Syst. Biol. 54(3):401–418, 2005

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150590947041

Joint Bayesian Estimation of Alignment and Phylogeny

BENJAMIN D. REDELINGS AND MARC A. SUCHARD

Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, California 90095-1766, USA; E-mail: msuchard@ucla.edu (M.A.S.)

Abstract.—We describe a novel model and algorithm for simultaneously estimating multiple molecular sequence alignments and the phylogenetic trees that relate the sequences. Unlike current techniques that base phylogeny estimates on a single estimate of the alignment, we take alignment uncertainty into account by considering all possible alignments. Furthermore, because the alignment and phylogeny are constructed simultaneously, a guide tree is not needed. This sidesteps the problem in which alignments created by progressive alignment are biased toward the guide tree used to generate them. Joint estimation also allows us to model rate variation between sites when estimating the alignment and to use the evidence in shared insertion/deletions (indels) to group sister taxa in the phylogeny. Our indel model makes use of affine gap penalties and considers indels of multiple letters. We make the simplifying assumption that the indel process is identical on all branches. As a result, the probability of a gap is independent of branch length. We use a Markov chain Monte Carlo (MCMC) method to sample from the posterior of the joint model, estimating the most probable alignment and tree and their support simultaneously. We describe a new MCMC transition kernel that improves our algorithm's mixing efficiency, allowing the MCMC chains to converge even when started from arbitrary alignments. Our software implementation can estimate alignment uncertainty and we describe a method for summarizing this uncertainty in a single plot. [Alignment bias; Bayesian phylogenetics; indel models; MCMC; statistical alignment; Tree of Life.]

Phylogenetic reconstruction using molecular sequences has become an invaluable tool in the study of evolution, driven by the growing wealth of available genetic information and the development of statistical methods for handling molecular data. Most modern statistical methods for estimating phylogenies from molecular sequence data rely on multiple sequence alignments that arrange the raw sequence data in a matrix to specify which residues are homologous (Holder and Lewis, 2003). However, this alignment is an inferred property of the sequences and cannot be directly observed, so the alignment must also be estimated. Current methods for phylogenetic reconstruction separate alignment reconstruction from phylogeny reconstruction. First, the alignment is estimated from the raw sequence data; some attempt must then be made to identify and remove ambiguous regions or handle them in some other appropriate manner (Lutzoni et al., 2000). After a full or partial alignment has been constructed, the phylogeny is then estimated from the alignment. This sequential approach works well when the alignment is well resolved, but can produce spurious or no results when the alignment contains uncertain regions (Lutzoni et al., 2000). An improved method would be helpful in reconstructing phylogenies of distantly related sequences because such sequences often have large ambiguous regions. To this end, we propose a model and algorithm that allows the estimation of the alignment and phylogeny simultaneously from unaligned sequence data. We operate in a Bayesian framework and use Markov chain Monte Carlo (MCMC) to sample from the joint posterior distribution of alignments and phylogenies given the simultaneous model.

Currently accepted methods for phylogeny reconstruction, such as maximum parsimony, maximum likelihood, and Bayesian estimation, use as input a single estimate of the alignment that is assumed to be correct. This assumption can lead to exaggerated support for inferred phylogenies if the alignment contains ambiguous regions, because near-optimal alignments are not taken into account (Lutzoni et al., 2000). In addition, the use of alignments constructed using progressive alignment methods can lead to inferred phylogenies that are biased towards the fixed guide tree assumed in generating the alignment (Lake, 1991; Thorne and Kishino, 1992; Sinsheimer, 1994). This bias can be extreme; our results include an example where the posterior probability for a particular partition of taxa rises from 0.553 to 0.998 when alignment uncertainty is ignored. Although there exist some methods for constructing multiple alignments without the use of a guide tree (Notredame et al., 2000), employing a tree estimate is responsible for several improvements in the accuracy of alignment software (Thompson et al., 1994). Thus, the availability of the phylogenetic tree during alignment construction leads to improved alignments, but accurate estimates of the tree are unavailable during the alignment construction phase of sequential estimation.

One common technique for dealing with this problem is to remove ambiguously aligned regions from the alignment before submitting the alignment to the phylogenetic reconstruction procedure. Simply removing ambiguous regions can result in the loss of a large fraction of informative sites (Lutzoni et al., 2000). One technique that mitigates this problem is to code information about ambiguity into the alignment itself (Geiger, 2002). A simple method to encode this information is to split ambiguous columns into groups of residues in which homology is unambiguous (Baldauf et al., 1996). Residues from these groups can then be placed in separate columns. This preserves unambiguous positional homology information within groups of taxa, while deleting ambiguous information about positional homology between groups. However, the process of determining which regions of the alignment are ambiguous is ad hoc and can be subjective.

Researchers have developed a number of techniques to make better use of the information in regions of

ambiguous alignment for phylogenetic reconstruction (Lee, 2001). One technique, known as elision, involves concatenating a set of near-optimal alignments into a larger alignment and then using the larger alignment as the basis for traditional phylogeny reconstruction techniques (Wheeler et al., 1995). The elision method weighs information about positional homology according to the fraction of near-optimal alignments that include the information. This allows the use of ambiguous positional homology information, as well as enabling the use of unambiguous information in ambiguous regions of the alignment. However, in this method, all near-optimal alignments are treated equally instead of being weighted according to alignment quality. Equal weighting becomes problematic if one wishes to include in the elision a large number of alignments, because high-quality alignments will be treated the same as medium quality alignments.

Another approach to using ambiguous regions, known as optimization alignment, involves simultaneous estimation of alignments and phylogenies within a parsimony framework. Instead of using a fixed alignment, Wheeler (1996) simultaneously estimates ancestral sequences and their pairwise alignment to neighboring sequences by minimizing the number of mutations, including both substitutions, and, optionally, insertion/deletion (indel) events. However, consideration of all possible internal sequences is extremely computationally expensive and so approximations are used in all proposed algorithms. For example, Lutzoni et al. (2000) and Wheeler (1999) have both independently developed a fixed-states algorithm in which only sequences that are observed at leaf nodes are considered as possibilities for sequences at internal nodes. This assumption is biologically unrealistic for deep divergences where alignment is often crucial. Because of the limitations of this method, Lutzoni et al. (2000) employ this method only when considering regions of ambiguous alignment and use standard techniques on the rest of the alignment. Research into further improvements in both the speed and the quality of optimization alignment continues (Wheeler, 2003). One problem that optimization alignment faces is that measures of uncertainty in both inferred alignments and inferred phylogenies are difficult to obtain. Because columns in the alignment are no longer independent, standard bootstrap techniques cannot be used, and there do not appear to be other well-motivated techniques to take the bootstrap's place. Although sensitivity analysis can reveal uncertainty arising from unknown optimization parameter values, it cannot characterize uncertainty in the tree and alignment given fixed parameter values and cannot reveal uncertainty in the tree resulting from uncertainty in the alignment (or vice versa). In addition, optimization alignment research has focused on maximum parsimony, although maximum likelihood analysis is possible as well. The likelihood-based approaches that we take in this paper are beneficial because they are model based, offer better use of phylogenetic information, and are statistically consistent (Felsenstein, 2003).

One approach that may resolve the above difficulties is to estimate alignment and phylogeny simultaneously in a Bayesian framework and assess confidence on inferred alignments and phylogenies using posterior probabilities. This joint estimation approach allows one to consider the myriad of near-optimal alignments when estimating phylogenies. These alignments are naturally weighted by their posterior probabilities that provide well-motivated and objective estimates of which parts of the alignment are reliable, as well as naturally taking into account information in ambiguous regions of the alignment. Because joint estimation requires no external guide tree, it addresses the problem that alignments constructed through progressive alignment are biased towards the guide tree (Lake, 1991). Instead of a fixed guide tree, a random internal estimate is constantly available. Joint estimation can safely make use of information from this tree when modeling the alignment without bias.

Joint estimation allows for more accurate substitution and indel models than is possible with sequential methods. Because a random tree estimate is available during alignment reconstruction, joint estimation enables the use of models that do not overcount single substitutions or indels shared between multiple taxa by common descent. For the substitution model, one is able to use stochastic models of substitution along a tree for alignment reconstruction as well as phylogeny reconstruction. Joint estimation can also take advantage of extended substitution models in scoring the alignment, such as models that take into account rate variation between sites. This contrasts with sequential methods, in which these models are not used during the alignment construction phase because the tree is not yet known. Further, joint estimation allows one to accurately model indels and to use information in shared indels to group taxa on the tree. Because the tree and the alignment are estimated simultaneously, the joint model can take into account the possibility that shared indels in sister taxa are homologous and result from a single indel event.

To sample from the joint posterior distribution of alignments and phylogenies, an indel model and an alignment-aware MCMC sampling method are necessary in addition to the traditional phylogenetic substitution model. The Thorne et al. (1991; TKF1) indel model is a reversible model for pairwise alignment and has been used as the basis for sampling from the posterior distribution of alignments conditional on a fixed tree (Holmes and Bruno, 2001; Lunter et al., 2002). The TKF1 model has the drawback of assuming that indels are always of unit length. When estimating alignment and phylogeny simultaneously, this is problematic not only because of the skewed gap distribution, but also because the model treats long indels that are shared between multiple taxa as multiple shared indels. This exaggerates the number of shared characters and, because indels are rare events, could significantly skew the posterior tree distribution. The Thorne et al. (1992; TKF2) alignment model extends the TKF1 model by allowing insertions and deletions

of sequence fragments containing several letters. However, the TKF2 model makes the assumption that inserted fragments are indivisible and that indels therefore are never nested or overlapping (Metzler, 2003). Thus, the TKF2 model forbids some possible positional homologies.

In developing an indel model, we make the simplifying assumption that indels occur on each branch of the tree in a manner independent of branch length. This allows us to model the alignment distributions on all branches using a single symmetrical pair-hidden Markov model (pair-HMM). The pair-HMM that we employ improves on the TKF1 model in allowing indels of multiple residues, and we improve on the TKF2 model by allowing for all homology structures. One consequence of our simplifying assumption is that we lose the nice TKF property of placing indels preferentially on longer branches. One way to side-step this loss would be to place a TKF2 model on each branch of the tree, but link alignments from adjacent branches through their shared sequence length instead of allowing fragment boundaries to be shared across branches. Despite these differences, our proposed model shares many properties with the TKF models. Indels occur independently on each branch, conditional upon model parameters and the lengths of sequences at internal nodes. In addition, insertions and deletions are equally likely and sequence lengths do not have a tendency to grow or shrink over time.

MCMC has previously been used to sample from the posterior distribution of model parameters and phylogenies conditional on a fixed alignment (Mau and Newton, 1997; Yang and Rannala, 1997; Li et al., 2000) or to sample from the posterior distribution of alignments conditional upon a fixed phylogeny and unaligned sequences (Allison and Wallace, 1994; Holmes and Bruno, 2001). Sampling from the posterior of a joint model requires combining these two approaches to sample from the posterior distribution of the alignment, phylogeny, and model parameters given only the unaligned sequence data. This requires the construction of new MCMC transition kernels to resample the topology and the alignment in tandem.

In this paper, we describe a novel approach to jointly estimating alignment and phylogeny in a Bayesian framework. In Methods, we begin by introducing extensions to traditional phylogenetic models to include a model of multiple alignments. We follow Holmes and Bruno (2001) in modeling pairwise alignments along each branch of the tree and in recording the presence and absence of characters at internal nodes. We introduce a new indel model that allows indels of multiple residues and also allows indels to nest or overlap if they lie on separate branches. We then describe our method of sampling from the posterior distribution using MCMC. We extend the approach of Holmes and Bruno (2001) by introducing novel MCMC transition kernels to sample topologies together with alignments and to improve the sampling of alignments. Finally, we describe methods of summarizing the posterior alignment distribution. Our methods enable researchers to conveniently identify which parts of the alignment are ambiguous and how much uncertainty there is in the exact location of gaps. In Results, we apply the joint model and estimation algorithm to the biological problem of inferring the early branching order on the Tree of Life. We analyze two data sets with a significant degree of sequence divergence: a 5S ribosomal RNA (rRNA) data set and an elongation factor 1α/Tu (EF-1α/Tu) data set. Because the sequences involved have diverged significantly, the alignments have ambiguous regions and the phylogenetic signal is not strong. By using joint Bayesian estimation, we hope to enhance this signal with indel information and information in ambiguous regions while simultaneously avoiding bias and overconfidence in inferred topologies. We conclude, in Remarks, with a brief discussion of merits and shortfalls of joint estimation and highlight several areas of future research.

METHODS

We start with the observed data $\mathbf{Y}$ consisting of a set of n homologous molecular sequences. Let the sequences in $\mathbf{Y}$ be indexed by $i = 1, \ldots, n$, with corresponding lengths $|\mathbf{Y}_i|$ and let the jth element of $\mathbf{Y}_i$ be denoted $Y_i[j]$. Element $Y_i[j]$ takes on values, called letters, from a set $\boldsymbol{\alpha}$ of possible values called an alphabet. For example, if the sequences are protein sequences, then $\boldsymbol{\alpha}$ is the set of amino acids, whereas if the sequences are DNA sequences, then $\boldsymbol{\alpha}$ is the set of DNA nucleotides {A, G, C, T}.

