Simultaneous Bayesian Estimation of Alignment and Phylogeny under a Joint Model of Protein Sequence and Structure

Joseph L. Herman,*[†,1,2] Christopher J. Challis,[†,3] Ádám Novák,[1] Jotun Hein,[1] and Scott C. Schmidler[3,4]

[1]Department of Statistics, University of Oxford, Oxford, United Kingdom

[2]Division of Mathematical Biology, National Institute of Medical Research, London, United Kingdom

[3]Department of Statistical Science, Duke University

[4]Department of Computer Science, Duke University

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: herman@stats.ox.ac.uk.

Associate editor: Rasmus Nielsen

Abstract

For sequences that are highly divergent, there is often insufficient information to infer accurate alignments, and phylogenetic uncertainty may be high. One way to address this issue is to make use of protein structural information, since structures generally diverge more slowly than sequences. In this work, we extend a recently developed stochastic model of pairwise structural evolution to multiple structures on a tree, analytically integrating over ancestral structures to permit efficient likelihood computations under the resulting joint sequence–structure model. We observe that the inclusion of structural information significantly reduces alignment and topology uncertainty, and reduces the number of topology and alignment errors in cases where the true trees and alignments are known. In some cases, the inclusion of structure results in changes to the consensus topology, indicating that structure may contain additional information beyond that which can be obtained from sequences. We use the model to investigate the order of divergence of cytoglobins, myoglobins, and hemoglobins and observe a stabilization of phylogenetic inference: although a sequence-based inference assigns significant posterior probability to several different topologies, the structural model strongly favors one of these over the others and is more robust to the choice of data set.

Key words: **structural alignment, Bayesian phylogenetics, statistical alignment, globin evolution, stochastic processes.**

Introduction

Early methods for inferring alignments and phylogenetic trees were based on combinations of carefully tuned heuristic procedures, designed to optimize certain types of scoring metrics. Such methods have yielded many valuable insights; however, the results are often highly sensitive to user-specified parameters, and the focus on a single alignment and tree ignores much of the uncertainty associated with the analysis.

With the development of probabilistic models of molecular evolution, it has become possible to quantify this uncertainty in a statistically meaningful fashion. Bayesian methods for phylogenetic inference, such as MrBayes (Huelsenbeck and Ronquist 2001) and BEAST (Drummond and Rambaut 2007), address the issue of tree uncertainty by generating a distribution over phylogenies given a fixed alignment, although the choice of alignment may still heavily bias the resulting distribution on trees (Lake 1991; Morrison and Ellis 1997; Lunter et al. 2008; Wong et al. 2008; Blackburne and Whelan 2013). A further set of methods have been developed to allow for joint sampling of alignments and trees, which allows this source of bias to be avoided (Lunter, Miklós, et al. 2005; Redelings and Suchard 2005; Miklós et al. 2008). Such approaches are more computationally intensive, and analyses to date have been limited to tens rather than hundreds of sequences; however, these analyses are less prone to the misleading conclusions that can result from analyzing a larger number of sequences under a biased model (Kumar et al. 2012).

Including Structural Information

However, for sequences that are highly divergent, there may be a significant degree of uncertainty associated with the resulting alignments and trees. One way of addressing this issue is to combine multiple different types of data into a joint, or mixed, evolutionary model (Ronquist and Huelsenbeck 2003). As well as offering a way of reducing uncertainty, this type of approach has the potential to lead to more robust and reliable results, because the resulting inference is based on multiple independent sources of information (cf. Kumar et al. 2012).

For protein-coding genes, additional information regarding evolutionary relationships can be obtained from protein structures. Because tertiary structure is typically much more highly conserved than sequence, even over large evolutionary distances (Panchenko et al. 2005; Illergård et al. 2009), structural similarity is therefore a more reliable way to infer homology in the so-called twilight zone of low sequence identity, leading to more accurate alignments (Eidhammer et al. 2000; Hasegawa and Holm 2009; Katoh and Standley 2013) and potentially also phylogenies (Johnson et al. 1990; Bujnicki 2000; Lundin et al. 2012).

Article

Fast Track

Recently, Challis and Schmidler (2012) introduced a probabilistic evolutionary model of the joint evolution of protein sequence and structure. In contrast to structurally constrained sequence models that modulate substitution rates based on a fixed structure (Robinson et al. 2003; Rodrigue et al. 2005; Choi et al. 2007; Kleinman et al. 2010), this approach includes an explicit model for the evolution of structure, allowing for structural information to be used to help infer evolutionary distances. Structural evolution is modeled according to a diffusion process with drift, which allows for tractable computation of the likelihood in the resulting joint sequence–structure model. Significant improvements were observed in the accuracy of inferred pairwise divergence times, especially for highly divergent sequences. In this work, we extend the model of Challis and Schmidler (2012) to a tree and explore the utility of incorporating structural information into joint estimation of multiple alignments and phylogenies. Because relatively little is known about structural evolutionary processes, we also introduce a model for heterogeneity in rates of structural evolution, which reduces the potential for conflict between structure- and sequence-based trees (Garau et al. 2005).

