Robust Inferences from Ambiguous Alignments

Benjamin D. Redelings and Marc A. Suchard

October 23, 2008

Contents

1 Introduction

1.1 Alignments, inferences, and uncertainty

Molecular sequence data has become an invaluable source of information for understanding evolutionary processes and for inferring evolutionary relationships between organisms. Molecular sequences provide a large number of separate characters (individual nucleotides, amino acids, or codons) that are easy to identify and distinguish. In addition, probabilistic models of how these molecular characters change over time allow us to estimate evolutionary process parameters and to quantify the evidence for evolutionary hypotheses. These parameters include phylogenetic trees, divergence times, insertion and deletion rates, and substitution rates. Probabilistic models of evolution also enable researchers to locate sequence motifs that are especially conserved or exhibit positive selection. These bioinformatic inferences depend not only on the observed molecular characters but also on the homology structure of the molecular characters, termed the "alignment" of the molecular sequences. This homology is not directly observed and can be difficult to assess, especially when the sequences have diverged for a long time so that sequence similarity is low. As a result, in many analyses there is substantial uncertainty about the alignment. In this chapter we examine methods for making robust inferences when the alignment is uncertain.

Traditional bioinformatic inference methods have usually assumed that the alignment was known with certainty. Ignoring alignment uncertainty when it is present can undermine bioinformatic inferences in several ways. For example, the use of only one plausible alignment estimate among many can lead to severely biased estimates of the phylogenetic tree (Lake, 1991) and other parameters (Thorne and Kishino, 1992), as well as unreliable or non-repeatable results (Morrison and Ellis, 1997). Furthermore, ignoring alignment uncertainty can lead to exaggerated measures of confidence in those results. Therefore, it is important for bioinformatic inferences to take alignment uncertainty into account. The scarcity of methods that are robust to uncertainty in multiple sequence alignments has been a significant obstacle to inferences based on highly divergent molecular sequences, and thus to answering questions about ancient divergences in the Tree of Life.

A number of bioinformatic techniques have been developed to avoid the bias and exaggerated confidence that typically result from conditioning on a single alignment estimate. This is a difficult and multifaceted problem, and an ideal technique faces many challenges. First, methods should characterize alignment uncertainty in a rigorous and objective fashion. This involves taking all sources of uncertainty into account, as well as avoiding subjective and non-reproducible judgments of homology. Second, it is beneficial to report fine gradations of uncertainty that indicate the weight of evidence for homology, instead of dividing the alignment into "ambiguous" and "unambiguous" regions. Third, an ideal method would make full use of the information in the data set, including information in ambiguous regions.

Although such techniques for handling alignment uncertainty have been developed for pairwise alignments, methods for handling uncertainty in multiple alignments have been slower to develop. Recent improvements in statistical methodology have made it feasible to rigorously assess the degree of uncertainty in multiple sequence alignments, to represent finer gradations of uncertainty, and to make use of information in ambiguously aligned regions without bias. In this chapter we will discuss several traditional methods of handling ambiguity in multiple sequence alignments, as well as new statistical approaches to this problem. We then illustrate how alignment uncertainty can affect phylogeny estimation by focusing on a small data example. This example demonstrates the negative effects of ignoring alignment uncertainty and also illustrates the practical benefits of the new statistical methods.

1.2 Multiple sequence alignments represent evolutionary history

1.2.1 Homology-based alignments versus function-based alignments

In this chapter we follow common practice in the field of molecular evolution in using alignments of molecular sequences to specify evolutionary homology. That is, multiple sequence alignments specify the homology of individual residues in a set of homologous sequences by arranging the residues in a matrix so that residues in a column (also called a "site") all descend from a single residue in the common ancestral sequence. Thus, the complete evolutionary history of individual residues in a set of homologous sequences is given by the combination of a multiple sequence alignment and a phylogenetic tree. However, we note that specifying homology by use of a matrix in which each row represents one sequence does not allow one to easily represent homologies within a single sequence. Thus, it is difficult to represent the within-sequence homology that occurs as a product of gene-internal duplications.

Researchers in other fields have sometimes conceived of alignment in different ways. For example, they might align residues that do not share a common ancestor, but that perform a common function or occupy similar positions in the three-dimensional structure of a protein (Mizuguchi et al., 1998; Thompson et al., 1999). While both kinds of information are useful, we note that functional and structural interpretations of alignments may not be well-defined, since residues may have multiple functions, and multiple residues may share one function. If a column corresponds to a function, then it could be necessary to place two adjacent residues from the same sequence in a single column, or to place a single residue in several columns. In this chapter, we therefore use alignments only to represent evolutionary homology.

1.2.2 Commonly observed violations of homology in alignment estimates

These different interpretations of multiple sequence alignments can lead to different alignment matrices in practice. For example, it sometimes occurs that two sequences in a multiple sequence alignment experience independent insertions in the same location. These two insertions are not homologous because they do not

descend from any sequence in their common ancestor. Thus, under the evolutionary homology interpretation, these insertions must not be aligned in the same columns, even though they may share similar functions. Similarly, it may happen that a residue in an ancestral sequence is deleted, and that one of its descendant sequences later experiences an insertion in the same location. Because the deleted residue is not ancestral to the inserted residue, these residues must not be aligned even though they may occupy similar positions in the protein's three-dimensional structure. These constraints, when observed, rule out many of the alignments that are commonly created by multiple sequence alignment programs today. However, without knowledge of the evolutionary tree relating the sequences and of the presence or absence of residues at ancestral sequences, it is difficult to determine conclusively if an alignment violates these constraints.

1.3 What is alignment ambiguity?

Before we discuss sources of alignment ambiguity and methods of handling this ambiguity, we first seek to clarify the meaning of alignment ambiguity. We begin by defining alignment ambiguity to mean that there are two or more plausible alternative alignments. We claim that this definition of ambiguity is the correct one, and we note that it also applies to the estimation of other parameters, such as phylogenetic trees. However, it is common to think of alignment ambiguity simply in terms of the degree of sequence conservation in multiple sequence alignments. This conception of alignment ambiguity may be implicitly used in identifying ambiguous regions by visual inspection, and is also explicitly used in some computerized procedures that exclude alignment regions based on the presence of gaps or of sequence variation between taxa. Although non-conserved regions are often a good indicator of alignment ambiguity, such indicators are successful only to the extent that they actually indicate the presence of plausible alternative alignments. Another class of methods seeks to characterize ambiguity in multiple sequence alignments by assessing the sensitivity of inferred alignments to perturbation in alignment parameters. Again, these methods are successful only to the extent that they indicate the presence of two or more plausible alternatives. Finally, we note that in considering alignment uncertainty, we are not addressing uncertainty about whether or not biological sequences are homologous. Instead, we are concerned with uncertainty about the homology of individual residues, but take as given that all sequences in a data set are homologous.

Second, homologies in an alignment are not simply "ambiguous" or "unambiguous", but can be of varying degrees of certainty, depending on the strength of evidence in the observed data for the homology[1]. Many methods have been developed to improve phylogenetic inferences by discarding portions of the aligned data matrix and keeping the rest, and it may be natural to speak of the discarded portions as "ambiguous regions". However, it is important to note that this does not mean that the resulting portions are "unambiguous", because complete certainty about homology is not possible.

Third, it is useful to consider alignment ambiguity for parts of the alignment as well as the whole. When considering partial alignments, it is important to define them in a coherent fashion. For example, it is natural to talk about ambiguous "regions" in an alignment, but this is problematic since the regions are defined by the alignment itself, which is not known with certainty. Indeed, if a "region" is ambiguous, then it may not exist. To avoid this problem, we define a partial alignment as a hypothesis about homology and note that the smallest possible hypothesis of homology is that two residues from different sequences are homologous or not homologous. Most homology hypotheses, including the hypothesis of a full alignment, can be decomposed into a collection of minimal homology hypotheses of this type. Now that we have more carefully defined what we mean by a partial alignment, we can extend the definition of alignment ambiguity to partial alignments. We note that many full alignments may conform to any partial alignment; for example, many full alignments satisfy the criterion that residue 1 of sequence 1 is homologous to residue 1 of sequence 2. With this in mind, we state that alignment ambiguity for partial alignments means that some plausible full alignments conform to the partial alignment and that some do not.

Fourth, plausibility is a relative term that always depends on the the degree of knowledge of the researcher. Thus, the degree of ambiguity in the alignment depends on what is known, so that it is not meaningful to simply speak of an alignment region as "ambiguous" without first specifying what facts about evolutionary parameters should be taken as given. For example, it may be the case that two alternative phylogenies

[1]The flip side of ambiguity is the strength of evidence. Ambiguity applies to a variable (e.g. the alignment) that has more than one plausible value. Strength of evidence applies not to a variable, but to a hypothesis, such as the hypothesis that the variable takes on a specific value.

support alternative alignment estimates. In this case, if the phylogeny is unknown, the alignment must be considered ambiguous, but if one of the phylogenies is known to be correct, then the alignment may be well supported. Similarly, different values of other alignment parameters such as the indel and substitution rates may lead to different estimates of the alignment. In some cases, precise knowledge of parameter values may result in substantially less uncertainty in the alignment. The general idea that knowledge of parameter values affects the degree of alignment ambiguity can be conveniently expressed in terms of conditional probabilities. For example, if $\mathbf{A}$ and τ refer to the unknown true alignment and tree while $\mathbf{A}_0$ and τ_0 are specific possibilities for the alignment and tree, then the probability $\mathrm{P}(\mathbf{A} = \mathbf{A}_0)$ that the alignment is $\mathbf{A}_0$ when the tree is unknown may differ substantially from the probability $\mathrm{P}(\mathbf{A} = \mathbf{A}_0 | \tau = \tau_0)$ that the alignment is $\mathbf{A}_0$ when the tree is known to be τ_0.

The interaction of alignment ambiguity with bioinformatic inferences is most important when alignment estimates are largely determined by initial assumptions about parameter values. In this case, alignment estimates will reflect and reinforce assumptions embodied in tunable parameter values such as the "guide tree" that is used to guide alignment construction in progressive alignment. In such cases, the alignment must be considered dangerously ambiguous, because the ambiguity may undermine the inference procedure. This observation may be clarified by noting that, when inferring an evolutionary parameter such as the tree or the insertion-deletion rate, it must not be treated as known during the inference[2]. For example, if we seek to infer the evolutionary tree topology, then we must not assume knowledge of a particular evolutionary tree during the estimation of the alignment. Nevertheless, this is exactly what is done when a guide tree is used. Likewise, when attempting to infer indel rates from an alignment, we must assume that there is a range of uncertainty about (at least) the gap-opening cost, because this corresponds to the indel rate.

1.4 Inferences from multiple sequence alignments

Multiple sequence alignments are a prerequisite for inferring a number of different biological properties. Therefore alignment ambiguity must be considered in each of these methods. Most of the properties inferred from molecular sequence alignments may be considered evolutionary parameters in that they specify the process of molecular sequence change or the historical relationships of the sequences.

First, modern methods for estimating phylogenetic trees relating molecular sequences require a multiple sequence alignment as input. By specifying which residues are homologous, multiple sequence alignments divide the set of sequences into a set of separate single-residue characters. This decomposition is a requirement for the most powerful and accurate modern phylogeny estimation methods, which infer clades based on the evidence of shared, derived characters. In this chapter, our central focus is on the estimation of phylogenetic trees, but we briefly consider implications of alignment uncertainty for other methods as well.

Second, alignments are necessary for the estimation of mutation rates, divergence times and evolutionary distances. This includes the estimation of the rates of insertion and deletion, as well as substitution rates and other other parameters of the evolutionary process.

Third, alignments are required in many methods for labeling sites as functionally divergent, conserved, positively selected, or hyper-variable. This is because the alignment defines the sites which are given functional labels, and so alignment uncertainty means uncertainty about what the sites are. In addition, most methods for annotating sites use the pattern of different character values at each site to assign meaningful labels. Thus, alignments are also required for motif-discovery methods if those methods rely on conservation patterns to find motifs.

Lastly, we note that alignments are also used to infer whether or not two sequences are homologous. However, we do not focus on this kind of analysis in this chapter, because questions about the existence of sequence homology are not about which alignment is correct, but about whether any alignment is correct. Furthermore, determinations of homology are often based on the distribution of various statistics under the null model of non-homology, whereas we are interested in alignment ambiguity under the assumption of homology. However, we note that improved methods that explicitly consider the alternative model of homology as well may improve the ability to detect homologs with very high sequence divergence (Csuros and Miklós, 2005).

[2]This observation has practical consequences for sensitivity analysis, which attempts to gauge alignment ambiguity by using a range of parameter values for uncertain parameters. Despite the fact that the topology is unknown, it seems uncommon for researchers to include a range of values for the guide tree when progressive alignment is used.

2 The structure of bioinformatic inferences

In order to characterize the quality of bioinformatic inference methods, we first describe the properties that we desire in a robust inference method. We then discuss the multi-stage pipeline structure that is common to many traditional estimation methods and discuss how estimation errors may propagate through this pipeline. We then overview two different approaches to bioinformatic inferences, the cost-based paradigm and the statistical paradigm. Finally, we detail the two primary sources of alignment uncertainty and compare the two paradigms in terms of their ability to make robust inferences in the presence of alignment uncertainty.

2.1 Robust inference

Inference methods must have a number of important properties in order to provide a sufficient basis for robust decisions. First, inference methods must avoid bias. Specifically, while inference methods do not always yield correct results because of insufficient information, we desire that results tend towards to the correct answer as the amount of information increases, and not towards any other value. Second, we desire an inference method to use make the efficient use of all information in the data set, decreasing the range of error and increasing power to detect things. Third, in order to be a useful basis for decisions, an inference method should provide accurate measures of confidence or precision that indicate the weight of evidence in favor of the inference or estimate. For example, if a researcher wishes to use molecular data to determine whether a clade is monophyletic, a tree estimate is of little value unless it is accompanied by the weight of evidence for clade monophyly. Bias and uncertainty in the alignment estimate naturally propagate to bias and uncertainty in estimates of trees, evolutionary rates, and other parameters, because these estimates are based on the alignment estimate. However, the way in which this happens depends on the way the inference is structured.

2.2 Sequential estimation

2.2.1 The structure of sequential estimation

Bioinformatic inference methods often consist of a series of chained estimation steps that are performed in a particular order. Instead of estimating the parameter of interest directly from unaligned sequence data, such "sequential estimation" methods are characterized by a pipeline structure in which the output of each estimation step may be used as input to the following step. For example, phylogenies are traditionally estimated in a two-step process. In the first step, a multiple sequence alignment is constructed from unaligned sequence data. This alignment is often manually edited or trimmed to remove estimation artifacts. Then, in the second step, this alignment estimate is used to construct a phylogeny estimate. Estimation of other evolutionary parameters may follow the same pattern. For example, a researcher might first construct a single alignment estimate, and then use that alignment to estimate branch lengths and the relative rates of indels and substitutions (Thorne et al., 1991). In both cases the error in inaccurate alignment estimates may propagate through the estimation pipeline and affect downstream estimates, including finally the parameter of interest.

In addition to unaligned sequence data that is used as input, the output of the estimation pipeline is influenced by several tunable parameters that must be specified by the researcher. For example, an alignment estimate may be influenced by the value of tunable alignment parameters such as the gap opening penalty, the gap extension penalty, and various mismatch penalties. When the alignment estimate is constructed by progressive alignment, these parameters additionally include a guide tree that determines the order in which pairs of partial multiple sequence alignments are combined to produce the full multiple sequence alignment estimate. Thus, alignment estimates may be influenced by the guide tree as well as by mutation costs. Although the guide tree may be specified by the researcher, it is frequently estimated from the unaligned sequence data using distance-based techniques such as neighbor joining (NJ). In this case, phylogeny estimation becomes a 3-step procedure in which the first step involves estimating the guide tree.

2.2.2 Sources of error in sequential estimation

Sequential estimation works well when the sequences are closely related and there is little ambiguity about the true alignment. However, when the sequences are more divergent, two sources of uncertainty may lead to uncertainty in the alignment estimate. First, there may be a myriad of near-optimal alignments. In this case, the best-scoring alignment cannot be confidently preferred over the second best alignment, because both alignments have similar quality. In this case, the choice of a single alignment estimate to submit to the next stage is somewhat arbitrary, and may result in an error that propagates through the estimation pipeline. We note that the degree of uncertainty here does not depend solely on prior beliefs about the alignment, but instead depends on the strength of evidence in the observed unaligned sequence data for each alignment as assessed by a probabilistic model or objective function. Therefore, this uncertainty can be categorized as *a posteriori* uncertainty.