We aim to estimate the unobserved evolutionary relationship among the sequences and the parameters of the evolutionary process that generated the sequences. The evolutionary relationship is specified by a multiple alignment $\mathbf{A}$ and a phylogenetic tree that relates the n molecular sequences. The multiple alignment, although separable from the data $\mathbf{Y}$, specifies how the data are arranged in an aligned data matrix $\mathbf{f}$. This matrix identifies which letters from the sequences are homologous to each other by arranging homologous letters into the same column. We define C to be the unknown number of columns in this matrix. Each column represents a homologous site and contains one letter, or possibly a missing value, per sequence. This process will be described in more detail later. The phylogenetic tree can be broken down into its unrooted topology τ and branch lengths $\mathbf{T}$. The topology τ is an acyclic graph in which all nodes have either one neighbor or three neighbors. The tree has n leaves, each of which corresponds to one of the n observed sequences. Sequences at the internal nodes correspond to ancestral sequences and can be estimated but are not observed. The total number of nodes in τ is $N = 2n - 2$ and the number of branches is $B = 2n - 3$.

The evolutionary process parameters include both the substitution process parameters $\boldsymbol{\Theta}$ and the indel process parameters $\boldsymbol{\Lambda}$. Substitution parameters $\boldsymbol{\Theta}$ contain the stationary frequencies of each letter in the alphabet and the transition rates between the letters, as well as parameters to describe among-site rate heterogeneity. Indel parameters $\boldsymbol{\Lambda} = (\delta, \epsilon, \zeta)$, where δ is the probability for an indel

along a branch, ϵ characterizes the geometric length distribution of the indels, and ζ characterizes the geometric length distribution of the alignment. Taken all together, the entire state space $\boldsymbol{\Omega}$ is composed of points $\boldsymbol{\omega} = (\mathbf{Y}, \mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda})$.

Probabilistic Model

Traditionally, phylogenetic reconstruction has implicitly conditioned on the alignment $\mathbf{A}$ when estimating the tree $(\tau, \mathbf{T})$ (Holder and Lewis, 2003). In order to contrast this traditional conditioning with the approach we use in this paper, we will make the conditioning explicit. This leads to the probability expression

$$\begin{aligned} P(\mathbf{Y}, \tau, \mathbf{T}, \boldsymbol{\Theta} \mid \mathbf{A}) = P(\mathbf{Y} \mid \tau, \mathbf{T}, \boldsymbol{\Theta}, \mathbf{A}) \times P(\tau, \mathbf{T} \mid \boldsymbol{\Theta}, \mathbf{A}) \\ \times P(\boldsymbol{\Theta} \mid \mathbf{A}), \end{aligned} \quad (1)$$

where, following a common abuse of notation, we write $P(\mathbf{X})$ to represent $P(\mathbf{X} = \mathbf{x})$ for any random variable $\mathbf{X}$ taking on a realized constant $\mathbf{x}$. The first term in Equation 1, $P(\mathbf{Y} \mid \tau, \mathbf{T}, \boldsymbol{\Theta}, \mathbf{A})$, is the model likelihood and is given by the substitution model. Priors on the trees and substitution parameters are usually chosen to be mutually independent and independent of the implied alignment, leading to the reduced expression

$$\begin{aligned} P(\mathbf{Y}, \tau, \mathbf{T}, \boldsymbol{\Theta} \mid \mathbf{A}) = P(\mathbf{Y} \mid \tau, \mathbf{T}, \boldsymbol{\Theta}, \mathbf{A}) \times P(\tau, \mathbf{T}) \\ \times P(\boldsymbol{\Theta}). \end{aligned} \quad (2)$$

In contrast, allowing the alignment to vary results in the following expression

$$\begin{aligned} P(\mathbf{Y}, \mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}) = P(\mathbf{Y} \mid \mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}) \\ \times P(\mathbf{A} \mid \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}) \times P(\tau, \mathbf{T} \mid \boldsymbol{\Theta}, \boldsymbol{\Lambda}) \\ \times P(\boldsymbol{\Theta} \mid \boldsymbol{\Lambda}) \times P(\boldsymbol{\Lambda}). \end{aligned} \quad (3)$$

Similar to before, we choose priors on the tree and substitution parameters $\boldsymbol{\Theta}$ that are mutually independent and independent of the alignment. We also assume that the tree and the substitution parameters are independent of the indel parameters $\boldsymbol{\Lambda}$. The prior that we assume on the alignment depends on the topology τ and the indel parameters, but not on the branch lengths $\mathbf{T}$. This leads to the reduced expression

$$\begin{aligned} P(\mathbf{Y}, \mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}) = P(\mathbf{Y} \mid \mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta}) \times P(\mathbf{A} \mid \tau, \boldsymbol{\Lambda}) \\ \times P(\tau, \mathbf{T}) \times P(\boldsymbol{\Theta}) \times P(\boldsymbol{\Lambda}). \end{aligned} \quad (4)$$

Notably, Equation 4 is identical to Equation 2 except for the inclusion of the term $P(\mathbf{A} \mid \tau, \boldsymbol{\Lambda})$, which is the prior on alignments, and the term $P(\boldsymbol{\Lambda})$, which is the prior on the indel model. Thus, our likelihood may be based on traditional substitution models such as reversible, continuous-time Markov chains, as described in the next section. We develop a prior on alignments based on a biologically realistic model of indel events. For the prior on trees, we use a Uniform prior across the finite number of topologies. We assume an Exponential prior with common mean μ on the length of each branch. We further assume that the hyperparameter μ is also exponentially distributed (Suchard et al., 2001).

Substitution model.—The probability that the substitution process results in the observed letters in the aligned data matrix $\mathbf{f}$ is given by the likelihood $P(\mathbf{Y} \mid \mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda})$. To express the likelihood, we must accomplish two tasks. First, we must describe how the alignment $\mathbf{A}$ groups the data $\mathbf{Y}$ into columns of homologous letters in $\mathbf{f}$. Second, we must specify the probabilistic model imposed on the columns of $\mathbf{f}$.

Matrix $\mathbf{f}$ is constructed from $\mathbf{A}$ and $\mathbf{Y}$ and consists of rows indexed by $i = 1, \ldots, N$ and columns indexed by $c = 1, \ldots, C$. The letters in row i all come from sequence i and must occur in order. Matrix $\mathbf{f}$ represents the hypothesis that the letters observed in each column c are all descended from a single residue in the common ancestor. If the value f_{ic} is missing because no letter of sequence i corresponds to position c, we fill the entry with '$-$', denoting a gap. Matrix $\mathbf{f}$ includes sequences at internal nodes using Felsenstein wildcards. Felsenstein wildcards are represented by '$\star$' and signify residues that are present but unobserved. To specify how $\mathbf{A}$ arranges the letters of $\mathbf{Y}$ into $\mathbf{f}$, we introduce the matrix $\mathbf{M}(\mathbf{A})$ with the same dimensions as $\mathbf{f}$ such that M_{ic} is the index of the letter in sequence i that is identified with c. Specifically, if sequence i is a leaf sequence, we have

$$f_{ic} = \mathbf{Y}_i[M_{ic}]. \quad (5)$$

If a column c has no letter in sequence i, then $M_{ic} =$ '$-$' and we define $\mathbf{Y}_i[\text{'}-\text{'}] =$ '$-$'. Figure 1 illustrates this formulation for a four-taxon example. One sequence has length three and three sequences have length four. The unobserved sequences at the two internal nodes are also of length four in this example.

To specify the likelihood, we recall that each column in $\mathbf{f}$ identifies which letters in each leaf sequence

$\mathbf{Y}$	$\mathbf{M(A)}$	$\mathbf{f}$
$Y_1 = (A, T, T, C)$	1 2 – 3 4	$A\ T\ -\ T\ C$
$Y_2 = (A, T, T, G)$	1 2 – 3 4	$A\ T\ -\ T\ G$
$Y_3 = (T, C, T, G)$	– 1 2 3 4	$-\ T\ C\ T\ G$
$Y_4 = (T, C, T)$	– 1 2 3 –	$-\ T\ C\ T\ -$
	1 2 – 3 4	$\star\ \star\ -\ \star\ \star$
	– 1 2 3 4	$-\ \star\ \star\ \star\ \star$

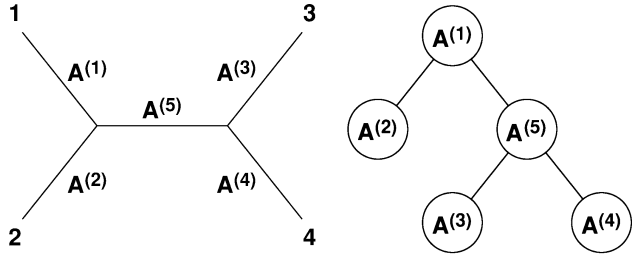
FIGURE 1. Construction of aligned data matrix $\mathbf{f}$ from data $\mathbf{Y}$ and alignment $\mathbf{A}$. The first column shows some example data consisting of unaligned sequences of various lengths. The second column shows $\mathbf{M}(\mathbf{A})$, a matrix that parameterizes the multiple alignment $\mathbf{A}$ by specifying where gaps appear in the aligned data matrix and which letters of $\mathbf{Y}$ appear in each column. Sequences at internal nodes are included in the multiple alignment. The last column shows $\mathbf{f}$, constructed by combining $\mathbf{Y}$ and $\mathbf{M}(\mathbf{A})$. Letters that are present at internal sequences are unobserved and are drawn as Felsenstein wildcards. Whereas the alignment is separable from the data, as shown in column 2, the aligned data matrix is not.

are homologous, descending from the same ancestral residue at the unobserved root node. Assuming that evolution is independent across columns in $\mathbf{f}$, the tuples of letters at the leaf nodes within a column are realizations from a multinomial distribution that depends on τ, $\mathbf{T}$, and $\boldsymbol{\Theta}$ (Goldman, 1993) and is generated by the following stochastic process. The ancestral letter at the root node is drawn from a distribution γ. Evolution then occurs independently along each branch according to a continuous-time Markov process (Lange, 1997). We consider only reversible Markov models and use the equilibrium distribution for the Markov process as the root distribution γ. This makes the location of the root unidentifiable (Felsenstein, 1981), so we use unrooted trees throughout this paper.

Assuming independence across columns of $\mathbf{f}$, the full likelihood is given by multiplying the column likelihoods together. We use the peeling algorithm introduced by Felsenstein (1981) to calculate the column likelihoods, easing the computational burden for large numbers of taxa. In the peeling algorithm, both Felsenstein wildcards and gaps are similarly treated as missing data and summed out. This behavior relies on the rule that the same residue cannot be deleted and reinserted, so that each column represents only one homologous feature. As a consequence, calculating the multinomial likelihood for the observed leaf letter tuples is unaffected by the presence or absence of gaps. This allows the separation of the substitution and indel processes into the model likelihood and alignment prior, respectively.

Gap model.—In the above description, the multiple sequence alignment $\mathbf{A}$ specifically includes information about the alignment of sequences at all nodes on the tree, including sequences at internal nodes. However, the data $\mathbf{Y}$ specifies the letters only for sequences at leaf nodes, as shown in Figure 1. Because sequences at internal nodes are included in the alignment, given a topology τ, $\mathbf{A}$ can be represented as a tuple of pairwise alignments $(A^{(1)}, \ldots A^{(B)})$ along each branch in τ, as illustrated in Figure 2a. The pairwise alignment along each branch specifies the homology of the sequences at either end of the branch. This convenient representation allows us to build up a distribution on $\mathbf{A}$ from a distribution ν on pairwise alignments generated from a pair-HMM with parameters $\boldsymbol{\Lambda}$. Including the alignment states for internal nodes also specifies on which branch each indel occurs and determines its length and position.