We also add a model of background (nonevolutionary) variability in structures, making use of prior information obtained from the X-ray crystallography experimental data, and drawing on aspects of other earlier probabilistic models of protein structure (Green and Mardia 2006; Schmidler 2006; Green et al. 2010; Wang and Schmidler 2014; Rodriguez A, Schmidler SC, unpublished data).

Probabilistic Evolutionary Models

In what follows, we deal with classes of probabilistic models on binary trees. Biologically, these trees define phylogenetic relationships between a set of organisms; probabilistically, given the sequence at a particular parent vertex, evolution along each of its child branches is assumed to proceed independently.

Sequence and Structure Data

Notation introduced in the following sections is summarized in Table A3, for reference. We consider a sequence evolving on a tree, Υ, with vertices $\mathcal{V}_\Upsilon$ and edges $\mathcal{E}_\Upsilon$, according to an evolutionary model with parameters (Φ, Λ, Θ), which describe rates of substitution, indel, and structural evolution processes, respectively. Associated with the K tips of the tree is a set of K homologous sequences $\mathcal{S} = \{S^{(1)}, \ldots, S^{(k)}\}$, with $S^{(k)}$ of length $L^{(k)}$, and corresponding three-dimensional (3D) structures, $\mathcal{C} = \{C^{(1)}, \ldots, C^{(K)}\}$, where $C^{(k)}$ is an $L^{(k)} \times 3$ matrix containing the Euclidean coordinates of the C_α atoms of structure k. To make use of the tree structure to permit tractable inference, each of the internal nodes of the tree is augmented with an associated sequence and structure, the corresponding sets denoted by $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{C}}$, respectively. The structural coordinates and characters associated with these internal sequences will eventually be marginalized out analytically.

Representation of a Multiple Alignment

A multiple alignment can be represented as a set of pairwise alignments along the branches of a tree, $\tilde{\mathcal{M}} = \{M^{(k,l)}\}$, with $(k, l) \in \mathcal{E}_\Upsilon$. Each pairwise alignment, of length $L^{(k,l)} \leq L^{(k)} + L^{(l)}$, can be thought of as a series of columns in a $2 \times L^{(k,l)}$ matrix, indicating homology between characters in $S^{(k)}$ and $S^{(l)}$, that is, the parent and child sequences along the branch. Each such column can take one of three possible states:

$$M_i^{(k,l)} \in \left\{ \begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} x \\ - \end{pmatrix}, \begin{pmatrix} - \\ y \end{pmatrix} \right\} \tag{1}$$

where $x \in \{1, \ldots, L^{(k)}\}$ and $y \in \{1, \ldots, L^{(l)}\}$ indicate the index of the characters aligned in the column and – indicates an insertion or deletion. We will also denote by $M^{(k)}$ the row corresponding to sequence k in $M^{(k,l)}$, with the zero elements removed, equal to the vector $(1, \ldots, L^{(k)})$; one of the requirements for a valid set of alignments, $\tilde{\mathcal{M}}$, is that all the pairwise alignments should be consistent in the sense that the mapping $M^{(k,l)} \mapsto M^{(k)}$ is the same for all l. Another requirement is that $L^{(k)}$ be equal to the length of $S^{(k)}$ when k is a leaf node. The full alignment, $\tilde{\mathcal{M}}$, can be projected down to a leaf alignment between the sequences at the leaves of the tree, $\mathcal{M}$, expressed in the familiar tabular format. We omit further notational details here for brevity.

Joint Model for Sequence and Structure

The first phylogenetic evolutionary models to be developed allowed only for substitution events, assuming the alignment of the sequences to be known and fixed (Kimura 1980; Felsenstein1981). However, work over the last 2 decades has shown that probabilistic modeling of insertion and deletion (indel) events can yield valuable additional information regarding evolutionary processes (Löytynoja and Goldman 2005; Dessimoz and Gil 2010), partly due to the rarity of such events (Lunter et al. 2003; Westesson et al. 2012). In this work, we build on these existing approaches, adding a probabilistic model of protein structure to yield a joint Bayesian model for substitutions, indels, and structural evolution on a tree.

