Are you my mother? Bayesian phylogenetic inference of recombination among putative parental strains

Janet S Sinsheimer,1,2,3 Marc A Suchard,2 Karin S Dorman,4,5 Fang Fang,3 Robert E Weiss3

1Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA; 2Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA; 3Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA, USA; 4Department of Statistics, Iowa State University, Ames, IA, USA; 5Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA, USA

Abstract: Reconstructing evolutionary relationships using Bayesian inference has become increasingly popular due to the ability of Bayesian inference to handle complex models of evolution. In this review we concentrate on inference of recombination events between strains of viruses when these events are sporadic, are rare relative to point mutations. Bayesian inference is especially attractive in the detection of recombination events because it allows for simultaneous inferences about the presence, nature and location of crossover points and the identification of parental sequences. Current frequentist recombination identification falls into a sequential testing trap. The most likely parental sequences and crossover points are identified using the data and then the certainty of recombination is assessed conditional on this identification. After briefly outlining basic phylogenetic models, Bayesian inference and Markov chain Monte Carlo (MCMC) computation, we summarise three different approaches to recombination detection and discuss current challenges in applying Bayesian phylogenetic inference of recombination.

Keywords: Bayes, recombination, phylogeny, MCMC

Introduction

Phylogeny, the science of reconstructing evolutionary histories, and Bayesian inference of phylogeny have been the subject of several excellent reviews. For a review of phylogenetic principles and models see, for example, Swofford et al (1996), and for a review of Bayesian approaches see, as an example, Huelsenbeck, Larget et al (2002). In this review, we briefly outline Bayesian phylogenetic inference and then discuss three solutions for inferring sporadic recombination in a Bayesian framework. The term sporadic means that recombination is rare, relative to point mutations. Here, we discuss only recombination, although with little modification these same ideas hold for sporadic horizontal gene transfer.

Molecular phylogeny uses nucleotide or amino acid data to infer evolutionary relationships among organisms (taxa), conditional on a model for nucleotide evolution. An unrooted tree, called a topology, with branch lengths indicating evolutionary distance between nodes, summarises evolutionary relationships among taxa. Figure 1 presents hypothetical data resulting from the recombination of genetically distinguishable subtypes, common, for example, in HIV-1. The figure includes nucleotide sequences from N taxa; a sequence UNK of unknown subtype, and representatives of three known subtypes A, B and C. The goal is to identify the subtype of the untyped sequence UNK. The sequences have two segments, labelled 1 and 2. If the data were restricted to either segment alone, it is not difficult to infer that the unknown sequence is closest to subtype A in segment 1 and closest to subtype C in segment 2. Virtually any of the many approaches to phylogenetic inference could be used. These topologies are given in Figure 1. The problem is that the position of the crossover point between segments is not known. Indeed, in general, we do not know if there are zero, one, or more than one crossover points, or their locations when they exist.

Nucleotide data and continuous time Markov chain nucleotide substitution models

The molecular phylogenetic methods that we consider here require data X consisting of N aligned nucleotide sequences,
one from each taxon, each of length \( L \). The nucleotide \( x_{ij} \) at site \( i (i = 1, \ldots, L) \) from organism \( n (n = 1, \ldots, N) \) is either A, G, C, or T/U, the nucleotides found in DNA/RNA. Insertions and deletions can be included, but for ease of presentation we do not consider them further. In the absence of recombination, a single topology \( \tau \) and estimates of evolutionary parameters that define the underlying nucleotide substitution model summarise the evolutionary history relating the sequences. The topology is an unlooped, connected graph \((V,E)\) with nodes \( V \) and branches \( E \). Each member \( b \) of set \( E \) has an associated branch length \( t_b \). Usually only bifurcating topologies are allowed, thus exactly 3 branches connect to each internal node and one branch to each external node. For \( N \) taxa, there are \( 2N-3 \) branches and \( H=(2N-5)(2N-3)/2\) possible unrooted topologies (Felsenstein 1978a).

Any appropriate probabilistic model can serve as the nucleotide substitution model in Bayesian phylogenetic inference, but continuous time Markov chain (CTMC) models (Huelsenbeck et al 2001) are most often used. The Markovian evolution assumption implies that the probability of observing a certain nucleotide at site \( i \) in the future depends only on the current nucleotide at that site. As a consequence, in the absence of recombination, the two descendants of an internal node (a bifurcation point) in the topology evolve independently of each other.

The CTMC model has a \( 4 \times 4 \) parameter matrix \( Q \) of instantaneous nucleotide substitution rates. In its most general form, \( Q \) contains 12 free off-diagonal parameters, with the \( ij \)th diagonal entry equal to negative the sum of the off-diagonal terms in the \( ij \)th row. The parameters include various instantaneous transition and transversion rates and the equilibrium nucleotide frequencies \( \pi \). We use \( \lambda \) to denote the set of parameters in \( Q \).

The nucleotide substitution probability matrix along a branch \( b \) of the topology is given by \( p(t_b) = \exp(t_b \lambda) \), a function of evolutionary time \( t_b \) and \( \lambda \). Time and the rate of nucleotide substitution are confounded in the CTMC, and it is necessary to introduce a constraint. Possible constraints include \( tr(Q) = -1 \) or \( tr[\text{diag}(\pi)Q] = -1 \), where \( tr(Q) \) is the trace of \( Q \) and \( \text{diag}(\pi) \) is a diagonal matrix with diagonal elements equal to the elements of the vector \( \pi \). Alternatively, one can fix a single, specified branch length \( t_b \) to 1. Actual divergence times, in years or generations, can not be determined without additional information.

**Bayesian inference**

For phylogenetic inference within a single segment, one seeks to draw conclusions about the topology \( \tau \), the branch lengths \( t = (t_b) \) and \( \lambda \), given the observed data \( X \). The Bayesian paradigm treats the parameters \( \omega = (\tau, t, \lambda) \) as fixed but unknown. It allows us to condition on either the data or the parameters, or neither, depending on the context and need, because they are all considered random variables. Uncertainty is quantified by specifying probability distributions for parameters both before and after data collection, and for data before they are observed. Bayesian inference supplies language and concepts for dealing with uncertainty and supplementary non-data information.

The main Bayesian inference is a posterior distribution of the parameters given the data. The posterior distribution is a high-dimensional distribution. To aid in understanding it, we summarise \( p(\omega | X) \) with: univariate and bivariate marginal distributions; posterior summaries including means, variances, quantiles and interval estimates for continuous parameters; and modes, intervals and probability distributions for discrete parameters.