To construct the distribution on $\mathbf{A}$ from a distribution ν on pairwise alignments, we first note that multiple pairwise alignments on adjacent branches cannot be completely independent because the alignments specify the length of the sequence at the node they share. However, after leaving the shared node, we assume evolution along each branch is then independent. Thus, pairwise alignments on a topology are conditionally independent given each alignment's neighbors. Arbitrarily choosing any branch alignment as the root alignment, we can construct a directed acyclic graph (DAG) to represent these dependencies (Fig. 2b). The DAG defines a parent branch $\rho(b)$ for every branch b. Labeling the root branch alignment



(a) Pairwise branch alignments (b) DAG

FIGURE 2. Pairwise alignments on a 4-taxon tree. (a) Each branch of the tree is labeled with the alignment along that branch. (b) Arbitrarily choosing the branch leading to taxon 1 as the root results in a directed acyclic graph (DAG) representation of dependence between the branches.

as 1 and applying standard conditioning arguments on DAGs yields

$$\mathrm{P}(\mathbf{A} \mid \tau, \boldsymbol{\Lambda}) = \mathrm{P}(A^{(1)} \mid \boldsymbol{\Lambda}) \prod_{b=2}^{B} \mathrm{P}\big(A^{(b)} \mid A^{(\rho(b))}, \boldsymbol{\Lambda}\big). \quad (6)$$

We simplify Equation 6 by recalling that dependence between neighboring alignments exists only through the sequence length at their shared node. We designate the node in τ shared by the branches b and $\rho(b)$ as $n(b)$ and we designate the length that is ascribed to the sequence at node $n(b)$ by $A^{(b)}$ as $|A^{(b)}_{n(b)}|$. This results in the expression

$$\mathrm{P}(\mathbf{A} \mid \tau, \boldsymbol{\Lambda}) = \mathrm{P}\big(A^{(1)} \mid \boldsymbol{\Lambda}\big) \prod_{b=2}^{B} \mathrm{P}\Big(A^{(b)} \Big| \big|A^{(b)}_{n(b)}\big| = \big|a^{(\rho(b))}_{n(b)}\big|, \boldsymbol{\Lambda}\Big), \quad (7)$$

where $a^{(\rho(b))}_{n(b)}$ is the realized value of random variable $A^{(\rho(b))}_{n(b)}$. The probabilities in Equation 7 now each depend only on the marginal distribution of one pairwise alignment. We can therefore choose a pairwise alignment distribution for each term. Given this freedom, we elect to use the same distribution ν for all terms, yielding

$$\begin{aligned}
&\mathrm{P}(\mathbf{A} \mid \tau, \boldsymbol{\Lambda}) \\
&\quad = \mathrm{P}_\nu\big(A^{(1)}\big) \prod_{b=2}^{B} \mathrm{P}_\nu\Big(A^{(b)} \Big| \big|A^{(b)}_{n(b)}\big| = \big|a^{(\rho(b))}_{n(b)}\big|\Big) \\
&\quad = \mathrm{P}_\nu\big(A^{(1)}\big) \prod_{b=2}^{B} \frac{\mathrm{P}_\nu\big(A^{(b)}\big) \times 1\big\{\big|A^{(b)}_{n(b)}\big| = \big|a^{(\rho(b))}_{n(b)}\big|\big\}}{\mathrm{P}_\nu\big(\big|A^{(b)}_{n(b)}\big| = \big|a^{(\rho(b))}_{n(b)}\big|\big)}, \quad (8)
\end{aligned}$$

where $1\{\cdot\}$ is an indicator function.

We note that the pairwise alignment distribution ν induces a length distribution on each of the two sequences at the alignment's ends. We restrict ourselves to distributions ν where these two length distributions are identical and label the single length distribution that results as ϕ. Multiplying all the indicator functions in Equation 8 together, we get an indicator function on a set, $S(\tau)$, in

which all pairwise alignments referring to the same specific node in τ ascribe the same sequence length to it. Given any multiple alignment $\mathbf{A} \in S(\tau)$, we simply refer to the length of the sequence at a node i as $|A_i|$ with realized value $|a_i|$, since all pairwise alignments agree on its length. This simplification yields

$$\mathrm{P}(\mathbf{A} \mid \tau, \mathbf{\Lambda}) = \mathrm{P}_v\left(A^{(1)}\right) \prod_{b=2}^{B} \frac{\mathrm{P}_v\left(A^{(b)}\right)}{\phi\left(|a_{n(b)}|\right)} \times 1_{S(\tau)}. \qquad (9)$$

Each internal node has three branches connected to it and the alignments on two of the branches condition on the alignment on the remaining parent branch in the DAG. Therefore, for each internal node i, the term $\phi(|a_i|)$ occurs twice in the denominator, resulting in the following expression

$$\mathrm{P}(\mathbf{A}|\tau, \mathbf{\Lambda}) = \frac{\prod_{b=1}^{B} \mathrm{P}_v\left(A^{(b)}\right)}{\prod_{i \in I} \phi\left(|a_i|\right)^2} \times 1_{S(\tau)}, \qquad (10)$$

where I is the set of internal nodes in τ. Given the insensitivity of Equation 10 to relabeling and choosing a different root node, it is clear that Equation 10 is independent of the choice of root alignment in the DAG.

It is impractical in an MCMC algorithm to directly sample from the full conditional posterior distribution of alignments when Equation 10 is used as the prior on alignments. To overcome this difficulty, we introduce a similar, but more convenient, alignment prior in order to create an approximate posterior distribution from which it is practical to use Gibbs sampling. We then use this modified posterior distribution as a proposal distribution in a Metropolis-Hastings (MH) transition kernel. Note that Equation 10 is not proportional to $\prod \mathrm{P}(A^{(b)})$ because the denominator depends on $\mathbf{A}$. Thus, we consider the following prior distribution

$$\mathrm{P}(\mathbf{A} \mid \tau, \mathbf{\Lambda}) = \frac{1}{K} \times \prod_{b=1}^{B} \mathrm{P}_v\left(A^{(b)}\right) \times 1_{S(\tau)}, \qquad (11)$$

where the new normalizing constant K does not depend on $\mathbf{A}$. Although we have not specified K explicitly, we can Gibbs sample from the modified posterior using dynamic programming (DP) when Equation 11 is used as the alignment prior. As described in the Appendix, we provide DP algorithms to sample from the modified posterior distribution of one, three, or five adjacent branch alignments conditional on all other parameters being fixed. When resampling only one branch at a time, however, the lengths $|a_i|$ remain constant and the acceptance probability for this proposal is always 1. However, when resampling three or five adjacent alignments, the lengths of the sequences at the internal nodes may change and we must use the full MH methodology.

We use a pair-HMM to specify the distribution on pairwise alignments. Our pair-HMM depends on three parameters, δ, ϵ, and ζ, and is depicted in Figure 3. Parameter ζ is the probability of transitioning from any state to the end state. Conditional on not transitioning to the end state, the parameter δ refers to the probability of an indel in either sequence, while the parameter ϵ refers to the probability of extending an existing gap. Gap lengths are geometrically distributed in this model with mean length $1/(1-\epsilon)$. Our pair-HMM leads to a model with affine gap penalties (Waterman, 1995), because the log probability of a single indel of length l is $\log \mathrm{P} = [\log(1-\epsilon) - \log\epsilon + \log\delta] + \log\epsilon \times l$, conditional on not transitioning to the end state.

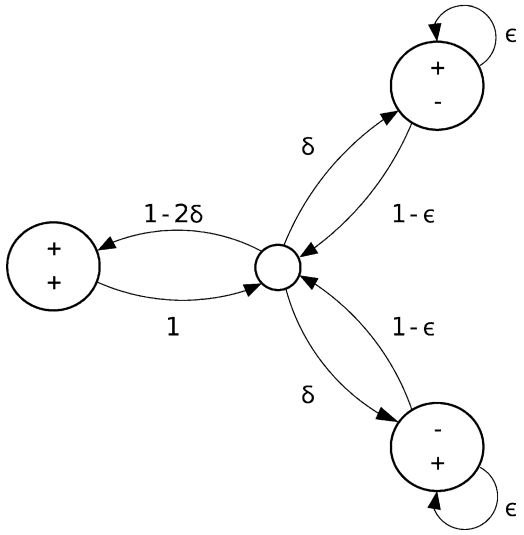


FIGURE 3. Hidden Markov model for pairwise alignments. The start and end states of the model are not shown. Every state emits (+) or does not emit (−) a residue in each of the two sequences. After a match (+/+) or a gap (+/− or −/+) ends, the chain returns to the silent state in the center. From there, a gap in either sequence opens with probability δ. Existing gaps are extended with probability ϵ, resulting in geometrically distributed gap lengths with mean length $1/(1-\epsilon)$. Transition probabilities shown are conditional on not moving to the end state. The silent state is shown here for clarity, but it can be removed, resulting in transitions only between nonsilent states.

To complete the description of the pair-HMM, we construct priors on its parameters. Because the posterior conditions on the known sequence lengths at the leaf nodes, the posterior is insensitive to reasonable changes in ζ. As a result, we fix $\zeta = 1/1000$ for our examples. In data sets with substantially longer sequences, smaller values of ζ may be appropriate. We assume a Double-Exponential distribution with mean -6 and standard deviation 0.5 on the approximate log odds of δ. This approximation results because δ's range in the pair-HMM is $(0, 1/2)$ instead of the unit-interval. The Double-Exponential distribution possesses longer tails than the Normal distribution and is, therefore, less informative. For ϵ, we note that indels must contain at least one residue and assume an Exponential distribution with mean $= 5$ on the expected length of the remaining residues. Prior medians for δ and ϵ are similar to estimates obtained from several independent data sets.

Posterior Sampling via MCMC

We sample from our model posterior $P(\mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}|\mathbf{Y})$ using MCMC. Our MCMC algorithm makes use of a number of reversible transition kernels and attempts to sample from every model parameter at least once during each iteration. The algorithm we propose employs a random-scan line (Liu et al., 1995) Metropolis-within-Gibbs (Tierney, 1994) approach. Specifically, we update each branch length, each pairwise branch alignment, and the sequence length and alignment at each internal node at least once each iteration. To update the topology, we propose a nearest-neighbor interchange (NNI) (Swofford et al., 1996) move across each internal branch at least once per iteration. Substitution parameters and hyperparameters are jointly sampled a Poisson number of times, with mean $1 + B/3$, where B is the number of branches. Thus, the substitution parameters are resampled after roughly three branch alignments have been resampled. Indel parameters are jointly sampled a Poisson number of times, with mean $1 + 2B$. Indel parameters are resampled more frequently than substitution parameters, reflecting the slightly slower mixing of the indel parameters. All proposals are executed in a random order. Random scanning improves mixing compared to a fixed order of proposals (Liu, 2001).

The proposals for sampling from branch lengths, substitution parameters, and indel parameters are straightforward. We employ MH updates using Gaussian proposal densities. For substitution parameters and external branch lengths, these proposals are reflected about their boundaries. If a negative internal branch length is proposed, this induces an NNI across that branch (Suchard et al., 2003b), resulting also in a topology update. Both the alignment and the topology require more complicated proposal distributions. In both cases, DP is used to consider an exponential number of possible states in polynomial time.

Topology sampling.—The topology τ is updated through a number of MH steps, each of which alters only part of the topology. Each internal branch of the tree is connected to four subtrees. Interchanging the four subtrees produces three possible topologies: the original topology and two alternatives. However, after an NNI topology change, the five pairwise alignments along the main branch and the four connected branches become undefined. This results because it may not be possible to maintain the pairwise alignment of all pairs of nodes on the 4-taxon subtree without violating the rule that the same residue cannot be deleted and reinserted. To solve this problem, we first integrate over all possible values of the five affected pairwise alignments given each topology, subject to the constraint that alignments between leaf nodes on the 4-taxon subtree around the branch do not change. In doing so, we marginalize from the state-space alignment information that may conflict with a new topology. Once a new topology is chosen, we reintroduce information consistent with this topology by resampling from the alignments over which we integrated. We refer to this integration and subsequent resampling as 5-way sampling. Given the constraint, we need to consider only the presence and absence of residues at the two internal nodes on the 4-taxon subtree. The resulting summation is a one-dimensional (1D) DP problem. We can then compute the partially marginalized probabilities of the three possible topologies based on the likelihood of each topology and the sum of the priors for all realizations of the five alignments. For our proposal distribution, we first choose between the topologies in proportion to their probability. After a topology is chosen, the internal nodes are resampled from the DP matrix for that topology. Note that the alignment $\mathbf{A}$ and the topology τ are both resampled after this step, although the alignments of leaf taxa to each other do not change.

This proposal distribution does not exactly match the target distribution, because our DP step uses (11) as the gap prior $P(\mathbf{A} \mid \tau, \boldsymbol{\Lambda})$. We control for the mismatch by rejection sampling in an MH step. Let n_1 and n_2 be the nodes at the ends of branch b across which the NNI move is proposed. Further, let $\ell^{(0)}_{n_1} = |a^{(b)}_{n_1}|$ and $\ell^{(0)}_{n_2} = |a^{(b)}_{n_2}|$ be the lengths of sequences ascribed to n_1 and n_2 before the proposal and let $\ell^{(1)}_{n_1}$ and $\ell^{(1)}_{n_2}$ be the corresponding lengths after the NNI and alignment resampling. Then, the MH acceptance probability (see Appendix) is

$$\min\left\{\frac{\phi\left(\ell^{(0)}_{n_1}\right)^2\phi\left(\ell^{(0)}_{n_2}\right)^2}{\phi\left(\ell^{(1)}_{n_1}\right)^2\phi\left(\ell^{(1)}_{n_2}\right)^2}, 1\right\}. \tag{12}$$

Unless the sequence lengths are extremely short, ϕ varies slowly with length. Thus, unless the difference between the lengths before and after resampling is large, the numerator and denominator will be very similar and the acceptance probability will be close to 1.

Alignment sampling.—Holmes and Bruno (2001) developed two MCMC transition kernels to resample alignments. When both moves are employed, the resulting chains are ergodic. The first move is to resample a single pairwise alignment along a branch, leaving other alignments constant. The second move resamples three pairwise alignments on the branches of a 3-taxon subtree, conditional on the implied pairwise alignments between sequences at the leaf nodes being fixed. This effectively resamples the presence or absence of the sequence at the internal node in each column of the alignment. Both of these updates can be accomplished through DP and are of computational orders $O(C^2)$ and $O(C)$ respectively.