For reasons of tractability, we focus attention on models where the joint posterior of the unknown parameters of interest, given the observed (leaf) and augmented (internal node) data, can be factored as the product of substitution and structural contributions, and a stochastic indel process:

$$P(\tilde{\mathcal{M}}, \Upsilon, \Phi, \Theta, \Lambda \mid \mathcal{S}, \tilde{\mathcal{S}}, \mathcal{C}, \tilde{\mathcal{C}}) \propto P(\Upsilon) \underbrace{P(\tilde{\mathcal{M}}, \Lambda \mid \Upsilon)}_{\text{indel}} \underbrace{P(\Phi, \Theta \mid \mathcal{S}, \tilde{\mathcal{S}}, \mathcal{C}, \tilde{\mathcal{C}}, \tilde{\mathcal{M}}, \Upsilon)}_{\text{substitution/structure}} \tag{2}$$

The above factorization will generally only be possible for independent-site models of substitution and structural evolution; insertions and deletions can change neighborhood relationships, such that substitution, structure, and indel processes are in general not separable in neighbor-dependent models.

In this work, we also make the further assumption of separability between the substitution and structural evolutionary processes, such that

$$P(\Phi, \Theta \mid \mathcal{S}, \tilde{\mathcal{S}}, \mathcal{C}, \tilde{\mathcal{C}}, \tilde{\mathcal{M}}, \Upsilon) = \underbrace{P(\Phi \mid \mathcal{S}, \tilde{\mathcal{S}}, \tilde{\mathcal{M}}, \Upsilon)}_{\text{substitution}} \underbrace{P(\Theta \mid \mathcal{C}, \tilde{\mathcal{C}}, \tilde{\mathcal{M}}, \Upsilon)}_{\text{structure}}$$

It should be noted that the branch lengths in the tree Υ are common to the substitution and structure components, such that the above separation still permits structural evolution to be expressed as a function of substitutions per site along each branch. Although it is also possible to formulate independent-sites models with a more explicit dependence between sequence and structure (e.g., by allowing for Θ to be a function of the amino acid content for a particular site), we leave such developments for future work.

Marginal Posterior

Ultimately, we are interested in the marginal posterior distribution over alignments, trees, and model parameters obtained by integrating over the unobserved internal node data

$$P(\tilde{\mathcal{M}}, \Upsilon, \Phi, \Theta, \Lambda \mid \mathcal{S}, \mathcal{C}) \propto P(\Upsilon)P(\tilde{\mathcal{M}}, \Lambda \mid \Upsilon) \times P(\Phi)P(\mathcal{S} \mid \Phi, \tilde{\mathcal{M}}, \Upsilon) \times P(\Theta)P(\mathcal{C} \mid \Theta, \tilde{\mathcal{M}}, \Upsilon)$$

We focus on cases where the observed data likelihoods $P(\mathcal{S} \mid \Phi, \tilde{\mathcal{M}}, \Upsilon)$ and $P(\mathcal{C} \mid \Theta, \tilde{\mathcal{M}}, \Upsilon)$ can be computed exactly by analytical summation and integration over ancestral characters and coordinates. Although, with some simplifying assumptions, certain indel models also allow for analytical summation over internal node alignments (Thorne et al. 1991; Lunter, Miklós, et al. 2005; Bouchard-Côté and Jordan 2013), for many models of interest this is not possible, yielding a problem of exponential complexity (Lunter, Drummond, et al. 2005); hence, we focus on the general case of inference for the full alignment $\tilde{\mathcal{M}}$ rather than directly targeting the marginal posterior for the leaf alignment $\mathcal{M}$.

Beyond the factorizability in equation (2), the statistical alignment framework we present here is not dependent on particular model choices for substitution and indel processes, but we will briefly describe the specific choices used in this work for the purposes of illustrating how they combine with the structural model. We introduce the structural model in more detail in the subsequent section, but note here that one of the key features of the approach we will present is that it allows the integration over unknown ancestral structures to be carried out analytically, greatly increasing the tractability of the resulting model.

Indel Model

In this work, we focus on the TKF92 model (Thorne et al. 1992) to generate the probability $p(\tilde{\mathcal{M}} \mid \Lambda, \Upsilon)$. This model is a birth/death process on fragments, each of which contains a contiguous run of characters (in our case amino acids). Fragments are inserted at rate λ and are deleted with rate μ; the length of each fragment is geometrically distributed according to a probability r. We adopt the scheme discussed by Miklós et al. (2008), whereby fragments are not inherited from parent to child branches; the contribution to the posterior for $\tilde{\mathcal{M}}$ from the indel model can then be factored over the branches of the tree

$$P(\tilde{\mathcal{M}}, \Lambda \mid \Upsilon) = \prod_{j \in \mathcal{V}_\Upsilon} P(M^{(j)}, \Lambda) \prod_{(k,l) \in \mathcal{E}_\Upsilon} \frac{P(M^{(k,l)}, \Lambda \mid \Upsilon)}{P(M^{(k)}, \Lambda)P(M^{(l)}, \Lambda)} = P(M^{(\text{root})}, \Lambda) \times \frac{\prod_{(k,l) \in \mathcal{E}_\Upsilon} P(M^{(k,l)}, \Lambda \mid \Upsilon)}{\prod_{j \in \text{an}(\Upsilon)} P(M^{(j)}, \Lambda)^2} \quad (3)$$

where $\mathcal{V}_\Upsilon$ and $\mathcal{E}_\Upsilon$ are the sets of vertices and, respectively, edges in the tree Υ, and $\text{an}(\Upsilon)$ is the set of ancestral (nonleaf) nodes of the tree. The vector $M^{(j)}$ is equal to one of the rows in the pairwise alignment $M^{(k,l)}$. The second line assumes that the tree is binary, which will be the case in all the examples we consider.

