting location along the 1000 bp sequence, while the horizontal blocks correspond to particular values of λ. On the right, the different shadings, representing each of the three topologies, are shown.

along the external branches until a fraction b of all nucleotide substitutions had occurred. Then a recombination event was generated, with a region of sequence 3 replacing the same region in sequence 1. Following that, the sequences were evolved along the remainder of the length of the branches ($x[1-b]$ substitutions per position). Values of b were 0.25 (an early recombination event, difficult to detect) and 0.75 (an easier event to detect). Recombinant regions of lengths 400, 200 and 100 bp were used.

For each set of conditions, five data sets were simulated. Here, however, only the results from one data set is reported, since these are generally representative of all the data sets. Full details of the results from an extended simulation study are contained in McGuire (1998). Six different priors were used for each data set, corresponding to λ taking the values 0.5, $0.\dot{6}$, 0.75, 0.8, 0.9 and 0.999. Since these data sets are homogeneous apart from the recombination event, the entire alignment is used to estimate the branch lengths. The key to the graphs and the shading representing each topology is shown in Figure 3. Topology 2 represents the branching pattern with sequences 1 and 2 clustering together, topology 3 has sequences 1 and 3 together and topology 4 has sequences 1 and 4 together. In the recombinant region, topology 3 (the hatching) is the underlying topology while topology 2 is the true topology elsewhere. Topology 4 (the solid filling) is valid nowhere. The results of the simulation study are shown in Figure 4. Note that the dotted lines in these graphs represent the exact recombination breakpoints.

In general, the Bayesian model succeeds at detecting evidence of recombination events for a wide range of conditions. Only the estimation of the shortest (100bp) and more distant ($b = 0.25$) recom-