To calculate the posterior distribution of \( \omega \), we apply Bayes' theorem. As a simple illustration, assume parameters \( t \) and \( \lambda \) are known, there is no recombination and we wish to identify the true topology given a particular dataset \( X \). Bayes' theorem requires as inputs: the sampling density, commonly referred to as the likelihood \( p(X | \tau) \) of the data \( X \) given the topology; and a prior distribution \( p(\tau) \). When we first begin research in an area, we often take priors that are vague or uninformative. For topologies, this prior could be, for example, the uniform distribution over all possible
topologies. As we gain experience in an area, we tend to develop and use priors that are informative; that is, the prior distribution favours those parameter values that our experience tells us are more probable.

The posterior probability of \( \tau \) is given by Bayes' theorem:

\[
p(\tau | X) = \frac{p(X | \tau) p(\tau)}{\sum_{\tau \in \text{all possible topologies}} p(X | \tau) p(\tau)}
\]

(1)

The denominator in equation (1) is equal to \( p(X) \), the marginal probability of the data; a sum over all possible topologies. Except when there are a modest number of taxa, \( p(X) \) is problematic if not impossible to calculate exactly, so approximate methods like Markov chain Monte Carlo (MCMC) are used to estimate \( p(\tau | X) \).

**Markov chain Monte Carlo**

Exact algebraic calculation of the posterior distribution \( p(\omega | X) \) or its summaries are usually impractical. Markov chain Monte Carlo is a method for numerical integration that can be employed in either frequentist or Bayesian inference (Gelfand and Smith 1990; Gilks et al 1996; Robert and Casella 1999). It is readily applied to complex posterior or otherwise intractable distributions. MCMC constructs a discrete time Markov chain with states in the parameter space of the statistical model in which the limiting distribution is equal to the posterior distribution \( p(\omega | X) \) of interest. MCMC is one of the most important algorithms of the 20th century (Cipra 2000) and has opened up an enormous number of previously unwieldy models to Bayesian analysis.

In applications of MCMC to phylogeny (eg Mau and Newton 1997; Yang and Rannala 1997; Larget and Simon 1999; Mau et al 1999; Li et al 2000; Suchard et al 2001), the states of the Markov chain \( \omega^{(m)} = (\tau^{(m)}, \theta^{(m)}, \lambda^{(m)}) \), \( m = 1 \ldots M \), are realisations of the parameter vector \((\tau, \theta, \lambda)\). If the Markov chain has the property that all states are connected, in the sense that from any starting state the Markov chain can eventually reach any other state, then it is irreducible. If there are no regular oscillations between different states then the chain is aperiodic. As long as the chain is irreducible and aperiodic, then by the ergodic theorem (Tierney 1994) the chain will converge to a unique stationary distribution. The chain is constructed using specially chosen transition kernels so that the desired posterior \( p(\tau, \theta, \lambda | X) \) is the stationary distribution. Functions \( g^{(m)} = g(\tau^{(m)}, \theta^{(m)}, \lambda^{(m)}) \) are calculated on the samples \((\tau^{(m)}, \theta^{(m)}, \lambda^{(m)})\) and, when averaged across samples, will approach their expected values \( E[g | X] \) under the stationary distribution as the length of the chain increases (for more details see, among many others, Tierney 1994; Robert and Casella 1999; Lange 2002). In our simple illustration of the previous section, we are interested in calculating \( p(\tau | X) \), particularly where it is high. MCMC provides samples \( \tau^{(m)} \) from the posterior \( p(\tau | X) \). We tabulate the fraction of \( \tau^{(m)} \) equal to a given topology \( \tau \), to estimate \( p(\tau | X) \).

The algorithm must start from an initial state, \( \omega^{(0)} = (\tau^{(0)}, \theta^{(0)}, \lambda^{(0)}) \), that is typically chosen in a non-stochastic manner. In practice, we do not use the first \( M_1 \) samples, called the burn in, to estimate features of the stationary distribution since these samples may be unduly influenced by the choice of initial state.

**Gibbs sampling**

The first general-purpose MCMC algorithm applied to Bayesian posteriors was the Gibbs sampler (Gelfand and Smith 1990; Gelfand et al 1990). For convenience of exposition, suppose the vector of parameters \( \omega = (\omega_1, \omega_2, \omega_3) \) has three components with posterior distribution \( p(\omega | X) \). Gibbs sampling proceeds from a starting value \( \omega^{(0)} = (\omega_1^{(0)}, \omega_2^{(0)}, \omega_3^{(0)}) \). Iteration \( m \) is created from iteration \( m-1 \) by sampling in an arbitrary order:

- \( \omega_1^{(m)} \) from density \( p(\omega_1 | \omega_2^{(m-1)}, \omega_3^{(m-1)}, X) \)
- \( \omega_2^{(m)} \) from density \( p(\omega_2 | \omega_1^{(m)}, \omega_3^{(m-1)}, X) \)
- \( \omega_3^{(m)} \) from density \( p(\omega_3 | \omega_1^{(m)}, \omega_2^{(m)}, X) \)

Each component \( \omega^{(m)} \)’s full conditional density depends on the latest samples of the other components, either \( \omega_1^{(m)} \) or \( \omega_2^{(m)} \) depending on the component update order. Together, the set of three conditional densities defines the Gibbs transition kernel. One iterates through all components \( M \) times.

The basic Gibbs algorithm depends on the ability to sample directly from the full conditional densities \( p(\omega_j | \omega_{\neq j}) \), where \( \omega_{\neq j} \) refers to all components of \( \omega \) except \( \omega_j \). In many models, including phylogenetic models, we are unable to sample from the needed densities. Either they are not densities of standard form or, in the case of sampling from the full conditional density of the topologies, the associated multinomial density has so many possibilities as to make sampling impossible. To allow for sampling from (conditional) densities when exact sampling is not possible, we use Metropolis–Hastings transition kernels (Metropolis et al 1953; Hastings 1970; Gilks et al 1996).

