BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 5 2009, pages 592–598
doi:10.1093/bioinformatics/btp015

Phylogenetics

A hierarchical model for incomplete alignments in phylogenetic inference

Fuxia Cheng[1,†], Stefanie Hartmann[2,5,†], Mayetri Gupta[3,*], Joseph G. Ibrahim[4] and Todd J. Vision[5]

[1]Department of Mathematics, Illinois State University, Normal, IL, USA, [2]Institute for Biochemistry and Biology, University of Potsdam, Potsdam, Germany, [3]Department of Biostatistics, Boston University, Boston, MA, [4]Department of Biostatistics and [5]Department of Biology, University of North Carolina at Chapel Hill, USA

Received on September 25, 2008; revised and accepted on January 5, 2009

Advance Access publication January 15, 2009

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Full-length DNA and protein sequences that span the entire length of a gene are ideally used for multiple sequence alignments (MSAs) and the subsequent inference of their relationships. Frequently, however, MSAs contain a substantial amount of missing data. For example, expressed sequence tags (ESTs), which are partial sequences of expressed genes, are the predominant source of sequence data for many organisms. The patterns of missing data typical for EST-derived alignments greatly compromise the accuracy of estimated phylogenies.

Results: We present a statistical method for inferring phylogenetic trees from EST-based incomplete MSA data. We propose a class of hierarchical models for modeling pairwise distances between the sequences, and develop a fully Bayesian approach for estimation of the model parameters. Once the distance matrix is estimated, the phylogenetic tree may be constructed by applying neighbor-joining (or any other algorithm of choice). We also show that maximizing the marginal likelihood from the Bayesian approach yields similar results to a profile likelihood estimation. The proposed methods are illustrated using simulated protein families, for which the true phylogeny is known, and one real protein family.

Availability: R code for fitting these models are available from: http://people.bu.edu/gupta/software.htm.

Contact: gupta@bu.edu

Supplementary information: Supplemantary data are available at *Bioinformatics* online.

1 INTRODUCTION

Advances in high-throughput sequencing and computation have enabled phylogenetic analyses on an unprecedented scale (de la Torre *et al*., 2006; Driskell *et al*., 2004; Sanderson and Driskell, 2003). Large-scale phylogenetic study of gene families can clarify organismal relationships (Philippe *et al*., 2005; Rokas *et al*., 2003) or gene evolution and function (Eisen, 1998; Sjolander, 2004). Such analyses are often restricted to genes with available full-length sequences, as partial sampling of a gene family may diminish the accuracy of downstream applications, such as orthology assignment (Storm and Sonnhammer, 2002; Zmasek and Eddy, 2001) and gene-tree reconciliation (Page and Cotton, 2001). Since the vast majority of publicly available sequence data from complex genomes is derived from large-scale partial gene sequencing projects, it is desirable in many applications to sample additional gene family members from the large number of available partial gene sequences. Here, we describe an approach for statistically modeling missing data before inferring phylogenetic trees from incomplete (MSAs), enabling the generation of phylogenies for more datasets than possible by restriction to alignments of full-length sequences.

Incomplete gene sequences derived from high-throughput DNA sequencing of random libraries of expressed genes (expressed sequence tags, or ESTs) are the predominant source of sequence data for many organisms (Rudd, 2003). Another source of partial sequence data is metagenomics, in which fragments from many different organisms in the same environmental sample are sequenced *en masse*, as was done with Sargasso Sea samples (Venter *et al*., 2004). In both cases, the missing data (gaps) for each sequence is spatially contiguous and corresponds to different columns of the MSA in different sequences. Gaps tend to be clustered at the beginning and/or the end of each unigene, and the missing positions often overlap but may not correspond between unigenes. This missing data pattern is different from gappiness in a superalignment (concatenated alignments), where some genes are missing from some taxa. In superalignments, boundaries of the missing data blocks strictly coincide among subsets of the sequences, while in EST-based alignments the gaps are staggered. Many studies have evaluated the effect of incomplete gene sampling when taking the superalignment approach to an incomplete multigene dataset (Philippe *et al*., 2004; Wiens, 2003a, b). It was previously believed that missing data does not pose a serious problem to the accuracy of phylogenetic inference, when sufficient data is present (Wiens, 2006). However, Hartmann and Vision (2008) recently showed that the pattern of missing data on using large amounts of EST data greatly compromises the accuracy of estimated phylogenies, especially if the incomplete alignments are used to infer a phylogeny using Neighbor Joining (NJ) or Maximum Parsimony. Approaches to improve accuracy of the trees obtained from incomplete, EST-based MSAs are thus critical for the

*To whom correspondence should be addressed.

†The author wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

application of techniques that rely upon large numbers of accurate gene trees, as in phylogenomics (Eisen, 1998; Philippe *et al*., 2005).

Gaps may reflect either technical limitations (i.e. the inability to sequence the full length of a gene) or a biological process (i.e. the insertion or deletion of residues from some, but not all, sequences in the alignment). Accordingly, gaps are variably treated in phylogenetics as missing data or as a different class of data that contains phylogenetic information (Young and Healy, 2003). For instance, with Maximum Parsimony, PHYLIP (Felsenstein, 2004) treats gaps as a binary (presence/absence) character by default, while PAUP (Swofford, 2000) treats them as missing data. The same choice is available for standard maximum likelihood and Bayesian methods. However, for all such methods, the appropriate relative weight of indels and sequence substitutions is open to debate. In the present case, where gaps are due to incomplete sequencing, it would clearly be inappropriate to treat gaps as having phylogenetic information.