Each pair term in the numerator of equation (3) can be computed via dynamic programming using the pair-HMM representation of the indel model (Miklós et al. 2008), allowing the augmented likelihood to be computed in time linearly proportional to the number of branches in the tree and the square of the average sequence length. The stationary probabilities for individual nodes are derived in Thorne et al. (1992), and take the form

$$P(M^{(k)} \mid \Lambda) \equiv P(L^{(k)} \mid \lambda, \mu, r) \quad (4)$$

$$= (1-m)m(1-r)[m(1-r)+r]^{L^{(k)}-1} \quad (5)$$

where $L^{(k)}$ represents the length of the kth sequence, equivalent to the length of $M^{(k)}$, and $m = \lambda/\mu$.

Substitution Model

Under the independent-sites assumption, the substitution process is modeled as a collection of independent processes on individual amino acids. This allows the marginal likelihood of the leaf sequences, given a particular alignment $\tilde{\mathcal{M}}$, to be calculated using the familiar sum-product algorithm of Felsenstein (1981), yielding the quantity $P(\mathcal{S} \mid \Phi, \tilde{\mathcal{M}}, \Upsilon) = \sum_{\tilde{\mathcal{S}}} (\mathcal{S}, \tilde{\mathcal{S}} \mid \Phi, \tilde{\mathcal{M}}, \Upsilon)$.

The analyses conducted here employ the Dayhoff et al. (1978) matrix of amino acid substitution to parameterize Φ, although other choices are possible.

Structural Drift Model

There is empirical evidence of correlation between evolutionary time and structural divergence, although the exact nature of this relationship has remained the source of much speculation (Chothia and Lesk 1986; Illergård et al. 2009). Chothia and Lesk (1986) famously observed an exponential relationship between structural divergence of core homologous residues as measured by the root-mean-square deviation (RMSD) and sequence divergence as measured by sequence

identity. This original relationship was proposed based on a small data set that was available at the time: 32 pairs of homologous proteins, as well as five instances of the same protein crystallized under different conditions. More recently, several authors have observed a linear relationship when sequence identity is converted to a measure of substitutions per site (Illergård et al. 2009), or if sequence identity and RMSD are replaced by approximate measures of significance (Wood and Pearson 1999), although in some families a nonlinear relationship may still be observed (Panchenko et al. 2005). In all cases, structural divergence is observed to increase as sequence similarity decreases.

Model Specification

To construct a model that allows for structural divergence to be a function of evolutionary distance, Challis and Schmidler (2012) introduced a diffusion-based model of structural drift. Whereas a probabilistic substitution model employs a continuous-time, finite-state Markov process, this structural model utilizes a reversible diffusion process in 3D space, modeling fluctuations in the amino acid positions (represented by their C_α coordinates). As discussed earlier, independence between atoms is assumed to retain tractability. Under this model, structural evolution is modeled using an Ornstein–Uhlenbeck (OU) process on each C_α atom. Unlike Brownian motion, the OU process has a well-defined stationary distribution and so is reversible, allowing the combined structural, indel, and substitution processes to form a reversible joint model.

With $C_{ij}(t)$ representing the jth coordinate of the ith C_α at time t, the structural drift model describes the change in coordinates over time according to the following stochastic differential equation:

$$dC_{ij}(t) = -\theta\, C_{ij}(t)\, dt + \sigma dB \tag{6}$$

where dB is standard Brownian motion, and θ is the rate at which a structure loses memory of its previous configuration, which we term the structural drift rate. The equilibrium distribution and conditional distributions of this process are Gaussians

$$C_{ij}(\infty) \sim N(0, \tau) \tag{7}$$

$$C_{ij}(t) \mid C_{ij}(s) \sim N\left(C_{ij}(s)e^{-\theta(t-s)}, \tau(1 - e^{-2\theta(t-s)})\right) \tag{8}$$

with the marginal variance $\tau = \sigma^2/(2\theta)$ proportional to the expected radius of gyration multiplied by the length of the structure. The quantity $\sigma^2/2$ can be thought of as a diffusion coefficient, with the expected mean square deviation after a time t approximately equal to $\sigma^2 t$ (see supplementary section S2, Supplementary Material online). As such, we will refer to σ^2 as the structural diffusivity.