**Metropolis–Hastings algorithm**

To sample \( \omega^{(m)} \) from an arbitrary density \( p(\omega) \), given a current sample \( \omega^{(m-1)} \), the Metropolis–Hastings algorithm works by
drawing a candidate \( \omega^* \) from known density \( q(\omega^* | \omega^{(n-1)}) \). Next, we set \( \omega^{(n)} = \omega^* \) with probability \( a \), or we set \( \omega^{(n)} = \omega^{(n-1)} \) with probability \( 1-a \). The acceptance probability \( a \) is:

\[
a = \min \left\{ 1, \frac{p(\omega^*) q(\omega^{(n-1)} | \omega^*)}{p(\omega^{(n-1)}) q(\omega^{(n-1)} | \omega^{(n-1)})} \right\}
\]

(2)

Since \( a \) depends on the ratio \( p(\omega^*)/p(\omega^{(n-1)}) \), the density \( p(\omega) \) need only be known up to a multiplicative constant. This is convenient as often \( p(\omega) \) is a posterior density. Often the proposal density \( q(\cdot | \cdot) \) is symmetric, that is \( q(\omega^* | \omega^{(n-1)}) = q(\omega^{(n-1)} | \omega^*) \), and the acceptance probability reduces to:

\[
a = \min \left\{ 1, \frac{p(\omega^*)}{p(\omega^{(n-1)})} \right\}
\]

(3)

which is the Metropolis–Hastings algorithm (Metropolis et al 1953).

The Metropolis–Hastings algorithm is more general than the Gibbs algorithm. In the Gibbs algorithm, the acceptance probability is always equal to one so that all proposed steps are taken. There are many ways to set up MCMC algorithms (Robert and Casella 1999; Liu 2001). In the Metropolis–Hastings algorithm, the vector \( \omega \) may consist of any number of parameters that are vectors, matrices, scalars, and are discrete or continuous in any combination as needed to make a convenient algorithm. The Metropolis–Hastings algorithm may be used inside the Gibbs algorithm to update the conditional density \( p(\omega | \omega_0, X) \), for an algorithm called Metropolis within Gibbs.

Reversible jump MCMC

When changing states means moving between models with different numbers of parameters, MCMC can be difficult to carry out. Models with more parameters have more diffuse distributions than models with fewer parameters, therefore leading to higher proposal probabilities for lower dimensional models and a biased sampling. A modification of the Metropolis–Hastings algorithm called reversible jump MCMC (rJMC) makes these multidimensional transitions easier (Green 1995). In rJMC the acceptance probability is:

\[
a = \min \left\{ 1, \frac{p(\omega^*) q(\omega^{(n-1)} | \omega^*)}{p(\omega^{(n-1)}) q(\omega^{(n-1)} | \omega^*)} \right\}
\]

(4)

The transition probabilities \( q(\omega^{(n-1)} | \omega^*) \) and \( q(\omega^* | \omega^{(n-1)}) \) can now propose births (additional crossover points in the case of recombination) and deaths (merging segments separated by a crossover point for recombination) as well as updates of other parameters while maintaining the same dimension, which in the case of recombination amounts to updating evolutionary parameters and topologies when holding the crossover points intact. These proposals include augmenting variables, \( u \), that compensate for the change in dimensions between states. The \( J \) term represents the Jacobian. Its argument transforms the current parameters \( \omega^{(n-1)} \) and \( u \) to the parameters \( \omega^* \) of the newly proposed state. This dimension matching between \( \omega^{(n-1)}, u \) and \( \omega^* \) allows for proposal models of higher dimensions to be accepted. For examples of rJMC applications, see DiMatteo et al (2001) on curve fitting by splines with an unspecified number of knots, and see Suchard et al (2002) and Suchard, Weiss, Dorman et al (2003) on phylogenetic detection of recombination.

Improving MCMC algorithms and MCMC diagnostics

Mixing is a qualitative characteristic of MCMC algorithms. A well mixing algorithm ensures that averages over the MCMC samples are good estimates of the corresponding posterior averages. Determining whether the chain is mixing well and the length of the burn-in period are troublesome aspects of using MCMC. Poorly mixing algorithms are not fatal, rather they indicate a need for better transition kernels. For continuous parameters, Metropolis–Hastings transition kernel mixing is often improved by choosing an appropriate variance for the proposal step and combining updates for correlated single parameters into a single multivariate update (Robert and Casella 1999; Liu 2001). Alternatively, other transitions more suited to the posterior may be required. For example, different transition kernels have been developed for sampling phylogenetic trees and topologies, including branch pruning and reattachment, node rearrangement and negative branch lengths (Kuhner et al 1995; Mau and Newton 1997; Larget and Simon 1999; Mau et al 1999; Li et al 2000; Suchard et al 2001; Suchard, Weiss, Dorman et al 2003; Suchard, Weiss, Sinsheimer 2003; Suchard, Weiss, Sinsheimer et al 2003).

The length of the burn-in and the total length of the run are monitored using diagnostics. Cowles and Carlin (1999) provide a thorough overview and critique of MCMC diagnostic methods. Continuous parameters such as \( t_h \) may be monitored using a time-series plot of \( t_h^{(m)} \) versus \( m \). Plots with strong deterministic trends over large fractions of \( M \) indicate poor convergence. Lagged correlation plots that drop to zero in a lag much smaller than \( M \) indicate good convergence. Particularly useful methods are Gelman and Rubin diagnostics (Gelman and Rubin 1992; Gelman 1996; Brooks 1999; Brooks and Giudici 1999) based on multiple
runs of the chain from diverse starting points. With these
diagnostics, an overestimate of the variance, based on the
between- and within-run variance, is divided by an
underestimate of the variance, the within-run variance, to
form a convergence statistic. If the chains have not converged,
the statistic will be greater than one. As the length of the
chains increases, the statistic should approach one. Suchard,
Weiss, Dorman et al (2003) used this diagnostic on the
posterior distribution of the number of segments, K.
Huson, Larget et al (2002) recommend comparing
independent chains using a measure of tree-to-tree distance
(Huson, Larget et al 2001; Suchard, Weiss, Sinsheimer et al
2003). To assess the adequacy of state mixing in rjMCMC,
Suchard and colleagues (Suchard et al 2002; Suchard, Weiss,
Dorman et al 2003) use scaled regeneration quantile (SRQ)
plots (see Myland et al 1995 and Li et al 2000).