There have been four main approaches for dealing with missing data in phylogenetics: omit, ignore, impute or model (Anderson, 2001; Diallo *et al*., 2006; Huelsenbeck *et al*., 1996; Kato *et al*., 2003; Kawakita *et al*., 2004; Landry *et al*., 1996; Levasseur *et al*., 2003; Makarenkov and Lapointe, 2004; Philippe *et al*., 2004; Waddell, 2005; Wiens, 2006). Hartmann and Vision (2008) examined the effect of two approaches specifically for EST-based sequence alignments, both of which improved the accuracy of phylogenies computed from incomplete alignments. In the first approach, alignment masking, potentially problematic columns and input sequences were excluded from the dataset. In the second, incomplete alignments were partitioned into subalignments with little missing data, a distance matrix computed for each subalignment, and a phylogenetic tree computed from a combined distance matrix estimated by scaling those from each subalignment. This approach succeeded in including almost all the input sequences. However, scaling factors for the subalignments were not estimated, but computed directly from the simulation parameters. Here, we develop a model-based method for estimating complete pairwise distances from fragmentary sequence alignments where some pairs of taxa have no sites in common, which allows estimation of scaling factors for different regions within the same gene. We devise profile likelihood and Bayesian approaches for efficient model fitting, and apply our method to simulated alignments from the study in Hartmann and Vision (2008) and a real dataset.

2 METHODS

Here, we describe our approach to compute a phylogeny from fragmentary alignment data—subdividing incomplete alignments, henceforth 'SIA'. Briefly, an incomplete MSA is partitioned into subalignments with little missing data, and a distance matrix computed for each subalignment (see Section 2.1). Submatrices of pairwise distances are combined into a single matrix, using linear weights estimated from a hierarchical model (see Section 2.2) and a phylogenetic tree is inferred from the combined distance matrix.

2.1 Incomplete alignments in phylogenetic inference

Figure 1 outlines the alignment subdivision procedure. Pairwise overlap between two sequences is calculated as the ratio of the number of common non-gap characters to the number of non-gap characters in the shorter sequence. An overlap graph is constructed in which each sequence is represented by a node, and undirected edges connect vertices with pairwise

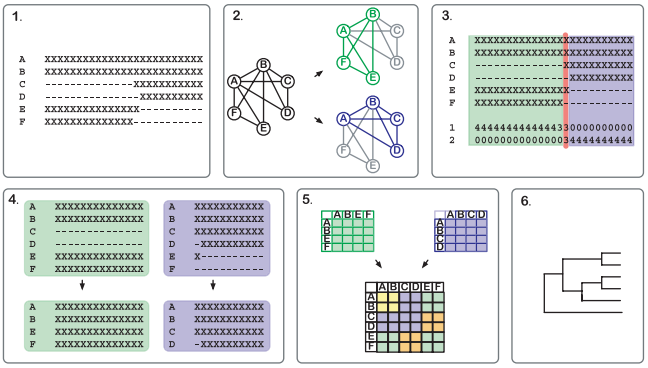

Fig. 1. Overview of the method for an incomplete alignment example of six sequences $(A,\ldots,F)$. 'X' represents any nucleotide (or amino acid), and '-' represents a gap (i.e. missing data): (**1**) input alignment; (**2**) overlap graph; (**3**) assignment of columns to cliques—columns 1–14 are assigned to the green clique, columns 16–25 to the blue clique. Column 15 (red) is tied between the two—it would be assigned to the blue clique; (**4**) concatenated columns (above) and masked subalignments (below); (**5**) combination of submatrices and imputation of missing values. Pairwise distances may have been estimated in only one or the other of the submatrices (green or blue), both (yellow) or neither (orange). The values of the yellow cells are estimated by the hierarchical model, while the values of the orange cells must be imputed; (**6**) the phylogeny inferred by Neighbor Joining.

overlap of at least some value. We used a value of 0.45, i.e. any two sequences have an alignment overlap of at least 45% with respect to the shorter sequence. Other thresholds have not yet been used. Using the Bron–Kerbosch algorithm (Bron and Kerbosch, 1973), maximal cliques of a predetermined size (here, at least 3) are identified in the overlap graph. Maximal cliques are subgraphs in which every node is connected to every other by an edge and cannot be extended further; these represent sets of sequences with sufficient pairwise overlap for computation of all pairwise distances. Each alignment column is assigned to the clique containing the largest number of sequences with non-gap characters in that column. Columns tied between two or more cliques are assigned to the clique with the fewest total columns. Columns assigned to a given clique are concatenated to generate a subalignment, which is then masked to remove sequences that are mainly gaps, using the software REAP (Hartmann and Vision, 2008). Cliques containing at least three columns are retained, and the evolutionary distance between each sequence pair is used to generate a submatrix, with the PROTDIST program within the PHYLIP package (Felsenstein, 2004).

The submatrices are then combined into one matrix by a linear model in which the distances in submatrix k are scaled by a factor that takes into account (i) the relative rate of substitution for the columns in each submatrix relative to the alignment as a whole and (ii) the relative uncertainty in that estimate as a function of the subalignment length. Only sets of sub-matrices having at least two sequences in common with another sub-matrix in the set are used, which are found by constructing a second graph in which the nodes are the submatrices and edges connect two nodes if they share two or more sequences. The largest connected components of this graph constitute the desired sets. If no connected components are found, the submatrix with the largest number of columns is used and the rest discarded. A single matrix of distances is now estimated from the set of submatrices. In combining pairwise distance values for sequences that are present in multiple cliques, each value is scaled based on the number of columns in the subalignment, which affects the error in the estimated pairwise distances, and the overall substitution rate. For submatrix k, a linear coefficient β_k that factors in the subalignment length as well as the substitution rate is estimated by the hierarchical model described in Section 2.2. Pairwise distance values that are not estimated in any individual subalignment are imputed using the procedure

of Landry *et al*. (1996), in which the sixth pairwise distance for a set of four sequences can be inferred provided that the other five pairwise distances are known, under the assumption that the distance matrix is additive. After estimating the β_k's, a phylogenetic tree is reconstructed from the combined distance matrix using Neighbor Joining (NEIGHBOR) within PHYLIP (Felsenstein, 2004).

2.2 A statistical model for incomplete alignments

For each subalignment for a sequence family computed in the previous step, we construct a linear model for the scaling factor in the form of a measurement error-type model, which involves the substitution rate, the subalignment length and the observed pairwise distances. For each set of sequences that are present in the possible combinations of subalignments (e.g., each set of sequences in Fig. 1 that are represented only in subalignment 1, only in 2 or in both 1 and 2), we compute a marginal likelihood function, and then a combined marginal likelihood function. Finally, the substitution rates are estimated using (i) the marginal posterior mean from a Bayesian approach or (ii) a maximum likelihood estimator from the profile likelihood. We next introduce the statistical framework for modeling incomplete sequence alignments, and describe the two estimation methods.