Structural Diffusion on a Tree

When extending this process to a set of structures related by a phylogeny, we must contend with an unknown ancestral structure at each internal node. Fortunately, the OU process allows for analytical integration over the unknown ancestral structure coordinates, such that the joint likelihood of the observed structures at the tips of the tree, $P(\mathcal{C} \mid \tilde{\mathcal{M}}, \Theta, \Upsilon)$, can be computed very efficiently. As discussed by Hansen and Martins (1996), for an OU process on a tree, the joint distribution for the data at the leaves is a multivariate Gaussian, in our case with a zero mean. The Markovian nature of the OU process means that the elements of the covariance matrix can be computed analytically, with $\Sigma_{kl}[\tau, \theta, \Upsilon] = \tau e^{-\theta d_{kl}(\Upsilon)}$, where $d_{kl}(\Upsilon)$ is the distance between leaves k and l along branches of Υ.

Denoting by $\mathcal{C}_j^{(\mathcal{M}_i)}$ the length$-|\mathcal{M}_i|$ vector obtained by taking the jth coordinate of each observed (leaf) structure containing a character at the ith column, the marginal likelihood of the observed structures is then given by a product over the L columns of the alignment and the three spatial dimensions:

$$P(\mathcal{C} \mid \tilde{\mathcal{M}}, \Theta, \Upsilon) = \prod_{i=1}^{L}\prod_{j=1}^{3} \mathrm{N}_{|\mathcal{M}_i|}\left(\mathcal{C}_j^{(\mathcal{M}_i)} \mid \mathbf{0}, \Sigma_{\mathcal{M}_i}[\tau, \theta, \Upsilon]\right) \tag{9}$$

where $\Sigma_{\mathcal{M}_i}$ is a submatrix of Σ of dimension $|\mathcal{M}_i|$ formed by selecting the columns and rows corresponding to ungapped positions in the alignment column $\mathcal{M}_i$.

Figure 1 illustrates a set of samples on a tree drawn from the structural drift model with $\sigma^2 = 0.7$ Å^2/substitution per site, and $\tau = 70$ Å^2, evolving from structure 2DN2 (human hemoglobin [Hb]) at the root.

Branch-Specific Structural Drift Rates

The model thus far assumes a constant structural diffusion coefficient, σ^2, throughout the phylogenetic tree. This assumes that structures respond to sequence mutations in a homogeneous fashion, leading to an approximately linear relationship between evolutionary time and mean-square-deviation (see supplementary section S2, Supplementary Material online). To allow for more general relationships between structural and sequence deviation, as well as reducing potential conflict between sequence- and structure-based trees, we relax this assumption and allow the structural diffusivity to vary over the tree. Following the approach of Thorne et al. (1998) and Aris-Brosou and Yang (2002) with respect to variable rates of sequence evolution, we allow σ^2 to vary by branch, which provides additional flexibility while allowing important properties such as infinite divisibility and reversibility to be maintained across the tree.

There are many ways in which this can be done; here, we consider a model formulation that limits the number of additional parameters required. Let $\mathcal{E}_\Upsilon$ be the set of branches of tree Υ, with $\{\sigma_k^2, \theta_k \mid k \in \mathcal{E}_\Upsilon\}$ the associated set of structural parameters. Allowing both σ_k^2 and θ_k to vary by branch does not preserve a common stationary distribution at each node of the tree, making the joint distribution difficult to specify. To solve this issue, we instead consider the alternative parameterization $\tau_k = \sigma_k^2/(2\theta_k)$ with $\tau_k = \tau$ for all k, such that τ represents the equilibrium variance common to all nodes of

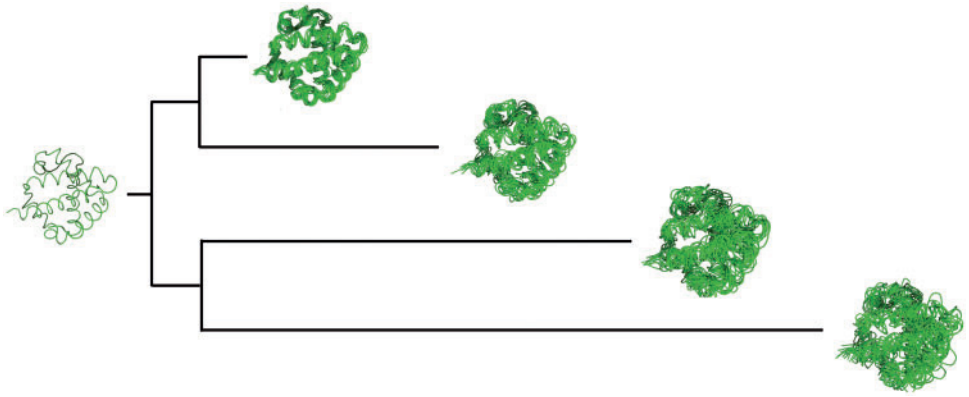


FIG. 1. Ten samples from the structural drift model on a tree, with $\sigma^2 = 0.7$ Å^2/substitution per site, and $\tau = 70$ Å^2. With σ^2 set to zero we would see equal variability at each leaf, whereas the structural drift model proposes that structural divergence will be larger over greater evolutionary distances, in accordance with empirical observations.