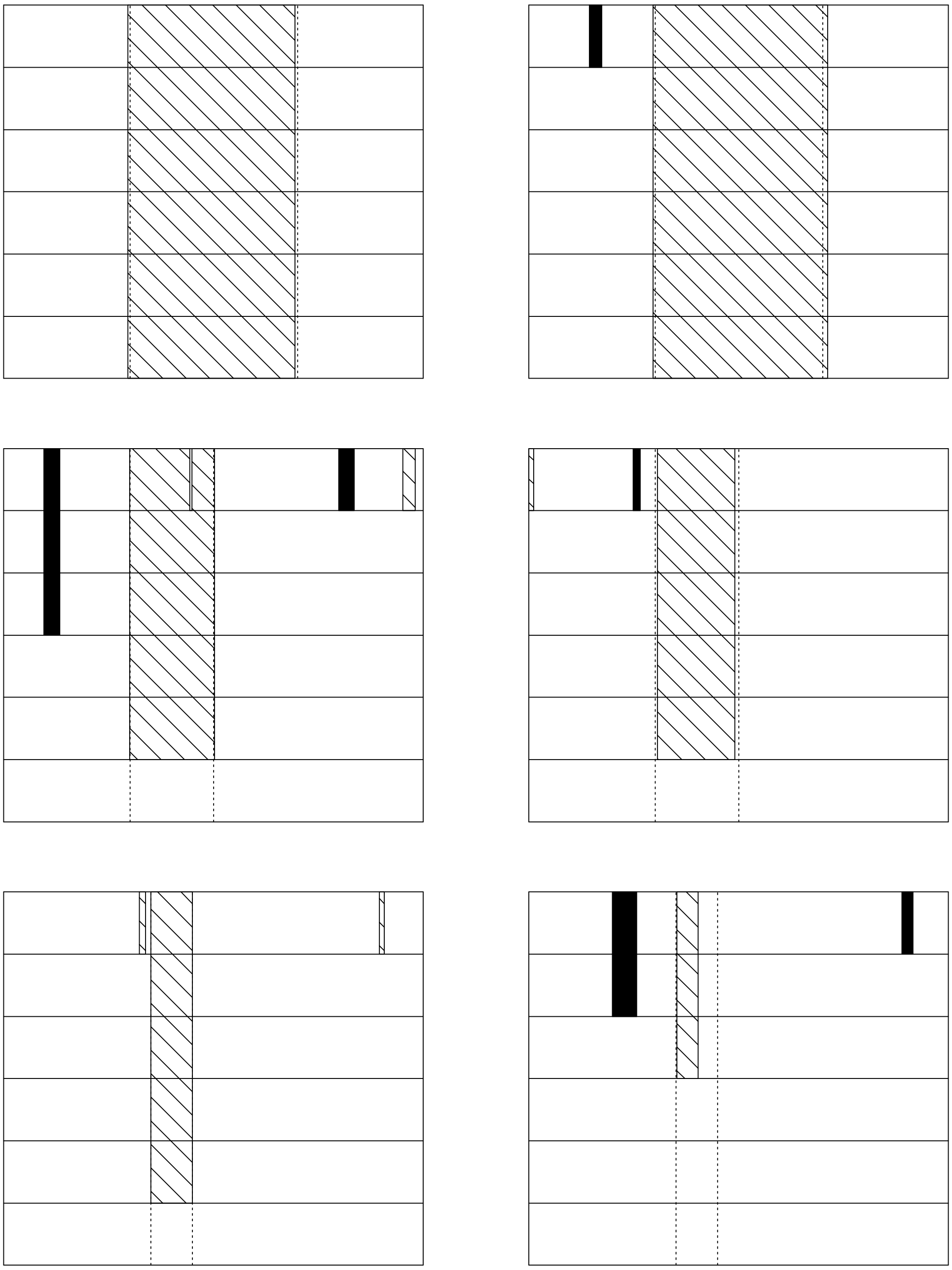

Figure 4: Top graphs: results for a recombinant region of 400 bp, occurring between the dotted lines. Middle graphs: results for a recombinant region of 20 0 bp, occurring between the dotted lines. Bottom graphs: results for a recombinant region of 100 bp, occurring between the dotted lines. The graphs on the left represent recent recombination events ($b = 0.75$), while those on the right correspond to distant events ($b = 0.25$).

bination event is poor; however, this is a very difficult event to detect so this is not surprising. The model is most successful for the longer and more recent events, detecting both the breakpoints and the correct resulting change in topology. Again, this is as expected.

The recombination events are inferred for a wide range of values of λ, suggesting that inferences are not unduly sensitive to the choice of the prior distribution. Low values of λ sometimes exhibit a 'patchy' effect (incorrect, short switches in topology); these arise due to the relative ease of changing topologies. For the shorter recombinant regions, higher values of λ can lead to no recombination event being detected, since there is not sufficient support from the likelihood to overcome the prior belief of topology changes being very rare. Thus, intermediate values of λ would appear to be the most useful. In practice, different values of λ should be used. Topology changes which are consistent over a range of values of λ suggest a recombination event, whereas other changes may be nothing more than the 'patchy' effect.

Application to a real data set

The model described above is now applied to a real data set, containing a known recombinant event. The data used are a subset of the *Neisseria argF* gene data set, described by Zhou and Spratt (1992). They use a data set containing eight *Neisseria* strains to detect the presence of recombination in *N. meningitidis*. Four of these sequences are used here, these being: *N. gonorrhoeae* (GenBank/EMBL accession number X64860); *N. meningitidis* (X64866); *N. cinerea* (X64869) and *N. mucosa* (X64873). The sequences were aligned using CLUSTAL W (Thompson et al., 1994) taking the default settings, yielding an alignment of length 787 bp. Following the numbering scheme of Zhou and Spratt (1992), the first nucleotide is labelled as 296 bp while the last is 1082 bp.

According to Zhou and Spratt (1992), there are two anomalous, or more diverged regions in the DNA alignment. These occur at positions 296–497 bp and 802–833 bp. In the rest of the alignment, *N. meningitidis* clusters with *N. gonorrhoeae* (later referred to as topology 1), while between 296 bp and 497 bp, they found that it is grouped with *N. cinerea* (topology 3). They were unable to determine the cause of the other diverged region (802–833 bp).

Before applying the Bayesian model, some parameters must be estimated. The Felsenstein 84 (Felsenstein and Churchill, 1996) model was used for the site likelihood calculations, with the stationary nucleotide frequencies (π_i, $i = A, C, G, T$) estimated as $\pi_A = 0.26$, $\pi_C = \pi_G = 0.28$ and $\pi_T = 0.18$. The transition-transversion ratio was estimated as 2.3 using the program PUZZLE (Strimmer and von Haeseler, 1996). The six previously mentioned values of λ were used. The branch lengths were estimated using subsets of 5 bp, since this data set appears to be quite heterogeneous (indeed using the entire data set to obtain the branch length estimates led to an incorrect recombination event being inferred; data not shown). The results are shown in Table 1.