We initially consider a simpler scenario with an alignment of six sequences that was divided into two subalignments. We then extend the model and estimation procedure to the general case of n aligned sequences $(S_1,\ldots,S_n)$, divided into K subalignments. Suppose there are six sequences, denoted by $S_1,\ldots,S_6$. The distance between the i-th sequence, S_i, and the j-th sequence, S_j, is denoted by Y_{ij}, $1\le i,j\le 6$. Because $Y_{ij}=Y_{ji}$ and $Y_{ii}=0$, we need only to consider those Y_{ij}'s with $i<j$. Due to the missing positions of the alignment, the full matrix of the Y_{ij}'s cannot be obtained for doing phylogenetic analysis, and thus with incomplete alignments, the Y_{ij}'s are quantities that are not observed. Further, we assume that for each of the subalignments, a matrix of pairwise distances has been computed.

Let D_{ijk} denote the distance between sequences S_i and S_j in the k-th subalignment ($k=1,2$), and l_k denote the number of columns in the k-th subalignment. D_{ijk} is completely measured as all positions in a subalignment are observed. Define the sets

$$\begin{aligned}
C_1 &= \{(ij)|i<j,\ S_i \text{ and } S_j \text{ are involved only in subalignment 1}\},\\
C_2 &= \{(ij)|i<j,\ S_i \text{ and } S_j \text{ are involved only in subalignment 2}\},\\
C_{12} &= \{(ij)|i<j,\ S_i \text{ and } S_j \text{ are involved only in subalignments 1 and 2}\}.
\end{aligned}$$

Under the assumption that S_1-S_4 are involved in the first subalignment while S_1, S_2, S_5 and S_6 are in the second one, we have $C_1=\{(13),\ (14),\ (23),\ (24),\ (34)\}$, $C_2=\{(15),\ (16),\ (25),\ (26),\ (56)\}$ and $C_{12}=\{(12)\}$. For the general case (with n sequences $S_1,\ldots,S_n$ and K subalignments, $1\le k_1<k_2<\cdots<k_l\le n$, $1\le l\le K$), we have the following general notation for the (ij)-th set: $C_{k_1k_2\cdots k_l}=\{(ij)|i<j,\ S_i \text{ and } S_j$ together are involved only in the k_1-th, k_2-th, $\cdots$, k_{l-1}-th and k_l-th subalignments$\}$. Denote $D_1=C_{12}\cup C_1$ and $D_2=C_{12}\cup C_2$. Now, in order to motivate our model framework, consider first that Y_{ij} being completely unobserved, we need to relate it to the observed distances within subalignments (D_{ijk}) accounting for possible differences in within-subalignment substitution rates (β_k). This requires making a distributional assumption for the Y_{ij}'s. Second, the definition of the substitution rate, a scaling factor required for combining distances in each submatrix to the whole, motivates our choice for the mean function of D_{ijk} as $\beta_k Y_{ij}$. Third, a Gaussian (normal) form for the distribution appears attractive both for reasons of simplicity in a hierarchical model framework as well as being supported by the observed bell-shaped histograms of the pairwise distances in simulated and real data (Supplementary Fig. S1). Finally, the variance of the Gaussian should allow uncertainty in the estimate of the true distance as a function of alignment length. Based on the above, we consider the following hierarchical model for the 'missing' alignment:

$$\begin{aligned}
D_{ijk} &\sim N\left(\beta_k Y_{ij},\ \frac{\sigma_\varepsilon^2}{w_k^2}\right),\\
Y_{ij} &\sim N(\mu,\sigma^2),\quad (ij)\in D_k;\ k=1,2
\end{aligned}\tag{1}$$

where $w_1=\sqrt{l_1}/(\sqrt{l_1}+\sqrt{l_2})$, $w_2=\sqrt{l_2}/(\sqrt{l_1}+\sqrt{l_2})$ and Y_{ij}'s are unobserved latent variables. The known weights w_1 and w_2 are determined from biological considerations, the intuition being that the subalignment with a larger number of columns should have less variation in distance between sequences. The main parameters of interest are (β_1,β_2) with (μ,σ^2) being treated as nuisance parameters. β_1 and β_2 can be interpreted here as the substitution rates for a given subalignment. By splitting sequences into subalignments, the taxa available within each subalignment are informative about relative rates in different parts of the sequences. Thus, the real advantage of our model is that it uses the taxa available in different segments of an alignment to estimate an underlying evolutionary rate for that segment, hence improving distance estimation. For example, the model improves the estimate of Y_{12} by using the different subalignments to estimate relative rates, instead of naively only using information in sequences 1 and 2.

Now we generalize the above. Assume the sequences $S_1,\ldots,S_n$ are in K subalignments. Let D_{ijk} denote the distance between sequences S_i and S_j in the k-th subalignment ($1\le k\le K$), and l_k denote the number of columns in the k-th subalignment. Then we have the K models:

$$D_{ijk}=\beta_k Y_{ij}+\varepsilon_{ijk},\quad (ij)\in D_k,\quad k=1,2,\ldots,K,$$

where, for any k $(1\le k\le K)$, $w_k=\sqrt{l_k}\Big/\sum_{l=1}^K\sqrt{l_l}$, D_k is the union of those sets $C_{k_1\cdots k_l}$ with one subscript k_m being k, the Y_{ij}'s are i.i.d. $N(\mu,\sigma^2)$, ε_{ijk}'s are i.i.d. $N\left(0,\sigma_\varepsilon^2/w_k^2\right)$, Y_{ij} and ε_{ijk} are independent. For each $(k_1\cdots k_l)$ with $1\le k_1<\cdots<k_l\le K$ and $C_{k_1\cdots k_l}\ne\emptyset$, let $N_{k_1\cdots k_l}$ be the cardinality of $C_{k_1\cdots k_l}$. It can be shown that the marginal likelihood function $L_{k_1\cdots k_l}(\beta_{k_1},\beta_{k_2},\ldots,\beta_{k_l},\sigma_\varepsilon^2,\mu,\sigma^2)$ based on the k_1-th, k_2-th,$\cdots$,k_l-th models, can be derived as