Testing in a Bayesian framework
using Bayes factors

In the Bayesian paradigm, multiple hypotheses can be
compared via Bayes factors and the posterior probabilities
of the hypotheses. For example, the Bayes factor, $BF$, for
comparing topology $\tau_i$ with topology $\tau_j$ is:

$$BF = \frac{p(\tau_i|X)}{p(\tau_j|X)} \frac{p(X|\tau_i)}{p(X|\tau_j)}$$

(5)

The Bayes factor is a ratio of marginal likelihoods and is akin
to a likelihood ratio. Since, equivalently, the Bayes factor is
the ratio of the posterior probabilities divided by the ratio of
the prior probabilities, we need only know the posterior
probabilities of the topologies up to a multiplicative constant.
In the special case in which the prior probabilities of $\tau_i$ and
$\tau_j$ are equal, the Bayes factor reduces to the ratio of the
posterior probabilities. $p(X|\tau_j)$ is a marginal likelihood of $\tau_j$
in which all the extraneous evolutionary parameters $\theta = (t, \lambda)$
have been integrated out:

$$p(X|\tau_j) = \int p(X|\tau_j, \theta)p(\theta)d\theta$$

(6)

When $BF >> 1$ there is strong support for $\tau_i$ over $\tau_j$, and
conversely when $BF << 1$ there is strong support for $\tau_j$ over
$\tau_i$. The Kass and Raftery (1995) guidelines for interpretation
of Bayes factors are presented in Table 1.

Bayes factor calculations require careful specification of
prior densities, particularly for parameters involved in the
null hypothesis. Explicit choice of priors depends on the
application. To go from the Bayes factor calculation to a
formal posterior probability of the hypothesis also requires

<table>
<thead>
<tr>
<th>Evidence for topology $1$ versus topology $2^*$</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z &gt; 2.00$</td>
<td>Decisive support for topology $1$</td>
</tr>
<tr>
<td>$2.00 &gt; Z &gt; 1.00$</td>
<td>Strong support for topology $1$</td>
</tr>
<tr>
<td>$1.00 &gt; Z &gt; 0.50$</td>
<td>Substantial support for topology $1$</td>
</tr>
<tr>
<td>$0.50 &gt; Z &gt; -0.50$</td>
<td>Barely worth mentioning</td>
</tr>
<tr>
<td>$-0.50 &gt; Z &gt; -1.00$</td>
<td>Substantial support for topology $2$</td>
</tr>
<tr>
<td>$-1.00 &gt; Z &gt; -2.00$</td>
<td>Strong support for topology $2$</td>
</tr>
<tr>
<td>$Z &lt; -2.00$</td>
<td>Decisive support for topology $2$</td>
</tr>
</tbody>
</table>

$*Z = \log_{10}(BF) = \log_{10}\frac{p(X|\tau_i)}{p(X|\tau_j)}$

a specification of the prior probabilities of the hypotheses.
We give specific examples of priors for the recombination
models in the section titled ‘Three general approaches to
inferring recombination’.

Inferring recombination

We are interested in detecting past recombination events
using the available sequence data. Recombination in a
sequence is often based on inference of topology because
recombination can lead to different topologies supported in
alternative segments of the alignment (Figure 1). The
representative descendant used as reference sequences are
often loosely referred to as parental sequences. Detecting a
change in topology along the sequence is the principle behind
most methods to detect recombination events (for example
see Crandall and Templeton 1999). These methods label a
sequence as recombinant when a model with different
topologies over the aligned sequence segments fits the
observed data significantly better than a model that imposes
a single topology across the entire alignment.

Crandall and Templeton (1999) have recently reviewed
frequentist methods for inferring HIV recombination.
Methods that employ frequentist inference have, for the most
part, adopted a two-stage approach. In the first stage, putative
recombinant sequences are screened to determine the
crossover points and parental sequences (e.g Smith 1992;
Robertson et al 1995; Grasdly and Holmes 1997), often using
a sliding window and assuming relatively simple models of
evolution. In the second stage, a more formal statistical
analysis is performed conditional on the estimated crossover
points and parental sequences (e.g Salminen et al 1996;
Crandall and Templeton 1999; Dorman et al 2002). This
approach suffers from a sequential testing trap by assessing
statistical significance conditional on a subset of previously
optimised parameters that, in addition, are often selected only
after multiple testing. It is difficult to determine the true level of statistical significance under these conditions. Further, uncertainty in the parental identities and uncertainty in the number of crossover points and their locations cannot be readily accommodated. Methods employing Bayesian inference can avoid this sequential testing trap by simultaneously estimating parental sequences as well as the number and location of crossover points (Suchard et al 2002). One integrates over all possible parents and crossover points weighted by their posterior probabilities when testing for significant evidence of recombination.

Three general approaches to inferring recombination

Directed acyclic graphs (DAGs)

Strimmer et al (2001) examine evidence for recombination among a set of aligned nucleotide sequences by using phylogenetic networks to model evolutionary relationships. Phylogenetic networks are generalisations of the phylogenetic tree model of evolution first developed by Felsenstein (1981). Although the Strimmer et al methods are presented in a likelihood framework, the networks employed are types of Bayesian networks (Russell and Norvig 1995; Cowell et al 1999; Jensen 2001) and so lead to clear Bayesian interpretations (Strimmer and Moulton 2000).

Strimmer and colleagues (Strimmer and Moulton 2000; Strimmer et al 2001) use two types of phylogenetic networks. The first (Strimmer and Moulton 2000), called planar split graphs, are graphical representations of circular collections of splits. A split is a partition of taxa into two disjoint groups. Each edge in a phylogenetic tree defines a split, and a collection of compatible splits uniquely defines a phylogenetic tree. The generalisation of phylogenetic trees to phylogenetic networks is accomplished by accommodating splits that are not compatible, i.e., splits that somehow contradict the tree structure. One less than general but still powerful way to include incompatible splits is to consider circular collections of splits. These are collections of possibly incompatible splits that frequently arise in biological datasets (Huson 1998). Specifically, a collection of splits is circular if there is at least one ordering of the taxa around a circle, such that the taxa within the groups formed by every split are adjacent. Conveniently, such networks can be drawn in the plane with no intersecting edges, hence the name planar split graphs.

Strimmer et al (2001) also use phylogenetic networks called ancestral recombination graphs (ARGs) borrowed from coalescent theory (Griffiths and Marjoram 1996, 1997). In these networks, $N$ extant taxa are followed backwards in time as they coalesce (two descendent, one parent) and recombine (one descendent, two parents) until only one most recent common ancestor (MRCA) is left. ARGs graphically summarise a finite collection of phylogenetic trees. These trees can be recovered by tracing back through the ARG and selecting only one parent at every recombination event; each recovered tree uses a different set of recombining parents and there can be up to $2^r$ phylogenetic trees encoded in a single ARG with $r$ recombination events. Strimmer et al (2001) assume a molecular clock, meaning the total length of all branches separating any two extant taxa from their most recent common ancestor is equal. For example, in Figure 1, segment 1’s topology’s branch length $t_{\text{line}}$ equals $t_r$. ARGs are designed specifically to model recombination, while split graphs are meant to capture any source of incompatibility in the model, including recombination (von Haeseler and Churchill 1993).