Apart from the patchiness in the results when $\lambda = 0.5$, the method finds the known recombination event successfully over a wide range of values of λ. It also correctly identifies the change in topology, with the sequence of topologies at each site starting with topology 3, then changing to topology 1. The method is not successful at identifying the shorter diverged region. This is not surprising as Zhou and Spratt (1992) were unable to determine the cause of this diverged region; if it is a recombination

Table 1: MAP estimates of recombination events for the *Neisseria* data set

λ	0.5	$0.\dot{6}$	0.75	0.8	0.9	0.999
	296–342(3)[a]					
	357–498(3)	296–498(3)	296–498(3)	296–498(3)	296–498(3)	296–498(3)
	827–864 (2)					

[a] Figures in brackets show the topology at this range of sites while unspecified sites have topology 1

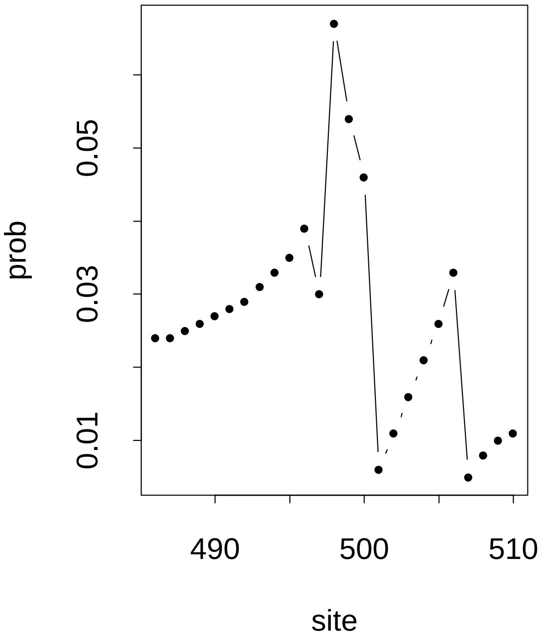

Figure 5: The probabilities that the topology changes from topology 3 to topology 1, given the data, at sites 486 to 510 along the *Neisseria* multiple alignment.

event, the recombinant DNA does not appear to originate from any of the strains in their data set. There is a change in topology towards the end of this diverged region when $\lambda = 0.5$. This may be picking up genuine information in the data, or it may be an artifact due to the low value of λ. Since it does not persist for some of the higher values of λ, a reasonable decision would be to ignore it.

The probabilities of the topology changing from state 3 to state 1 (calculated using eq. 13) at sites in a neighbourhood of the 498th site (the estimated location of the change of topology) are investigated. A value of 0.8 is used for λ. The probabilities at all sites from 486 to 510 are shown in Figure 5. The probability of change at site 498 is considerably higher than all of the surrounding sites apart from sites 499 and 500, suggesting a reasonable level of confidence in the estimated location of the recombination breakpoint. Note that the apparently low value of the probability of change at site 498 (0.066) is due to the very low transition probability of changing topologies in the Markov chain prior. When $\lambda = 0.8$, this probability is $0.0\dot{6}$. The near equivalence in the two probability values stems from the fact that product of the the terms in (13) apart from $p_{497,498}$ is close to one as the likelihood at site 498 given topology 1 contributes the most to the posterior distribution at that site.

Discussion

This paper proposes a Bayesian model for changes in the topology along an alignment caused by recombination events. It has the potential to be a useful tool for inference. Currently, the maximum *a posteriori* (MAP) estimate is used for inferences, with the probabilities of a change in the underlying topology given the data providing an adhoc way of estimating a range of values for the location of a putative recombination breakpoint. However, it is possible that a Monte Carlo Markov Chain approach may be used to simulate from the posterior distribution allowing more scope for inferences. This is being investigated. If such an approach is feasible, it may remove the need for the model to be structured as a Hidden Markov model since the MAP estimate could also be found using MCMC. This would allow a greater range of choices of prior distribution, and thus would permit a more elaborate dependency structure to be incorporated.