$$\begin{aligned}
&L_{k_1\cdots k_l}(\beta_{k_1},\ldots,\beta_{k_l},\sigma_\varepsilon^2,\mu,\sigma^2)\\
&\quad=\left(\frac{2^{-(l+1)/2}\pi^{-l/2}\prod_{t=1}^l w_{k_t}}{\sigma_\varepsilon^l\sigma\left(\sum_{t=1}^l w_{k_t}^2\beta_{k_t}^2/2\sigma_\varepsilon^2+1/2\sigma^2\right)^{1/2}}\right)^{N_{k_1\cdots k_l}}\\
&\qquad\times\exp\left\{-\sum_{(ij)\in C_{k_1\cdots k_l}}\left(\sum_{t=1}^l w_{k_t}^2 d_{ijk_t}^2/2\sigma_\varepsilon^2+\mu^2/2\sigma^2\right)\right\}\\
&\qquad\times\frac{\exp\left\{\sum_{(ij)\in C_{k_1\cdots k_l}}\left(\sum_{t=1}^l\beta_{k_t}w_{k_t}^2 d_{ijk_t}/2\sigma_\varepsilon^2+\mu/2\sigma^2\right)^2\right\}}{\left(\sum_{t=1}^l w_{k_t}^2\beta_{k_t}^2/2\sigma_\varepsilon^2+1/2\sigma^2\right)}.
\end{aligned}$$

The detailed derivations of the likelihood and other formulae are given in the Supplementary Materials. In order to preserve identifiability, we assume that $\sigma_\varepsilon^2=c\sigma^2$, where c is a specified positive scalar. Since the value of c in unknown (and is non-identifiable from the likelihood), we recommend in practice that sensitivity analyses be conducted for several values of c in order to ensure the robustness of the estimates of $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_K)$.

2.2.1 Fully Bayesian inference. The Bayesian approach for inference in this model, based on Markov chain Monte Carlo (MCMC) sampling is well-suited for large K. To complete the model, we specify priors for μ and σ^2, integrate them out from the full posterior distribution and do MCMC sampling from the induced marginal posterior to obtain estimates of $\boldsymbol{\beta}$. We use standard conjugate priors for the parameters, $p(\beta_1,\beta_2,\ldots,\beta_K)\propto 1$, $\mu\sim N(\mu_0,\tau_0\sigma^2)$, and $\sigma^2\sim$ Inverse-Gamma(δ_0,γ_0), i.e. $p(\sigma^2)\propto(\sigma^2)^{-\delta_0-1}e^{-\gamma_0/\sigma^2}$. [As previously, the symbol $N()$ refers to the Normal distribution, μ and σ^2 are the mean and variance parameters that are

also used in the model specification in Equation (1), and $\delta_0, \gamma_0, \mu_0$ and τ_0 are a set of prior 'hyper-parameters' that are discussed more in Section 2.2.2.] Let us denote the set of all distances as $D=((d_{ijk}))$. Then, the full posterior distribution of all parameters $(\beta_1, \beta_2, \ldots, \beta_K, \mu, \sigma^2)$ can be written as follows:

$$p(\beta_1, \beta_2, \ldots, \beta_K, \mu, \sigma^2|D)$$
$$\propto \prod_{C_{k_1 \cdots k_l} \neq \emptyset} L_{k_1 \cdots k_l}(k_1, \ldots, k_l, \sigma_\varepsilon^2, \mu, \sigma^2) p(\beta_1, \beta_2, \ldots, \beta_K) p(\mu|\sigma^2) p(\sigma^2).$$

MCMC sampling is carried out from the marginalized posterior distribution $p(\beta_1, \beta_2, \ldots, \beta_K|D)$ after integrating out the nuisance parameters μ and σ^2 (details in Supplementary Material). We use the Adaptive Rejection Metropolis Sampling (ARMS) procedure (Gilks *et al*., 1995) implemented in the `HI` package in the statistical software R. MCMC-based Bayesian inference not only yields estimates of posterior means of the parameters, but also posterior SDs, quantile estimates and kernel density estimates. An alternative to MCMC-based inference is to obtain posterior modal estimates of the $\boldsymbol{\beta}$ by maximizing the posterior distribution $p(\beta_1, \beta_2, \ldots, \beta_K|D)$. Since this posterior does not allow direct analytical solution to find the modal estimate of $\boldsymbol{\beta}$, numerical optimization tools must be used. It turns out that maximizing $p(\beta_1, \beta_2, \ldots, \beta_K|D)$ yields estimates that are nearly identical to the profile likelihood approach, discussed in Section 2.2.3.

2.2.2 Elicitation of hyperparameters. For the parameter $\boldsymbol{\beta}$, it suffices to take a uniform improper prior, mimicking likelihood-based procedures. Based on the context, we need to carefully elicit the hyperparameters μ_0, τ_0, δ_0 and γ_0. μ_0 is interpreted as the prior average pairwise distance—one possible choice is to take μ_0 to be the median (or mean) of all pairwise distances based on the fully observed data. For τ_0, we can specify it as $1/\tau_0 = (1-f)\alpha_0$, where f corresponds to the missing data fraction ($0 \leq f \leq 1$), and α_0 is a scalar multiple. Thus, a non-informative prior for μ would correspond to small values of α_0, such as $\alpha_0 \leq 10^{-4}$. Note that as $f \to 0$, implying no missing data, then α_0 is just the prior precision for μ, whereas if $f \to 1$, this implies that the prior precision goes to 0, leading to a uniform improper prior for μ. These strategies for specifying (μ_0, τ_0) are attractive, semiautomatic, and have worked well in practice. For (δ_0, γ_0), our experience shows that values of $(\delta_0, \gamma_0) = (10^{-3}, 10^{-3})$ lead to non-informative priors and inference for $\boldsymbol{\beta}$ is robust with respect to these hyperparameters as long as (δ_0, γ_0) are small.