Second, when the input sequences are divergent, uncertainty about tunable parameter values may also lead to error. This is because small changes in the values of tunable parameters may result in a different estimate. For example, different gap penalties in the alignment stage may lead to different alignment estimates, which in turn lead to substantially different phylogeny estimates (Morrison and Ellis, 1997). Thus, when the true value of these parameters is unknown or not precisely known, uncertainty in the parameters leads to uncertainty in downstream estimates that depend on the alignment estimate. In the sequential estimation paradigm, tunable parameter values are initially chosen based on prior knowledge or belief, instead of being estimated from the observed data that is used as input. Therefore this uncertainty can be categorized as *a priori* uncertainty.

2.2.3 Circular dependencies in sequential estimation

Sequential estimation can lead to biased inferences because accurate knowledge of tunable parameters for the alignment estimation process is not available during the alignment construction step. To demonstrate this, we note that tunable parameters of the alignment construction process in fact correspond to the parameters of the evolutionary process. For example, progressive alignment algorithms require phylogenetic information in the form of a guide tree in order to yield high-quality alignment estimates (Thompson et al., 1994). Likewise, the gap-opening penalty is a proxy for the indel rate, the gap-extension penalty corresponds to the mean indel length, and mismatch penalties depend on the percent identity of the sequences. Therefore, the phylogeny, indel rate, and branch lengths should be known in advance before the alignment is estimated. This observation makes intuitive sense, in that knowledge of the evolutionary process should lead to better estimates of the alignment. However, it leads to problematic circular dependencies. For example, traditional approaches to phylogeny estimation require the alignment to be known in advance before the phylogeny can be estimated. However, high-quality estimates of the alignment require the phylogeny to be known in advance before the alignment can be estimated. Clearly, both of these conditions cannot be satisfied: in a sequential estimation framework either the phylogeny or the alignment must be estimated first.

Circular dependencies lead to biased estimates when the alignment is ambiguous because alignment estimates may simply reflect and artificially reinforce the initial guesses for tunable parameters in this case. For example, alignments constructed by progressive alignment tend to support phylogenies that are similar to the guide tree, which is a tunable parameter (Lake, 1991; Thorne and Kishino, 1992). Likewise, the number of indels in an alignment estimate may be largely determined by the gap-opening penalty, so that estimates of indel rates strongly depend on that parameter. One possible way to handle circular dependencies is to alternate between estimating the alignment and the parameter of interest. However, the tendency of the alignment estimates to reinforce bad initial guesses when the alignment is ambiguous can lead to mutually reinforcing, but incorrect, estimates for the alignment and other parameters. In such a case, iteration will not step away from local optima to achieve global convergence.

2.3 Joint estimation

Bioinformatic inference methods may avoid the bias and overconfidence introduced in sequential estimation by estimating mutually dependent parameters simultaneously. This "joint estimation" approach has two primary attractions. First, it eliminates the circular dependencies that plague the sequential estimation paradigm. This is done by means of a joint score function for all mutually dependent parameters, including

the alignment, phylogeny, indel rates, substitution rates, and other parameters. For example, if the alignment A and phylogeny τ are jointly estimated, then a joint score function of the form $f(\mathbf{A}, \tau)$ is necessary. Because an alignment is always available when scoring the tree, a previous alignment estimation step is unnecessary. Likewise, because a phylogeny is always available when scoring the alignment, a separate guide tree is unnecessary. We note that this score function may be either a cost function as in Wheeler (1996) or a probability function based on a joint probabilistic model as in Redelings and Suchard (2005).

Second, joint estimation allows all possible alignments to be considered when scoring each evolutionary parameter. For example, Thorne et al. (1991) estimates indel rates by maximum likelihood in a way that is not dependent on the choice of a single alignment. This involves computing the likelihood of parameters by summing the probability of all alignments conditional on those parameter values. Similarly, a joint score function can be used to estimate phylogenies in a way that considers all alignments. We also note that when the alignment is the parameter of interest, all possible values of other parameters may be considered.

2.3.1 Types of joint estimation

There are a number of different ways that a joint score function can be used (Wheeler, 2006), and we return to our previous example in which the alignment A and topology τ are to be estimated to illustrate this point.

First, the researcher may estimate the optimal parameter combination by finding the parameter values that optimize the joint score function. Wheeler (1996) defined a cost-based score function and followed this optimization approach to estimate phylogenies and alignments. Thus the estimates $\hat{\tau}$ and $\hat{\mathbf{A}}$ are constructed such that

$$(\hat{\tau}, \hat{\mathbf{A}}) \quad = \quad \arg\min_{\tau, \mathbf{A}} f(\tau, \mathbf{A}),$$

where the mathematical notation $\arg\min_{\tau,\mathbf{A}}$ denotes the value of the arguments $(\tau, \mathbf{A})$ at which the function $f(\tau, \mathbf{A})$ takes its minimum. Non-optimal parameter combinations make no contribution to parameter estimates in this approach. We also note that each topology τ can be compared to other topologies based on a reduced score function $\tilde{f}(\tau)$ in which the alignment argument $\mathbf{A}$ is optimized out:

$$\begin{aligned} \tilde{f}(\tau) &= \min_{\mathbf{A}} f(\tau, \mathbf{A}) \\ &= f(\tau, \arg\min_{\mathbf{A}} f(\tau, \mathbf{A})) \end{aligned}$$

Thus, each topology is scored using a different alignment that is optimally adapted to it. In contrast, the sequential estimation approach compares topologies based on a score function $f_{A_0}(\tau)$, where the alignment A_0 was computed in a previous step and remains the same for all topologies instead of being adapted to each topology.

Second, the researcher might construct the maximum likelihood estimate of τ. Here the joint score function $f(\tau, \mathbf{A})$ is the probability $\Pr(\mathbf{Y}, \mathbf{A}|\tau)$ of the data $\mathbf{Y}$ and alignment $\mathbf{A}$ given topology τ. The reduced score function $\tilde{f}(\tau) = \Pr(\mathbf{Y}|\tau)$ that is used to compare topologies is constructed by *summing* over the alignments instead of maximizing or minimizing over them:

$$\begin{aligned} \tilde{f}(\tau) = \Pr(\mathbf{Y}|\tau) &= \sum_{\mathbf{A}} \Pr(\mathbf{Y}, \mathbf{A}|\tau) \\ &= \sum_{\mathbf{A}} f(\tau, \mathbf{A}). \end{aligned}$$

This process of summing over all possible values of a variable is known as marginalization. Marginalization leads to less biased estimates than simple maximization because it does not score a topology τ only against the alignment that are optimally adapted to it, but also considers near-optimal alignments. This makes allowances for the fact that, conditional on a topology τ being correct, the optimal alignment for τ may still not be the correct alignment. For example, Thorne et al. (1991) showed that alignments must be

summed out in order to give unbiased estimates of indel rates and evolutionary distances when the alignment is ambiguous. We note that in the maximum likelihood paradigm the phylogeny, indel rates, and other evolutionary parameters are not summed out, because they are parameters and not missing data like the alignment.

Lastly, the researcher might use a Bayesian approach to estimate τ. In this case the joint score function $f(\tau, A)$ is the posterior probability $\Pr(\tau, \mathbf{A}|\mathbf{Y})$. As in the maximum likelihood case, estimation of the topology is made by marginalizing over the alignment:

$$\begin{aligned} \tilde{f}(\tau) = \Pr(\tau|\mathbf{Y}) &= \sum_{\mathbf{A}} \Pr(\tau, \mathbf{A}|\mathbf{Y}) \\ &= \sum_{\mathbf{A}} f(\tau, \mathbf{A}). \end{aligned}$$

However, unlike the maximum likelihood case, inference on the alignment can be made in the same way by summing out the topology. This is made possible in the Bayesian paradigm by placing a prior distribution over the topology and other parameters in order to treat them as random variables. Thus, inferences about the alignment may take into account uncertainty in the topology and other parameters.

2.3.2 Joint estimation and bootstrap fractions

One of the most common ways of characterizing the strength of evidence for a phylogenetic hypothesis is the bootstrap fraction. We note that the bootstrap fraction should not be interpreted as a probability that a clade is correct. For example, a bootstrap fraction of greater than 0.7 is often considered to represent strong support. Nevertheless, the bootstrap approach is attractive because it can assess how sensitive conclusions are to data selection even when an estimator does not provide this information itself.

However, the bootstrap fraction cannot be used to characterize uncertainty of phylogenetic hypotheses when the alignment is co-estimated with the phylogeny. This follows from the fact that the bootstrap approach assumes that the data can be decomposed into a series of separate columns or sites that can be resampled in the bootstrap procedure. However, when the alignment is unknown, the sites are no longer defined, and so the bootstrap fraction cannot be computed as usual.

2.3.3 Joint Bayesian estimation

Joint Bayesian estimation is our preferred method of analysis for bioinformatic inferences. One benefit of the Bayesian approach and the maximum likelihood approach is that all mutually dependent parameters may be jointly estimated. In contrast, score-based approaches such as Wheeler (1996) are able to co-estimate the alignment and phylogeny but allow indel and substitution costs to remain as tunable parameters. Thus, circular dependencies still exist between the alignment and these costs. Another benefit of the Bayesian and maximum likelihood approaches is that alignments may be summed out, which is necessary to avoid bias. As noted by Wheeler (2006), optimization based approaches may optimize either a cost-based score function or a probability expression. In the second case, a statistical model of the insertion-deletion process is necessary, but unobserved internal sequences may be maximized over instead of averaged over (Wheeler, 2006). In other approaches, the letters of the internal sequences may be averaged over, but their homology may be maximized over (Fleissner et al., 2005). Both approaches lead to less accurate estimates that may be biased, but can improve computational efficiency. Finally, we recall that Bayesian approaches are able to incorporate parameter uncertainty into posterior distributions of latent variables such as the alignment, while it is not clear how to do this in the maximum likelihood framework.

Joint Bayesian estimation of pairwise alignments and evolutionary process parameters is not a new development (Allison and Wallace, 1994). However, it is only recently that statistical techniques for estimating multiple sequence alignments have been developed (Holmes and Bruno, 2001; Holmes, 2003), leading to joint Bayesian estimation of alignments and phylogenies (Lunter et al., 2005; Redelings and Suchard, 2005; Fleissner et al., 2005). Because many bioinformatic analyses require multiple sequence alignments, these new techniques have created new opportunities for robust statistical inference in the presence of alignment uncertainty. However, for the moment, the extreme amounts of computation time required for these analyses limits the size of data sets that they can be applied to.

2.4 Cost-based methods versus statistical estimation

2.4.1 Estimation in the cost-based paradigm

Both pairwise and multiple sequence alignments have traditionally been estimated by finding the alignment that optimizes some score function. In the cost-based paradigm, the score function is computed as a sum of penalties, or "costs", for each observed sequence change in the alignment. In many pairwise alignment algorithms each aligned residue pair may incur a mismatch penalty if the residues are different, and each indel incurs a gap-opening penalty (GOP) as well as (in some algorithms) a separate gap-extension penalty (GEP) for each additional deleted or inserted character. The maximum parsimony method of inferring phylogenetic trees is an example of a cost-based method because its score function is the sum of a number of penalty terms for observed sequence differences. In addition, some early (Sankoff et al., 1973; Sankoff, 1973; Sankoff et al., 1976) and some later (Wheeler, 1996, 2006) methods for multiple sequence alignment seek to minimize a score function that explictly accounts for insertions and deletions on internal branches of the tree by including in the alignment unobserved ancestral sequences at internal nodes on the phylogeny.

However, most newer methods are willing to sacrifice some of this biological realism in order to substantially increase speed. First, most commonly used methods for multiple sequence alignment rely on an objective function that does not explicitly consider substitutions or indels occuring on each branch of the tree. Instead, they might use a less biologically motivated score function, such as a sum-of-pairs score (Edgar, 2004) or a tree-based weighted-sum-of-pairs score (Thompson et al., 1994). Because the sum-of-pairs score does not depend on the evolutionary tree, it may therefore double-count the cost for shared, derived changes that are observed in more than one leaf sequence. Second, most of these alignment algorithms rarely succeed in optimizing their score function, preferring instead to use progressive alignment in order to quickly discover a relatively high-scoring alignment.

Despite the variation in these methods, they all share a common drawback of the cost-based paradigm: cost parameters cannot be estimated by minimizing a cost function, since this would simply result in setting all the costs to zero. This is primarily a problem when attempting to determine gap penalties, since substitution costs may be determined from data using probabilistic methods, as in the PAM and BLOSSUM matrices. However, even in this case, cost-based methods to not allow the cost parameters to be tuned to the data set at hand through optimization of the cost function. Lastly, we note that cost parameters do not represent a biological property, so it is unclear what it means for a cost parameter value to be called correct or incorrect.

2.4.2 Estimation in the statistical paradigm

Alignment estimation in the statistical paradigm requires a probabilistic model of insertion and deletion. The alignment can then be inferred either by maximizing the likelihood (a frequentist approach) or by calculating the probability distribution of the alignment, given the observed sequence data (a Bayesian approach). If the frequentist approach is used, then the likelihood becomes a score function, although it is maximized instead of minimized like cost-based score functions. We note that log probabilities of accepted mutations are similar in meaning to the penalties for these mutations in the cost-base paradigm. This is because the log probabilities of independent events are added to compute the total probability, similar to the way penalties are added in the cost-based paradigm to compute the total penalty.

One benefit of the statistical paradigm is that statistical models of evolution allow the cost parameter for each type of mutation to be replaced with a biologically interpretable parameter that measures the rate of accepted mutations. For example, the gap-opening penalty may be replaced by an insertion-deletion rate, and mismatch penalties may be replaced by substitution rates. One exception to this rule is that the gap-extension cost is not replaced with a rate, but with an extension probability that controls the mean length of gaps. This is because each gap extension is not a separate mutation, but simply a penalty for longer gaps that is separate from the penalty for gap creation. We note that rate parameters play the same role as penalties or costs - a high rate corresponds to a low cost, and a low rate corresponds to a high cost. However, modelling evolution in terms of mutation rates more naturally accounts for multiple changes on a branch of the tree and for the occurrence of more changes on branches of longer duration.

A second benefit of the statistical paradigm is that it is possible to estimate rate parameters from the data. This differs from cost-based estimation in that increasing the mutation rates (which corresponds

to decreasing mutation costs) decreases the penalty for a mutation but also increases the penalty for not mutating. Therefore, the likelihood does not always increase with increasing mutation rates, allowing rates to be estimated by maximizing the likelihood. Because all parameters in a statistical model can be estimated from data, it is not necessary to specify the rate parameters based only on prior belief. For example, the relative rates of transitions and transversions can be estimated from the observed sequence data, as can the relative rates of indels and substitutions. This yields empirically driven parameter values, instead of parameter values chosen based on subjective or heuristic choices. Thus, the relative weight of different types of mutational events can be driven by the data.

2.5 Sources of alignment uncertainty

Alignment uncertainty comes primarily from two sources: uncertainty in parameter values and uncertainty due to near-optimal alignments. These sources of uncertainty are dealt with quite differently in the cost-based paradigm and the statistical paradigm. Additionally, there are smaller but important differences between maximum likelihood estimation and Bayesian estimation within the statistical paradigm.

2.5.1 Alignment ambiguity from parameter uncertainty

Parameter uncertainty leads to alignment ambiguity because different values of tunable alignment parameters lead to different alignment estimates. This is to be expected, since these parameters characterize the evolutionary process, and therefore determine how plausible each alignment should be. In the cost-based paradigm, such parameters include a cost for each type of mutation in addition to the guide tree, whereas in the statistical paradigm cost parameters for each type of mutation are replaced with mutation rates. Parameter uncertainty is almost always *a priori* uncertainty in the cost-based paradigm and *a posteriori* uncertainty in the statistical paradigm. That is, uncertainty about cost parameters in the cost-based paradigm is based on prior belief and not on the data. This is because it is not clear how to estimate cost parameters from the observed data. Therefore, when determining alignment ambiguity via sensitivity analysis, it maybe difficult to justify any particular range of parameters as being large enough. In the statistical paradigm, on the other hand, the amount of parameter uncertainty depends primarily on how much data is collected and how informative it is. Parameters may be estimated via maximum likelihood and parameter uncertainty may be characterized in terms of confidence intervals. If a Bayesian approach is taken, prior information or belief may be incorporated into parameter estimates, but this *a priori* information has decreasing influence on estimates as the amount of data increases. In contrast, maximum likelihood estimates and confidence intervals do not incorporate prior information.

2.5.2 Alignment ambiguity from near-optimal alignments

Near-optimal alignments indicate alignment ambiguity because they indicate the presence of plausible alternatives to the optimal alignment. Taking into account near-optimal alignments is difficult to do within a cost-based framework because differences in cost scores have no intrinsic meaning. The scores can be multiplied by any positive scaling factor without changing the optimum, and so it is unclear how close an alignment must be to optimal before it is consider to be "near" the optimum. In addition, if one seeks to down-weight near-optimal alignments that are further away from the optimum, it is unclear how much a sub-optimal alignment should be down-weighted. In contrast, when inference is carried out under a probabilistic model it is possible to incorporate alignment uncertainty into a further analysis by weighting each alignment according to its probability. In addition, it is possible to determine (for example) a 95% probable set of alignments if a cutoff is needed.