the tree, while σ_k^2 is the local structural diffusivity, which is allowed to vary by branch. Since $\sigma_k^2 = 2\tau\theta_k$, the diffusivity of a branch is proportional to its structural drift rate, hence when describing heterogeneity across the tree, we will refer to these quantities interchangeably. The joint distribution of leaf nodes under this model remains simple and easy to obtain. The marginal distribution for each coordinate is then $N(0, \tau)$ as before, while the covariance between coordinates of leaves k and l becomes

$$\Sigma_{kl}[\tau, \theta, \Upsilon] = \tau \exp\left\{ \sum_{m \in \pi(k,l \mid \Upsilon)} t_m(\Upsilon) \frac{\sigma_m^2}{2\tau} \right\} \qquad (10)$$

where $\pi(k, l \mid \Upsilon)$ represents the set of branches lying on the unique shortest path from leaf k to leaf l, and $t_m(\Upsilon)$ is the length of branch m in tree Υ.

Nonevolutionary Sources of Structural Variability

With sequence data, sequencing errors are relatively rare, such that any differences between sequences can generally be attributed to mutation events. However, for structural data, other sources of variability in the coordinates arise from factors such as flexibility of polypeptide chains, variable conformations, and measurement error (Gutin and Badretdinov 1994; Grishin 1997; Illergård et al. 2009). Moreover, this uncertainty may vary across the protein, with surface residues and loops exhibiting increased flexibility over buried core positions.

Information about this uncertainty for high-resolution structures solved by X-ray diffraction is contained in crystallographic *B*-factors for each atomic coordinate. These values, reported by the crystallographer, are intended to summarize a combination of experimental uncertainty and thermal fluctuations and are often strongly correlated with intrinsic structural flexibility measured by nuclear magnetic resonance and molecular dynamics simulations (Rueda et al. 2007). *B*-factors can be converted to units of coordinate uncertainty using approximate formulae such as the diffraction-component precision index (Cruickshank 1960, 1999). This can be combined with additional assumptions (Schneider 2000) to obtain a linear relationship between the *B*-factor and the standard deviation of the coordinates for each atom. We therefore model the variance for the ith atom of structure k (with *B*-factor B_{ki}) as

$$\epsilon_{ki} = \epsilon \frac{B_{ki}^2}{\left(\sum_j B_{kj} \right)^2} \qquad (11)$$

where ε is a global scale parameter for background variance, to be estimated from the data. For the ith column, we compute the expected variance for the column as the average over the atoms aligned to the column

$$\epsilon_i = \frac{1}{|\mathcal{M}_i|} \sum_{k \in \mathcal{M}_i} \epsilon_{k\mathcal{M}_{ik}} \qquad (12)$$

Incorporating this into the structural drift model leads to a variance components model, with column i having covariance $\Sigma^{(i)} = \Sigma_{\mathcal{M}_i} + \epsilon_i I_{|\mathcal{M}_i|}$.

Uncorrelated Structural Perturbations (Nonphylogenetic Structural Model)

In the limiting case as $\sigma_k^2, \theta_k \to 0$, keeping the ratio $\frac{\sigma_k^2}{2\theta_k} = \tau$ fixed, all structural deviation is explained via ε, and the marginal distribution of the observed data in the ith column is

$$C_{ij}^{(\mathcal{M}_i)} \mid \mathcal{M}, \tau, \epsilon, \Upsilon \sim N_{|\mathcal{M}_i|}(0, \Sigma^{(i)}), \qquad (13)$$

where $\Sigma_{kl}^{(i)} = \tau$ if $k \neq l$, and $\Sigma_{kk}^{(i)} = \tau + \epsilon_i$. This is similar to the nonevolutionary Bayesian structure alignment models described above (Wang and Schmidler 2014), where structural perturbations are independent of evolutionary distance. In this limiting model, the structural likelihood does not depend on the tree nor on the evolutionary parameters, and structural information only indirectly affects the distribution over trees via the effect on the alignment.

Rotations and Translations

Up to this point, we have assumed that the data consist simply of a set of 3D coordinates. However, the coordinates of each structure are recorded with respect to an arbitrary reference frame, and the likelihood is not invariant to transformations of the coordinate system. This can be addressed without compromising the reversibility of the model by introduction of auxiliary rotation and translation random variables for each structure, as discussed in Challis and Schmidler (2012). Since the OU process is symmetric and hence invariant to rotations of the coordinate system, we can omit the rotation for an arbitrarily chosen reference protein; this reference protein still has an associated translation, such that the likelihood is independent of the choice of reference. With independent uniform priors over rotations and translations, reversibility is maintained (see supplementary section S3, Supplementary Material online), and the resulting posterior is proper.