For both the planar split graph and ARG networks, Strimmer and colleagues convert the network to a DAG by applying a direction to each edge in the graph. ARGs are naturally ordered from the MRCA down to the extant taxa. In contrast, split graphs are not naturally directed, but one can choose any node as the root and assign directions to the edges such that no cycles arise (Strimmer and Moulton 2000). Even with reversible models of evolution, the likelihood is affected by the root location in split graphs, so the root should be considered an additional parameter of the model (Strimmer and Moulton 2000).

To calculate the likelihood of data at aligned site $i$ in DAGs where, unlike the case for phylogenetic trees, a sequence can derive from multiple parents, we need the probability of descendent nucleotide $x$ given parent nucleotides $y$ and $z$:

$$p(x|y,z) = p_x(i)p(x|y) + p_x(i)p(x|z)$$

(7)

where $p_x(i)$ is the probability that site $i$ derived from parent $i$; and $p(x|y) = p_x(i)$, $i = (y,z)$, are the finite time transition probabilities generated under the CTMC described earlier. In the breakpoint model (Strimmer and Moulton 2000; Strimmer et al 2001), a breakpoint location parameter $d$ is introduced such that all sites less than $d$ are derived from one parent, say $y$, and all sites greater than $d$ are derived from the other parent, $z$. Then:
In the mixture model (Strimmer and Moulton 2000), the functions \( p_i(l) = p_{i,j} \) are constant throughout the alignment. Here, the precise ancestry of site \( l \) remains unknown, but a proportion \( p_i \) of sites in descendents \( x \) derive from parent \( y \).

As a specific example, consider the DAG in Figure 2, in which 4 taxa are aligned: one putative recombinant, UNK; and three possible parental representatives, A, B and C. We arbitrarily assume that the tree is rooted in the branch leading to B. When recombination is suspected, as depicted in Figure 1 and as a DAG in Figure 2, the probability of nucleotide pattern \( X \), at any site \( l \) under the mixture model can be expressed as:

\[
p(X_l) = \sum_{x_1} \sum_{x_2} \sum_{x_3} p(x_1 | x_2, x_3) p(x_2 | x_3) p(x_3 | x_2, x_1) [v p(x_1 | x_2) + (1 - v) p(x_1 | x_3)] p(x_{UNK} | x_2)
\]

where \( x_i \) represents the nucleotide at the root; \( x_j, i = 1, 2, 3 \), represents the nucleotides at the internal nodes; \( x_j, j = A, B, C, UNK \), represents the nucleotides at the external nodes; and \( v = p_i \) is the probability that a site has parent 1. The null hypothesis of no recombination is satisfied when \( v \) is 0 or 1.

Each site is treated as independent of all others, so the probability of the sequence alignment is the product of the site probabilities.

Although calculating the likelihood of the data at one site is manageable for the network induced by only four taxa and one recombination event, direct calculation becomes infeasible as the number of taxa and recombination events rises. To approximate \( p(X) \), Strimmer and Moulton (2000) use a Gibbs sampler. In the sampler, the nucleotide state at each internal node in the network becomes a component in the state space vector. To initiate the sampler, the states are randomly assigned and then updated along with the remaining model parameters. Recombination parameters, breakpoints \( d \) or parental proportions \( p_p \), can be estimated as part of the model or fixed to constant values based on prior information. Simultaneous inference of DAG structure seems more difficult. Strimmer and Moulton (2000) infer a split graph using SplitsTree (Huson 1998) and then maximise the branch lengths and recombination parameters conditional on a simplified split graph. Similarly, Strimmer et al (2001) use a separate analysis to infer the phylogenetic trees and breakpoints of a candidate ARG structure and then optimise branch lengths. The molecular clock assumption utilised in Strimmer et al (2001) greatly eases computations by reducing the number of free parameters, but the assumption is not always valid (for example see Suchard et al 2001), and its violation could affect the inference about the existence of recombination.

**Hidden Markov models of recombination**

Husmeier, Wright and colleagues (McGuire et al 2000; Husmeier and Wright 2001, 2002; Husmeier and McGuire 2002) approach the detection of recombination events by modelling multiple site evolution. They assume that the underlying recombination events form a Markov chain and that some or all of the states of this Markov chain are not directly observable, called a hidden Markov model (HMM).

For clarity, we include a diagram similar to Figure 1 of McGuire et al (2000) as our Figure 3. The observed data are the nucleotide patterns \( X_i \) observed at each site along the aligned sequence. The hidden states that form the Markov chain are the topologies \( T_i \) for each site.

Husmeier and Wright (2002) use Bayes factors to select among submodels that have varying numbers and locations of crossover points. Between crossover points, topologies

\[
X_{i-1} \xrightarrow{\tau_{i-1}} X_i \xrightarrow{\tau_i} X_{i+1}
\]

(underlying topologies)

\[
\cdots \xrightarrow{\tau_{i-1}} X_i \xrightarrow{\tau_i} X_{i+1} \xrightarrow{\tau_{i+1}} \cdots
\]

(aligned sequence data)

**Figure 3** A hidden Markov model of recombination. Adapted from McGuire et al (2000).
are made equal from site to site and the approximate locations are first determined with other methods. Husmeier and Wright (2002) compute:

$$BF = \frac{p(X|C_j)}{p(X|C_2)}$$  \hspace{1cm} (10)

where $C_j$ represents one set of crossover points and $C_2$ another set of crossover points (often the null hypothesis of zero crossover points), and they use MCMC and approximation techniques to estimate these marginal likelihoods. They place a Markov prior on the set of topologies $\tau = (\tau_1, \tau_2, \ldots, \tau_k)$ such that $p(\tau) = p(\tau_1|\tau_{k+1}) \cdots p(\tau_k|\tau_1)p(\tau_1)$, where $\tau_1$ is equally likely to be any of the $H$ possible topologies and $K$ is the number of segments generated by the crossover points. For subsequent segments, the probability that $\tau_i$ matches $\tau_{i+1}$ is $\chi$. If the topology differs (with probability $1-\chi$), the topology is chosen from the remaining $H-1$ topologies with equal probability.