2.5.3 Effects of under-estimating ambiguity

Simultaneously accounting for both parameter uncertainty and near-optimal alignments is an important feature of any bioinformatic analysis. If one of these sources of uncertainty is ignored then exaggerated confidence may be ascribed to the resulting estimates. Mathematical readers will appreciate that this is similar to the common ANOVA formula about the proportion of the variance in X that is explained by Y

$$\mathrm{Var}(X) \quad = \quad \mathrm{Var}\left[\mathrm{E}(X|Y)\right] + \mathrm{E}\left[\mathrm{Var}(X|Y)\right],$$

where the first term is the variance in X that is explained by variation in Y, and the second term is the variation in X that is not explained by variation in Y. Ignoring either of these contributions to the uncertainty may lead to under-estimation of the effects of alignment ambiguity, and hence overconfidence in a bioinformatic inference. However, it is not clear how one can consider both sources of uncertainty when using cost-based models.

2.5.4 Alignment ambiguity in the frequentist paradigm

While both maximum likelihood and Bayesian methods attempt to simultaneously estimate model parameters $\mathbf{\Theta}$ and alignment $\mathbf{A}$ from the data $\mathbf{Y}$, the methods differ substantially in their handling of uncertainty in the alignment. In the maximum likelihood framework, a common method is to first construct an estimate $\hat{\mathbf{\Theta}}$ of $\mathbf{\Theta}$ by summing over all possible alignments in proportion to their probability:

$$\begin{aligned} \hat{\mathbf{\Theta}} &= \arg\max \mathrm{P}(\mathbf{Y}|\mathbf{\Theta}) \\ &= \arg\max \sum_{\mathbf{A}} \mathrm{P}(\mathbf{Y}, \mathbf{A}|\mathbf{\Theta}). \end{aligned}$$

Approximate confidence intervals for $\mathbf{\Theta}$ can then be obtained by assuming the asymptotic normality of $\hat{\mathbf{\Theta}}$ and estimating the inverse of the Fisher information matrix. These confidence intervals then account for uncertainty in both the alignment and parameter values because the alignment is summed out. However, uncertainty in the alignment is usually determined under the assumption that $\mathbf{\Theta} = \hat{\mathbf{\Theta}}$, by considering the alignment distribution $\mathrm{P}(\mathbf{A}|\mathbf{Y}, \hat{\mathbf{\Theta}}) = \frac{\mathrm{P}(\mathbf{A},\mathbf{Y}|\hat{\mathbf{\Theta}})}{\sum_{\mathbf{A}} \mathrm{P}(\mathbf{A},\mathbf{Y}|\hat{\mathbf{\Theta}})}$. Unfortunately, confidence intervals obtained from this distribution ignore uncertainty in the parameter estimates $\hat{\mathbf{\Theta}}$, and therefore do not take into account parameter uncertainty. Thus, in estimating alignments, the common maximum likelihood method accounts for alignment uncertainty in parameter estimates but does not account for parameter uncertainty in alignment estimates.

2.5.5 Alignment ambiguity in the Bayesian paradigm

The Bayesian approach to statistical estimation naturally incorporates both sources of uncertainty simultaneously. Bayesian inference is based on the joint posterior distribution of the alignment $\mathbf{A}$ and parameters $\mathbf{\Theta}$ given the data $\mathbf{Y}$. This distribution represents the posterior uncertainty in both $\mathbf{A}$ and $\mathbf{\Theta}$, as well as representing the correlation between them. The posterior distribution for $\mathbf{A}$ is obtained by integrating over possible values of $\mathbf{\Theta}$

$$\mathrm{P}(\mathbf{A}|\mathbf{Y}) = \int d\mathbf{\Theta}\, \mathrm{P}(\mathbf{A}, \mathbf{\Theta}|\mathbf{Y}).$$

Thus, credible intervals for $\mathbf{A}$ that are based on this posterior alignment distribution take parameter uncertainty into account. As a result, joint Bayesian estimation is an attractive method for estimating alignments and other parameters because it simultaneously accounts for uncertainty due to near-optimal alignments and also parameter uncertainty.

3 Alignment uncertainty: pairwise alignments

Before we consider methods of handling alignment ambiguity in multiple sequence alignments, we first summarize the progress made for pairwise alignments. Methods for handling alignment uncertainty in pairwise alignments preceded methods for handling uncertainty in multiple sequence alignments by a significant time period. Nevertheless, the development and improvement of methods is parallel, so that many improvements can be illustrated with pairwise alignments and then extrapolated to multiple sequence alignments.

3.1 Probabilistic models and pairwise alignment uncertainty

3.1.1 Models of the insertion/deletion process

Probabilistic approaches to pairwise alignment rely on a probabilistic model of the insertion/deletion process and the substitution process. For example, Bishop and Thompson (1986) specified a probability distribution on pairwise alignments in terms of the probabilities of gap opening and gap extension. Under such a model, parameters may be estimated via maximum likelihood without relying on the choice of a single alignment estimate. Instead, the likelihood for any set of parameter values is computed by using dynamic programming to sum the probabilities of all pairwise alignments, conditional on the given parameter values. Given an estimate of the evolutionary process parameters, a probabilistic model makes it possible to measure confidence for homology of two residues in terms of probabilities, by summing over all alignments that display the homology and weighting each alignment by its probability given the parameters. However, this maximum likelihood approach does not take parameter uncertainty into account when considering alignment uncertainty. A Bayesian approach to estimation allows the incorporation of prior beliefs about parameters and also incorporates posterior parameter uncertainty into measures of posterior alignment uncertainty (Allison and Wallace, 1994).

A further advance was the construction of stochastic process models for the insertion-deletion process (Thorne et al., 1991, TKF1). The TKF1 model specifies not only a distribution on pairwise alignments, but also describes how insertion and deletion events accumulate along over time to create a pairwise alignment between ancestor and descendant sequences. For example, previous probabilistic models did not distinguish between two adjacent deletion events and one long deletion because they model only "gaps" and not the indel events that create them. Additionally, the TKF1 model replaces the gap probability of the Bishop and Thompson model with an insertion rate and a deletion rate that are biological meaningful parameters. However, the TKF1 model has the drawback of assuming that all indels are of unit length. The Thorne et al. (1992, TKF2) model extends the TKF1 model by allowing indel lengths to follow a geometric distribution. The TKF2 model therefore adds an additional parameter to specify the extension probability of this distribution.

As noted above, the use of probabilistic models can decrease bias in parameter estimates by performing a weighted average over all possible alignments instead of considering only the optimal alignment. Especially for divergent sequences, estimates of the number of substitutions are biased upwards and estimates of the number of indels are biased downwards when using the optimal alignment (Thorne et al., 1991; Yee and Allison, 1993). Thus, when there is significant alignment uncertainty, the optimal alignment may not be typical of the set of plausible alignments. By instead averaging over all pairwise alignments, the bias is much decreased.

3.1.2 Extrapolation to multiple sequence alignments

Probabilistic models of insertion and deletion allow bioinformatic inferences to take alignment ambiguity into account by summing over all alignments. However, the number of possible pairwise alignments is astronomically large, growing faster than exponentially as sequence length increases. Sums over all pairwise alignments are computationally tractable because of the use of dynamic programming, an approach later formalized in terms of Hidden Markov Models (HMMs) (Durbin et al., 1998). The amount of computation time required by dynamic programming algorithms for pairwise alignments grows only as the square of sequence length for TKF models and the Bishop and Thompson (1986) model. This makes it possible to analyze sequence lengths of several thousand letters without approximations, and even greater lengths with approximations are made.

Unfortunately, there are several reasons that dynamic programming algorithms cannot be practically applied to directly sum over the possible alignments between three or more sequences. Firstly, the computation time and memory requirements increase as $O(L^n)$, where L is the sequence length and n is the number of sequences. Thus, dynamic programming can be performed for three sequences, provided the sequences are kept quite short, but it is impractical to extend to more than three sequences. Secondly, computation time and memory requirements grow exponentially in the number of observed sequences because the number of states in the HMM increases exponentially. Thus, even if it were a simple matter to describe a dynamic programming algorithm for aligning many sequences simultaneously, this algorithm would be too compu-

tationally burdensome to carry out. Therefore, sums over all possible multiple sequence alignments must be approximate. Using MCMC techniques, these sums can be approximated without visiting every possible multiple sequence alignment. However, the development of practical techniques for performing MCMC on multiple sequence alignments was delayed until Holmes and Bruno (2001) introduced new strategies for sampling alignments on a fixed topology. These new developments were of course assisted by the fact that the speed of desktop computers has continued to increase.

4 Alignment uncertainty: multiple sequence alignments

In contrast with the advanced statistical methods for dealing with uncertainty in pairwise alignments, common procedures for dealing with uncertainty in multiple sequence alignments have until recently been much less developed. Methods to handle uncertainty in multiple sequence alignments must handle many problems. First, they must be able to detect uncertainty in the alignment. In doing so, they must incorporate uncertainty from both near-optimal alignments and from parameter uncertainty including the phylogeny.

4.1 Identification of ambiguous regions based on a single alignment

When phylogenies are inferred by sequential estimation, alignment uncertainty is often handled by labeling some regions of the single alignment estimate as "ambiguous", leaving the remainder as presumably "unambiguous". The ambiguous regions are then thrown out, and the remaining alignment columns are then submitted as input for further analysis. However, identification of ambiguously alignment regions is in itself a challenging task. The success of this approach therefore depends critically on how well the researcher is able to identify unambiguous columns in the alignment. If the researcher fails to identify incorrectly aligned columns, then this failure may bias the rest of the analysis. On the other hand, the removal of correctly aligned columns decreases the power to distinguish between alternative hypotheses.

Unfortunately, alignment ambiguity is commonly identified by a subjective and *ad hoc* "visual inspection". Subjective determination of alignment ambiguity can lead to conflicting phylogeny estimates when researchers make different choices about including or excluding alignment regions (Lutzoni et al., 2000). In addition, subjective determination of alignment ambiguity makes it very difficult to reproduce reported results. Therefore, a significant challenge has been to identify ambiguous sites in an objective and repeatable way (Gatesy et al., 1993).

One method for avoiding this subjectivity is to use a computer program to remove all alignment columns that are near gaps and are not part of conserved blocks according to specified rules (Castresana, 2000, GBLOCKS). This approach codifies some of the intuition that is commonly used in removing alignment columns to prepare for phylogenetic analyses. However, this approach does not fully address the problem of identifying ambiguity and so has a number of drawbacks which make it difficult to apply to highly divergent sequences. First, the GBLOCKS method relies on a single alignment estimate to locate ambiguously aligned regions, and so the method is sensitive to the specific placement of gaps in this alignment, which may be incorrect. Thus, the automatic censoring of alignments created using different methods can retain different columns (Talavera and Castresana, 2007).

Second, the method attempts to assess ambiguity without any knowledge of evolutionary parameters such as the frequency of indels or the phylogenetic tree. Because the tree is not known, columns are identified as "conserved" based on the frequency of the majority character value in the column, instead of the predicted number of substitutions in the column, which may be small or large. Third, the GBLOCKS method uses a stringent and conservative test for retaining alignment columns, and so it may lead to the removal of a large number of phylogenetically informative characters. This results partly from the fact that the method is completely agnostic about alignment parameters, and bases confidence in alignments on conservation. Thus, it may throw out blocks that would be unambiguously aligned if the guide tree or indel costs were known. The removal of phylogenetically informative characters is also partly by design, because the removal of phylogenetic informative but rapidly changing sites is sometimes considered to be a positive feature when the phylogenetic inference method does not account for rate heterogeneity between sites. However, if the phylogenetic inference method can handle rate heterogeneity, then removal of these sites may be significant drawback.

Whether the approach is repeatable or not, the approach of categorizing columns as ambiguous or unambiguous has a few more drawbacks. First, the categorization is binary, but it would be preferable to have a degree of certainty in the homology of a column, because all alignment regions have some degree of uncertainty, no matter how small. Second, in identifying columns as ambiguous, the side knowledge of the researchers is not taken into account. For example, it is possible that a region could be ambiguous if the phylogeny is unknown, but relatively unambiguous if the phylogeny is known. Third, methods that remove gaps or alignment columns are common in phylogenetic inference, where it is common to view each column as giving independent evidence about the phylogeny. However, methods that remove columns with gaps are inherently unable to estimate some interesting biological parameters, such as indel rates. Methods that remove ambiguously aligned columns may also be detrimental to motif detection because they might remove the columns that contain a motif, when one of the sequences is ambiguously aligned to the other sequences. However, despite these issues, censoring simulated alignments via the GBLOCKS algorithm has been shown to improve the accuracy of phylogenies inferred from `Clustal W` alignments using maximum likelihood, maximum parsimony, and neighbor joining (Talavera and Castresana, 2007).

4.2 Multiple sequence alignment ambiguity resulting from parameter uncertainty

Many researchers have observed that the outcome of multiple alignment estimation depends on which values are chosen for parameters that characterize the evolutionary process (Gatesy et al., 1993). These parameters include the phylogeny used as a guide tree in many alignment methods, and the relative rates or costs assigned to insertions, deletions, and different types of substitutions such as transitions and transversions. When different values of evolutionary parameters are used, the estimated alignment may change. Therefore, uncertainty about the correct values of the parameters leads to uncertainty about the correct alignment.

4.2.1 Identifying ambiguous and unambiguous alignment regions

Gatesy et al. (1993) introduce a procedure called "culling" in order to objectively divide alignments into ambiguous and unambiguous regions. They propose to generate a collection of alignment estimates from a range of different costs for gaps and substitutions. Alignment columns which are not present in all of the resulting alignments are considered to be ambiguous, and are removed or "culled" from the alignment. The remaining columns are considered to be unambiguously aligned, and are retained for further analysis. This kind of method is called sensitivity analysis because it attempts to identify alignment regions that are not "sensitive" to parameter values[3]. This method is one of the only methods available that can be used to account for alignment uncertainty in a *non-statistical* (e.g. cost-based) framework.

Although this procedure is objective in the sense that it is repeatable, the range of parameter values is chosen based on the researcher's prior knowledge or subjective beliefs. This range includes all plausible values for these parameters, and it is the range of parameter values that determines which alignment columns are considered certain or uncertain. Thus, the degree of ambiguity in an alignment does not depend only on the sequences to be aligned, but also on prior knowledge about the evolutionary process. For example, Gatesy et al. (1993) varied transition, transversion, and gap costs. For the gap costs, they used values of 2/3, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 300 times the cost of a transition. However, if the gap cost was known to be between 2 or 3, then a much narrower range could be used, leading to a greater number of columns labeled as unambiguous. Also, if the alignment of a region is sensitive to a guide tree parameter, which is the case in progressive alignment, the region would be considered "ambiguous" if the phylogeny is not known, but unambiguous if the phylogeny is considered known. Clearly, if the resulting alignment estimates will be later used to estimate the phylogeny, the phylogeny must not be considered known. Instead, the range of parameters should include all plausible phylogenies, as well as all plausible values for substitution and indel process parameters. However, many sensitivity studies do not account for this dependence (Morrison and Ellis, 1997).

[3]We note that alignments may also be sensitive to which taxa are used in the analysis. This kind of sensitivity checking is oriented towards finding results that are robust to internal inconsistencies of the method, and is not discussed here.

4.2.2 Incorporating ambiguous information

The culling method has the unfortunate downside that it throws out a large fraction of all informative features. For example, in the two data-sets considered by Gatesy, only 91/250 and 12/250 sites were considered unambiguous. In addition, the culling method allows only two levels of confidence, ambiguous or unambiguous, instead of allowing various degrees of certainty about a column. Wheeler et al. (1995) address these concerns by a technique called "elision". Similar to culling, this technique also consists of generating a collection of alignments using a range of parameter values. However, instead of removing columns, the collection of alignments is concatenated end-to-end. When used as input for phylogeny estimation in a parsimony framework, this effectively weights each alignment column by the fraction of the total alignments that it occurs in.

We note that the use of several alternative alignments should not be seen as a contradiction (Lutzoni et al., 2000) but as an *ad hoc* way of considering several equally-weighted alternatives. If the parsimony score for the concatenated alignment is divided by the number of alternative alignments, then the elision method simply maximizes the average of the parsimony scores over all alternative alignments. However, it is not usually necessary to carry out this division, because it simply scales the score function and therefore does not affect the optimal phylogeny estimate. We note that this approach could in principle be applied in a statistical framework by maximizing either the average likelihood or the average log-likelihood, where the average is taken over all alignment alternatives. We recommend the first approach, since it corresponds to the hypothesis that all generated alignments to be equally likely *a priori*, whereas it is not clear to what hypothesis the second approach corresponds. However, the second approach is the direct analogue of the parsimony-based technique of Wheeler et al. (1995), because each alignment has equal influence in selecting the optimum tree, even though some alignments may lead to a lower score than others.