2.2.3 Marginal profile likelihood estimation. We define the marginal likelihood function of $(\beta_1, \ldots, \beta_K, \sigma_\varepsilon^2, \mu, \sigma^2)$ as

$$L(\beta_1, \beta_2, \ldots, \beta_K, \sigma_\varepsilon^2, \mu, \sigma^2) = \prod_{C_{k_1 \cdots k_l} \neq \emptyset} L_{k_1 \cdots k_l}(k_1, \ldots, k_l, \sigma_\varepsilon^2, \mu, \sigma^2). \quad (2)$$

To estimate $\beta_1, \ldots, \beta_K$ using the profile likelihood approach, we first estimate (μ, σ^2) as a function of $(\beta_1, \ldots, \beta_K)$, then substitute these estimators back into (2) and then maximize (2) with respect to $(\beta_1, \ldots, \beta_K)$. We have implemented this using an available routine for a constrained quasi-Newton algorithm (Byrd *et al*., 1995) in the statistical package R (R Development Core Team, 2004) for arbitrary values of the number of subalignments K.

2.2.4 Comparison of inference methods. Full Bayes based on MCMC is generally a superior method since it deals with the nuisance parameters more naturally; also, MCMC samples can be used to construct various summary measures of location and dispersion. Our experience shows that full Bayes is not more computationally demanding than maximizing the posterior or profile likelihood. For profile likelihood, the modal estimate of $\boldsymbol{\beta}$ does not have a closed form and thus one must use numerical optimization tools, such as a quasi-Newton algorithm. If K is small, such optimization techniques work well; however, with large K (i.e. >6), a combination of optimization/sampling tools may be necessary to ensure computational stability. With small K, in simulation studies and applications both methods produce very similar results; full Bayes is computationally superior when K is large. By using weakly informative priors, we essentially recover the exact same estimates for $\boldsymbol{\beta}$ as in profile likelihood (in cases where the profile method finds the correct estimate). The only benefit of the profile method would be when sufficiently non-informative priors are difficult to specify for (μ, σ^2) due to weak identifiability.

2.2.5 Estimating the pairwise distances from $\boldsymbol{\beta}$. The most rigorous way to estimate Y_{ij} is to derive the marginal distribution of Y_{ij} from $p(\beta_1, \beta_2, \ldots, \beta_K|D)$ (details in Supplementary Material) and use the mean of this distribution. Although such an approach may seem more formally attractive, the marginal distribution of Y_{ij} is intractable, not having a closed form. Here, instead, we propose a useful and simple plug-in estimator for Y_{ij}, based on the commonly used method of moments as

$$\hat{Y}_{ij} = \frac{1}{K} \sum_{k=1}^{K} \frac{D_{ijk}}{\hat{\beta}_k} \quad (3)$$

where $\hat{\beta}_k$ is the estimate of β_k from the inference methods discussed in Section 2.2. This plug-in estimator is effective and easily programmable.

3 IMPLEMENTATION

3.1 Simulation study with EST alignments

We applied our methods in a simulation study with incomplete protein alignments. The software Rose (Stoye *et al*., 1998) was used to simulate aligned protein sequence families, in which each sequence was derived from a common ancestor along a defined evolutionary tree (the *true tree*). To simulate missing data patterns comparable to those with EST unigenes, gap patterns were chosen from MSAs in the Phytome plant comparative genomics database (Hartmann *et al*., 2006). A total of 5400 simulated alignments were used in this study—differing with respect to sequence length, number of sequences, substitution rate, tree topology and amount of missing data. Hartmann and Vision (2008) evaluated in detail the effect of missing data in alignments on phylogenetic accuracy and compared the SIA approach to alignment masking. $\boldsymbol{\beta}$-values for the SIA approach were, however, not estimated from the data but directly computed from the simulation parameters. In the present study, three alignments were used from the larger set of simulated families from Hartmann and Vision (2008), and $\boldsymbol{\beta}$-values were estimated from the data as described in Section 2.2. The selected families each contain 60 sequences, were 200 or 500 amino acids long, and based on a different gap pattern.

For each simulated alignment, we computed three different phylogenetic trees. To measure the accuracy of estimated trees, we used quartet distance (QD) (Estabrook, 1992), implemented in the software QDist (Christiansen *et al*., 2006). QD measures the number of quartets (sets of four sequences), that differ in topology (placement of the internal branch) between an estimated tree and the true tree. To remove the effect of the number of quartets on QD, we calculated standardized QDs (SQDs), dividing QD by the total number of possible quartets $\mathrm{SQD} = \mathrm{QD}/\binom{n}{4}$, where n is the number of sequences in common between the two phylogenies. SQD ranges from 0 (no topological disagreement with the true tree) to 1 (no quartets correctly inferred). The first set of trees was computed directly from incomplete alignments. Phylogenetic accuracy appears compromised when missing data is ignored, especially for alignments with many gaps or where it is not possible to compute all pairwise distances due a lack of sequence overlap (i.e. families B and C in Table 1). This agrees with analyses of the larger dataset of simulated alignments (Hartmann and Vision, 2008),

Table 1. Accuracy of estimated phylogenies for three simulated protein families

Family	k	Bayes	Profile	Computed	Incomplete	Matrix	Alignment
A	3	0.134	0.122	0.011	0.058	0.0	0.21
B	6	0.355	0.377	0.413	0.546	0.153	0.55
C	9	0.395	0.403	0.426	0.546	0.198	0.58

Shown are values of SQD relative to the true tree. Computed: β-values computed from simulation parameters directly; incomplete: distance matrix without pretreatment; matrix: proportion of sequence pairs having no overlap in the alignment; alignment: proportion of gap characters in the alignment.