Using only the two Holmes and Bruno (2001) updates can result in inefficient MCMC mixing in some circumstances. On a 3-taxon subtree, if a residue is present in two leaf sequences but missing in the third, then moving between aligned and unaligned states requires movement through an intermediate state where the residues are unaligned, but one of them is present at the central node. This intermediate state contains an extra indel over the unaligned state and two extra indels over the aligned state (Fig. 4), resulting in slow mixing between these two states. As a consequence, MCMC chains must be run for

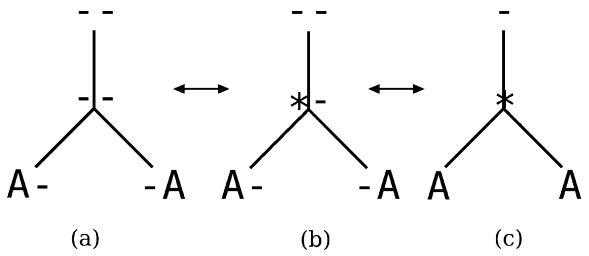


FIGURE 4. Unlikely intermediate results in inefficient mixing. Here we assume that only the two transition kernels of Holmes and Bruno (2001) are available. If two characters present in two sister taxa are not present in the third taxon (a), then in order to reach a state where they are aligned (c), one of the characters must become present at the internal node (b). Only then can the other feature be aligned to it. However, this path introduces an unnecessary indel, which means that the intermediate state (b) is less likely than either (a) or (c). Thus, transitions between (a) and (c) will be inefficient.

much longer times to assess questionable alignment or unalignment rather than to assess other uncertainties, such as gap positions.

We first note that simultaneously resampling the entire 3-way alignment solves this difficulty, resulting in greatly increased mixing (Jensen and Hein, in press). However, this is an extremely expensive operation at $O(C^3)$. In addition, the amount of computer memory needed for this operation makes it difficult to analyze sequences of even medium length.

We describe a method by which the mixing problem can be alleviated and some of the benefit of sampling the entire 3-way alignment can be gained, at the cost of only an $O(C^2)$ algorithm. Notably, the alignment along a branch and the sequence at an internal node at one end of the branch are resampled in the same step, decreasing coupling between adjacent points in the MCMC chain (Roberts and Sahu, 1997). The transition kernel that we propose consists of sampling from the entire 3-way alignment subject to the constraint that the alignment of two of the three leaf sequences is held constant, including the ordering of the columns that are not strictly ordered. When using DP to draw samples, 3-way alignments correspond to paths through a cubic DP matrix. Our approach is equivalent to sampling only from the set of paths that result in the equivalent 2D path produced by projecting down onto a specific face of the matrix. This results in a 2D DP problem as described in the Appendix.

Alignment Uncertainty (AU) Plots

As posterior sampling of alignments using MCMC is a relatively unexplored topic, little work has focused on methods to summarize such posterior samples in an easily interpretable form. We describe a method to annotate a point-estimate $\hat{\mathbf{A}}$ of the alignment from the posterior with a measure of uncertainty about the point-estimate. The method uses a coloring scheme. This scheme marks each letter with the approximate probability that the letter is homologous to the ancestral residue in its column. The scheme also marks each gap with the approximate probability that no letter from the gapped sequence is homologous to the ancestral residue in the gap's column. For the alignment to annotate $\hat{\mathbf{A}}$, we use the alignment from the most probable point in our complete model space. The ability to annotate individual entries, letters or gaps, in $\hat{\mathbf{A}}$ versus entire columns is important because the number of columns that are completely homologous necessarily decreases as the number of taxa in the multiple alignment increases. Although a naive approach might annotate each entry independently, it often happens that one subset of taxa aligns poorly with another subset of taxa, but the sequences in each subset align strongly to each other. Thus we seek a method of summarizing the posterior which captures these groupings.

Let $\hat{f}_c$ represent column c in $\hat{\mathbf{A}}$. For each column $\hat{f}_c$ separately, we approximate the posterior probabilities that individual entries in $\hat{f}_c$ align with the remaining letters. To accomplish this, we fit a simple phylogenetically based Poisson model to the posterior sample $\mathbf{A}^{(p)}$ for $p = 1, \ldots, P$. Across all posterior samples, we count the number of times U_{ij} that pairs of entries in $\hat{f}_c$ from taxon i and j are no longer found in the same alignment column. We consider gaps in $\hat{f}_c$ to reside in the first column of $\mathbf{A}^{(p)}$ that contains a gap in the same row and at least one other letter of $\hat{f}_c$ in other rows. When no such column is found, we place the gap in a special column. We convert the counts of unalignment events into pairwise distances D_{ij} by assuming a Poisson model, such that

$$D_{ij} = -\log\left(1 - \frac{U_{ij}}{P}\right). \tag{13}$$

Under this measure, a long distance between a pair of entries indicates a high probability that the pair is unaligned across the posterior sample. Based on the maximum a posteriori (MAP) topology, we then reconstruct unalignment branch lengths from the pairwise distances using least-squares (Cavalli-Sforza and Edwards, 1967). Finally, for each entry in $\hat{f}_c$, we estimate the probability that no unalignment event occurred between the leaf node corresponding to the entry and the mid-point root (Farris, 1972). On the AU plot, we assign a color or grayscale to each entry based on this probability.

Computation

We developed a new software program in C++ named BAli-Phy to sample from the joint posterior of our alignment and phylogeny model. BAli-Phy is available to interested readers at our Web site (http://www.biomath.ucla.edu/msuchard/bali-phy/). To draw inference given the sequence data used in this paper, we used this software to generate multiple MCMC samples. For the 5-taxon problems, we discarded the first 1000 samples from each chain as burn-in and then collected 80,000 to 150,000 further samples for analysis. For the 12-taxon problem, we discarded the first 5000 samples as burn-in and collected 40,000 further samples. We ran BAli-Phy

on standard PC hardware: an AMD Athlon 1.7 GHz processor with 1 Gb of RAM. Generating posterior samples required 1 to 8 days of CPU time for our smallest to largest problems.

RESULTS

Early Branching in the Tree of Life: Resolving the Archaea

Resolution of the early branching order in the Tree of Life remains controversial because support for these deep branches is often low and depends on the data and phylogenetic methods used (Baldauf et al., 1996; Roger et al., 1999; Brown and Doolittle, 1997). These issues are exacerbated by the difficulty in aligning distantly related sequences; estimated phylogenies can be strongly affected by biased multiple alignments generated using guide trees (Lake, 1991) and by the common practice of throwing out ambiguous regions in the alignment that can decrease resolution (Lee, 2001).

One important question about deep branches in the Tree of Life that remains unresolved is whether the Archaea form a monophyletic group (Brown and Doolittle, 1997). Archaea were initially classified as Eubacteria because both groups lack nuclei, but Woese et al. (1990) later separated the Archaea from other prokaryotes based on a phylogeny reconstruction using the 16S rRNA. Woese et al. (1990) divided all living organisms into the three domains: Archaea, Bacteria, and Eucarya. This division has been further supported by research into the molecular biology of Archaea. Archaea have been found to have many traits formerly thought to be exclusive to the Bacteria, e.g., lack of organelles, and exclusive to the Eukaryotes, e.g., histone-like proteins and specific transfer RNA (tRNA) introns (Brown and Doolittle, 1997). Although initial 16S rRNA analyses suggest that Archaea form a monophyletic group, some recent analyses based on proteins such as EF-1α/Tu suggest that the Crenarchaeotes branched separately from the remaining Archaea and are sister taxa to the Eukaryotes (Rivera and Lake, 1992). This discrepancy between results presents two alternative hypotheses about the early branching order in the Tree of Life, each represented by a different phylogenetic tree. The first tree contains a single Archaea clade; we refer to this tree as the archaeal tree (Fig. 5a). The alternative tree places the Crenarchaeotes, also called Eocytes after Lake (1991), as sister taxa to the Eukaryotes; we refer to this tree as the eocyte tree (Fig. 5b). We apply our methodology of joint Bayesian estimation of alignment and phylogeny to this problem in hopes of improving the consistency and certainty of results.

TABLE 1. Species contributing sequences to either example. The 12 taxa in this table make up the 12-taxon data set for the EF-1α/Tu example. Taxa marked with a * make up the 5-taxon data set for the 5S rRNA and EF-1α/Tu examples.

Taxa	Domain	Order	Note
Homo sapiens (HS)*	Eukaryotes	Metazoa	Human beings
Nicotiana tabacum	Eukaryotes	Plantae	Tobacco plant
Euglena gracilis	Eukaryotes	Protista	Photosynthetic single cell
Giardia lamblia	Eukaryotes	Diplomonadida	Intestinal protist
Sulfolobus acidocaldarius (SA)*	Archaea	Crenarchaeota	High pH thermophile
Aeropyrum pernix	Archaea	Crenarchaeota	Anaerobic thermophile
Pyrococcus woesei (PW)*	Archaea	Euryarchaeota	Thermophile
Halobacterium salinarum (HA)*	Archaea	Euryarchaeota	Methanogen
Methanococcus vannelli	Bacteria	Aquificae	Thermophile
Thermotoga maritima	Bacteria	Thermotogales	Thermophile
Anacystis nidularans	Bacteria	Cyanobacteria	Photosynthetic, aquatic
Escherichia coli (EC)*	Bacteria	Proteobacteria	Intestinal symbiont

We consider two different types of molecular sequences present throughout the Tree of Life: the short nucleotide sequences of the 5S rRNA and the longer amino acid sequences of the EF-1α/Tu protein. For each type of sequence, we examine data sets containing five taxa. In addition, we will examine a 12-taxon data set for EF-1α/Tu, to show that our method can handle a larger number of taxa if need be. Table 1 lists all 12 taxa used in our analyses.

Example 1: 5S rRNA

The 5S rRNA is found in Archaea, Bacteria, and Eukaryotes and has a highly conserved secondary structure (Barciszewska et al., 2001). At the present time, only a few organisms have been found to lack 5S rRNA; these include *Giardia* (Edlind and Chakraborty, 1987) and some mitochondria (Gray et al., 1999). The 5S rRNA forms part of the large ribosomal subunit and is thought to play a role in stabilizing that subunit. The 5S rRNAs we examine vary in length from 120 to 126 nucleotides, and their sequence identity runs as low as 46%. Phylogenetic studies have generally focused on much larger rRNAs such as

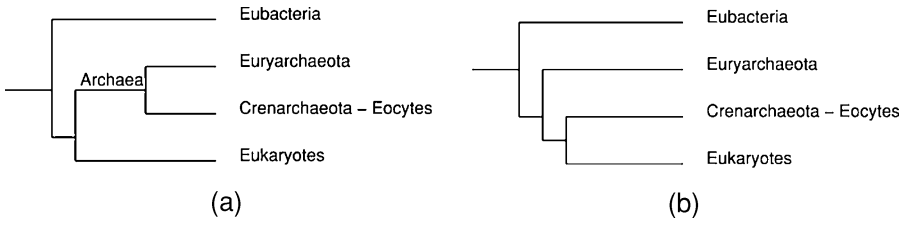


FIGURE 5. Two possibilities for early branching in the Tree of Life. (a) The archaeal tree implies that Archaea form a monophyletic group. (b) The eocyte tree implies that Archaea are paraphyletic (Eocytes and Euryarchaeota) and that the eocyte Archaea are more closely related to Eukaryotes than to other Archaea.

the 16S and 25S subunits that contain considerably more characters. The 5S rRNA has generally been considered too small to give reliable phylogenetic signals for such divergent trees as the Tree of Life.

Model and priors.—We use the Hasegawa et al. (1985; HKY85) model of nucleotide substitution. We fix the nucleotide frequencies in the model to their empirical frequencies, as the posterior frequency distribution rarely departs considerably from the empirical frequencies (Li et al., 2000). The HKY85 model is reversible with one free parameter, κ, that is equal to the ratio of the transition to transversion rates among nucleotides. Because κ is a ratio, we assume a diffuse log-Normal prior with median $\kappa = 2$ to reflect the belief that transition rates are generally higher than transversion rates. We describe our prior on the indel process parameters $\mathbf{\Lambda}$ in "Gap model." We further assume a Uniform prior over the 15 topologies that are possible with five taxa and assume the Exponential prior on the hyperparameter μ has mean 0.5.

Alignment uncertainty.—Figure 6 shows the AU plot inferred for the 5S rRNA example. As seen in the figure, homology is well resolved in the first half of the alignment, illustrated by the dark background shading. The second half of the alignment is less well resolved, especially the alignment of *S. acidocaldarius* with respect to the other sequences. In addition, uncertainty in the exact positions of gaps is clearly visible because the letters on either side of the gaps are shaded lightly, indicating that their positions are not well resolved. For example, note that the position of the gap marked with a + in the *H. sapiens* sequence is not well resolved. The three letters to the left are more lightly shaded, indicating that their presence in their respective columns is uncertain. This is an indication that the gap may move across these columns.