Because the elision method maximizes a score function that is obtained by averaging over a collection of alternative alignments, this collection can be seen as representing a probability distribution on alignments that results from a probability distribution over alignment parameters. From this perspective, the range of parameter values used in the sensitivity analysis then represents a prior distribution on parameter values. Unlike the culling procedure, multiple inclusion of a single parameter value leads to higher *a priori* confidence in that value. Although a discrete range of separate values may be used in practice, we note that denser sampling of values from a region leads to increased prior confidence that the true value lies in that region. For example, the choice of gap-opening costs used for culling means that 10/15 values lie between 1 and 10, corresponding to a weight of 2/3 on this possibility. In contrast, only three values like between 10 and 300, indicating a sparser sampling, and a lower prior weight on this region.

4.2.3 Shortcomings of sensitivity analysis

Compared to statistical methods, there are several shortcomings of sensitivity analysis, especially as applied to cost-based methods. First, the sensitivity analysis methods described above do not consider uncertainty resulting from near-optimal alignments, but only consider alignment uncertainty that results from parameter uncertainty. Because near-optimal alignments are ignored, the culling method may fail to identify some ambiguous regions, and the elision method may fail to sufficiently down-weight them. In addition, ignoring near-optimal alignments may exaggerate the effect of changing parameter values because it is possible that the set of near-optimal alignments remains almost the same under new parameter values, but a different alignment is chosen from this set. However, we note that this shortcoming seems to be primarily an accident of implementation instead of an essential property of sensitivity analysis. We also note, that if multiple equally optimal alignments are found, then the culling and elision approaches do consider these alternative alignments. While this does not seem to us to go far enough, does seem to be helpful feature and a step in the right direction.

Second, the prior information or subjective beliefs used in sensitivity analysis are not informed by the data. As a result, conducting a sensitivity analysis with a very broad range of values is likely to result in the detection of few unambiguous columns. Unfortunately, it is unclear how cost parameters in cost-based alignment methods can be informed by the sequences that they are aligning. In addition, it is unclear what it would mean for a value for a cost parameter to be "correct", since it does not correspond to a biological quantity. In contrast, given a statistical model of the evolutionary process, a Bayesian approach allows incorporation of prior knowledge or belief about parameter ranges in the form of a prior distribution, and

allows this information to be updated by the data. Thus, the posterior distribution may have a significantly narrower range, representing decreased uncertainty. Unlike sensitivity analysis, broad or "diffuse" prior distributions on evolutionary parameters do not necessarily lead to extreme uncertainty about the alignment. Therefore jointly estimating alignments and mutation rates via statistical methods should commonly result in less alignment ambiguity than a sensitivity analysis, and is less dependent on prior knowledge or subjective beliefs about parameter ranges.

5 Statistical inference under alignment uncertainty

Sound statistical methods for inferring evolutionary parameters from molecular sequence data in the presence of alignment uncertainty all take as given only the observed, unaligned sequence data. These methods require a joint probabilistic model which describes how unobserved evolutionary parameters may combine to generate the observed sequences. These parameters include the alignment, the tree, and the rate of occurrence of different kind of mutations. We note that when inferring the tree or mutation rates, the alignment is considered a nuisance variable which may be summed out. However, the same statistical model may be used when the alignment is the parameter of interest, to sum out uncertainty in the tree. In this section we give a brief introduction to the model, notation, and methods described in (Redelings and Suchard, 2007) because this paper describes how we analyze the data example presented in section 7. Then, using this approach as a reference we describe other recent developments in statistical estimation of alignments, trees and other parameters. We then discuss shortcomings and possible improvements.

5.1 Joint Statistical Model

5.1.1 Variables and definitions

We begin by describing the data and unobserved evolutionary parameters in the model. We consider a collection $\mathbf{Y}$ of n homologous molecular sequences. The individual sequences are labeled $\mathbf{Y}_i$, indexed by $i = 1 \dots n$, and have lengths $|\mathbf{Y}_i|$. Each sequence $\mathbf{Y}_i$ has elements $Y_i[j]$ indexed by $j = 1 \dots |\mathbf{Y}_i|$ that take on values in a set α called the alphabet. Each letter in the alphabet represents a monomer in the molecular sequences $\mathbf{Y}$. For example, if the sequences are DNA sequences, then the alphabet consists of the nucleotides {A, T, G, C}, whereas if the sequences are protein sequences, then the alphabet is the set of amino acids. We note that the alphabet does not contain a gap letter "-" because gaps are not part of the observed data, but must be inferred.

The data $\mathbf{Y}$ are generated from an evolutionary process that is characterized by a number of unobserved parameters that we seek to estimate. The alignment $\mathbf{A}$ and the phylogenetic tree combine to specify the complete evolutionary relationship of the sequences in $\mathbf{Y}$. The alignment $\mathbf{A}$ is separable from the observed letters in $\mathbf{Y}$ because it specifies the homology of these letters without mentioning their values. The alignment therefore specifies how the letters in $\mathbf{Y}$ may be arranged to form the aligned data matrix $\mathbf{f}$. This matrix indicates which letters are homologous to each other by arranging groups of homologous letters into a single column. Each row of $\mathbf{f}$ contains one of the sequences of $\mathbf{Y}$, so that each column contains one letter from each sequence, or possibly a missing value. We denote the unknown number of columns C. The phylogenetic tree may be separated into its topology τ and its branch lengths $\mathbf{T}$. We assume that the tree is unrooted, and define the topology as an undirected acyclic graph in which all internal nodes have 3 neighbors. The topology contains exactly n leaf nodes and $n-3$ internal nodes, and each leaf node corresponds to one of the n observed sequences. The total number of nodes, which we denote by N is therefore $2n-3$. Each branch b is associated with a branch length $T^{(b)}$.

Evolutionary parameters also include parameters $\mathbf{\Theta}$ that characterize the substitution process, and parameters $\mathbf{\Lambda}$ that characterize the insertion-deletion process. The substitution parameters $\mathbf{\Theta}$ include the rates of nucleotide or amino acid replacement and also describe how these rates vary between different sites. The insertion-deletion parameters $\mathbf{\Lambda}$ include not only the rates of accepted insertions and deletions, but also specify the length distribution of accepted indels. We assume that insertions and deletions have the same rate λ, and that the length of indels follows a geometric distribution with extension probability ϵ, so that $\mathbf{\Lambda} = (\lambda, \epsilon)$. Taken together, the entire state space Ω is composed of points $\omega = (\mathbf{Y}, \mathbf{A}, \tau, \mathbf{T}, \mathbf{\Theta}, \mathbf{\Lambda})$.

5.1.2 Probability expression

Traditional methods for estimating the tree or other bioinformatic parameters have assumed that the alignment was known with certainty. These approaches therefore implicitly condition on the alignment, leading to the probability expression

$$\mathrm{P}(\mathbf{Y}, \tau, \mathbf{T}, \mathbf{\Theta}|\mathbf{A}) \quad = \quad \mathrm{P}(\mathbf{Y}|\mathbf{A}, \tau, \mathbf{T}, \mathbf{\Theta}) \times \mathrm{P}(\tau, \mathbf{T}) \times \mathrm{P}(\mathbf{\Theta}) \tag{1}$$

where, following a common abuse of notation, we write $\mathrm{P}(\mathbf{X})$ to represent $\mathrm{P}(\mathbf{X} = \mathbf{x})$ for any random variable $\mathbf{X}$ taking on a realized constant value $\mathbf{x}$. The first term $\mathrm{P}(\mathbf{Y}|\mathbf{A}, \tau, \mathbf{T}, \mathbf{\Theta})$ in equation (1) is the likelihood for the model, and is determined by the model of the substitution process. The other terms represent prior distributions on trees and on substitution process parameters, respectively.

In contrast with this traditional approach, a joint probability model allows the alignment to vary, yielding a modified probability expression

$$\mathrm{P}(\mathbf{Y}, \mathbf{A}, \tau, \mathbf{T}, \mathbf{\Theta}, \mathbf{\Lambda}) \quad = \quad \mathrm{P}(\mathbf{Y}|\mathbf{A}, \tau, \mathbf{T}, \mathbf{\Theta}) \times \mathrm{P}(\mathbf{A}|\tau, \mathbf{T}, \mathbf{\Lambda}) \times \mathrm{P}(\tau, \mathbf{T}) \times \mathrm{P}(\mathbf{\Theta}) \times \mathrm{P}(\mathbf{\Lambda}). \tag{2}$$

We note that equation (2) is identical to equation (1) except for the addition of two new terms. We may therefore base the likelihood $\mathrm{P}(\mathbf{Y}|\mathbf{A}, \tau, \mathbf{T}, \mathbf{\Theta})$ on traditional substitution models such as reversible, continuous-time Markov chains. The first new term, $\mathrm{P}(\mathbf{A}|\tau, \mathbf{T}, \mathbf{\Lambda})$, is the prior distribution on alignments and is based on the insertion-deletion process. We describe a prior distribution on alignments below that is biologically realistic and penalizes alignments with more indels. The second new term, $\mathrm{P}(\mathbf{\Lambda})$, is the prior distribution on insertion-deletion process parameters. We note that the likelihood and the alignment prior are separable in equation (2) because the substitution process and the insertion-deletion process are separate and operate independently. This is possible because the alphabet does not include a "gap" letter "-", so that the substitution process is not responsible for insertions and deletions.

5.1.3 Substitution model

The substitution model determines the likelihood $\mathrm{P}(\mathbf{Y}|\mathbf{A}, \tau, \mathbf{T}, \mathbf{\Theta})$ that the letters $\mathbf{Y}$ are observed. This probability may be expressed in terms of the aligned data matrix $\mathbf{f}$ that depends on both the data $\mathbf{Y}$ and the alignment $\mathbf{A}$, as

$$\mathrm{P}(\mathbf{Y}|\mathbf{A}, \tau, \mathbf{T}, \mathbf{\Theta}) \quad = \quad \mathrm{P}(\mathbf{f}|\tau, \mathbf{T}, \mathbf{\Theta}). \tag{3}$$

This approach is useful because we assume that observations in each column of the aligned data matrix are independent realizations from the substitution process, so that the full likelihood is the product of the likelihood of each column in the aligned data matrix $\mathbf{f}$. We follow common practice in molecular phylogenetics by using reversible continuous-time Markov chain (CTMC) models to describe the process of substitution in each column (Goldman, 1993) and will not describe them here. Therefore, in order to express this probability we must define the aligned data matrix $\mathbf{f}$ and describe how to construct it from the data $\mathbf{Y}$ and the alignment $\mathbf{A}$.

The matrix $\mathbf{f}$ consists of rows indexed by $i = 1 \ldots N$ and columns indexed by $c = 1 \ldots C$. The letters in row i of $\mathbf{f}$ all come from sequence i and must occur in order. The matrix $\mathbf{f}$ represents the hypothesis that all the letters in each column c descend from a single residue in the sequence of the common ancestor. If no letter in sequence i is homologous to other residues in column c, then we place the a gap "-" at f_{ic} to indicate a missing value. In addition to the observed leaf sequences, the matrix $\mathbf{f}$ also includes unobserved sequences at internal nodes as missing data. These sequences are composed of Felsenstein wildcards that are represented by "$\star$" to indicate that a letter is present but its value is unknown.

We introduce the matrix $\mathbf{M}(\mathbf{A})$ to specify how the alignment $\mathbf{A}$ arranges the data $\mathbf{Y}$ into the matrix $\mathbf{f}$ while remaining separable from $\mathbf{f}$. The matrix $\mathbf{M}(\mathbf{A})$ has the same dimensions as $\mathbf{f}$, and specifies which letter of sequence Y_i belongs in column c through the equation

$$f_{ic} \quad = \quad \mathbf{Y}_i[M_{ic}]. \tag{4}$$

If no element of Y_i belongs in column c, then $M_{ic} = \text{`}-\text{'}$ and we define $\mathbf{Y}_i[\text{`}-\text{'}] = \text{`}-\text{'}$. Figure 1 illustrates the relationship of $\mathbf{Y}$, $\mathbf{M}(\mathbf{A})$, and $\mathbf{f}$ for a 4-taxon example.

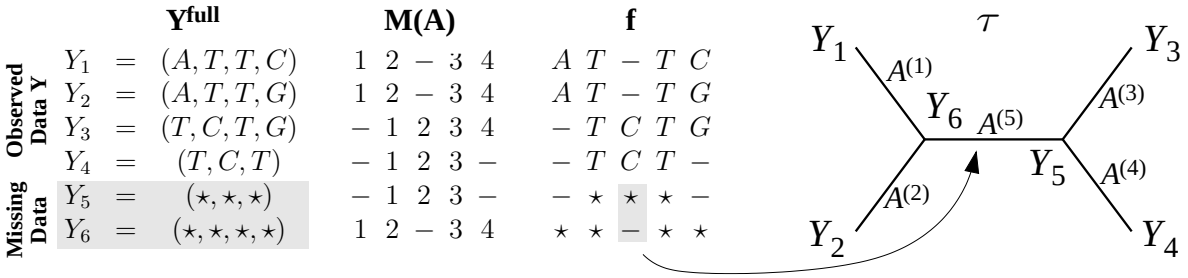


Figure 1: Construction of aligned data matrix $\mathbf{f}$ from data $\mathbf{Y}$ and alignment $\mathbf{A}$. The first column shows some unaligned sequences $\mathbf{Y}^{\text{full}}$ of various lengths. These sequences include the observed leaf sequence data $\mathbf{Y}$ as well as unobserved sequences Y_5 and Y_6 at internal nodes of the tree. The lengths of the sequences Y_5 and Y_6 are unknown and all possibilities must be considered; one possibility is shown. The second column shows $\mathbf{M}(\mathbf{A})$, a matrix that parametrizes the multiple sequence 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 sequence alignment. The third column shows the aligned data matrix $\mathbf{f}$, 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. While the alignment is separable from the data, as shown in column 2, the aligned data matrix $\mathbf{f}$ is not. The fourth column contains the evolutionary tree topology τ. Each branch b is associated with a pairwise alignment $A^{(b)}$ between the two sequences at the endpoints of the branch. This is made possible by the inclusion of the missing data Y_5 and Y_6 in the alignment $\mathbf{A}$, and allows each indel to be localized to a particular branch. For example, the shaded indel in column 3 of $\mathbf{f}$ must occur on the internal branch of the tree.

5.1.4 Alignment prior

As described above, the multiple sequence alignment $\mathbf{A}$ includes alignment information for internal node sequences as well as leaf sequences. However, the observed data $\mathbf{Y}$ specifies the letters only for leaf sequences (Figure 1). The alignment $\mathbf{A}$ also specifies the length of all sequences in addition to homology information, so that $\mathbf{A}$ must agree with $\mathbf{Y}$ on the length of observed sequences, but the length of unobserved sequences at internal nodes of the tree is unknown.

Augmenting the alignment to include homology about sequences at internal nodes is beneficial because it allows us to decompose the multiple sequence alignment $\mathbf{A}$ into a collection of pairwise alignments (Holmes and Bruno, 2001). Given a topology τ with B branches, $\mathbf{A}$ can be decomposed into a tuple of pairwise alignments $(A^{(1)}, \ldots, A^{(B)})$. Each branch b of the tree is associated with a pairwise alignment $A^{(b)}$ that specifies the homology of the sequences at each end of the branch. Representing the alignment $\mathbf{A}$ in this way allows us to construct a distribution on multiple sequence alignments from a distribution $\nu_{\lambda t_b}$ on pairwise alignments that is parametrized by indel rate λ and branch length t_b for branch b. Pairwise alignments on each branch will be independent conditional on the lengths of the sequences at the internal nodes because evolution occurs independently on each branch of the tree. This leads to an alignment prior of the following form

$$\mathrm{P}(\mathbf{A}|\tau, \mathbf{T}, \mathbf{\Lambda}) \quad = \quad \frac{\prod_{b=1}^{B} \mathrm{P}_{\nu_{\lambda t_b}}(\mathbf{A}^{(b)})}{\prod_{i \in I} \phi(|\mathbf{A}_i|)^2}, \tag{5}$$

where $\phi(\cdot)$ indicates the time-independent sequence length distribution (Redelings and Suchard, 2007). We note that augmenting the alignment to include internal sequences is beneficial also because it specifies exactly on which branch of the tree each indel occurs, making it possible to model indel events instead of modelling gaps (Figure 1).

5.2 Inference under a statistical model

Three methods for joint estimation of alignment and phylogeny have been proposed to date, and we compare and contrast these methods here. Redelings and Suchard (2005, RS05) and (Lunter et al., 2005, LMDJH05) draw inference in a Bayesian framework using Markov chain Monte Carlo (MCMC) integration, whereas Fleissner et al. (2005, FMH05) employ a maximization approach using simulated annealing. Both MCMC and simulated annealing are guaranteed to compute exact solutions given sufficient processing time. However, in practice, both approaches sacrifice some degree of precision in order to avoid considering all possible combinations of alignments, trees, and other parameters, which would be computationally prohibitive.