One possible improvement to the prior might be to place a hyper-prior on λ to remove the subjective step of selecting an appropriate value. Assuming MCMC methods (for example) would make this model tractable, it is unclear whether such an approach is sensible.

To explain this, consider the maximisation of the posterior probability in (6) over λ (this is equivalent to putting a uniform hyper-prior on λ). The object now is to find the combination of topologies and the value of λ which maximises the posterior probability. To investigate the consequences of this approach, three data sets with different recombination events were generated as described above. The branch length value x in Figure 2 was again 0.05, while the Kimura two parameter model of nucleotide substitution was used (transition-transversion ratio of 2). The recombination events occurred three quarter of the way along the exterior branches and were of lengths 400 bp, 200 bp and 100 bp. The branch lengths for the site likelihoods calculations were estimated using subalignments of 50 bp.

For each of the three data sets, the maximum posterior probability was found for values of λ ranging between 0 and 1. The results are shown in Figure 6. In all cases, the posterior probability is highest when $\lambda = 1$. For the highest values of λ, no recombination event is found for any of the data sets, although λ gets very close to one before this happens for the 400 bp recombinant data set. This occurs because the increase in site likelihoods caused by allowing for the recombination event does not offset the very small transition probabilities of change when λ takes on values close to one. A sufficiently high value of λ will mean that the recombination event is not detected by the MAP estimate. Hence, many choices of hyper-prior are likely to lead to a value of $\lambda \approx 1$ being estimated and correspondingly, no recombination event will be detected. It might be possible to obtain sensible results by using a hyper-prior which places a very small probability on λ being high, but it is questionable whether this is worth the effort given the ease of finding the MAP estimate over a sensible range of λ. In addition, choosing such a hyper-prior is subjective, so that problem still remains.

The application has only been described for data sets of four sequences in this paper, since this reduces the computational burden as the model has only three hidden states (topologies). It should be possible to extend the model to deal with more sequences by restricting the possible topologies considered at a site to the set of possible topologies which could arise from the existing topology T following a recombination event. Hein (1993) uses this approach in his parsimony based method, considering only those topologies which could have arisen from the existing one following one recombination event