Table 2. Estimated values of β for three simulated alignments using the Bayes (first row in group) and profile likelihood (second row) methods: Family A (three subalignments), Family B (six subalignments), Family C (nine subalignments)

$\hat{\beta}_1, \ldots, \hat{\beta}_K$	MAB	SIM
0.851, 0.820, 0.821	0.169	0.0003
0.732, 0.741, 0.716	0.270	0.0002
0.521, 3.225, 0.265, 1.024, 0.509, 0.435	0.753	1.256
1.409, 1.404, 1.525, 1.323, 1.484, 1.504	0.442	0.0058
0.93, 0.95, 0.68, 1.73, 0.96, 0.98, 1.13, 2.21, 0.94	0.29	0.234
1.90, 1.91, 2.22, 1.43, 1.96, 1.65, 1.80, 1.94, 2.70	0.95	0.127

The true value of β is 1 in all cases. MAB: mean absolute bias; SIM: similarity measure for β, defined as the variance of β_i over subalignments.

emphasizing that the particular gap pattern found in alignments of partial gene sequences needs to be considered carefully. The second set of trees was computed from subdivided alignments in which the true β were computed directly from the simulation parameters (Hartmann and Vision, 2008), and the third set of trees were computed from distance matrices in which the β were estimated using the hierarchical model (see Section 2.2). Substitution rates were estimated using the Bayesian and profile approaches, with c ranging between 0 and 1. For each family, SQDs were essentially identical for all values of c between 0.01 and 0.07 (the range of c giving the most robust and accurate estimates), and branch lengths differed negligibly (data not shown)—so here, we only report results for $c = 0.03$. Computed (true) β-values are all 1, and the estimated βs are shown in Table 2. The βs were used to compute combined matrices and phylogenies, and the corresponding SQDs (Table 1).

For the Bayesian approach, the results in Table 2 are reported for hyperparameter settings $\mu_0 = 0$, $\tau_0 = 100$, $\delta_0 = \gamma_0 = 0.1$, which are slightly informative priors. We conducted sensitivity studies for hyperparameters τ_0 in the range from 10 to 1000, and δ_0, γ_0 from 0.001 to 10. The β estimates varied slightly, but the SQD of computed trees were identical, so we used the set of values stated above also for the application to the protein family (see Section 3.2). The MCMC was found to converge well within 2000 iterations. Figure 2 shows that posterior estimates of β cluster around the true value (i.e. 1), with the performance deteriorating as missing data increases, but not affected much by increasing numbers of subalignments.

Families B and C had the most subalignments (6 and 9) and the largest percentage of non-overlapping sequences (15.3% and 19.8%). Phylogenetic accuracy is improved for

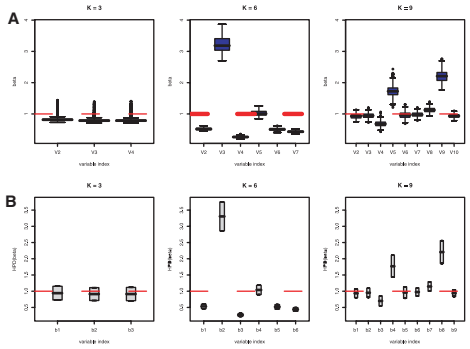

Fig. 2. (**A**) Boxplots representing the posterior distribution of β. (**B**) The 95% highest posterior density intervals for β. The estimates get closer to the truth as the fraction of missing data decreases.

these two families for both new approaches compared with trees computed directly from the incomplete alignments. The Bayesian approach performs slightly better than the profile approach when the number of subalignments and percentage of missing data is larger. For family A, which has three subalignments, the most accurate phylogeny was computed directly from the incomplete alignment. These results suggest that our method improves phylogenetic accuracy for sequence families with large amounts of missing data and/or larger numbers of subalignments, but may not be useful in cases with less missing data. This is consistent with the findings from a larger study using β_k^c-values only (Hartmann and Vision, 2008). Pairwise evolutionary distances from which these estimates are made are reflections of a fixed underlying tree structure. Aberrant estimates of pairwise distances are not uncommon, as is seen in the distribution of pairwise distances within each subalignment (Supplementary Fig. S1). To test robustness in the presence of outliers, we reran the analysis excluding pairwise distances greater than 5 (values as high tend to be inaccurate). The results did not change much, so we discontinued this line of analysis. A more robust approach may be implemented through the use of heavy-tailed error distributions, such as a Student's t. We repeated the simulation study for scenarios where the true βs are not 1; the results were almost identical. The profile likelihood method was observed in general to take slightly less CPU time. In simulated data with $K = 3$ cliques, the Bayes method required 335 s in comparison to 309 s for the profile method.

3.2 Application to a real protein family

We applied both methods on a serine hydroxymethyltransferase protein family from Panther, an online database of protein families (http://www.pantherdb.org; family ID PTHR11680). To obtain the full alignment, we downloaded 58 of the 67 training sequences that were used to generate its profile hidden Markov model (HMM) from GenBank (Benson *et al.*, 2006); 9 of the 67 sequences could not be found in GenBank. From the 58 sequences we excluded 6 for being too short. We aligned the remaining 52 sequences to the profile HMM from Panther using the software hmmalign (http://hmmer.janelia.org/). Low-confidence regions at the beginning (228 positions) and at the end (24 positions) were manually removed, and the final alignment of 581 amino acids was used for analysis and considered the 'complete alignment'.

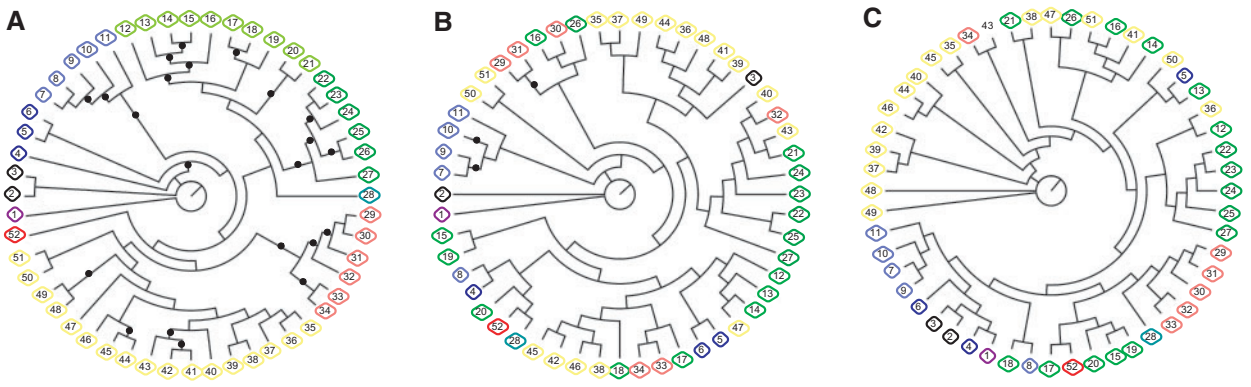

Fig. 3. Unrooted NJ phylogenies estimated for an alignment of a real protein family (serine hydroxymethyltransferase sequences). (**A**) The phylogeny computed from the complete alignment without missing data. (**B**) The tree computed from the incomplete alignment without pretreatment. (**C**) The phylogeny computed from the subdivided alignment using the Bayesian method. Sequences sharing recent common ancestry in (A) are color-coded identically in all trees for easy comparison of major differences in tree topology. For each of the trees shown in (A, B), 100 bootstrap datasets were analyzed. Nodes with support $>95\%$ are marked with a black circle. Tree bootstrapping cannot be done for the tree in (C), where the 'EST-like' alignment was pretreated with SIA.