MCMC differs from the simulated annealing approach of FMH05 in two major ways. First, MCMC naturally leads to measures of uncertainty, such as posterior confidence intervals for continuous parameters and posterior probabilities for tree topologies, alignments, and other discrete parameters. As a maximization approach, the FMH05 method instead generates a single parameter estimate, but does not produce a measure of confidence in the estimate that would reflect the degree of evidence for it in the data. However, the maximization approach may be significantly faster, since it only needs to find the optimum tree and alignment, instead of visiting a large number of other points to yield measures of confidence.

Second, the Bayesian approach involves summing over all possible alignments in order evaluate the relative probability of evolutionary tree topologies and other parameters. In contrast, the FMH05 approach does not sum over all possible alignments, but instead maximizes over them. This approach differs from the Bayesian approach and also from the standard maximum likelihood procedure, which involves maximizing the likelihood $\mathrm{P}(\mathbf{Y}|\tau, \mathbf{T}, \mathbf{\Theta}, \mathbf{\Lambda})$ of the observed data. Computation of this likelihood naturally involves a sum over all alignments, since

$$\mathrm{P}(\mathbf{Y}|\tau, \mathbf{T}, \mathbf{\Theta}, \mathbf{\Lambda}) \quad = \quad \sum_{A} \mathrm{P}(\mathbf{Y}, \mathbf{A}|\tau, \mathbf{T}, \mathbf{\Theta}, \mathbf{\Lambda}).$$

The FMH05 approach avoids computing or approximating this sum and instead maximizes the likelihood $\mathrm{P}(\mathbf{Y}, \mathbf{A}|\tau, \mathbf{T}, \mathbf{\Theta}, \mathbf{\Lambda})$ of the observed data and an unobserved alignment. The approach therefore maximizes over the alignment $\mathbf{A}$ even though it is a latent variable representing missing data and should instead be summed out, perhaps using the expectation-maximization (EM) algorithm. The choice to maximize over the alignment may therefore lead to biased estimates of branch lengths and other parameters, as mentioned above. However, because it may lead to faster computation, it may allow analysis of larger data sets.

5.2.1 Computational efficiency: MCMC and simulated annealing

In order to be computationally efficient both MCMC and simulated annealing require well-designed "transition kernels" that propose new alignments and trees. Proposed new alignments and trees should frequently be of similar or higher probability than the previous value, or they will be rejected, leading to wasted computation time. In addition, proposed values must sometimes be substantially different that current values. If this is not the case, then simulated annealing methods will move only slowly away from the starting point, or fail to explore the full parameter space, and thus fail to find the optimum point in a short time. Similarly, MCMC proposal functions that fail to propose substantially different points may lead to a failure to converge to the equilibrium distribution, or a high autocorrelation between adjacent samples and a low effective sample size. MCMC samplers may be used as simulated annealing search algorithms by raising the posterior probability density to successively higher powers in each iteration. This approach is roughly followed by FMH05, with the exception that certain details required for correct sampling from the posterior distribution were ignored because they were not necessary to find the optimum.

Bayesian sampling of multiple sequence alignments via MCMC was pioneered by Holmes and Bruno (2001, HB01). The HB01 approach samples the alignment under the TKF1 model given a fixed tree topology. HB01 introduced the idea of augmenting the alignment to include homology information about internal node sequences, and also introduced two novel transition kernels that require this augmentation in order to resample parts of the multiple sequence alignment. While the augmentation enables the use of the new HB01 transition kernels, it also makes it difficult to change the tree topology. This is because the augmented alignment makes sense only on a given tree topology, since it specifies a pairwise alignment for each branch. Therefore, when changing the tree topology, a new indel history must be proposed that is consistent with the new topology.

In order to relax the constraint of a fixed tree, each of the three joint-estimation methods takes a different approach that has unique advantages and disadvantages. The RS05 and FMH05 methods retain alignment augmentation, while the LMDJH05 approach dispenses with it. The LMDJH05 approach is able to dispense with alignment augmentation because it uses an "indel peeling algorithm" to evaluate the probability of unaugmented alignments on a tree by summing over all possible augmentations (Lunter et al., 2003). The indel peeling algorithm is currently restricted to the TKF1 model, and so the LMDJH05 approach assumes that all gaps have unit length. Summing out the augmentation is beneficial to the LMDJH05 approach because it means that new tree topologies can be proposed without being constrained to be consistent with the current indel history. In addition, summing over missing data generally leads to decreased autocorrelation between adjacent MCMC samples, and thus improves mixing efficiency compared to augmenting with missing data if the time per sample is not increased excessively by the summation.

However, the lack of alignment augmentation has the drawback that the HB01 transition kernels for alignment sampling are not available. The second of these transition kernels (HB01-TK2) simply updates the alignment augmentation, and therefore is not needed in the LMDJH05 approach. However, the loss of the first of the transition kernels (HB01-TK1) is an important drawback. The HB01-TK1 transition kernel divides the alignment into two sub-alignments along a branch of the tree, and re-aligns the two sub-alignments with respect to each other. This proposal is an efficient method for resampling alignments, and is never rejected because the proposal is proportional to the target distribution. However, this is achieved by resampling the pairwise alignment of two (possibly unobserved) sequences at either end of the tree branch. LMDJH05 is not able to re-align sub-alignments in this way because it does not store alignment information for internal sequences, nor can it sample this information from the correct distribution to use temporarily. Thus the LMDJH05 approach is instead forced to propose alternative alignments that may be rejected. Because the probability of rejection grows with sequence length, only blocks of the alignment of randomly chosen lengths have their alignment resampled in this fashion.

In contrast, the RS05 approach extends the HB01 method to sample topologies without removing the alignment augmentation for internal node sequences. By summing out only the alignment of sequences at internal nodes that lose definition during a nearest-neighbor interchange (NNI), the RS05 approach is able to compare nearby topologies. The RS07 method extends this approach to subtree-prune-and-regraft (SPR) changes to tree topology as well. Because these summations are carried out using dynamic programming, the alignment information for internal node sequences that is summed out in order to change topologies may be resampled to be compatible with the topology that is chosen, thus reconstituting the alignment augmentation. In summary, traditional transition kernels for topologies, such as NNI and SPR, are modified in the RS05 approach to resample the indel history as well as the topology in order to maintain compatibility between the two. Unfortunately, this kind of procedure may be too computationally inefficient to use for larger topology changes, such as those induced by tree-bisection-and-reconnection (TBR). Thus, the RS05 approach is constrained in the topologies that it may propose compared to the LMDJH05 approach.

RS05 introduces two novel alignment transition kernels can substantially improve mixing efficiency. The first of these (RS05-TK1) resamples alignment information about unobserved sequences at two adjacent internal nodes which keeping fixed implied the alignment between all other sequences. This transition kernel is helpful for efficient mixing of MCMC, and is also required for summing out local alignment augmentation to allow NNI proposals. The second of these novel alignment transition kernels (RS05-TK2) divides the alignment into two sub-alignments along a branch of the tree, and simultaneously resamples the alignment of the two sub-alignments and the alignment information about an internal node sequence at one end of the branch. This transition kernel helps to avoid improbable intermediate states, and may improve mixing efficiency dramatically. For example, this transition kernel improved convergence speed by more than 70-fold in a simple 12-taxon example (Redelings and Suchard, 2005). The RS05-TK2 transition kernel is used to sum and resample alignment augmentation for SPR topology changes. We also note that the RS05-TK2 transition kernel subsumes both HB01-TK1 and HB01-TK2 in the sense that it resamples the alignment with fewer constraints than either of HB01-TK1 or HB01-TK2.

5.2.2 Insertion/deletion models

The three approaches to joint estimation of alignment and phylogeny make use of three different indel models. The FMH05 approach makes use of the TKF2 indel model. The TKF2 model has the benefit of

allowing multiple residue indels, as well as the benefit of being a continuous-time stochastic process that specifies how indel events occur along a branch. We note that the TKF2 model makes the biologically unrealistic assumption that each sequence is composed of unbreakable fragments, which are inserted or deleted as a unit, thus disallowing overlapping insertions and deletions. The drawbacks of this approach may be largely removed by allowing the fragment boundaries to be different and independent on each branch of the tree, so that indels on different branches may indeed overlap. However, in the FMH05 approach, these fragments boundaries are not allowed to differ across branches of the tree. This choice results in simpler and faster dynamic programming algorithms, but leads to unrealistically low predictions for the probability of overlapping indels.

In contrast to the FMH05 approach, the other two approaches each give up one of the advantages of the TKF2 model. For example, the LMDJH05 approach uses the TKF1 model. Like the TKF2 model, the TKF1 model is a continuous-time stochastic process, and so it describes the generating of individual indel events along each branch of the tree. However, it allows only single-residue indels, and therefore may fail to cluster gaps in its alignment estimates. In addition, by treating a deletion of several residues as several independent deletions of one residue, the TKF1 model may over-weight the phylogenetic evidence in shared indels of multiple residues.

The RS05 approach uses a model that allows indels of multiple residues, but is based on an HMM instead of a continuous-time stochastic process. As a result, the RS05 model places a distribution directly on pairwise alignments and does not describe the dynamics by which indel events accumulate along a branch of the tree to yield a pairwise alignment between ancestor and descendant sequences. Additionally, in the RS05 model the probability of an indel occurring on a branch is independent of the branch length. This drawback was remedied in Redelings and Suchard (2007, RS07), which introduces an indel rate parameter, instead of just specifying the probability that an indel occurs on a branch. As a result, the RS07 model has biologically interpretable parameters similar to a TKF2 model in which the insertion and deletion rates are equal, and leads to a distribution on pairwise alignments that is approximately the same as the distribution generated by the TKF2 model for short times.

5.3 Improved statistical models of insertion and deletion

Statistical methods for inferring alignments or for inferring other parameters in the presence of alignment uncertainty all rely on stochastic models of the insertion-deletion process to correctly down-weight alignments with more indel events. In order to yield high-quality inferences, such models must be biologically realistic in specifying the frequency of occurrence for different kinds of indel events. Unfortunately, there is often a trade-off between biological realism and computational efficiency, so researchers must seek a balance between the quality of inference and speed of inference, instead of employing the most realistic models that are known.

Since the initial development of the TKF1 model, several extensions have been proposed. The TKF1 model has just two additional parameters: the insertion rate λ and the deletion rate μ. The TKF1 model assumes an exponential distribution of sequence lengths at equilibrium with mean $\lambda/(\mu - \lambda)$. Only single residues may be inserted or deleted, leading to "linear gap penalties". To remedy this problem, the TKF2 model applies the TKF1 model to unbreakable sequence "fragments" containing multiple residues. The number of residues in a fragment is random and is distributed according to a geometric distribution with parameter ϵ and mean $1/(1 - \epsilon)$. This leads to "affine gap penalties." Thus, insertion and deletion of fragments lead to multiple-residue insertions and deletions. The downsides of this approach are that (a) imaginary fragment boundaries remain after fragments are inserted (b) it is impossible to delete parts of a fragment (c) indel rates are now *per-fragment* instead of *per nucleotide*. Lastly, another difficulty that is less important, but still real, is that (d) probably neither indel mutations nor accepted indels have a geometric length distribution (Cartwright, 2006). A geometric length distribution is significantly easier to deal with computationally, but tends to underestimate the probability of seeing long indels. There have therefore been several attempts to improve on these models. The models are used at *equilibrium* so that the total forward and backward rates are equal. Length distribution is geometric (TKF1) or roughly geometric (TKF2). In both the TKF1 and TKF2 models, pairwise alignments can be estimated in $O(NM)$ time for two sequences of length N and M. One approach that avoids some problems with the TKF2 models is simply to assume that fragment boundaries can be different on each branch of the evolutionary tree. This allows a fragment that is inserted on one branch to be partially deleted on a different branch of the tree. This technique

is attractive because it does not substantially increase the computational burden. However, other, more accurate models have been proposed as well.

Miklos et al. (2004) introduce a new model that improves on the TKF1 model by allowing insertions and deletions of multiple residues without relying on unobserved fragments as the TKF2 does. Furthermore, this model allows an arbitrary length distribution for indels instead of requiring a geometric length distribution, and is therefore called the "long indel model." However, conducting inference under this model is significantly more computationally burdensome than the TKF1 or TKF2 models. Firstly, the dynamic programming algorithm for this model is $O(M^2N^2)$ when no approximations are made, instead of $O(MN)$. Secondly, the dynamic programming recursion involves terms whose value cannot currently calculated analytically. Instead, these terms are estimated using MCMC methods to approximately calculate "trajectory likelihoods", where all intermediate events on a branch must be explicitly considered. Additionally, certain reasonable assumptions about the number of overlapping indel events were made to ensure computational tractability. While this long indel model does yield improved estimates of pairwise alignments, it may be too computationally expensive to incorporate in multiple sequence alignment estimation. Despite this fact, the long indel model has been used to improve pairwise alignment accuracy.

One statistical issue that is not addressed by current models is the sequence length distribution. Most statistical models assume that over evolutionary time the equilibrium distribution on sequence length will converge to a geometric distribution whose mean is the same for all genes. The use of a geometric distribution implies that a shorter sequence length is always more likely than a longer sequence length. Although this assumption is mathematically convenient, it seems more biologically realistic that each protein family would have a separate mean length over evolutionary time, and that the most likely length would not be zero. When analyzing a collection of homologous genes in a Bayesian framework, it would be desirable to use an informative prior on the mean sequence length that is based on an empirical distribution of protein family lengths in curated databases. However, the degree of variation around this mean could be estimated from the homologous genes themselves. Unfortunately, the failure of current models to separate the distribution of lengths within and between protein families prevents informative priors from being used.

We note that equilibrium sequence length distributions are determined entirely by mutation pressure, and thus imply completely neutral evolution on sequence lengths. This unrealistic assumption could be replaced in two separate ways. First, it is possible to add to an indel model selection for sequence length *per se*. Second, it is possible to consider that some "conserved" residues are much less tolerant of deletion over evolutionary time, and that the number of such residues within a protein remains roughly constant within the protein family. This second option would tend to counteract the assumption within current indel models that over a long evolutionary time period, all ancestral residues should be deleted and replaced with newly inserted residues. It would also be able to model heterogeneity in the indel process and handle "indel hot-spots" similar to site-heterogeneity models for substitution rates. As an extreme example of this type, one could consider a model in which some unknown fraction of residues may never be deleted, similar to the invariant sites assumption for substitution models (Thorne et al., 1992).

5.4 Limitations of statistical methods

Although statistical techniques are now available for making robust inferences in the presence of alignment ambiguity, these methods have important limitations. First, these techniques may require vast amounts of computer processing time, severely limiting the size of data sets that can be analyzed in practice. For example, an MCMC analysis under the model described in equation (2) of 12 protein sequences with lengths of about 450 amino acids required seven days in 2005 (Redelings and Suchard, 2005). While such analyses are feasible, they indicate that current approaches to joint estimation of alignment and phylogeny are not able to cope with either long sequences or a large number of taxa. In the current manuscript we analyze a 25-taxon data-set, but we note that the length of the sequences is extremely short at about 130 nucleotides, indicating a trade-off between sequence length and number of taxa.

Second, the statistical models of indel formation are limited in the types of biological processes that they consider. For example, current models assume that insertion and deletion rates are independent of the DNA or amino-acid sequence where they occur, and are distributed uniformly across the DNA or protein molecule. In addition, current models do not allow duplication or within-sequence homology, but assume that inserted sequences are random and independent of the parent sequence. These assumptions can lead to inaccurate

estimates of phylogenetic trees and other parameters when the sequence data has evolved according to indel processes that are not present in the model. For example, in phylogeny estimation, indel events are all assigned the same weight, regardless of whether or not they occur in an indel hot-spot. Expansion and contraction of simple sequence repeats may be more rapid than other types of indels due to slipped-strand mispairing, but these changes will not be appropriately down-weighted.

6 Representing alignment ambiguity

Alignment ambiguity is more difficult to represent than ambiguity in continuous parameters, such as branch lengths or indel rates. This follows from the fact that alignments are discrete and unordered, so that one cannot summarize their distribution by a mean and variance, or by a median and a confidence interval of the form (a, b). Alignment ambiguity must be represented either for visual display, or for use as input to an inference procedure. When representing alignment ambiguity visually, it is often more important that the alignment distribution be summarized in a single figure than that all the information in the distribution is preserved.

6.1 Alignment ambiguity for use in further analysis

The simplest approach is to divide an alignment into certain and uncertain regions. This has the downside of assuming that homology is known with complete confidence (which is unlikely) or else not known at all. That is, there are no degrees of confidence in homology statements. We note that the issue of confidence is theoretically separate from how the confidence measure is calculated. Even the best method for assessing confidence will be of marginal usefulness if it is limited to declaring alignment regions "certain" or "uncertain." However, the results of this method are easy to visualize. This representation of alignment ambiguity is used by the culling method and the GBLOCKS approach.