Phylogeny estimation.—Under the HKY85 model, the posterior mean of the transition/transversion rate ratio is 1.90, and the 95% Bayesian credible interval (BCI) is (1.18, 2.87). The posterior distribution of κ varies only modestly from its prior, reflecting the lack of information available given the length of the sequences is small and the alignment is ambiguous. The posterior mean of $\log \delta$ is -4.98 and its 95% BCI is $(-5.76, -4.29)$. The posterior mean of $\log \epsilon$ is -0.57 and its 95% BCI is $(-1.61, -0.18)$. This interval corresponds to expected indel lengths ranging from 1.25 to 6.07 and is informative because the range implies that both the single-residue and multiple-residue indel regimes are supported by the limited amount of information available in the data.

Given the data, the MAP topology clusters *E. coli* with *H. sapiens* and *H. salinarum* with *P. woesei* with a posterior probability (PP) of only 0.308 (Table 2). No single topology is strongly supported. Marginal clustering of *E. coli* and *H. sapiens* supports the hypothesis of archaeal monophyly (PP = 0.553). Clustering of *H. salinarum* with *P. woesei* has PP = 0.494. The support for these partitions is not high, with a posterior odds ratio of only 1.23 for the partition with the strongest support. This lack of resolution results because much of the alignment is uncertain (Fig. 6) and several near-optimal alignments support different topologies. As such, we believe that this lack of topological resolution yields an accurate portrayal of the limited phylogenetic information in these sequences.

To demonstrate the ability of our methodology to avoid exaggerated confidence in inferred topologies by considering multiple alignments, we consider model restrictions that bring our methodology more in line with standard phylogenetic reconstruction techniques. First, we fix the alignment to that estimated by ClustalW and rerun our sampler with this constraint. This type of analysis can be strongly biased towards the guide tree used by ClustalW because progressive alignment algorithms insert shared gaps into taxa clustered on the guide tree when these clusters are aligned as a group against other clusters. These shared gaps are interpreted by the joint model as strong evidence for common descent because indels are less frequent than substitutions. To avoid this extreme bias towards the guide tree, we also consider the fixed ClustalW alignment under the traditional Bayesian phylogenetic model described in Equation 2 and implemented in software such as BAMBE and MrBayes (Larget and Simon, 1999; Huelsenbeck and Ronquist, 2001). As opposed to the joint model introduced in Equation 4, which uses information in shared gaps, the traditional

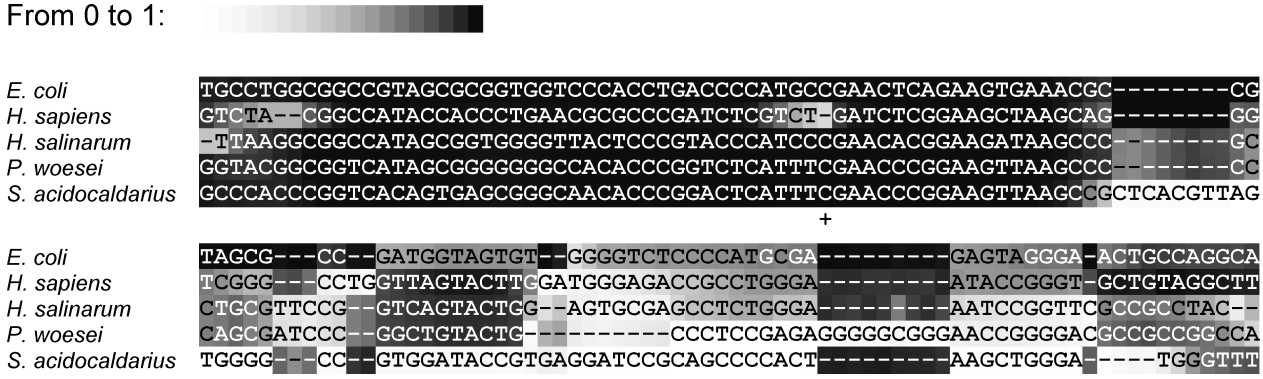


FIGURE 6. Alignment uncertainty plot for the 5S rRNA. Dark shading indicates that an entry is well resolved; light shading indicates that the position is not well resolved. The first half of the alignment is fairly well resolved, but the second half is much less well resolved. Uncertainty in the exact position of gaps is visible as light shading in adjacent letters. *S. acidocaldarius* does not align well with the other sequences.

TABLE 2. Support for the most probable 5-taxon 5S rRNA topologies and partitions. Columns report the posterior probability (PP) and $\log_{10}$ odds (LOD) with their 95% Bayesian credible interval (BCI) in favor of each hypothesis under three different models. These models are the joint estimation model presented in this paper (JE), a joint model constrained to a fixed alignment (Indels), and a traditional model based on a fixed alignment (NoIndels). Fixed alignments are estimated using ClustalW. Although alignments are fixed, the second model does make use of indel information while inferring the phylogeny, whereas the latter model does not. Taxa abbreviations are given in Table 1 and BCIs are estimated using a block bootstrap (Suchard et al., 2003b).

	JE			Indels			NoIndels		
Topologies and partitions	PP	LOD	95% BCI	PP	LOD	95% BCI	PP	LOD	95% BCI
((EC,HS),(SA,(PW,HA)))	0.308	−0.35	(−0.38, −0.32)	0.996	+2.38	(+2.33, +2.43)	0.172	−0.68	(−0.69, −0.67)
((EC,HS),((SA,PW),HA))	0.208	−0.58	(−0.63, −0.54)	0.002	−2.71	(−2.77, −2.65)	0.700	+0.37	(+0.36, +0.38)
(EC,((HS,SA),(PW,HA)))	0.120	−0.86	(−0.90, −0.83)	0.001	−2.89	(−3.02, −2.80)	<0.001	−4.88	< −4.70
((EC,PW),((SA,HS),HA))	0.088	−1.02	(−1.08, −0.97)	<0.001		<−5.17	<0.001		<−5.17
((EC,SA),(HS,(PW,HA)))	0.066	−1.15	(−1.21, −1.10)	0.001	−3.12	(−3.30, −3.11)	<0.001	−4.03	(−4.40, −3.86)
((EC,HS),(PW,(SA,HA)))	0.037	−1.42	(−1.47, −1.37)	<0.001	−3.47	(−3.62, −3.37)	0.127	−0.84	(−0.84, −0.83)
EC,HS | HA,PW,SA	0.553	+0.09	(+0.06, +0.13)	0.998	+2.72	(+2.64, +2.82)	>0.999	+3.78	(+3.50, +4.06)
EC,HS,SA | PW,HA	0.494	−0.01	(−0.04, +0.02)	0.998	+2.64	(+2.59, +2.70)	0.173	−0.68	(−0.69, −0.67)

model gains information about topologies only from substitutions and ignores information in shared gaps.

The first restricted model supports the same MAP topology as the full joint model, while the second restricted model supports a different MAP topology (Table 2). Unlike the joint model, the first restricted model strongly supports its MAP topology (PP = 0.996), which is the same as the topology of the guide tree used by ClustalW in its estimate of the multiple alignment. This high PP supports the hypothesis that the use of information in shared gaps for inferring phylogeny is problematic when based on multiple alignments generated using progressive alignment. Although the second restricted model does not support a single topology very strongly, it does support the marginal clustering of *E. coli* and *H. sapiens* with a PP > 0.999. We note that this clustering occurs in the guide tree and is evidence of bias. These findings illustrate that short sequences can indeed yield strong posterior support for specific partitions of taxa, but that this support can almost entirely result from alignment bias when the alignment signal is not strong.

Example 2: EF-1α/Tu

For our second example, we turn to EF-1α/Tu. EF-1α is a highly conserved protein found in Eukaryotes, Archaea, and Bacteria where it is called EF-Tu. It is a monomeric G protein, and its functional role is to load tRNAs onto the ribosome during translation. The sequences we examine vary in length from 394 to 462 amino acids. This protein is at least 26% conserved between all 12 taxa in our data set. This value is near the threshold of 20% to 25% conservation below which homology becomes difficult to detect (Rost, 1999). Other researchers have attempted to make inferences about the Tree of Life based on EF-1α/Tu (Rivera and Lake, 1992; Baldauf et al., 1996; Roger et al., 1999).

Model and priors.—We assume a Uniform prior over all possible topologies for 5 or 12 taxa, depending on the data set examined. Our priors on branch lengths and alignments remain as described in the previous 5S rRNA example. We use the Whelan and Goldman (2000; WAG) substitution model, a reversible amino acid model estimated from many proteins. After fixing the amino acid frequencies to their empirical frequencies, there are no free parameters in WAG model.

We also extend the basic substitution process by introducing rate heterogeneity across sites (Yang, 1996). Specifically, we incorporate both an invariant-sites (INV) approach (Adachi and Hasegawa, 1995) and model rate variation in the remaining sites according to a Gamma distribution (Yang, 1994). We approximate the Gamma distribution (Γ_4) using four bins of equal probability. These rate variation models are standardly employed in traditional phylogenetic reconstruction, but have not previously been invoked during the alignment process. We assume a Beta prior on the fraction of invariant sites with a mode at 0.05 and 95% of its mass less than 0.20. We further characterize the Gamma distribution by its coefficient of variation α and place a diffuse Double-Exponential prior on $\log \alpha$.

Alignment uncertainty.—The AU plot for the EF-1α/Tu alignment is presented in Figure 7. Most of the alignment is shown with a dark background indicating that the alignment is well resolved under our model. Note that most of the uncertainty comes from the weak alignment between the *E. coli* sequence and the other sequences. The *E. coli* sequence is much shorter than the other sequences and is more distantly related to the others than they are to each other. The distant relationship between *E. coli* and the other sequences decreases the alignment signal; the alignment uncertainty is further increased by the substantial number of gaps that must be introduced to account for the length difference in the sequences. For example, in the first row in Figure 7, there is an ambiguous section; this section is ambiguous because the position of the *E. coli* sequence may be aligned in multiple near-optimal ways.

We also note that several shared insertions are depicted in Figure 7. *H. sapiens* and *S. acidocaldarius* share a 7 to 10 amino acid insertion marked +. This insertion has been previously reported (Rivera and Lake, 1992). Although the exact size of this insertion is uncertain, its presence is highly supported, given the dark shading of the seven amino acids in the interior of the

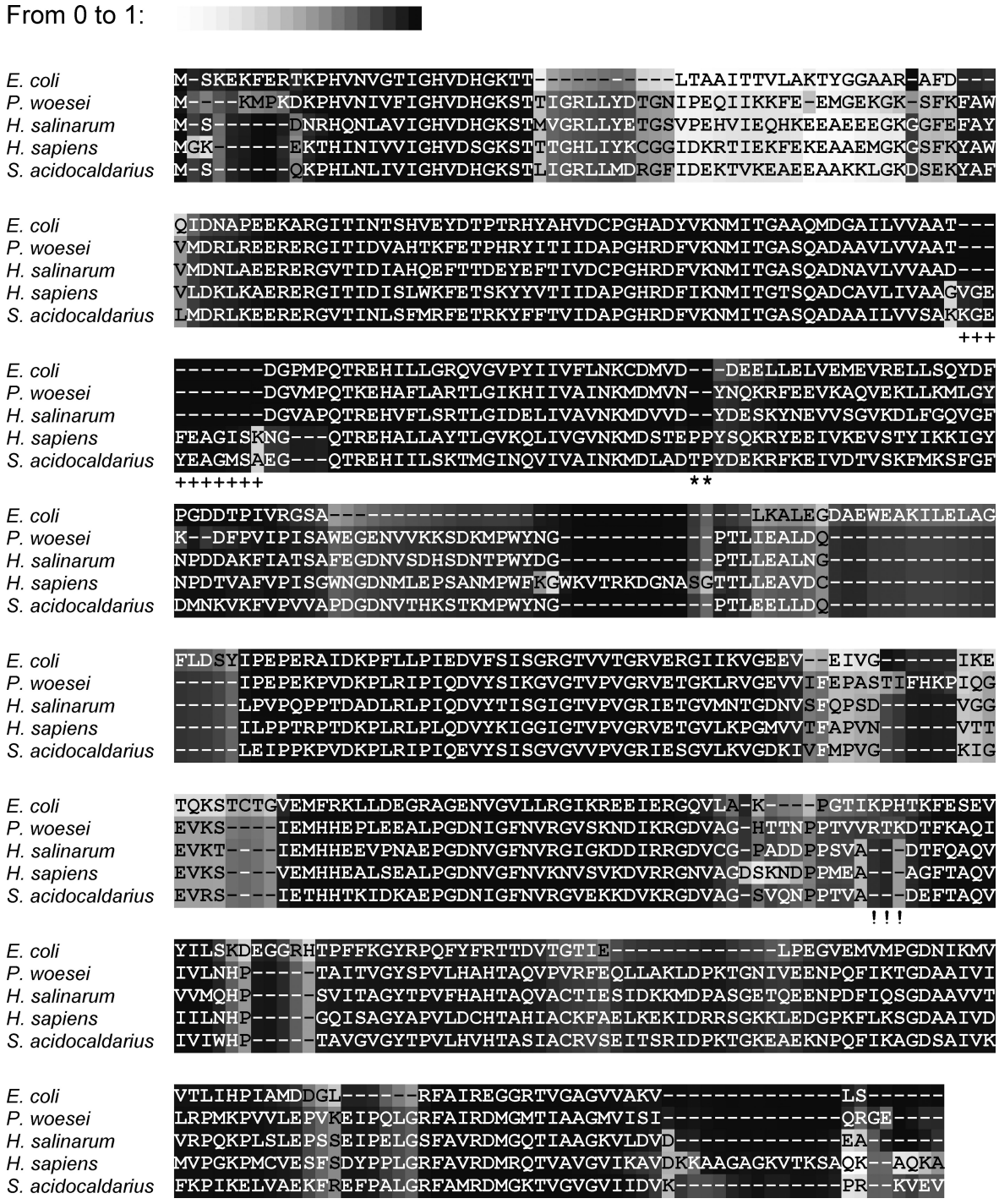


FIGURE 7. Alignment uncertainty plot for EF-1α/Tu. Well-resolved entries have a dark background, whereas less well-resolved entries are given a light background. Three shared insertions are present in this alignment: the 10–amino acid insertion in *H. sapiens* and *S. acidocaldarius* marked with +, the 2–amino acid insertion in the same species marked with *, and the 3–amino acid insertion in *P. woesei* and *E. coli* marked with !. The first and last insertions have uncertain size and location as the lightly shaded adjacent letters indicate.

insertion. *H. sapiens* and *S. acidocaldarius* also share a 2–amino acid insertion marked *. Finally, *E. coli* and *P. woesei* share a 3–amino acid insertion marked ! whose presence is not as highly supported. The exact location and length of this last insertion remains the most variable.