The substitution rates and true tree for these data are unknown. To simulate missingness, we applied to this complete alignment an EST-like gap pattern from Phytome that consisted of 46% gaps with missing data concentrated at one of the two ends of most of the sequences (Hartmann and Vision, 2008). We calculated pairwise distances from the complete alignment of 52 sequences and 581 columns as well as from the alignment with 46% missing data, using Protdist with the Jones-Taylor-Thornton model (JTT) substitution matrix (Jones *et al*., 1992) and applied the Neighbor Joining algorithm to the matrix of pairwise distances. In addition, we used SIA to subdivide the incomplete alignment into six subalignments and estimated β-values for the corresponding submatrices, which were used to compute a combined distance matrix and a phylogeny. The two approaches resulted in trees with 20–21% fewer topologically incongruent quartets than the NJ tree in which the distance matrix is computed without pretreatment. The greatest improvement was observed for the Bayes method (SQD of 0.204 versus 0.275 relative to the complete alignment tree, and 0.213 for profile likelihood). The distribution of pairwise distances within each subalignment is shown in Supplementary Figure S1; some cases show considerable non-normality. As in the simulations, SQD was unaffected by the value of c (data not shown). Estimated phylogenetic trees from the different methods are shown in Figure 3. Some within-tree relationships are very robust and can be seen in all three topologies (e.g. sequences 7–11, 22–27). Other groups of related sequences that are observed in the complete-alignment phylogeny (e.g. 2–3, 29–34) are almost completely broken up in the NJ tree computed from the incomplete alignment, but recovered with alignment subdivision.

4 DISCUSSION

ESTs and other partial gene sequences are the predominant source of sequence data for a large and taxonomically diverse set of species. These sequences are valuable for gene discovery, genome annotation, comparative genomics or marker development (Bouck and Vision, 2007; Rudd, 2003). However, for studies of gene family evolution or for large-scale analyses of gene families, one must contend with large amounts of missing data in alignments derived from partial sequences. Of the ≈27 000 families in the Phytome database (Hartmann *et al*., 2006) for which there are three or more sequences, the average proportion of alignment gaps is 37%. It was recently shown that the pattern of gappiness in MSAs derived from partial gene sequences substantially compromises phylogenetic accuracy, even in the absence of alignment error (Hartmann and Vision, 2008), and beyond what is expected based on the amount of missing data. This is particularly dramatic for Neighbor Joining and Maximum Parsimony, demonstrating that partial gene sequences and gappy MSAs can pose a major problem for phylogenetic analysis. Different approaches, however, can improve the accuracy of trees obtained from a gappy MSA. Approaches include removing potentially problematic columns and input sequences from the dataset, as well as a combination of modeling and imputing missing data. Here, we describe an approach to statistically model missing data in order to retain as many sequences as possible for phylogenetic analysis. Our two methods developed for fitting this model, a profile likelihood and a Bayesian method, improve phylogenetic accuracy, and are highly comparable in performance, with the Bayes method slightly outperforming when there are large numbers of subalignments. Both outperform approaches where a phylogeny is computed directly from an incomplete alignment ignoring the missing data.

The choice of c is important in our model; c is taken to be fixed for model identifiability. Future work may include specifying a prior (such as a gamma) on c, examining the sensitivity of the inference to hyperparameter choice. One possible shortcoming is that though distances are non-negative, the model assumes normality on the real line. In cases where the histograms of distances are concentrated near zero, using a model with a positive support may lead to more accurate inference. Additional extensions are (i) to consider general error distributions and (ii) to consider models that transform the Y_{ij}'s leading to better approximations to normality. Other approaches for combining subalignments can also be tested. Our current implementation imputes all pairwise distances that

cannot be computed from the submatrices using a four-point metric (Landry *et al*., 1996; Lapointe *et al*., 1999). Implementations could be improved by incorporating a three-point metric or a weighted least-squares imputation (De Soete, 1984; Landry *et al*., 1996; Makarenkov and Lapointe, 2004). Approaches that model the missing alignment data probabilistically or by imputation would allow more accurate likelihood or Bayesian phylogenetic techniques to be applied while retaining all the input sequences. Another approach would be to infer phylogenies separately for each subalignment and then calculate a supertree for the full dataset (Bininda-Emonds, 2004).

In conclusion, our model-based approach shows potential for improving the accuracy of trees obtained from gappy alignments. In modeling missing alignments, we estimated substitution rates for different regions of the same gene. The use of different parameters for different regions within an alignment has also been addressed in the context of different genes (Huelsenbeck *et al*., 1996; Seo *et al*., 2005), and other approaches for combining data from different partitions of a phylogenetic dataset have recently been developed (Bevan et al., 2007; Criscuolo *et al*., 2006). Additional studies can evaluate how combining submatrices and modeling can be optimally implemented. Comparing the performance of our method with others dealing with incomplete alignments (Diallo *et al*., 2006; Makarenkov and Lapointe, 2004) will be critical for the application of techniques relying upon large numbers of accurate gene trees, as in phylogenomics (Eisen, 1998; Philippe *et al*., 2005).

ACKNOWLEDGEMENTS

The authors would like to thank Jack Snoeyink and Craig Falls for assistance with the SIA method.

Funding: National Institutes of Health (GM070335 to J.G.I. and M.G.); National Science Foundation (DBI-0227314 to T.J.V.).