The next method of representing alignment confidence is by a list of alignments of equal weight, as in the elision method. For analyses that treat columns independently, columns that occur in many alignments effectively have a higher weight than columns that do not. If alignments themselves may be repeated multiple times, then we can replace the list by a set of unique alignments associated with integer weights that indicate the number of times these alignment occur in the previous list. This list may then indicate a collection of plausible alignments, along with their weights. However, we note that the elision method explicitly concatenates the alignment end-to-end, and that the weights lead to a weighted sum of parsimony scores for each tree. We note that a list of alignments is difficult to visualize.

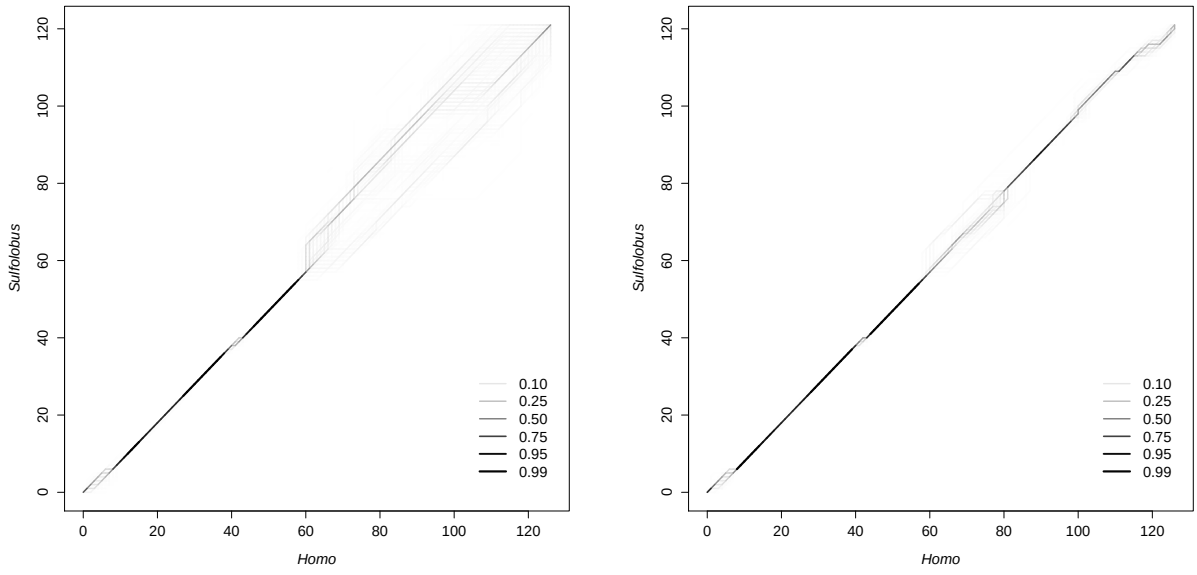
Finally, a general framework for dealing with certainty and uncertainty about alignments is to consider a probability distribution on alignments. This distribution may be represented by a sample, and this sample will most likely be unweighted since the number of possible alignments makes it unlikely a single alignment will recur. However, the sample conceptually represents a weighted collection of alignments, as opposed to an unweighted collection. This conception naturally arises in the Bayesian statistical framework, in which case the posterior alignment distribution must be represented.

6.2 Graphically representing alignment ambiguity

6.2.1 Pairwise alignments

A probability distribution on alignments between two sequences can be summarized on a 2-dimensional sheet of paper or computer screen using a path graph representation (Naor and Brutlag, 1994). Given two sequences of length M and N, each pairwise alignment corresponds to a path through an integer lattice from $(0, 0)$ to (M, N) such that diagonal edges represent match columns and vertical or horizontal segments represent gap columns. For example, the edge $(i-1, j-1) \rightarrow (i, j)$ indicates that character i or the first sequence and character j of the second sequence are homologous. Such paths are known as path graphs, and correspond to a route through the dynamic programming matrix.

One method for using path graphs to indicate alignment ambiguity is to color each possible edge according to the probability that it occurs in the alignment (Figure 2). For example, edges may be colored black if they are certain to occur, white if they are certain not to occur, and shades of gray for intermediate degrees



(a) 6 sequences

(b) 25 sequences

Figure 2: Weighted path graph representation allows easy comparison of two pairwise alignment distributions. While only the alignments between the *Homo* and *Sulfolobus* sequences are shown, the full analyses are based on (a) 6 sequences and (b) 25 sequences respectively. The pairwise alignment distribution based on 25 sequences contains substantially less alignment ambiguity, indicating that alignment ambiguity between a pair of sequences may decrease when additional sequences are included in the analysis. Line segments with high posterior probability are darker and thicker, while edges with lower posterior probability are colored with lighter shades of grey. The posterior probabilities in this figure are based on the 25-taxon 5S rRNA data set (b) described in the Results section and a 6 sequence subset of those sequences (a). The RS07 indel model was used, along with the GTR+gwF+log-Normal$_8$ model (Section 7.1).

of certainty. We note that representation of the alignment distribution for 2 sequences may not capture the full information in the distribution because it does not represent correlations in support for adjacent columns. Additionally, when an insertion is adjacent to a deletion, multiple paths may correspond to the same homology structure, leading to trouble with path graph interpretation. However, drawing weighted path graphs in this manner does successfully convey which regions of the alignment contain plausible alternatives, and also conveys how plausible these alternatives are.

Although this method could in theory be used to represent alignment distributions of n sequences where $n > 2$, such representations are impractical because they would require an n-dimensional plotting surface. Even a 3-dimensional version of this technique is not practical because the most probable edges may be hidden behind less probable edges from every viewing angle. Thus, 2-dimensional projections of a 3-dimensional image are not sufficient. Another alternative for using path graphs to represent a distribution on multiple alignments is to draw all projected pairwise alignments separately. This technique can be useful when n is small, although information about the correlation between sequences is lost. However, the number of such graphs grows quadratically with n, and so quickly becomes unmanageable for large n.

6.2.2 Selecting a representative multiple sequence alignment

When summarizing multiple sequence alignment distributions, one common approach is to construct a representative alignment and then annotate it to indicate the degree of confidence in various regions. Selection of a representative multiple sequence alignment is itself a challenging task. One criterion would be to select the multiple sequence alignment with the highest posterior probability. However, this criterion has two drawbacks - one theoretical, and one practical. First, the most probable alignment may not be very probable, and so that individual columns may have low posterior probability. Second, it is common for each alignment

sampled via MCMC to be unique. This indicates that the most probable alignment may not have been sampled and that it is not possible to determine the posterior probability of the alignments that have been sampled. One simple way around this problem is to select the multiple alignment from the sampled point with the higher posterior probability (Redelings and Suchard, 2005). Although such estimates are always at hand in an MCMC approach and give usable results, they are *ad hoc*, fail to marginalize properly, and are rarely repeatable. One way around these difficulties is to select a representative alignment by maximizing the sum of column probabilities, or "posterior decoding" (Durbin et al., 1998; Lunter et al., 2005). This approach is practical because each column often occurs enough that its posterior probability may be estimated from MCMC samples, although a full alignment does not occur sufficiently often. However, it is not clear how such an alignment should be found if the researcher desires to look beyond the MCMC samples that have been generated in order to find the maximum. Dynamic programming may be used to find a maximum posterior decoding alignment for two or perhaps three sequences, but the speed and memory requirements grow exponentially in the number of sequences in a multiple sequences alignment, making such approaches impractical.

6.2.3 Annotating a representative multiple sequence alignment with column probabilities

Once a representative alignment is found, this alignment must be annotated to indicate the degree of confidence in various regions. For maximum posterior decoding alignments, a natural measure of confidence is the posterior probability of each column, which may be plotted above the chosen alignment. This method clearly indicates when a column is strongly aligned. However, when a column has a low posterior probability, there are a number of possible reasons. A low column probability could result because two sub-groups of characters are weakly aligned to each other, but strongly aligned within groups. Alternatively, a low column probability could result because all characters are weakly aligned to each other. Thus, the use of only one value per column is not capable of capturing the confidence in partial columns. We also note that, if one sequence such as an outgroup is weakly alignment to all other sequences, then both alignment selection by posterior decoding and alignment annotation using column probabilities may be undermined.

6.2.4 Alignment uncertainty (Au) plots

Alignment uncertainty plots offer an alternative annotation method to the use of column probabilities (Redelings and Suchard, 2005). The underlying idea behind Au plots is to annotate each letter separately by shading or coloring it to indicate the posterior probability that it is placed in its correct column. Thus, when annotating a multiple sequence alignment with n sequences, an Au plot may use n values instead of just the one value of column probability. We note that Au plots do not depict exact probabilities. Instead, they approximate the posterior probabilities for the $n(n-1)/2$ pairs of characters in each column using a tree structure with only $2n-3$ branch lengths. Each character is then annotated with the probability that it aligns with a hypothetical character at the root of the tree, and shaded accordingly (Figure 3). This procedure can be easily depict confidence in a partial column even when a few characters in the column are weakly aligned. Thus, when the exact position of a gap within a sequence is ambiguous, adjacent characters in the same sequence may be lightly shaded to indicate ambiguity.

6.2.5 Multidimensional scaling

Multidimensional scaling offers an alternative to Au plots and provides a starting point from which to assess convergence in Bayesian samplers that explore alignment uncertainty. Multidimensional scaling is a statistical technique commonly used for high-dimensional data visualization (Young and Hamer, 1987; Borg and Groenen, 1997; Cox and Cox, 2001). Multidimensional scaling algorithms start with a sample-to-sample distance (or similarity) matrix $\mathbf{D} = \{D_{ij}\}$ and assign a location $\vec{x}_i$ in a low-dimensional, visualizable space to each sample. Optimal assignments proceed via minimizing a stress function, such as the Kruskal-1 function. Hillis et al. (2005) recommend multidimensional scaling to explore phylogeny distributions. Setting $D_{ij} = m(\mathbf{A}^{(i)}, \mathbf{A}^{(j)})$, where $m(\cdot,\cdot)$ is an arbitrary distance metric between two multiple sequence alignment samples $\mathbf{A}^{(i)}$ and $\mathbf{A}^{(j)}$, allows for multidimensional scaling projections of multiple sequence alignment distributions. In low-dimensional spaces, visual comparisons can assess differences between distributions. This is useful for assessing convergence and interactive displays.

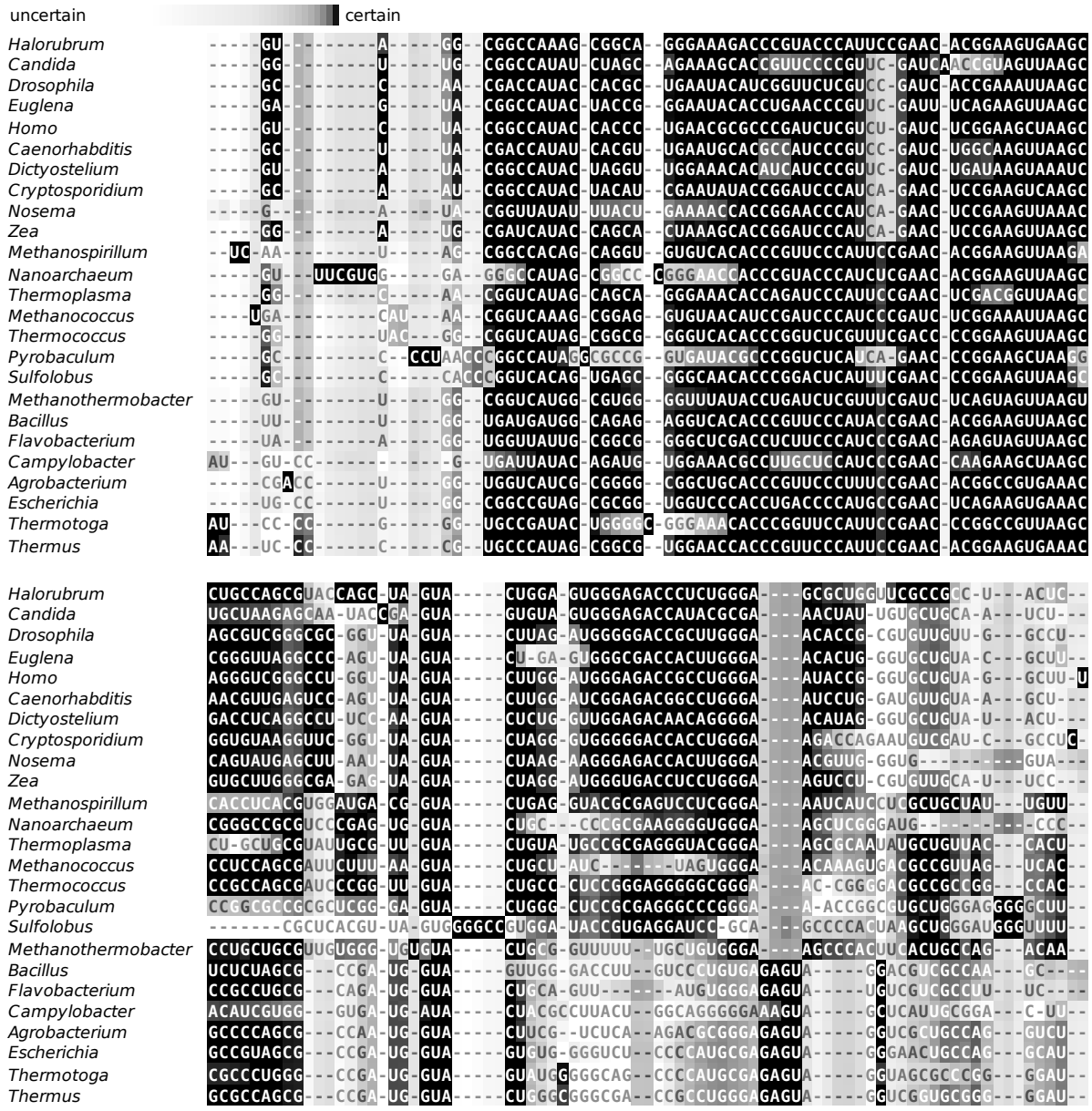


Figure 3: Au plot showing alignment uncertainty for 5S rRNA sequences. Each residue is shaded to indicate the probability that it is correctly aligned with other residues in its column. The probability distribution that is summarized in this figure is the posterior distribution on multiple sequences alignments that is calculated in Section 7.

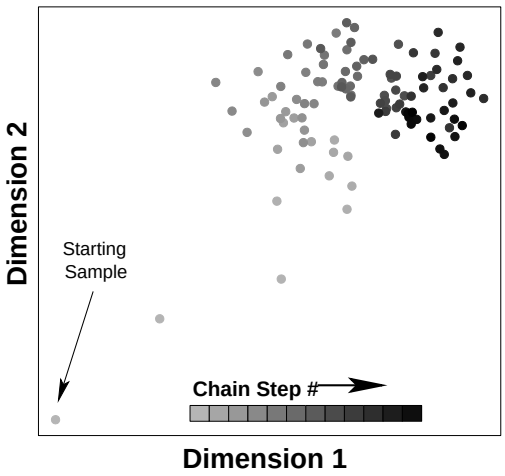


Figure 4: Alignment burn-in of a Bayesian sampler set visualized via multidimensional scaling. The data set consists of 12 taxa that span the tree of life. The data set contains sequence of EF-1α and EF-Tu (its bacterial homolog). Sequences are about lengths are about 450 amino acids in length.

Distance metrics for multiple sequence alignments remain under-developed. Well-studied, however, are metrics on pairwise sequence alignments. Schwartz et al. (2005) provide a biologically interpretable metric for pairwise alignments. The metric counts the number of homology statements on which two pairwise alignments disagree. Schwartz et al. (2005) also suggest how to construct a metric $m(\cdot, \cdot)$ on multiple sequence alignments as the sum of all possible pairwise alignment metrics; while this metric overcounts homology disagreements because the metric ignores the evolutionary correlation between pairs, the metric warrants consideration as the underlying phylogeny is unknown.

Figure 4 demonstrates the use of multidimensional scaling to visualize convergence of alignment samples. The 100 samples depicted in the figure are drawn from a posterior simulation involving 12 taxa that span the Tree of Life. After approximately 80 steps, the alignment samples appear to reach convergence.

7 Example - 5S ribosomal RNA

As an illustration of the above techniques, we examine a data set consisting of 25 5S ribosomal RNA gene sequences that displays substantial alignment uncertainty. The 5S ribosomal RNA is present throughout the Tree of Life except in a few basal eukaryotes such as *Giardia*. We therefore include 5S sequences from organisms spanning the Tree of Life, including 7 Eubacteria, 9 Archaea, and 9 Eukaryotes. The 5S rRNA is relatively short, ranging from 111 nucleotides to 131 nucleotides in this data set. Although we have previously shown that alignment uncertainty is so high that there is little phylogenetic information in a data set of only 5 sequences (Redelings and Suchard, 2005), we seek to show that this data set containing 25 sequences does indeed contain substantial phylogenetic information. However, this data set must be analyzed using joint estimation instead of sequential estimation in order to avoid conclusions with strongly supported phylogenetic errors.