Phylogeny estimation.—The posterior mean of $\log \delta$ for the 5-taxon data set is -5.38 and its BCI is $(-5.80, -4.99)$. The probability of a gap along a branch is only about 50% higher in the 5S rRNA example than here. However, the posterior mean of $\log \epsilon$ is -0.19 with a BCI of $(-0.30, -0.10)$, signifying that the expected gap length

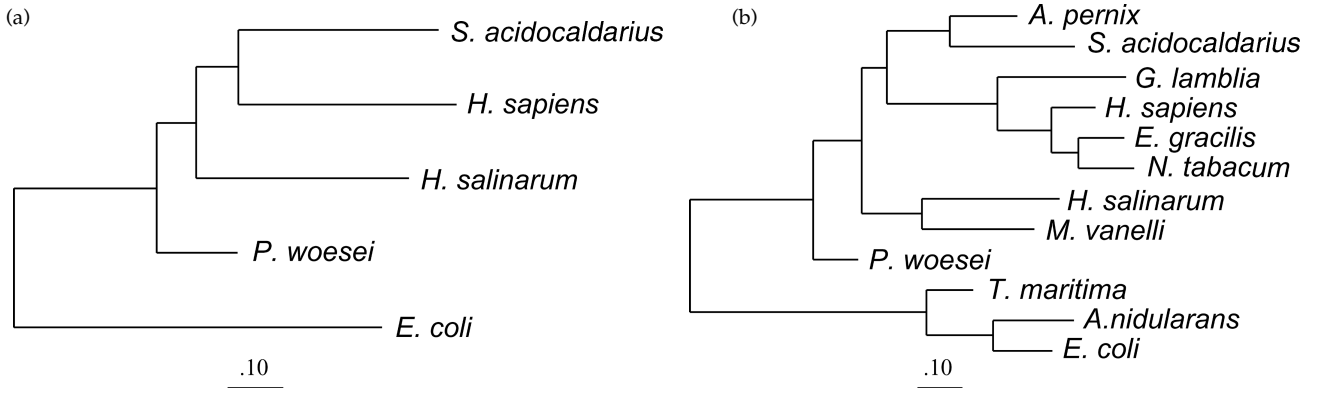


FIGURE 8. MAP topologies for EF-1α/Tu 5- and 12-taxon data sets. Reported branch lengths are posterior means. The 12-taxon tree (b) is consistent with the 5-taxon tree (a). Both trees support the eocyte hypothesis. Both trees also place *P. woesei* closer to the root and in a separate clade from *H. salinarum*.

is longer than for the 5S rRNA and a model that can generate multiple-residue indels is strongly preferred. Substitution models with rate heterogeneity give very similar estimates. For the 12-taxon example, $\log \delta$ has a posterior mean of -5.91 and $\log \epsilon$ has a posterior mean of -0.25. As expected, $\log \delta$ is slightly smaller for the 12-taxon data set, because branches are shorter (posterior mean of μ: 0.200 vs 0.348). For the same reason, gaps tend to be shorter, although the difference between examples is not as pronounced.

The MAP topology for the 5-taxon data set is drawn in Figure 8a. In the topology, *H. sapiens* and *S. acidocaldarius* are nearest neighbors with PP > 0.999. Thus, this data set strongly supports the eocyte hypothesis under our model. Part of this strong support stems from our methodology's ability to use the information from the 10–amino acid insertion shared by *H. sapiens* and *S. acidocaldarius* (Fig. 7) to support the clustering of these two taxa. In addition, *P. woesei* is not grouped with *H. salinarum* on the MAP topology, as would be expected if the remaining Archaea were monophyletic. Instead, *P. woesei* is grouped with *E. coli* with PP $= 0.974$ (0.964–0.983, 95% BCI). This grouping is supported by the common insertion in *E. coli* and *P. woesei* marked with ! in Figure 7. Our findings for this gene imply that the Archaea may have branched off independently several times and are polyphyletic. However, we note that this finding is dependent on the rooting of the Eukaryote-Archaea subtree by the Eubacteria outgroup tree and that a more flexible substitution model may be necessary to accurately handle such questions (Phillipe and Forterre, 1999; Van de Peer et al., 2000).

Incorporating both rate heterogeneity extensions leads to three additional models: INV, Γ_4 and INV+Γ_4. These three models have 1, 1, and 2 additional free parameters, respectively. Using an importance sampling estimator (Newton and Raftery, 1994; Suchard et al., 2003a), we calculate that all extensions significantly increase the marginal likelihood of the data, with the INV+Γ_4 model leading to the most improvement of roughly 19 log units over a model without rate variation. Under the INV+Γ_4 model, posterior support for clustering *E. coli* with *P. woesei* rises to 0.996 (0.992–0.999). This equates to an increase in the odds ratio from 37.5 to 249.0, providing further support for polyphyly in this gene.

Figure 8b shows the MAP topology for the 12-taxon data set. The topology here is consistent with the MAP topology for the 5-taxon data set and retains similarly high support with a PP > 0.999. The only observed uncertainty lies in the placement of *P. woesei* in a separate clade from other Euryarchaeota. This branching is consistent with the placement of the Thermococcales in EF-1α/Tu phylogenies reconstructed by Keeling et al. (1998). We note also that, in the MAP topology, *E. gracilis* finds itself in the same clade as plants, as suggested by Van de Peer et al. (2000). Finally, the MAP topology places *G. lamblia* as an early branching Eukaryote, consistent with previous studies (Keeling et al., 1998; Baldauf et al., 1996).

Convergence and Efficiency

When estimating posterior probabilities via MCMC sampling, it is common practice to throw out samples from the beginning of the chain. By discarding this burn-in period, we ignore initial samples that tend to be correlated with the starting point of the chain and not representative of the probability distribution of the model we are simulating. A chain with a smaller burn-in period requires less computation time to produce accurate estimates of the posterior distribution of parameters such as the alignment $\mathbf{A}$, the tree $(\tau, \mathbf{T})$, the substitution parameters $\mathbf{\Theta}$, and the indel parameters $\mathbf{\Lambda}$. We show that our MCMC algorithm converges to its equilibrium distribution more quickly than is possible given previously available MCMC transition kernels. This speed-up allows us to start from randomly chosen alignments and trees, instead of starting from a position estimated using other software such as ClustalW. In addition, if two successive samples from the MCMC chain are highly correlated, then one requires a large number of samples and, therefore, a large amount of computer time in order to produce accurate estimates from the samples remaining after the burn-in period. We show that the number of

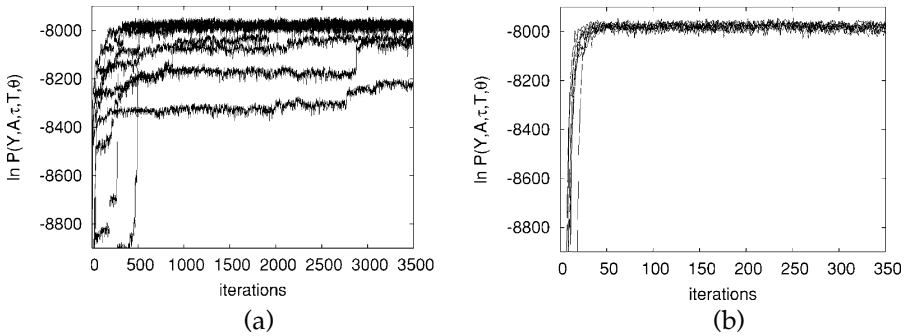


FIGURE 9. Enabling the 3-way sampling MCMC transition kernel decreases burn-in time and autocorrelation. (a) The log probability of the data and parameters takes greater than 3500 iterations to converge to its equilibrium distribution when 3-way sampling is disabled. (b) The log probability converges to its equilibrium distribution within about 50 iterations when 3-way sampling is enabled. Thus, 3-way sampling decreases burn-in time by a factor of at least 70 for the 12-taxon EF-1α/Tu data set used here. Time-series for 10 runs are shown in each plot.

samples obtained from our MCMC algorithm are sufficient to yield accurate estimates.

Decreased burn-in time.—In our MCMC algorithm, we extend the work of Holmes and Bruno (2001) in sampling alignments by adding a new transition kernel. This new routine allows a section of sequence with questionable homology to be unaligned or realigned to a section of another sequence without going through an unfavorable intermediate that temporarily introduces an extra gap. MCMC chains that include this transition kernel have substantially decreased burn-in times and less autocorrelation. We plot the log probability of the data and parameters for 3500 iterations of 10 chains without the new transition kernel (Fig. 9a) and for 350 iterations of 10 chains with the new kernel (Fig. 9b). The chains that include the new kernel require at most 50 iterations to reach equilibrium values. However, the chains that do not include this kernel take greater than 3500 iterations to converge to equilibrium values of the log probability, at least 70 times more iterations. Because the chains that contain the new transition kernel run only about 25% slower, using this transition kernel provides substantial improvement in computational efficiency. Further inspection of the samples from the slowly converging chains confirms our hypothesis that the low log probabilities and high correlation observed in the chains in Figure 9a result from alignments containing unaligned homologous regions (data not shown).

Convergence and mixing.—To assess convergence for continuous parameters, we compute Gelman-Rubin R statistics based on 10 chains for sampled leaf branch lengths and substitution and indel parameters. Internal branch lengths do not necessarily retain definition across topologies and can not be used. For our largest example involving the EF-1α 12-taxon data set, all statistics are $<$1.01, suggesting that each chain converged to the same distribution. To assess mixing, we estimate the number of effectively independent samples based on parameter autocorrelation. Effective sample sizes range from 2874 to 16,876 depending on the parameter for chains of length 40,000. Further, we report credible intervals for estimates of PPs and log odds based on a block bootstrap (Suchard et al., 2003b) (Table 2). These intervals give a quantitative measure of the accuracies of the PPs we report and account for the effects of long-distance correlation between all parameters, including discrete parameters such as the topology and the alignment.

REMARKS

In this paper, we present a Bayesian method of simultaneously estimating alignments and phylogenies directly from unaligned sequence data. The joint estimation approach increases the resolving power of the data by making use of information in shared indels while avoiding bias and overconfidence in inferred topologies resulting from inaccurate alignments. In developing our approach, we construct an MCMC algorithm to sample from the joint posterior distribution of alignments and phylogenies. The algorithm introduces a new transition kernel that significantly improves the speed and quality of posterior alignment samples. We describe a method to summarize these posterior samples into a single plot that clearly identifies regions of alignment interest and ambiguity.

Simultaneous construction of alignment and phylogeny enables one to incorporate a more realistic indel model than is traditionally used in alignment reconstruction algorithms. The availability of an internal estimate of the tree during alignment avoids overcounting indels that are shared between taxa by common descent. Furthermore, the availability of the tree allows for a more accurate alignment model by counting indel events instead of counting gaps. Alignment models based on gap counting are biased against insertions in only one or a few taxa, because these insertions are counted as independent gaps in all other taxa. Finally, the joint model uses shared indels in multiple taxa as evidence for common descent in phylogenetic reconstruction. This information appears important in the EF-1α/Tu example. Accounting for shared indels contributes to the strong posterior support of the eocyte tree. The ability to incorporate the evidence of shared indels into a statistical framework allows such evidence to be properly weighed against evidence of shared substitutions.

Joint estimation further avoids bias toward particular phylogenies as well as exaggerated confidence in inferred phylogenies. Although accurate point-estimates of phylogenies are important, knowledge about the variability of estimates produced is equally important in verifying phylogenetic hypotheses. The estimation framework presented in this paper improves accuracy by decreasing bias towards the external guide trees used in progressive alignment. These external guide trees are usually estimated using very simple models and methods that do not adequately capture the rich information in the data. In addition, operating in a Bayesian framework allows us to consider all possible alignments weighted by their posterior probabilities, such that all near-optimal alignments are taken into account automatically. Both of these improvements come to light in the analysis of the 5S rRNA example. In this example, failure to consider near-optimal alignments produces exaggerated support for the inferred tree. Also, the (*H. sapiens*, *E. coli*) partition that receives increased support when using a sequential approach already exists in the ClustalW guide tree, indicating bias toward this tree.