To estimate the Bayes factor, Husmeier and Wright (2002) first partition the log of the marginal likelihood $p(X|C)$ into an entropy term, $S(C)$, and a negative internal energy term, $U(C)$, for model C. These terms are:

$$S(C) = -\int p(\omega|X,C)\ln p(\omega|X,C)d\omega$$ \hspace{1cm} (11)

$$U(C) = \int p(\omega|X,C)(\ln p(X|\omega,C) + \ln p(\omega|C))d\omega$$ \hspace{1cm} (12)

To calculate $U(C)$ and $S(C)$, Husmeier and Wright (2002) assume that sites and segments between crossover points evolve independently so that the true topologies $\tau_s$, the branch lengths $l_{bs}$, and the evolutionary parameters $\lambda_s$ are independent conditional on the crossover points and data. Then they can conveniently estimate $U(C)$ using standard Bayesian phylogenetic software (Lartet and Simon 1999). $S(C)$ requires further approximation. Fixing optimal crossover point locations, they use Laplace's method to approximate the expectations of $l_{bs}$ and $\lambda_s$. They can introduce uncertainty into the crossover points by taking another Gaussian approximation to the posterior centred at the previously optimised locations. Husmeier and Wright (2002) then find the marginal distribution over all crossover locations by assuming that the perturbations of the crossover points have a negligible effect on the posterior distribution of the other parameters. By first determining the crossover points however, this procedure is liable to the sequential testing trap.

Many of the approximations used by Husmeier and Wright (2002) were avoided by Husmeier and McGuire (2002), albeit at the price of increasing the computational complexity and effectively limiting the number of taxa to 5.

In McGuire et al.'s (2000) HMM approach, each site in the aligned sequence $X$, $i \in \{1 \ldots L\}$ is produced by an unobservable state $\tau_i$ that takes on one of the $H$ possible topologies. If sites $i$ and $j$ have the same topology, $\tau_i = \tau_j$, then they have the same set of branch lengths $t_i = t_j$ and evolutionary parameters $\lambda_i = \lambda_j$. Sites with different topology states have different sets of branch lengths and evolutionary parameters. Let $\tau \in \{\tau_1, \ldots, \tau_k\}$, the set of topology states for the entire sequence alignment. Husmeier and McGuire (2002) calculate the posterior probability $p(\tau|X)$ with multidimensional integration:

$$p(\tau|X) = \int p(\tau, I, \lambda, \chi|X)d\lambda d\beta d\gamma$$ \hspace{1cm} (13)

The prior probability that site $I$ has a different topology from its neighbours is a first-order Markov chain dependent on $\chi$ in which:

$$p(\tau_I|\tau_{I-1}, \ldots, \tau_1) = p(\tau_I|\tau_{I-1})$$

$$= \chi^{I}_{(\tau_{I-1}, \tau_I)} + \frac{(1-\chi)(1-\chi)^{I-1}}{H-1}$$ \hspace{1cm} (14)

where $I$ is the indicator function. A priori parameters $\chi, I$, and $\tau$ are independent so $p(\tau, I, \lambda) = p(\tau)p(I|\tau)p(\lambda)$. $\chi$ has a beta distribution, and branch lengths are uniform on the interval 0 to 1. To model the evolutionary parameters, $\lambda$, Husmeier and McGuire (2002) use a version of the Felsenstein 84 model (Felsenstein and Churchill 1996) in which each state has its own transition bias parameter but nucleotide frequencies are shared across all states. The prior for each transition bias parameter is uniform on the interval 0 to 2 and the nucleotide frequency prior is Dirichlet(1,1,1,1).

The integral (equation 13) cannot be solved analytically, so Husmeier and McGuire (2002) use an MCMC algorithm to sample from $p(\tau, I, \lambda, \chi|X)$. The Gibbs sampler updates the parameters as follows:

$$r^{(n)} \text{ from density } p(r^{(n)}, \lambda^{(n-1)}, \chi^{(n-1)}, X)$$

$$\lambda^{(n)} \text{ from density } p(\lambda^{(n)}, \tau^{(n-1)}, \chi^{(n-1)}, X)$$

$$\chi^{(n)} \text{ from density } p(\chi^{(n)}, \tau^{(n)}, \lambda^{(n)}, X)$$

The conditional distribution of $\chi$ is beta and so can be easily sampled. To sample $\tau$, Husmeier and McGuire (2002) exploit the first-order Markov nature of the topology states at each site and use a series of univariate Gibbs updates to sample each topology state $\tau_i$ conditional on the neighbouring sequence topology states $\tau_{i-1}$ and $\tau_{i+1}$. To sample $\tau$ and $\lambda$, they use a Metropolis within Gibbs sampler.
The results are depicted as recombination profiles that plot the posterior probabilities of each of the possible topologies at each site of the aligned sequence (see Figure 4b for an example). In these profiles, the posterior probability is calculated by summing over the MCMC sampled states, thereby removing any conditioning on t or A. The recombination detection results with both simulated and actual sequence data are impressive, outperforming commonly used recombination detection methods like PLATO (Grassly and Holmes 1997).

The hidden states in Husmeier and McGuire’s HMM represent the topologies only. All sites sharing the same topology state are identically distributed. This restriction can lead to false inference of recombination due to rate variation, as rates may vary without a change in topology (Dorman et al 2002; Husmeier and McGuire 2002). The following example better accommodates rate variation by allowing changes in rate as well as topology.

Multiple change-point models of recombination
Suchard and colleagues (Suchard et al 2002; Suchard, Weiss, Dorman et al 2003) use a multiple change-point model for crossover points and rate variation change points. The model assumes that sites along the aligned sequences separate into an unknown number of contiguous, non-overlapping segments. Sequence segments can have different topologies, branch lengths or nucleotide transition parameters (as in Figure 1). Within a segment, sites are independently and identically distributed. The ability to vary the evolutionary parameters independent of the topology avoids confounding of topology change with rate changes (Dorman et al 2002; Suchard et al 2002). The number of distinct segments K can range between 1 and L, the total length of the aligned sequences. The number of crossover points is less than the number of change points.

Nucleotide substitution is modelled as a CTMC using the HKY (Hasegawa, Kishino and Yano) model (Hasegawa et al 1985; Suchard et al 2002; Suchard, Weiss, Dorman et al 2003) for the infinitesimal rate matrix Q. In the HKY model:

\[
\begin{align*}
A & \to \beta \bar{x}_L + [\beta(x_L + x_R)] \\
G & \to \alpha \bar{x}_L - [\alpha x_L + \beta(x_L + x_R)] \\
C & \to \beta \bar{x}_L - [\beta x_L + \alpha(x_L + x_R)] \\
T & \to \alpha \bar{x}_L - [\alpha(x_L + x_R)]
\end{align*}
\]
Suchard and colleagues (Suchard et al 2002; Suchard, Weiss, Dorman et al 2003) set \( \mu/[1 - \mu] = -\beta \) such that \( \beta = (1 - \alpha)/2 \). The nucleotide frequencies, \( \pi \), are fixed to the sample averages within a sequence segment.