7.1 Model and priors

For the nucleotide substitution process we employ the GTR+gwF+log-Normal$_8$ model. We also consider the simpler GTR+gwF model that assumes no rate heterogeneity between sites. The GTR+gwF model is a reversible continuous-time Markov chain model of nucleotide substitution. Nucleotide frequencies are specified by parameters $\boldsymbol{\pi} = (\pi_A, \pi_T, \pi_G, \pi_C)$. This yields 3 degrees of freedom, because of the constraint that the sum of the frequencies must be 1. The GTR+gwF model also contains parameters $\boldsymbol{\psi} = (\psi_{AT}, \psi_{AG}, \psi_{AC}, \psi_{TG}, \psi_{TC}, \psi_{GC})$ specifying the symmetric exchangeability of each pair of nucleotides independently from their frequencies. Because only the relative rates can be estimated from data, we follow

common practice in scaling the exchangeabilities so that the mean substitution rate is 1. Because of this constraint, these 6 parameters yield only 5 degrees of freedom. Finally, the GTR+gwF model contains an additional parameter $f \in [0, 1]$ that specifies whether common letters occur more frequently because they are more highly conserved (low f) or because they are more often proposed as replacements for other letters (high f). The +gwF formulation (Goldman and Whelan, 2002) may be reduced to the more common +F formulation (Cao et al., 1994) by fixing $f = 1$. We model rate heterogeneity between sites using a log-Normal distribution that is approximated by a discrete distribution with 8 bins. We set the mean of the distribution to 1, and parametrize it by its standard deviation V, so that there is no heterogeneity when $V = 0$. The substitution model is therefore fully characterized by $\mathbf{\Theta} = (\boldsymbol{\pi}, \boldsymbol{\psi}, f, V)$ with 10 degrees of freedom.

For the insertion-deletion process we use the RS07 indel model from Redelings and Suchard (2007). This model contains 2 parameters: the insertion-deletion rate λ, and the gap extension probability ϵ. Therefore, the insertion-deletion model is characterized by $\mathbf{\Lambda} = (\lambda, \epsilon)$.

We place a uniform Dirichlet prior on $\boldsymbol{\pi}$. We place a non-uniform Dirichlet prior on $\boldsymbol{\psi}$ with weight 4 on transversions and weight 8 on transitions (Zwickl and Holder, 2004). We place a Uniform(0,1) prior on f, and we place a Laplace prior on $\log V$ with scale 1 that is centered at -3. This leads to a prior median for V of about 0.05. We place a Laplace prior on $\log \lambda$ that is centered at -5 with scale 1.5 and an Exponential(5) prior on the mean indel length minus one. For the prior on phylogenies, we place a uniform distribution on tree topologies and an Exponential(μ) prior on branch lengths, where μ is a hyper-parameter. We then place an Exponential(1) prior on μ.

7.2 Results

Inference under the model was conducted using the MCMC sampling program `BAli-Phy`. We ran two chains from different starting positions and obtained 400,000 samples from each chain. This analysis took about two weeks to complete on a Pentium 4 processor. We discarded the first 40,000 samples from each chain as burn-in. We additionally analyzed the same data with the alignment held constant to either the `Clustal W` estimate or the `Muscle` estimate. For these latter analyses, we used the same burn-in period and the same number of samples. However, we did not make use of an indel model, thus ignoring the phylogenetic evidence of shared indels and considering only shared substitutions.

We report estimates of model parameters for both models in terms of the posterior median and a 95% Bayesian credible interval. As shows in Table 1, the log indel rate $\ln \lambda$ was estimated as -4.24 with rate heterogeneity, and as -3.41 without. However, we note that because the branch lengths are defined in terms of substitutions, the indel rate here is defined relative to the substitution rate, and so in this case the difference in indel rate seems to indicate only that the substitution rate scale has changed. This is indicated by the fact that the total tree length $|T|$ is estimated as 16.7 with heterogeneity, but 7.57 without. Additionally, the mean branch length μ was estimated as 0.365 with rate heterogeneity and 0.165 without. Thus $\lambda \cdot \mu$ remains roughly constant.

7.3 Bias and Alignment uncertainty

The 5S rRNA data set exhibits alignment uncertainty in two major ways. First, the posterior alignment distribution is diffuse, placing similar support on a large number of distinct alignments. This is illustrated in the Au-plot in figure 3. Because the posterior topology distribution is also diffuse, it is possible that the diffuseness of the posterior alignment distribution results from the diffuseness of the posterior topology distribution. We therefore selected the MAP topology from the joint analysis with heterogeneity and re-ran the analysis with the topology fixed to this MAP topology. However, fixing the topology increases the fraction of residues aligned at the 0.5, 0.75, 0.95, and 0.99 levels only slightly (data not shown). Therefore we conclude that alignment uncertainty in the posterior distribution is not primarily a result of topological uncertainty.

Second, the alignment is ambiguous enough to be biased by the guide tree used during the progressive alignment procedure, so that the use of a single alignment estimate constructed using progressive alignment leads to substantial bias in phylogeny construction. For example, when the alignment is fixed to the `Clustal W` alignment estimate, the posterior probability (PP) that Eubacteria are monophyletic is only 0.283 with a posterior log odds (PLOD) score of -0.405. This is because the Eubacterium taxon *Campylobacter* is

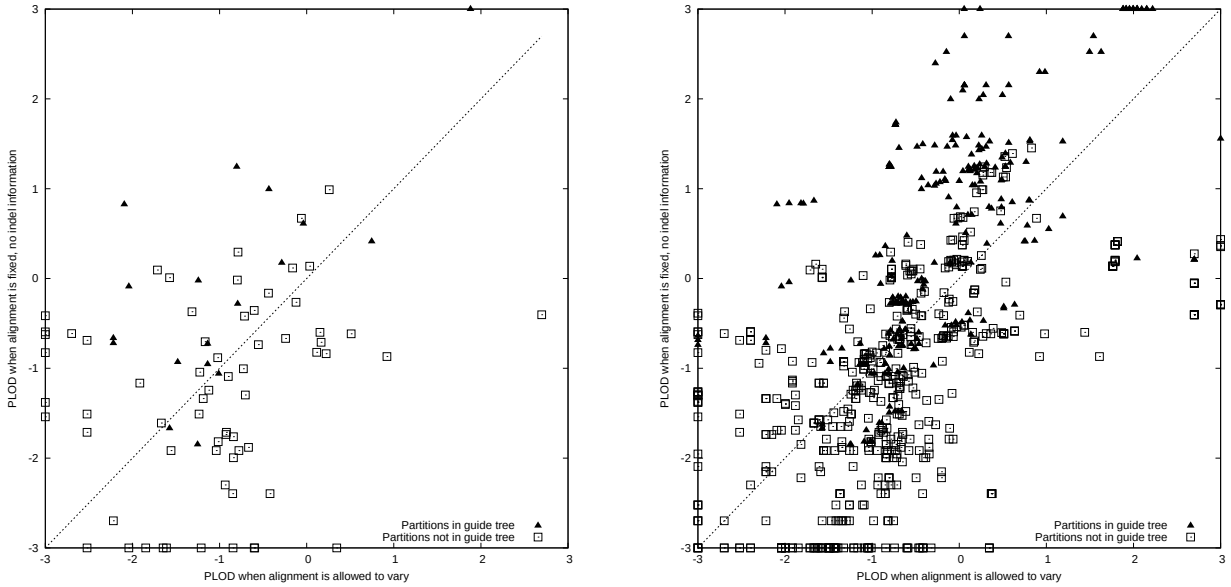
Table 1: Parameter Estimates

Parameter estimates from the joint model with and without rate heterogeneity. The posterior median and a 95% Bayesan credible interval are reported for each parameter. Each row contains parameter estimates under a different model. The first and second row refer to models in which indel information is not used, and the alignment is fixed to the `Clustal W` and `Muscle` alignment estimates respectively. The third and fourth rows refer to models in which alignment is allowed to vary, and substitution rate heterogeneity is absent/present, respectively. Note the differences in alignment length and in the parsimony score when the alignment is allowed to vary. This may indicate that `Clustal W` and `Muscle` align residues too readily, resulting in shorter alignments with more mismatches.

Model	$\log \lambda$	$\log \epsilon$	f
no indel model / Clustal	–	–	$0.382\ (0.0170, 0.956)$
no indel model / Muscle	–	–	$0.412\ (0.0218, 0.962)$
no rate variation	$-3.41\ (-3.77, -3.07)$	$-0.716\ (-1.00, -0.492)$	$0.407\ (0.181, 0.962)$
GTR+gwF+log-Normal$_8$	$-4.24\ (-4.83, -3.74)$	$-0.634\ (-0.881, -0.437)$	$0.250\ (0.00994, 0.892)$

Model	μ	$\lvert T\rvert$	V
no indel model / Clustal	$0.244\ (0.172, 0.359)$	$11.7\ (9.13, 14.4)$	$0.999\ (0.750, 1.37)$
no indel model / Muscle	$0.285\ (0.193, 0.434)$	$12.9\ (10.3, 17.9)$	$1.26\ (0.935, 1.83)$
no rate variation	$0.165\ (0.124, 0.226)$	$7.57\ (6.90, 8.28)$	-
GTR+gwF+log-Normal$_8$	$0.365\ (0.226, 0.650)$	$16.7\ (11.4, 27.8)$	$2.20\ (1.44, 3.83)$

Model	#indels	|indels|	#subst	$\lvert\mathbf{A}\rvert$
no indel model / Clustal	–	–	$745\ (737, 756)$	135 –
no indel model / Muscle	–	–	$755\ (746, 768)$	135 –
no rate variation	$58\ (49, 70)$	$111\ (93, 137)$	$701\ (681, 719)$	$166\ (155, 181)$
GTR+gwF+log-Normal$_8$	$57\ (50, 66)$	$118\ (101, 145)$	$711\ (693, 729)$	$172\ (160, 194)$



(a) Bi-partitions of full leaf taxa

(b) Bi-partition of subsets of leaf taxa

Figure 5: Fixed alignment leads to bias towards guide tree. The data set consists of 5S ribosomal RNA sequences from 25 taxa across the Tree of Life. In both panels (a) and (b) each point represents the support for a bi-partition when near-optimal alignments are considered using joint Bayesian estimation of phylogeny and alignments (x-axis) and when the alignment is fixed and indel information is ignored (y-axis). Support is indicated by the posterior $\log_{10}$ odds (PLOD). Filled triangles represent bi-partitions that are implied by the guide tree used by `Clustal W` to estimate the fixed alignment; open squares represent bi-partitions that contradict this guide tree. Note that triangles tend to fall above the line $y = x$, while open squares tend to fall below it. This indicates that that under the fixed alignment, partitions have increased support if they are in the guide tree, and decreased support otherwise. This illustrates bias towards the guide tree of a progressive alignment estimate when the alignment is fixed. Points in panel (a) represent bi-partitions of the leaf taxa, but points in panel (b) represent bi-partitions of subsets of the leaf taxa. Because the posterior topology distribution contains many wandering taxa, bi-partitions of the full leaf taxon set do not reveal the full amount of information in the distribution.

placed among the Archaea, in accordance with the `Clustal W` guide tree. In contrast, when the alignment is allowed to vary during phylogeny estimation, the posterior support for monophyly of the Eubacteria rises to 0.998, with a posterior LOD score of PLOD=2.74. This indicates that the support for clustering *Campylobacter* with the Archaea is an artifact of the `Clustal W` and `Muscle` guide trees. This observation is further evidence for the view that sequential estimation does not yield robust inferences in the presence of alignment uncertainty.

The bias in phylogeny inference that results from conditioning on an alignment is not limited to the placement of the single taxon *Campylobacter*. In Figure 5 we compare the posterior topology distributions when the alignment is fixed to the `Clustal W` estimate and when it is allowed to vary. Splits which are present in the `Clustal W` guide tree tend to have higher support when the alignment is constructed using the guide tree. Because some taxa may plausibly attach at several places on the tree under both distributions, not many full splits are strongly supported (Figure 5a). The exception is that the monophyly of Bacteria is strongly supported when the alignment is not fixed to the `Clustal W` estimate. However, many *partial* splits do show strong support (Figure 5b), and partial splits that occur in the guide tree tend to be strongly supported only when the fixed alignment is used that was created using the guide tree.

7.4 Rate heterogeneity

Substitution rate heterogeneity was strongly supported by the data set and substantially decreases the amount of posterior alignment uncertainty. To determine whether rate heterogeneity was supported by the data, we estimated the marginal likelihood of the data set under the GTR+gwF and GTR+gwF+logNormal$_8$ models using the stabilized harmonic mean estimator (Newton and Raftery, 1994; Suchard et al., 2003). The marginal likelihood estimates for these two models are -2829.2 ± 0.2 and -2739.6 ± 0.4 on the $\log_e$ scale. Placing equal prior weight on both models yields a log Bayes factor of 89.6 in favor of the model with rate heterogeneity. The strength of evidence here is surprisingly high, given that the longest sequence length is only 131 nucleotides. However, this may be explained by the extreme degree of rate heterogeneity since V is estimated at 1.93 $(1.28, 3.24)$ far from $V = 0$.

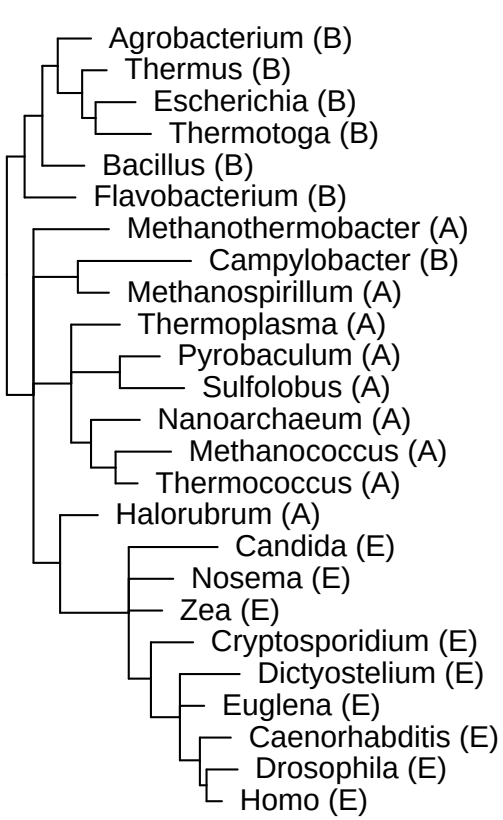
Including rate heterogeneity in the substitution model substantially decreases the degree of alignment ambiguity in the posterior distribution. This is indicated in comparison of the Au plots of the two models (not shown) and in Figure 7. Figure 7 shows that the fraction of aligned residues in each pair of sequences tends to be higher under the GTR+gwF+log-normal$_8$ model than under then GTR+gwF model with no rate variation. This may be explained by the observation that the first half of the 5S rRNA sequence exhibits greater conservation than the second half. By allowing different substitution rates in each site, the GTR+gwF+logNormal$_8$ can improve the likelihood by increasing the observed substitutions in this region. This explanation is with in increased posterior median parsimony score (714 versus 702) under the model with rate heterogeneity. Since Table 1 indicates that the substitution rate does indeed vary across the alignment, this trade-off should be biologically realistic and lead to more accurate alignments and trees.