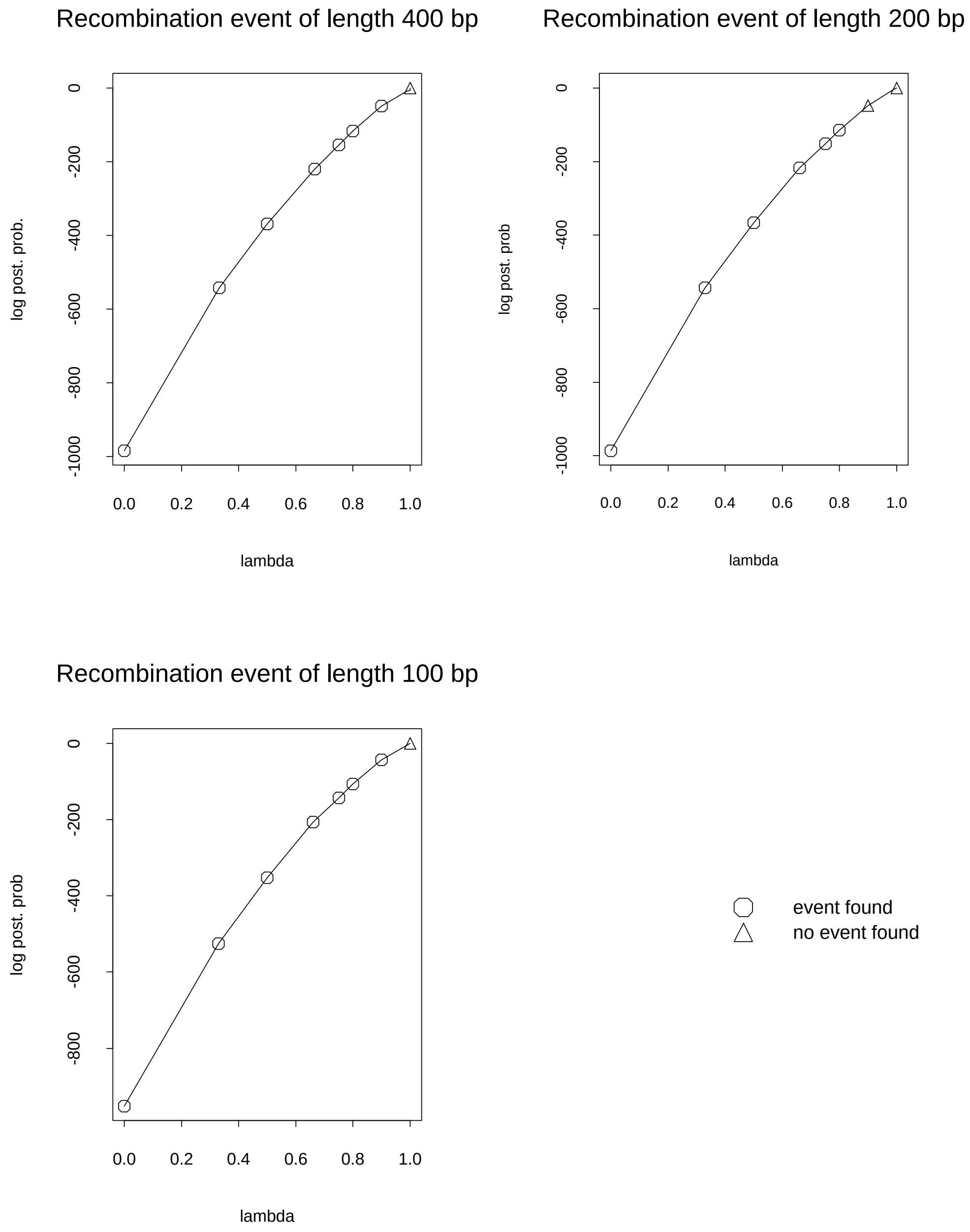

Figure 6: The values of the log posterior probability for different values of λ. A triangle means that the MAP estimate did not find any recombination event for this value of λ.

for larger numbers of sequences. This is the subject of current work.

Initial exploration suggests that this Bayesian model may be applied to short data sets (e.g., 300-400 bp long) with a reasonable amount of success at detecting recombination events. This is potentially a very useful application (e.g., analysis of short HIV sequences) and is being investigated further.

Acknowledgements

We would like to thank Nick Goldman for useful discussion and Joe Felsenstein for explaining the workings of DNAML. Alan Bleasby also provided useful computing advice. This work was funded by the Scottish Office Agriculture, Environment and Fisheries Department and the Isaac Newton Institute for Mathematical Sciences, Cambridge.

References

Bandelt, H. and Dress, A. W. M. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phyl. Evol.*, **1**, 242–252.

Felsenstein, J. (1978). Cases in which parsimony or compatability methods will be positively misleading. *Syst. Zool.*, **27**, 401–410.

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

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Ann. Rev. Genet.*, **22**, 521–565.

Felsenstein, J. (1993). Phylip. Version 3.5c, University of Washington, Seattle. http://evolution.genetics.washington.edu/phylip.html.

Felsenstein, J. and Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, **13**, 93–104.

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

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

Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov Models for speech recognition. *Technometrics*, **33**, 251–272.

Lawrence, J. G. and Hartl, D. L. (1992). Inference of horizontal genetic transfer from molecular data: an approach using the bootstrap. *Genetics*, **131**, 753–760.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman and Hall, London.

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

Maynard Smith, J. and Smith, N. H. (1998). Detecting recombination from gene trees. *Mol. Biol. Evol.*, **15**, 590–599.

McGuire, G. (1998). *Statistical Methods in DNA Sequence Analysis: Detection of Recombination and Distance Estimation*. PhD thesis, University of Edinburgh.

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

Salminen, M. O., Carr, J. K., Burke, D. S., and McCutchan, F. E. (1995). Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res. Hum. Retrov.*, **11**, 1423–1425.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Mol. Biol. Evol.*, **6**, 526–538.

Stephens, J. C. (1985). Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol. Biol. Evol.*, **2**, 539–556.

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

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. In Hillis, D. and Moritz, C., editors, *Molecular Systematics*, pages 407–514. Sinauer Associates, Sunderland, Mass., second edition.

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

Thorne, J. L., Goldman, N., and Jones, D. T. (1996). Combining protein evolution and secondary structure. *Mol. Biol. Evol.*, **13**, 666–673.

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