Priors are chosen to be vague yet proper. Specifically, \( K \) is distributed as a truncated Poisson where \( E[K - 1] = \varepsilon \), the expected number of change points. Users should consider the application, data and their beliefs to set the prior mean number of change points \( \varepsilon \). Suchard and colleagues have used two different priors for the other parameters, though in each case they integrate over branch lengths assuming \( \nu_1 \) are exponentially distributed with mean \( \mu \). In Suchard et al (2002), within each segment \( k \), the prior is distributed as uniform \([0,1]\) and \( \mu_1 \) is distributed as exponential with mean 1. In Suchard, Weiss, Dorman et al (2003), the authors use hierarchical priors on \( \alpha \) and \( \nu_1 \):

\[
\begin{align*}
\alpha_k &\sim \text{Beta}(u,v) \\
\frac{\mu}{u + \nu} &\sim \text{Beta}(20,10) \\
\frac{u + \nu}{u + v} &\sim \text{Gamma}(1,1) \\
\mu_1 &\sim \text{Exponential}(\xi) \\
\xi &\sim \text{Exponential}(\theta)
\end{align*}
\]

(16)

Suchard, Weiss, Dorman et al (2003) reduce the number of possible topologies by limiting consideration to 5 parental at most. For more than 5 parents, the approach of Suchard et al (2002) reduces the number of possible topologies by assuming the topology of the \( T \) parental sequences is known with certainty. Like Husmeier and colleagues (Husmeier and McGuire 2002; Husmeier and Wright 2002), Suchard et al (2002) then place a Markov prior on the set of topologies. The user selects the probability that the topologies differ between segments \( \chi \), again adjusting to the specific application and data. In the absence of previous information concerning recombination, Suchard et al (2002) suggest choosing \( \chi \) and \( \varepsilon \) so that the probability of recombination is approximately 0.5.

Suchard and colleagues (Suchard et al 2002; Suchard, Weiss, Dorman et al 2003) use \( \eta \)MCMC to make transitions between submodels with differing numbers of segments. The details of the corresponding proposal steps can be found elsewhere (Suchard, Weiss, Dorman et al 2003).

A disadvantage of the multiple change-point model is the time needed to run the MCMC sampler. The computational demand is partially alleviated by using a hierarchical prior on branch lengths and then removing these parameters through analytical integration, yet the increased speed comes at a cost. To integrate out the branch lengths, the authors have implicitly assumed that, within a segment, there is prior independence of branch lengths across branches and sites. This integration may therefore make the model more susceptible to a parsimony-like long-branch attraction bias (Felsenstein 1978b) than a model that incorporates across-site dependence.

Like Husmeier and Wright (2002), Suchard and colleagues assess the statistical significance of recombination through Bayes factors. In addition, they simultaneously determine the 95% Bayesian confidence intervals (BCIs) for the change-point locations and present profiles of the parameters \( \tau \), \( \alpha \), and \( \mu \) (Figure 4). In the recombination profiles, the posterior probability of recombination is not conditional on the value of \( \alpha \) or \( \mu \). These parameters are integrated out by the MCMC sampling used to estimate the posterior probability. The evolutionary parameter profiles plot the marginal posterior mean and the 95% BCI for \( \alpha \) and \( \mu \) along the aligned sequences. Figure 4 shows sample profile plots for the full-length Rwandan HIV-1 sequence RW009, previously identified as a recombination between subtype parents A and C (Gao et al 1998). For this analysis, RW009 is compared with the consensus sequences of several East African subtype A and C sequences (data kindly provided by Dr Jean Carr).

**Extensions of the change-point model**

Phylogenetic-based methods for detecting recombination suffer from low power as diversity decreases (Crandall and Templeton 1999; Posada and Crandall 2001; Posada 2002). In these methods, parental type is often summarised using the modal nucleotide at each site, called the consensus sequence. Although consensus sequences can summarise multiple sequences of the same parental type (Hahn et al 1995; Suchard et al 2002) and may partially offset the power loss of using a single representative sequence, the consensus ignores minority sequences, may not represent a biologically meaningful sequence and could lead to biased conclusions. Here we describe an extension of the multiple change-point model of recombination (Suchard et al 2002; Suchard, Weiss, Dorman et al 2003) that accommodates more flexible summaries of parental types.

When multiple sequences are available from the same parent type, each site of the parent can be summarised as a vector of nucleotide frequencies. The frequencies may be the observed proportions in the sampled sequences or more complicated, and possibly more biologically meaningful, quantities. Here, as an example, we also report results using the estimated conditional ancestral probabilities for each
observed or predicted nucleotides at the extant parental nodes.

To test the extended model, we use Seq-Gen (Rambaut and Grassly 1997) to simulate datasets consisting of 78 aligned 2400-nucleotide-long sequences according to a topology structure inferred from HIV env sequences (Figure 5); 19 each from two subtypes (B, D) and 20 each from another two subtypes (A, C). Diversity was decreased by proportionally shrinking the topology, and 100 alignments with 98%-99%, 97%-98% and 95%-96% average pairwise sequence similarity were generated. We construct a ‘recombinant’ sequence from the left half of one A subtype sequence and the right half of one C subtype sequence, so that the crossover point occurs between sequence positions 1200 and 1201. The parents are selected as likely to pose difficulties for recombination detection. The C sequence is deeply rooted, while the A sequence comes from the under-represented A2 sub-subtype (5 A2 of 20 A sequences). The parental sequences were removed from the dataset and the remaining 76 aligned sequences plus the recombinant query were examined with the multiple change-point model (Suchard et al. 2002) using three alternative summary measures of the parental sequences: consensus (CONS), observed nucleotide frequencies (FREQ) and inferred conditional ancestral probabilities (ANCES) (Table 2).

Parental data using estimated nucleotide probabilities (FREQ and ANCES) can increase the power to detect recombination. In particular, the power to detect the correct parents and number of crossover points increases, especially for ANCES (see row 1 of Table 2). Since conditional ancestral nucleotide probabilities perform well, we are investigating whether jointly inferred single best ancestral parent sequences could perform equivalently without the

Figure 5 The tree used to simulate recombinant HIV env sequences data included 19 each from two subtypes (B, D) and 20 each from the other two subtypes (A, C). Some branches are too small to observe at this resolution. Parent P1 is from subtype A and parent P2 is from subtype C. The left half of P1 is combined with the right half of P2 to form the recombinant query sequence. Node P2 is a deep branch in the subtype C cluster and P1 is a part of the A2 sub-subtype cluster in the subtype A cluster.