Conflict of Interest: None declared.

REFERENCES

Anderson,J. (2001) The phylogenetic trunk: maximal inclusion of taxa with missing data in an analysis of the lepospondyli. *Syst. Biol.*, **50**, 170–193.

Benson,D. *et al*. (2006) Genbank. *Nucleic Acids Res.*, **34**, D16–D20.

Bevan,R. *et al*. (2007) Accounting for gene rate heterogeneity in phylogenetic inference. *Syst. Biol.*, **56**, 194–205.

Bininda-Emonds,O.R. (2004) The evolution of supertrees. *Trends Ecol. Evol.*, **19**, 315–322.

Bouck,A. and Vision,T.J. (2007) The molecular ecologist's guide to expressed sequence tags. *Mol. Ecol.*, **16**, 907–924.

Bron,C. and Kerbosch,J. (1973) Algorithm 457; finding all cliques of an undirected graph [h]. *Commun. ACM*, **16**, 575–577.

Byrd,R.H. *et al*. (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, **16**, 1190–1208.

Christiansen,C. *et al*. (2006) Fast calculation of the quartet distance between trees of arbitrary degrees. *Algorithms Mol. Biol.*, **1**, 1–16.

Criscuolo,A. *et al*. (2006) SDM: a fast distance-based approach for (super)tree building in phylogenomics. *Syst. Biol.*, **55**, 740–755.

de la Torre,J. *et al*. (2006) ESTimating plant phylogeny: lessons from partitioning. *BMC Evol. Biol.*, **6**.

De Soete,G. (1984) Ultrametric tree representations of incomplete dissimilarity data. *J. Classif.*, **1**, 235–242.

Diallo,A.B. *et al*. (2006) A new effective method for estimating missing values in the sequence data prior to phylogenetic analysis. *Evol. Bioinformatics*, **2**, 127–135.

Driskell,A. *et al*. (2004) Prospects for building the tree of life from large sequence databases. *Science*, **306**, 1172–1174.

Eisen,J. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.

Estabrook,G.F. (1992) Evaluating undirected positional congruence of individual taxa between two estimates of the phylogenetic tree for a group of taxa. *Syst. Biol.*, **41**, 172–177.

Felsenstein,J. (2004) Phylip (phylogeny inference package). Department of Genome Sciences, University of Washington, Seattle.

Gilks,W.R. *et al*. (1995) Adaptive rejection Metropolis sampling. *Appl. Stat.*, **44**, 455–472.

Jones,D.T. *et al*. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Hartmann,S. and Vision,T.J. (2008) Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evol. Biol.*, **8**, 95.

Hartmann,S. *et al*. (2006) Phytome: a platform for plant comparative genomics. *Nucleic Acids Res.*, **34**, D724–D730.

Huelsenbeck,J.P. *et al*. (1996) Combining data in phylogenetic analysis. *Trends Ecol. Evol.*, **11**, 152–157.

Kato,M. *et al.* (2003) An obligate pollination mutualism and reciprocal diversification in the tree genus glochidion (euphorbiaceae). *Proc. Natl Acad. Sci. USA*, **100**, 5264–5267.

Kawakita,A. *et al.* (2004) Cospeciation analysis of an obligate pollination mutualism: have glochidion trees (euphorbiaceae) and pollinating epicephala moths (gracillariidae) diversified in parallel? *Evolution*, **58**, 201–2214.

Landry,P. *et al.* (1996) Estimating phylogenies from lacunose distance matrices: additive is superior to ultrametric estimation. *Mol. Biol. Evol.*, **13**, 818–823.

Lapointe,F. *et al.* (1999) Total evidence, consensus, and bat phylogeny: a distance-based approach. *Mol. Phylogenet. Evol.*, **11**, 55–66.

Levasseur,C. *et al.* (2003) Incomplete distance matrices, supertrees and bat phylogeny. *Mol. Phylogenet. Evol.*, **27**, 239–246.

Makarenkov,V. and Lapointe,F. (2004) A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics*, **20**, 2113–2121.

Page,R.D.M. and Cotton,J.A. (2001) Vertebrate phylogenomics: reconciled trees and gene duplications. In *Proceedings of the Pacific Symposiun on Biocomputing*. World Scientific Publishing, Singapore, pp. 525–536.

Philippe,H. *et al*. (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol. Biol. Evol.*, **21**, 1740–1752.

Philippe,H. *et al.* (2005) Phylogenomics. *Annu. Rev. Ecol. Syst.*, **36**, 541–562.

R Development Core Team (2004) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rokas,A. *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–804.

Rudd,S. (2003) Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci.*, **8**, 321–329.

Sanderson,M.J. and Driskell,A.C. (2003) The challenge of constructing large phylogenies. *Trends Plant Sci.*, **8**, 374–379.

Seo,T. *et al.* (2005) Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. *Proc. Natl Acad. Sci. USA*, **102**, 4436–4441.

Sjolander,K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **220**, 170–179.

Storm,C.E.V. and Sonnhammer,E.L.L. (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, **18**, 92–99.

Stoye,J. *et al.* (1998) Rose: generating sequence families. *Bioinformatics*, **14**, 157–163.

Swofford,D.L. (2000) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, MA.

Venter,J. *et al.* (2004) Environmental genome shotgun sequencing of the sargasso sea. *Science*, **304**, 66–74.

Waddell,P. (2005) Measuring the fit of sequence data to phylogenetic model: allowing for missing data. *Mol. Biol. Evol.*, **22**, 395–401.

Wiens,J.J. (2003a) Incomplete taxa, incomplete characters, and phylogenetic accuracy: is there a missing data problem? *J. Vertebr. Paleontol.*, **23**, 297–310.

Wiens,J.J. (2003b) Missing data, incomplete taxa, and phylogenetic accuracy. *Syst. Biol.*, **52**, 528–538.

Wiens,J. (2006) Missing data and the design of phylogenetic analyses. *J. Biomed. Inform.*, **39**, 34–42.

Young,N.D. and Healy,J. (2003) GapCoder automates the use of indel characters in phylogenetic analysis. *BMC Bioinformatics*, **4**, 6.

Zmasek,C. and Eddy,S. (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, **17**, 821–828.