Priors

To complete the specification of the full Bayesian model, it is necessary to assign prior distributions to each unknown parameter. In general, we opt for diffuse priors, reflecting our lack of strong prior knowledge regarding the parameters, using standard conjugate priors where available (specific prior choices are described in supplementary section S1, Supplementary Material online).

Shrinkage Prior for Branch-Specific Diffusivity

With a separate drift rate for each branch, there might be concern that the structural drift model could be overparameterized (Dutheil et al. 2012; Groussin et al. 2013). To address this possibility, we adopt a shrinkage-favoring mixture prior for the branch-specific σ_k^2 parameters:

$$\sigma_k^2 \mid \sigma_g^2, \nu \sim \gamma\delta(\sigma_k^2 - \sigma_g^2) + (1-\gamma)\mathrm{LogN}(\log \sigma_g^2, \nu), \quad (14)$$

with $\sigma_g^2 \sim \mathrm{Gamma}(a_g, b_g)$ and $\nu \sim \mathrm{Gamma}(a_\nu, b_\nu)$. This setup allows for pooling of information about σ_g^2 from all branches, while maintaining the flexibility of individual rates for each branch, as well as allowing for some degree of variable selection when appropriate. We set $a_g = 1, b_g = 2$ and $a_\nu = 1, b_\nu = 6$.

When $\gamma = 1$, all σ_k parameters are shrunk to the global mean, whereas $\gamma = 0$ yields the fully branch-specific model. For $0 < \gamma < 1$, the σ_k parameters that lie close to the global mean are shrunk strongly to σ_g. This additional shrinkage beyond the basic hierarchical prior is useful in larger trees where the internal branch drift parameters may have high uncertainty, particularly when the corresponding branches are very short.

For smaller trees, we fix $\gamma = 0$; for larger trees, γ is inferred from the data, using a $\mathrm{Beta}(a_\gamma, b_\gamma)$ prior. We set $a_\gamma = 1.35$ and $b_\gamma = 1.1$, which favors shrinking most of the σ_k^2 parameters to the global σ_g^2 but is weak enough to allow for the prior to be overruled when strong evidence exists for heterogeneity among the diffusivity parameters.

To carry out inference under this prior for γ, we employ a standard data augmentation scheme, with indicator variables z_k for inclusion of σ_k^2. To improve mixing, we can integrate out γ from this augmented model, yielding a beta-binomial prior for z

$$P(z \mid a_\gamma, b_\gamma) = {}^n\mathbb{C}_m \frac{B(a_\gamma + m, b_\gamma + n - m)}{B(a_\gamma, b_\gamma)}$$

where n is the number of branches in the tree, $m = \sum_k z_k$ is the number of free σ_k^2 parameters, and $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function.

Markov Chain Monte Carlo Inference

Calculations of posterior distributions are performed by Markov chain Monte Carlo (MCMC) sampling. Since the joint posterior over alignments, topology, and parameters can be complicated, careful design of the MCMC algorithm is essential, and we have developed a number of specialized moves to increase the efficiency of convergence and mixing. Continuous parameters (Θ, Φ, Λ) plus the branch lengths of the tree, are updated using random walk Metropolis updates after appropriate transformations, and tree topologies are proposed using a combination of stochastic nearest-neighbor interchanges and the LOCAL move of Larget and Simon (1999) with the acceptance ratio given in Holder et al. (2005). Alignments are resampled using a window-based progressive dynamic programming scheme to generate proposals, correcting the acceptance ratio by the ratio of likelihoods under the full model. The scheme is similar to the approach outlined in Miklós et al. (2008), augmented to include the structural likelihood. Although the rotations and translations would ideally be integrated out of the model analytically, this typically leads to marginal likelihoods that are complicated functions of the unknown ancestral structures, even for uncorrelated Gaussian noise models (Goodall and Mardia 1993). Hence, we sample rotations and translations using the scheme described in Challis and Schmidler (2012). We also make use of joint proposals that combine the various moves above, to help improve convergence (cf. supplementary section S2, Supplementary Material online).

Monitoring Convergence

All MCMC simulations reported used four independent chains with randomized initial conditions. The overall likelihood and all scalar parameters were monitored for convergence using Gelman–Rubin potential scale reduction factors. For tree topologies, we monitored the stability of clade probabilities in the consensus tree, computing the average standard deviation of split frequencies (ASDSFs) as an overall measure; for alignments, we monitored convergence of alignment length and stabilization of the maximum posterior decoding/minimum-risk summary alignment (Satija et al. 2009; Herman JL, Novák Á, Lyngsø R, Szabó A, Miklós I, Hein J, unpublished data) and associated probabilities for each column.