To summarize the posterior distribution of alignments, we introduce AU (pronounced "gold") plots. AU plots consist of an alignment point-estimate annotated to indicate alignment variability. These plots are highlighted in Figures 6 and 7. AU plots are constructed from a sample of alignments drawn from their posterior distribution. To produce the plots, we assume a simplistic Poisson model over the posterior sample. The model approximates the probability that each letter in the point-estimate is aligned to the ancestral residue of the letter's depicted column. This measure prevents AU plots from being sensitive to the addition of duplicate or near-duplicate sequences; such sequences are not necessarily shown as well aligned globally just because a larger fraction of sequences aligns closely with them. For example, if one subset of taxa aligns weakly to another subset, but the sequences within each subset align strongly to each other, both subsets will be depicted as weakly aligned unless the root taxon lies within one subset. In this case, the subset containing the root taxon will be shown as well-aligned, and the other subset will be shown as weakly aligned. In combination with our MCMC algorithm that produces samples from the posterior distribution, these plots provide a valuable tool for assessing alignment ambiguity.

Locating ambiguous regions within sequence alignments is an important process for most phylogenetic reconstruction methods, but remains a difficult task. Waterman et al. (1992) and Gatesy et al. (1993) have suggested varying the gap opening and extension penalties well beyond supported values during alignment and then marking regions that change as uncertain. This approach is able to find regions of the alignment that are so greatly homologous that they remain aligned under unlikely gap models, but does not characterize the range of ambiguity under one set of alignment parameters. Lutzoni et al. (2000) have developed a method for delimiting ambiguous regions based on sliding gaps laterally until alignment quality decreases. Because the MCMC approach considers all possible alignments, it has a broader scope than the Lutzoni et al. (2000) algorithm. In addition, we estimate indel parameters instead of fixing them to predefined values. Currently, we are forced to employ diffuse priors on these parameters because biologically based informative priors are not yet available. We consider the development of such priors to be important future work.

Although AU plots yield a gestalt impression of which parts of an alignment estimate are certain, they are not intended to fully summarize the posterior distribution of alignments. One approach to more fully visualize this distribution is to represent individual alignment samples as connected, increasing paths through an n-dimensional lattice (Waterman, 1995), where n is the number of aligned sequences. This approach then plots the entire distribution using histograms on an alignment path graph. Zhu et al. (1998) provide such an example for a pairwise alignment. Unlike AU plots that annotate only a single alignment, it is possible to plot all posterior realizations simultaneously with their posterior support encoded either by color or histogram height. One potential drawback to the alignment path graph approach is its dimensionality. For small n, it is possible to create all $\binom{n}{2}$ marginal pairwise alignments for plotting on paper. For larger n, the number of pairs increases rapidly and higher dimensional visualization techniques become necessary.

Although we garner many strengths from the joint estimation model, there exist several limitations to our implementation. For example, the indel process we describe uses the same alignment parameters in the pairwise alignment along every branch. This contrasts with ClustalW. ClustalW varies gap penalties on branches by its length determined by the substitution model (Thompson et al., 1994). Making gap penalties a function of branch length favors indels on longer branches and is biologically justifiable (Thorne et al., 1992). To extend our joint estimation model such that each branch b has a different pairwise alignment distribution $\nu(T_b)$, it is necessary for all the $\nu(T_b)$ to induce the same sequence length distributions. Although the TKF models have this property at equilibrium, it remains an open problem to make the alignment distribution used in this paper depend on branch length while keeping the sequence length distribution constant. On the other hand, to fit the TKF models into our joint estimation framework, it becomes necessary to define a reversal operator $r(\cdot)$ that takes any pairwise alignment $A^{(b)}$ into another pairwise alignment $r(A^{(b)})$ with the same homology structure and the first and second sequences interchanged. The probability of the reversed pairwise alignment $\mathrm{P}_\nu(r(A^{(b)}))$ must be identical to the probability of the original pairwise alignment $\mathrm{P}_\nu(A^{(b)})$. When ν is based on a symmetric pair-HMM like the one in this paper, then $r(\cdot)$ can be computed simply by interchanging the first and second sequences. However, when ν comes from a TKF1 model, $r(\cdot)$ is more complicated (Holmes and Bruno, 2001) and no $r(\cdot)$ has been published for the TKF2 model.

Another possible modeling extension counts indels at the beginning and end of a pairwise alignment as more probable than those in the interior, as is expected when sequence lengths differ (Gribskov and Devereux, 1991). However, even without these extensions, our methodology provides a solid starting point, producing reasonable estimates of both alignments and phylogenies simultaneously, as demonstrated through the 5S rRNA and EF-1α/Tu examples.

ACKNOWLEDGMENTS

We wish to thank Kenneth Lange for his encouragement and mentorship. B.D.R. is supported by NSF training grant DGE9987641 and NIH training grant GM008185. M.A.S. is supported in part by NIH grant GM068955 and USPHS grant CA16042. Major computing for this paper was performed on the UCLA Bioinformatics 126AMD1800 Cluster, also sponsored by NSF training grant DGE9987641.

REFERENCES

Adachi, J., and M. Hasegawa. 1995. Improving dating of the human/chimpanzee separation in the mitochondrial DNA tree: Heterogeneity among amino acids. J. Mol. Evol. 40:622–628.

Allison, L., and C. S. Wallace. 1994. The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and the optimisation of multiple alignments. J. Mol. Evol. 39:418–430.

Baldauf, S. L., J. D. Palmer, and W. F. Doolittle. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci., USA 93:7749–7754.

Barciszewska, M. Z., M. Szymański, V. A. Erdmann, and J. Barciszewski. 2001. Structure and functions of 5S rRNA. Acta Biochim. Polon. 48:191–198.

Brown, J. R., and W. F. Doolittle. 1997. Archaea and the prokaryote-to-eukaryote transition. Microbiol. Mol. Biol. Rev. 61:456–502.

Cavalli-Sforza, L., and W. Edwards. 1967. Phylogenetic analysis: Models and estimation procedures. Evolution 21:550–570.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. *Biological sequence analysis: Probabilistic model of protein and nucleic acids*. Cambridge University Press, Cambridge, UK.

Edlind, T., and P. Chakraborty. 1987. Unusual ribosomal RNA of the intestinal parasite *Giardia lamblia*. Nucleic Acids Res. 15:7889–7901.

Farris, J. 1972. Estimating phylogenetic trees from distance matrices. Am. Nat. 106:645–668.

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

Felsenstein, J. 2003. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Gatesy, J., R. DeSalle, and W. Wheeler. 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Mol. Phylogenet. Evol. 2:152–157.

Geiger, D. L. 2002. Stretch coding and block coding: Two new strategies to represent questionably aligned DNA sequences. J. Mol. Evol. 54:191–199.

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Gray, M. W., G. Burger, and B. F. Lang. 1999. Mitochondrial evolution. Science 283:1476–1481.

Gribskov, M., and J. Devereux. 1991. Sequence analysis primer. Stockton Press, New York.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 12:160–174.

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

Holder, M., and P. Lewis. 2003. Phylogeny estimation: Traditional and Bayesian approaches. Nat. Rev. Genet. 4:275–284.

Holmes, I., and W. J. Bruno. 2001. Evolutionary HMMs: A Bayesian approach to multiple alignment. Bioinformatics 17:802–820.

Huelsenbeck, J., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Jensen, J., and J. Hein. In press. Gibbs sampler for statistical alignment. *Statistica Sinica*.

Keeling, P. J., N. M. Fast, and G. I. McFadden. 1998. Evolutionary relationships between translation initiation factor eIF2-γ and selenocysteine-specific elongation factor SELB: Change of function in translation factors. J. Mol. Evol. 47:649–655.

Lake, J. A. 1991. The order of sequence alignment can bias the selection of tree topology. Mol. Biol. Evol. 8:378–385.

Lange, K. 1997. Mathematical and statistical methods for genetic analysis. Springer-Verlag, New York.

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

Lee, M. S. Y. 2001. Unalignable sequences and molecular evolution. Trends Ecol. Evol. 16:681–685.

Li, S., D. Pearl, and H. Doss. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. J. Am. Stat. Assoc. 95:493–508.

Liu, J. 2001. Monte Carlo strategies in scientific computing. Springer, New York.

Liu, J., W. H. Wong, and A. Kong. 1995. Covariance structure and convergence rate of the Gibbs sampler with various scans. J. R. Stat. Soc. B 57:157–169.

Lunter, G. A., I. Miklos, Y. S. Song, and J. Hein. 2002. An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. J. Comput. Biol. 10:869–889.

Lutzoni, F., P. Wagner, V. Reeb, and S. Zoller. 2000. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. Syste. Biol. 49:628–651.

Mau, B., and M. Newton. 1997. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. J. Comput. Graph. Stat. 6:122–131.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092.

Metzler, D. 2003. Statistical alignment based on fragment insertion and deletion models. Bioinformatics 4:490–499.

Newton, M., and A. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc., B 56:3–48.

Notredame, C., D. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for multiple sequence alignments. J. Mol. Biol. 32:205–217.

Phillipe, H., and P. Forterre. 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. 49:509–523.

Rivera, M. C., and J. A. Lake. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74–76.

Roberts, G., and S. Sahu. 1997. Updating schemes, correlation structure, blocking and parameterization of the Gibbs sampler. J. R. Stat. Soc., B 59:291–317.

Roger, A. J., O. Sandblom, W. F. Doolittle, and H. Philippe. 1999. An evaluation of elongation factor 1α as a phylogenetic marker for eukaryotes. Mol. Biol. Evol. 2:218–233.

Rost, B. 1999. Twilight zone of protein sequence alignments. Protein Eng. 12:85–94.

Scott, S. 2002. Bayesian methods for hidden Markov models, recursive computing in the 21st century. J. Am. Stat. Assoc. 97:337–551.

Sinsheimer, J. S. 1994. Extensions to evolutionary parsimony. Ph.D. thesis, University of California, Los Angeles.

Suchard, M., C. Kitchen, J. Sinsheimer, and R. Weiss. 2003a. Hierarchical phylogeneic models for analyzing multipartite sequence data. Syst. Biol. 52:649–664.

Suchard, M., R. Weiss, J. Sinsheimer, K. Dorman, M. Patel, and E. McCabe. 2003b. Evolutionary similarity among genes. J. Am. Stat. Assoc. 98:653–662.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18:1001–1013.

Swofford, D., G. Olsen, P. Waddell, and D. Hillis. 1996. Phylogenetic inferences. Pages 407–514 in Molecular systematics 2nd edition (D. Hillis, C. Moritz, and B. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

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

Thorne, J. L., and H. Kishino. 1992. Freeing phylogenies from artifacts of alignment. Mol. Biol. Evol. 9:1148–1162.
Thorne, J. L., H. Kishino, and J. Felsenstein. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33:114–124.
Thorne, J. L., H. Kishino, and J. Felsenstein. 1992. Inching towards reality: An improved likelihood model of sequence evolution. J. Mol. Evol. 34:3–16.
Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). Ann. Stat. 22:1701–1762.
Van de Peer, Y., S. Baldauf, W. Doolittle, and A. Meyer. 2000. An updated and comphrehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. J. Mol. Evol. 51:565–576.
Waterman, M. 1995. Introduction to computational biology, maps, sequences and genomes. CRC Press, Boca Raton, Florida.
Waterman, M., M. Eggert, and F. Lander. 1992. Parametric sequence comparisons. Proc. Nat. Acad. Sci. USA 89:6090–6093.
Wheeler, W. 1999. Fixed character states and the optimization of molecular sequenced data. Cladistics 15:379–385.
Wheeler, W., J. Gatesy, and R. DeSalle. 1995. Elision: A method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. Mol. Phylogenet. Evol. 4:1–9.
Wheeler, W. C. 1996. Optimization alignment: The end of multiple sequences alignment in phylogenetics? Cladistics 12:1–9.
Wheeler, W. C. 2003. Iterative pass optimization of sequence data. Cladistics 19:254–260.
Whelan, S., and N. Goldman. 2000. A general empirical model of protein evolution derived from multiple protein families using a maximum-liklihood approach. Mol. Biol. Evol. 18:691–699.
Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. Towards a natural system of organisms: Proposal for the domains *Archaea*, *Bacteria*, and *Eucarya*. Proc. Nat. Acad. Sci. USA 87:4576–4579.
Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.
Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.
Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.
Zhu, J., J. Liu, and C. Lawrence. 1998. Bayesian adaptive sequence alignment algorithms. Bioinformatics 14:25–39.

First submitted 13 April 2004; reviews returned 7 July 2004; final acceptance 5 November 2004
Associate Editor: Paul Lewis

APPENDIX

Here, we describe algorithms for MCMC sampling alignments based on DP. We begin with a general approach to sampling the entire alignment simultaneously. We then restrict our attention to sampling the three pairwise alignments that share a common internal node and the five pairwise alignments on or adjacent to one internal branch. For these latter procedures, we discuss how to improve computational efficiency by making additional constraints in the DP matrix.