Table 2 Multiple change-point model performance with alternative parental data

<table>
<thead>
<tr>
<th></th>
<th>High diversity (95%-96% similar)</th>
<th>Median diversity (97%-98% similar)</th>
<th>Low diversity (98%-99% similar)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CONS</td>
<td>FREQ</td>
<td>ANCES</td>
</tr>
<tr>
<td>$B_{sp} &gt; 10^6$</td>
<td>50%</td>
<td>63%</td>
<td>98%</td>
</tr>
<tr>
<td>Median 95% BCP</td>
<td>284</td>
<td>312</td>
<td>311</td>
</tr>
<tr>
<td>MSE $&lt; 10^4$</td>
<td>1.3</td>
<td>1.4</td>
<td>1.8</td>
</tr>
<tr>
<td>Number$^4$</td>
<td>48</td>
<td>48</td>
<td>55</td>
</tr>
</tbody>
</table>

$^4$ The proportion of simulations with Bayes factor $B_{sp} > 10^6$, ie strong support, for the hypothesis that the sequence is correctly classified as an AC recombinant (A-like on left and C-like on right).

$^a$ The median difference between the upper and lower limits of the 95% Bayesian confidence interval (BCI) for the crossover point.

$^b$ Mean square error (MSE), calculated by finding the mean squared distance of the median posterior crossover point from the true crossover point.

$^c$ The number of simulations qualifying in each group, where MSE and 95% BCI are calculated using only simulations where all methods resulted in Bayes factors $> 10$.

Abbreviations: CONS – consensus, FREQ – observed nucleotide frequencies, ANCES – inferred conditional ancestral probabilities.
computational cost of using conditional ancestral nucleotide
probabilities.

Discussion
As with most Bayesian methods of inference, there is a need
for more relevant priors that translate current scientific
knowledge into parametric and hierarchical priors, more
efficient MCMC samplers, and better diagnostics to assess
the adequacy and efficiency of the MCMC samplers. Rather
than elaborate on all of these issues, we briefly outline one
possible approach to improve the efficiency of the rjmMCMC
sampler used in the change-point model discussed above.

Low acceptance probabilities for moves between states of
the chain result in an inefficient sampler, and acceptance
probabilities for moves between submodels having different
dimensions in the current multiple change-point model
implementation can be low. Hurn et al (2002) propose a
general method to improve acceptance probabilities for
rjmMCMC jumps between submodels having different
dimensions, by breaking jumps into two steps. Starting in
state $\omega^{(r)}$ in a submodel with $r$ dimensions, the first step for
a birth jump is the usual proposal of $\omega^+$ from a submodel
with $r > r$ dimensions. Intermediate state $\omega^{(r)}$ is then updated
multiple times using a fixed dimension sampler. The resulting
state $\omega^{(r)}$ is accepted or rejected by comparison with $\omega^{(r)}$.
The additional updates are designed to bring the proposed
step closer to the posterior mode in $x$-dimensional space,
leading to higher acceptance probabilities. Death proposals
are likewise modified to retain the reversibility of the sampler.

The models discussed in this manuscript are applicable to
data in which recombination is a rare event relative to
point mutations. When multiple recombination events are
suspected or diversity is very low, it may be more appropriate
to use recombination detection based on a population genetic
model, the coalescent (Crandall and Templeton 1999;
Rodrig and Felsenstein 1999; Rosenberg and Norberg
2002). In coalescent recombination models however, the
result is an estimate of the overall rate of recombination in
the population rather than identification of specific
recombinant sequences or crossover points.

Bayesian phylogenetic inference of recombination events
has distinct advantages over frequentist inference. The
number and locations of crossovers points can be
simultaneously estimated along with the inference of
recombination, thus avoiding the sequential testing common
in frequentist approaches. In addition, Bayesian inference is
better able to test complex biological hypotheses, such as
whether a crossover point exists anywhere within a particular
gene or region.

We have illustrated the flexibility and utility of Bayesian
phylogenetic inference using recombination detection as an
example. Bayesian phylogenetic inference has also been used
successfully to examine a variety of questions including the
validity of the molecular clock assumption (Suchard et al
2001; Suchard, Weiss, Sinrheimer 2003), inference of
copulation (Huelsenbeck et al 2000), estimation of
ancestral nucleotide frequencies (Huelsenbeck and Bollback
2001), inference of the tree root position (Huelsenbeck,
Bollback et al 2002) and identification of positively selected
sites (Yang and Nielsen 2002), to name only a few
applications. The versatility of Bayesian phylogenetic
inference is limited only by the imagination of the researcher.

Acknowledgements
The authors thank Dr. Jean Carr for providing the aligned
East African subtypes A and C HIV sequence data used in
the Bayesian profile plots. JSS and MAS are supported in
part by USPHS grant CA16042. MAS is supported in part
by the UCLA AIDS Institute, Center for AIDS Research
(AL2697). KSD and MAS are supported in part by NIH
grant GM068955.

References
The Statiscician, 48:99–100.
Bayesian statistics 6, New York: Oxford University Press, p 733–42.
Cipra BA. 2002. The best of the 20th century: editors name top 10
algorithms [entire], SIAM News, Volume 33, Number 4, Accessed 7
Nov 2003. URL: http://www.siam.org/siamnews/05-00/current.html
Cowles K, Carlin B. 1996. Markov chain Monte Carlo convergence
Crandall KA, Templeton AR. 1999. Statistical approaches to detecting
recombination. In Crandall KA, ed. The evolution of HIV. Baltimore:
Johns Hopkins Univ. Press, p 153–76.
Dilmaa T, Geno C, Kass RE. 2001. Bayesian curve fitting with
Felsenstein J. 1980. Cases in which parsimony or compatibility methods
will be positively misleading. Syst Zool, 27:899–100.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum
Felsenstein J, Churchill GA. 1996. A hidden Markov model approach to
variation among sites in rate of evolution, J Mol Evol, 13:93–104.
Gar F, Robertson J, Carruthers CD, Morrison SG, Jian B, Chen Y, Barré-
comprehensive panel of near-full-length clones and reference
sequences for non-subtype B isolates of human immunodeficiency