8 Discussion

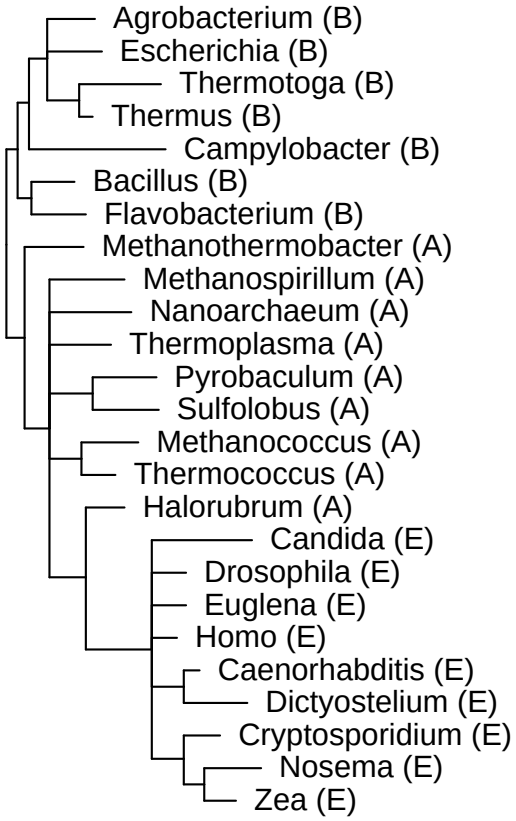
In this chapter we described how alignment ambiguity can undermine bioinformatic inference methods based on sequential estimation, preventing the robust inference of phylogenies and other evolutionary parameters from collections of distantly related sequences. We then described bioinformatic inference methods that make use of homology information in multiple sequence alignments, and yet are robust to alignment uncertainty. In doing so, we emphasized that in order to yield robust inferences and accurate measures of confidence, an inference method must take into account both of the two sources of alignment uncertainty: parameter uncertainty and near-optimal alignments. We also emphasized the importance of jointly estimating the alignment and any other parameters that are mutually dependent on the alignment. When the alignment is ambiguous, joint estimation is necessary in order to prevent circular reasoning. Bayesian inference methods that fulfill these conditions have long been available for pairwise alignments, but have only recently become feasible for multiple alignments. In addition to robust inference, these Bayesian methods make it possible to assess alignment uncertainty in an objective and repeatable fashion, and can safely make use of information in ambiguous regions of the alignment. Bayesian methods naturally asses fine gradations in support for homologies, and also allow characterizing support for homology statements that range in size from an entire alignment to a single pair of residues.

8.1 Speed and Sequential Estimation

Despite the many benefits of the joint estimation approach, consideration of near-optimal alignments during estimation of other parameters comes at the cost of increased run time when the data set contains multiple sequences. For example, when the software `BAli-Phy` is used, considering near-optimal alignments increases the computation time for each iteration by about a factor of 2.2 over a standard Bayesian analysis for the example in Section 7, where the sequence lengths are about 120 letters. For longer sequences of about 500 letters, joint estimation is about 10 times slower than the use of a fixed alignment per iteration. In addition, more iterations are often required when the alignment is allowed to vary. Therefore, the benefits of increased power through the use of indel information and the use of substitution information in ambiguous regions must be balanced against the limited size of data sets that may currently be analyzed in the joint estimation paradigm. Because of the computational cost we predict that sequential estimation based on the use of a single censored alignment will continue to be used by commonly biologists for at least a decade, whether this is advisable or not. Of course, when a rough estimate of the alignment is required for the purpose of

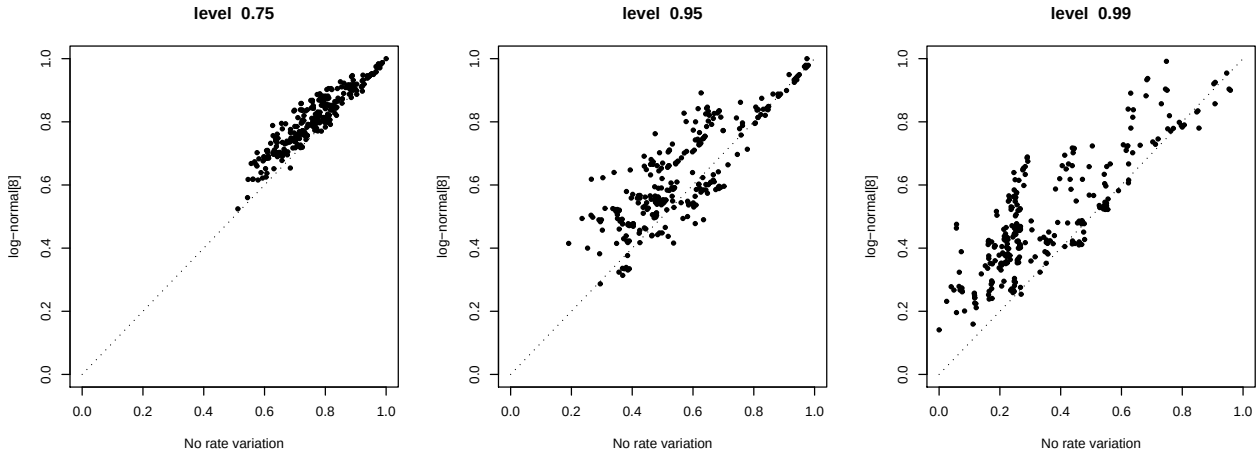


(a) Clustal W alignment



(b) Variable alignment

Figure 6: Majority consensus trees summarize the posterior phylogeny distribution for the 5S rRNA data set described in Section 7. Partitions with posterior probability greater than 0.5 are displayed in each tree, and branch lengths are posterior mean branch lengths conditional on the topology shown. (a) The tree on the left summarizes the posterior distribution when the alignment is fixed to the `Clustal W` alignment estimate. (b) The tree on the right summarizes the posterior distribution under the RS07 model when the alignment is allowed to vary. Note that the taxon *Campylobacter* is in correctly placed among the Archaea when the `Clustal W` is used, in accordance with the `Clustal W` guide tree. Each taxon is labelled B, A, or E, to indicate the Bacteria, Archaea, or Eukaryotes.



(a) Alignment uncertainty with and without rate variation

Figure 7: Modelling site-to-site variation in substitution rate leads to decreased alignment ambiguity. Each point represents a pair of sequences. Both axes represent the fraction of residues in the shorter sequence that are aligned to a gap or specific residue in the other sequence with a posterior probability greater than the cutoff value. The horizontal axis represents the aligned fraction under the GTR+gwF model, and the vertical axis represents the aligned fraction under the GTR+gwF+log-normal$_8$ model. In each plot the aligned fraction is substantially greater under the model with rate heterogeneity; this trend increases as the threshold level of posterior probability increases. Additionally, the separate plots illustrate the fact that as the threshold increases, fewer residues in each pair are aligned with the required posterior probability.

visualization, speed will always be the predominant concern. In these cases progressive alignment and other heuristic techniques may be preferred indefinitely. However, it is still important in such cases to indicate when the alignment should be trusted and when it might be unreliable. It is therefore worth considering how much the speed of joint estimation might be increased, and what improvements in robustness might be made for methods that attempt to handle alignment uncertainty through censoring.

8.1.1 Speeding up joint estimation of alignment and other parameters

Joint Bayesian estimation requires additional computation time compared to traditional Bayesian analysis for two primary reasons: first, because the alignment must continually be resampled and, second, because the Markov chain requires more iterations to give equally precise estimates. The time required for resampling using DP is linear in the number of genes used, but is quadratic in the length of each gene. Therefore, just as in other applications of DP, the cost of resampling the alignment can be decreased by ignoring the corners of the DP matrix or by constraining the alignment to contain specific homologies that are considered certain. The second cause for slowness can be improved by new transition kernels that increase mixing efficiency, but it is difficult to characterize how much mixing efficiency is decreased by allowing the alignment to vary.

A number of factors may combine to make joint Bayesian estimation quicker in several common cases. First, note that if a rough estimate of the alignment without measures of confidence is all that is needed, then the MCMC analysis may require many fewer iterations, since the analysis can be stopped when burn-in is reached. For example, we estimate that the analysis in section 7 would require only about an hour or two to produce an estimate, instead of about two weeks to produce detailed measures of confidence as well. This amount of time may be further decreased by using a starting tree and alignment that are computed from a quicker method. However, one problem with stopping directly after burn-in is complete is that it is difficult to know the burn-in time without running a longer analysis.

Secondly, in many cases the branching order of the taxa in the data set is already known. In such cases, the use of a fixed topology should greatly decrease the number of iterations required for convergence of the Markov chain, as well as decreasing the number of samples that need to be collected after convergence. Furthermore, in many such cases each gene may be analyzed independently in order to infer substitution rates, insertion and deletion rates, the degree of positive selection, or other parameters. In this case, a large

computing cluster should enable whole-genome analysis of coding regions.

8.1.2 Improvements to traditional alignment methods

Several new consistency-based methods for multiple sequence alignment hold out the possibility of improving the accuracy of multiple sequence alignments when evolutionary parameters are not known in advance. For example, the `ProbCons` (Do et al., 2005) approach mitigates the bias induced by the use of a guide tree through iterative refinement and the 3-way consistency transformation, while sequence annealing as implemented in `AMAP (Schwartz and Pachter, 2007)` does not rely on progressive alignment and so does not need a guide tree. While these methods can be slower than `Clustal W`, they are still quite fast.

Many of these methods additionally compute some measure of reliability for homologies in a multiple alignment. For example, `ProbCons` provides a column reliability score based on posterior probabilities, as do several other consistency-based methods. The program `AMAP` takes a different approach, and provides a series of alignments at various levels of specificity. However, this specificity applies only to homologies, and not to gaps. Thus `AMAP` leaves more residues unaligned which increased reliability is required.

8.2 Testing Alignment Reliability Measures

Alignment estimation methods are often ranked by the number of correctly predicted homologous pairs or columns as measured against curated alignment databases such as BAliBase (Thompson et al., 1999). However, even if an alignment method correctly predicts 40% of aligned residues, this degree of accuracy may not be very useful unless the program also indicates which residues in the alignment estimate are correct aligned. In addition to statistical alignment programs, which indicate confidence in alignment regions using posterior probabilities, a number of other programs, such as `T-Coffee` (Notredame et al., 2000), `ProbCons`, and `AMAP` also provide estimates of reliability. Thus, in order to improve or rank these approaches, improved alignment benchmarks that compare sensitivity at each level of specificity are needed.

8.3 Verifying Bioinformatic inferences by simulation

When developing new bioinformatic inference methods, researchers are hampered by the fact that the accuracy of such methods is often difficult to check. This is because phylogenies and many other evolutionary parameters cannot be directly measured. One way around this problem is to simulate a collection of sequences conditional on a specific set of evolutionary parameters so that the true values for the parameters are known (Ogden and Rosenberg, 2006). However, this approach does have a few difficulties. First, it is important that conclusions and generalizations about the accuracy of a method as a whole are carefully based on results for a large collection of different parameter settings. Second, inference methods that are based on a specific evolutionary model often perform well on data sets simulated from that model, but may perform worse on real data if the model is inaccurate. Inaccuracies on real data often stem from the fact that inference under a model that captures all relevant biological phenomena is computationally prohibitive. For example, all current alignment algorithms fail to account for slipped-strand mispairing that causes insertions and deletions of tandem repeats.

However, it is often computationally feasible to simulate under biologically realistic models that contain complex phenomena, even when it is not feasible to use such models for inference. By using data simulated from biologically realistic models, it should be possible to conduct more tests of bioinformatic inference methods whose results can be extrapolated to real data with greater confidence. In addition, it would be possible to ascertain which aspects of the more complex simulation model must be included in the inference model to improve accuracy. Thus, we believe that the development of more complex and realistic models of sequence evolution would well repay any effort that is put in to developing them. Models for simulation would ideally include both substitution and indel rate heterogeneity, and this heterogeneity would have biologically realistic spatial patterns, perhaps based on structures of known proteins.

9 Acknowledgments

The authors wish to thank Jeff Thorne and Eric Stone for helpful comments and stimulating discussions. Ben Redelings was supported by NIH grant ??? and Marc Suchard was supported by ???.

References

Allison, L. and Wallace, C. S. (1994). The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and the optimisation of multiple alignments. *Journal of Molecular Evolution*, 39:418–430.

Bishop, M. and Thompson, E. (1986). Maximum likelihood alignment of dna sequences. *J. Mol. Biol.*, 190(2):159–65.

Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition.

Cao, Y., Adachi, J., Janke, A., Pääbo, S., and Hasegawa, M. (1994). Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *J Mol Evol*, 39(5):519–527.

Cartwright, R. A. (2006). Logarithmic gap costs decrease alignment accuracy. *BMC Bioinformatics*, 7:527.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *J. Mol. Biol. Evol.*, 17:540–552.

Cox, M. F. and Cox, M. A. A. (2001). *Multidimensional Scaling*. Chapman & Hall, New York.

Csuros, M. and Miklós, I. (2005). Statistical alignment of retropseudogenes and their functional paralogs. *Mol Biol Evol*, 22(12):2457–2471.

Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. (2005). Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, 15(2):330–340.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Model of Protein and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Fleissner, R., Metzler, D., and von Haeseler, A. (2005). Simultaneous statistical multiple alignment and phylogeny reconstruction. *Systematic Biology*, 54(4):548–561.

Gatesy, J., DeSalle, R., and Wheeler, W. (1993). Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Molecular Phylogenetics and Evolution*, 2:152–157.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, 36:182–198.

Goldman, N. and Whelan, S. (2002). A novel use of equilibrium frequencies in models of sequence evolution. *Molecular Biology and Evolution*, 19(11):1821–1831.

Hillis, D. M., Heath, T. A., and John, K. S. (2005). Analysis and visualization of tree space. *Syst Biol*, 54(3):471–482.

Holmes, I. (2003). Using guide tree to construct multiple-sequence evolutionary hmms. In *Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology*, pages 147–157, Menlo Park, CA. AAAI Press.

Holmes, I. and Bruno, W. J. (2001). Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, 17(9):802–820.

Lake, J. A. (1991). The order of sequence alignment can bias the selection of tree topology. *Molecular Biology and Evolution*, 8:378–385.

Lunter, G., Miklós, I., Drummond, A., Jensen, J. L., and Hein, J. (2003). *Bayesian Phylogenetic Inference under a Statistical Insertion-Deletion Model*, volume 2812 of *Lecture Notes in Computer Science*, chapter 19, pages 228–244. Springer Berlin / Heidelberg.

Lunter, G., Miklós, I., Drummond, A., Jensen, J. L., and Hein, J. (2005). Bayesian co-estimation of phylogeny and sequence alignment. *BMC Bioinformatics*, 6(83).

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

Miklos, I., Lunter, G. A., and Holmes, I. (2004). A "long indel" model for evolutionary sequence alignment. *Mol. Biol. Evol.*, 21(3):529–40.

Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998). Homstrad: a database of protein structure alignments for homologous families. *Protein Sci*, 7(11):2469–2471.

Morrison, D. and Ellis, J. (1997). Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. *Molecular Biology and Evolution*, 14:428–441.

Naor, D. and Brutlag, D. L. (1994). On near-optimal alignments of biological sequences. *J Comput Biol*, 1(4):349–366.

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, 56:3–48.

Notredame, C., Higgins, D., and Heringa, J. (2000). T-Coffee: a novel method for multiple sequence alignments. *Journal of Molecular Biology*, 32:205–217.

Ogden, T. H. and Rosenberg, M. S. (2006). Multiple sequence alignment accuracy and phylogenetic inference. *Syst Biol*, 55(2):314–328.

Redelings, B. D. and Suchard, M. A. (2005). Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology*, 54(3):401–418.

Redelings, B. D. and Suchard, M. A. (2007). Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evo. Bio.*, 7. 40.

Sankoff, D. (1973). A test for nucleotide sequence homology. *J. Mol. Biol.*, 77:159–164.

Sankoff, D., Cedergen, R. J., and Lapalme, G. (1976). Frequency of insertion-deletion, transversion, and transition in the evolution of 5s ribosomal rna. *J. Mol. Evol.*, 7:133–1499.

Sankoff, D., Morel, C., and Cedergren, R. J. (1973). Evolution of 5s rna and the non-randomness of base replacement. *Nat New Biol*, 245:232–234.

Schwartz, A. S., Myers, E. W., and Pachter, L. (2005). Alignment metric accuracy.

Schwartz, A. S. and Pachter, L. (2007). Multiple alignment by sequence annealing. *Bioinformatics*, 23(2):e24–e29.

Suchard, M., Kitchen, C., Sinsheimer, J., and Weiss, R. (2003). Hierarchical phylogenetic models for analyzing multipartite sequence data. *Systematic Biology*, 52:649–664.

Talavera, G. and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*, 56(4):564–577.

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

Thompson, J. D., Plewniak, F., and Poch, O. (1999). Balibase: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1):87–88.

Thorne, J. L. and Kishino, H. (1992). Freeing phylogenies from artifacts of alignment. *Molecular Biology and Evolution*, 9:1148–1162.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, 33:114–124.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1992). Inching towards reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution*, 34:3–16.

Wheeler, W. C. (1996). Optimization alignment: the end of multiple sequences alignment in phylogenetics? *Cladistics*, 12(1):1–9.

Wheeler, W. C. (2006). Dynamic homology and the likelihood criterion. *Cladistics*, 22(2):157–170.

Wheeler, W. C., Gatesy, J., and DeSalle, R. (1995). Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. *Molecular Phylogenetics and Evolution*, 4(1):1–9.

Yee, C. N. and Allison, L. (1993). Reconstruction of strings past. *Comp. Appl. Biosci.*, 9(1):1–7.

Young, F. W. and Hamer, R. M. (1987). *Multidimensional Scaling: History, Theory and Applications*. Lawrence Erlbaum Associates.

Zwickl, D. J. and Holder, M. T. (2004). Model parameterization, prior distributions, and the general time-reversible model in bayesian phylogenetics. *Systematic Biology*, 53(6):877–888.