Basic Algorithm

To sample from the posterior distribution of a multiple alignment $\mathbf{A}$ given the data $\mathbf{Y}$, the tree $(\tau, \mathbf{T})$, substitution model parameters $\mathbf{\Theta}$ and indel process parameters $\mathbf{\Lambda}$, we construct a multiple-HMM whose paths correspond one-to-one to multiple alignments. We can efficiently Gibbs sample paths from the multiple-HMM using the forward-backward algorithm (Scott, 2002) based on DP. As the distribution of paths under the multiple-HMM given in Equation 11 varies slightly from the distribution over $\mathbf{A}$ that we seek in Equation 10, we first draw a new sample path from the multiple-HMM and then perform an accept-reject step on the corresponding multiple alignment to control for these differences. Rejections are quite rare.

Multiple-HMM.—We begin construction of the multiple-HMM by considering its set of emission states. Each state s in this set emits either an observed letter or gap for each of the n sequences at the external nodes and either a Felsenstein wildcard or gap for each of the $n-2$ internal nodes of the phylogeny. Each state therefore corresponds to a column in a multiple alignment and is characterized by whether the state contains a residue (+) or gap (−) at each node. For example, the state of the first column in the alignment in Figure 1 is (+,+,–,–,+,–). Not all combinations of residues and gaps are allowed. Any state that places a gap at a node between two residues is forbidden, because these states imply that an inserted residue is homologous to a previously deleted residue. The emission probabilities for the residues in a state are equal to their phylogenetic column likelihood and are given by the substitution model described in Methods.

There exists a natural relationship between the multiple alignment $\mathbf{A}$ and the pairwise alignments $A^{(b)}$ which comprise it. Columns in a pairwise alignment $A^{(b)}$ correspond to one of three states in the pair-HMM that generated the column: the M state (+,+), the G1 state (−,+) or the G2 state (+,−). The state of each column in $A^{(b)}$ then projects to the column in $\mathbf{A}$ that contains the same residues, carrying with it its emission properties. For example, the fourth column of A_{15} in Figure 1 is in the M state and corresponds to the fifth column of the multiple alignment. However, columns of $\mathbf{A}$ that do not emit any residues in a pairwise alignment $A^{(b)}$ do not correspond to any column of $A^{(b)}$. The emission properties of such columns are the same as (−,−) when projected down to $A^{(b)}$ and we say that $A^{(b)}$ is not active in such columns. For example, column 3 of $\mathbf{A}$ in Figure 1 emits no residues in either sequence 1 or sequence 5 and does not correspond to any column of A_{15}.

We augment each state s in the multiple-HMM with additional values $m(s, b)$ for each branch b. Values $m(s, b)$ keep track of the last active state of each pairwise alignment $A^{(b)}$. If $A^{(b)}$ is active in s, then $m(s, b)$ is simply the state of $A^{(b)}$ in s. However, if $A^{(b)}$ is not active, then the value of $m(s, b)$ carries forward from the previous column. To accomplish this, we choose the transition matrix for the multiple-HMM to assign probability 0 to paths which do not set $m(s, b)$ to the last active state of $A^{(b)}$. As an example, consider the third column of the alignment in Figure 1. The alignment A_{15} is not active in this column, but it was active in the previous column with state M. So, the last active state for A_{15} is M in column 3.

We also impose a constraint on paths to ensure that their correspondence to multiple-alignments is one-to-one. If, for any two adjacent states in a path, there is a pairwise alignment $A^{(b)}$ that is active in both states, then interchanging the state order changes at least one of the pairwise alignments. However, if this is not the case, then interchanging the state order does not change any of the pairwise alignments and, thus, does not change the multiple alignment. In this situation, we say that the two states are not strictly ordered with respect to one another. This means that two or more paths through the multiple-HMM map to the same multiple alignment. To impose a one-to-one correspondence, for any two not strictly ordered states s_1 and s_2, we forbid the transition $s_1 \to s_2$ unless $s_1 \le s_2$ according to some external ordering that we choose. Our choice must be a total ordering over all states to provide a single best order for adjacent sets of more than two not strictly ordered states. Paths in which all not strictly ordered states follow the total ordering are considered legal.

Given this description of the states, we now propose a transition kernel. We define $L(s_1, s_2) = 1$ if s_1 and s_2 are legally ordered or 0 otherwise. We define $s(b) = 1$ if $A^{(b)}$ is active in s or 0 otherwise. We define the transition matrix $\mathbf{P} = \{P_{i,j}\}$ for the multiple alignment in terms of the transition matrix $\mathbf{Q} = \{Q_{i,j}\}$ for the pairwise alignments, such that

$$P_{s_1,s_2} = L(s_1, s_2) \times \prod_{b=1}^{B} Q^{s_2(b)}_{m(s_1,b),m(s_2,b)} \times \prod_{b=1}^{B} 1\{m(s_1, b) = m(s_2, b)\}^{(1-s_2(b))}. \quad (14)$$

The first term assigns probability 0 to all illegal paths. The exponent $s_2(b)$ in the second term simply removes terms from the product when $A^{(b)}$ is not active in s_2 and leaves other terms unchanged. Thus, the second term contributes the transition probability from the previous active state $m(s_1, b)$ to the current state $m(s_2, b)$ for all active pairwise

alignments $A^{(b)}$ in s_2. The last term enforces the condition that inactive pairwise alignments in s_2 carry forward their last active state.

To see that transition matrix $\mathbf{P}$ yields the desired distribution on multiple alignments, consider a path $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_{C+1})$ through the multiple-HMM corresponding to an alignment $\mathbf{A}$ with C columns. We note that if $\boldsymbol{\pi}$ correctly records the last active state of pairwise alignment $A^{(b)}$ in each state π_i then

$$\mathrm{P}_v(A^{(b)}) = \prod_{i=1}^{C+1} Q^{\pi_i(b)}_{m(\pi_{i-1},b),m(\pi_i,b)}. \tag{15}$$

If the path $\boldsymbol{\pi}$ is legal as well, then

$$\begin{aligned}\mathrm{P}(\boldsymbol{\pi}) &= \prod_{i=1}^{C+1} P_{\pi_{i-1},\pi_i} = \prod_{i=1}^{C+1}\prod_{b}^{B} Q^{\pi_i(b)}_{m(\pi_{i-1},b),m(\pi_i,b)} = \prod_{b}^{B}\prod_{i=1}^{C+1} Q^{\pi_i(b)}_{m(\pi_{i-1},b),m(\pi_i,b)} \\ &= \prod_{b}^{B} \mathrm{P}_v(A^{(b)}),\end{aligned} \tag{16}$$

where $\mathrm{P}_v(A^{(b)})$ is the distribution induced by the pair-HMM.

Computing issues.—In addition to the set of emitting states described above, the multiple-HMM contains a start, end, and absorbing state. We choose the M state as the starting position for each pair-HMM (Durbin et al., 1998; Metzler, 2003). This induces the start state of the multiple-HMM to equal the state in which all nodes are present. The end state occurs when all pair-HMMs become inactive. The addition of an absorbing state is a formal device to aid in normalizing the multiple-HMM. The state accounts for our restriction that pairwise alignments on adjacent branches must agree on the length of the sequence at their shared node.

In calculating the emission probabilities for the multiple-HMM, we sum out all Felsenstein wildcards at the internal nodes using the peeling algorithm. This procedure results in the emission of at most n observed leaf sequence letters, substantially reducing the dimension of the DP matrix. This integration can produce additional silent states. These occur when the multiple-HMM emits only in residues at internal nodes, as these residues have been summed out. Sampling algorithms using DP require that the HMM have no cycles of silent states. We overcome this difficulty by blocking cycles of silent states into a single silent-block state. We Gibbs sample a path from the blocked HMM using the forward-backward algorithm on the DP matrix and then re-expand the path by sampling the lengths of the silent blocks given the blocked path.

Metropolis-Hastings acceptance probabilities.—We present here a derivation of the acceptance probabilities necessary for the accept-reject steps. Let ρ_{12} be the probability of proposing a new multiple alignment $\mathbf{A}''$ given the current multiple alignment $\mathbf{A}'$, the data $\mathbf{Y}$ and all other modeling parameters τ, $\mathbf{T}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\Lambda}$ and let ρ_{21} be the probability of the reverse proposal. Using the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), the acceptance probability α_{12} for moving from $\mathbf{A}'$ to $\mathbf{A}''$ is

$$\alpha_{12} = \min\left\{\frac{\mathrm{P}(\mathbf{A}'' \mid \mathbf{Y}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}) \times \rho_{21}}{\mathrm{P}(\mathbf{A}' \mid \mathbf{Y}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}) \times \rho_{12}}, 1\right\}. \tag{17}$$

By dividing the two priors defined in Equations 10 and 11, we determine that Gibbs sampling yields

$$\begin{aligned}\rho_{12} &= \frac{1}{K} \times \mathrm{P}(\mathbf{A}'' \mid \mathbf{Y}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}) \times \prod_{i\in I} \phi(|a_i''|)^2, \text{ and} \\ \rho_{21} &= \frac{1}{K} \times \mathrm{P}(\mathbf{A}' \mid \mathbf{Y}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}) \times \prod_{i\in I} \phi(|a_i'|)^2,\end{aligned} \tag{18}$$

where $|a_i'|$ and $|a_i''|$ are the lengths ascribed to the internal nodes of τ by $\mathbf{A}'$ and $\mathbf{A}''$ respectively. Substituting Equation 18 into Equation 17 results in

$$\alpha_{12} = \min\left\{\frac{\prod_{i\in I} \phi(|a_i'|)^2}{\prod_{i\in I} \phi(|a_i''|)^2}, 1\right\}. \tag{19}$$

The acceptance probability (19) reduces down to that given in (12) when resampling the alignments associated with a NNI topology proposal in which the sequence lengths of only two nodes possibly change.

Sampling a Subset of the Pairwise Alignments

The algorithm provided above can sample the entire multiple alignment $\mathbf{A}$ simultaneously. However, for all but the smallest number of taxa, it is computationally prohibitive to construct the necessary DP matrix. As an alternative, we recommend sampling $\mathbf{A}$ through a series of local proposals, each updating a subset of the pairwise alignments along the branches of 3- or 4-taxon subtrees within τ. When sampling the local alignment along a subtree, the only change required to the basic algorithm concerns how the data $\mathbf{Y}$ specify the subtree leaf sequences emitted by the multiple-HMM. A leaf node in the subtree may correspond to an internal node of τ. In this case, the sequence associated with the new leaf node has a fixed length, but the letters emitted at each site are unobserved. The conditional probability that each site in the sequence emits each possible letter in the alphabet is determined by applying the peeling algorithm to the internal node's descendants in the full tree.

Constrained sampling.—We propose a constrained procedure under which it is computationally practical to sample a subset of the three pairwise alignments on 3-taxon subtrees (3-way sampling) and to sample a subset of the five pairwise alignments on 4-taxon subtrees (5-way sampling). Using this procedure, the pairwise alignments between some leaf sequences in the subtree are fixed. We require both that match states and gaps are preserved between sequences whose alignments are fixed and that the order of columns in the pairwise alignment remains unaltered when the columns are not strictly ordered. For 3-way sampling, we fix the alignment between two of the three leaf sequences in the subtree and, for 5-way sampling, we fix the alignment between all pairs of leaf sequences in the subtree.

Under 3-way sampling, paths through the 3D DP matrix visit grid points $\{i, j, k\}$. Our constraint requires that the projection of paths into the subspace $i = 0$ remains constant, equivalent to fixing the alignment between the second and third sequences. After projection, the remaining variable subspace can be uniquely represented by grid points $\{i, c(j, k)\}$ where $c(j, k)$ is the column in the pairwise alignment of sequences 2 and 3 that contains residue j of sequence 2 aligned to residue k of sequence 3. The resulting DP matrix is 2D and can be Gibbs sampled from in $O(C^2)$. Although the dimensionality of this constrained DP problem is the same as sampling a single pairwise alignment, the number of possible states at each grid point is larger because we must also consider whether the internal node in the subtree emits a residue or not. The additional states increase computational space requirements and run-time compared to sampling a single pairwise alignment. However, the increased space and time enable us to sample both a pairwise alignment and the sequence at the internal node. This substantially improves the MCMC efficiency of the sampler without having to resort to unconstrained sampling at $O(C^3)$.

We employ 5-way sampling in conjunction with NNI proposals on 4-taxon subtrees. Under 5-way sampling, all pairwise alignments between the leaf nodes are fixed. As a result, the path through the DP matrix is held constant, whereas the realized state at each grid point may change, generating a 1D DP matrix. This procedure samples from all possibilities of emitting or not emitting a residue for the sequences at both internal nodes in each column of the multiple alignment. The procedure may also introduce or remove columns that contain only residues in the internal sequences of the subtree, thus changing the length of the multiple alignment. We note that if we instead fixed the pairwise alignments between only three of the four leaf sequences, this would allow us to simultaneously sample the alignment of one of the leaf sequences and propose an NNI topology change across the internal branch. Simultaneously proposing such alignment and topology updates may further improve MCMC mixing and offers a direction for future research.