Results and Model Comparison

To investigate the benefits of the structural model, we focused on data sets with highly divergent sequences, for which sequence-based analysis leaves significant uncertainty. We devote particular attention to the well-studied globins as a test case (table A1); previous attempts to reconstruct the evolutionary history for this family using sequence data have yielded trees with high uncertainty. We also examine a set of cysteine proteinases (table A2), which further demonstrate the utility of structural information in reducing uncertainty in alignments and topologies, while also providing insight into patterns of structural divergence.

To assess the accuracy of parameter estimation (including topologies and alignments), data were simulated from the structural drift model, with the modification that inserted residues were placed at the midpoint of their two neighbors, to avoid unrealistic bond lengths. The structure at the root was set to be equal to the human Hb 2DN2, and model parameters were chosen based on typical values observed on test runs on small globin data sets: $\sigma_k^2 = 0.7$, $\lambda = 0.03$, $\mu = 0.0305$, and $r = 0.67$. All B-factors were set to be equal to 1 for simplicity, and ε was varied over the set $\{0, 0.5, 1.0, 2.0\}$. Three different tree topologies were used, with 6, 8, and 10 leaves, respectively, and for each topology, branch lengths were multiplied by two different scale factors (1.0 and 2.0) to yield varying levels of divergence. Each parameter combination was simulated ten independent times and results averaged over the ten replications.

For each data set, we perform analysis using the sequence-only model, and the phylogenetic (ε, σ^2) and nonphylogenetic (ε-only) variants of the structural model, to assess the effect of including structural information.

Structural Information Improves Alignments

For the simulated data sets, the true multiple alignment is known, and we can measure the distance of the posterior alignment samples to this known alignment using the column score (proportion of correct columns) and the sum-of-pairs score (proportion of correct pairwise homology statements—see supplementary section S2, Supplementary Material online). The alignment accuracy metrics are averaged over the ten repetitions for each tree. Under the sequence-only model alignment, accuracy decreases markedly as branch lengths increase; in contrast, with the structural models, alignment accuracy remains high (fig. 2).

On the 5-globin and cysteine proteinase data sets, alignment accuracy was measured with respect to the alignments contained in the homstrad database (Mizuguchi et al. 1998), which were constructed using 48 (globin) and 13 (cysteine proteinase) structures. In each case, the addition of structural information results in a consistent improvement in alignment accuracy and decreased variability (fig. 2), as with the simulated data.

Structure Reduces Topological Uncertainty

The 5-globin data set was chosen as a simple test case to explore the effect of structural information on topology uncertainty. Results were generated from four independent runs of 100,000 samples, thinned from 10 m iterations, after a 5 m burn-in. For sequence-only, on average around 80,000 topology switches were observed during the 10 m iterations. With the nonphylogenetic structural model included, around 2,200 switches were observed, and with the phylogenetic structural drift model around 700. The ASDSF values for the consensus trees were 0.009, 0.000, and 0.000, respectively (supplementary fig. S3, Supplementary Material online).

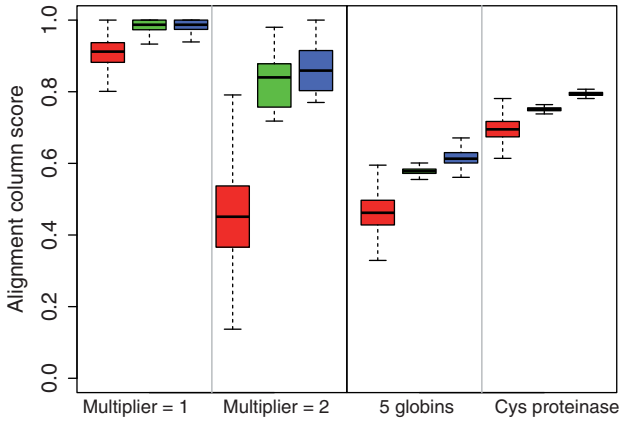
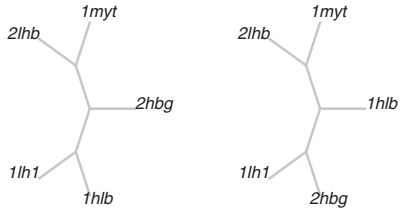


FIG. 2. Alignment accuracy on simulated data (left two panels) for short branches (multiplier = 1) and long branches (multiplier = 2), and on the 5-globin and cysteine proteinase data sets (right panels). Shown are posterior distributions of distance to true alignment (simulated data) or homstrad alignment (globins and cysteine proteinases) obtained under the sequence-based model alone (red), and after combining with the nonphylogenetic (green) and phylogenetic (blue) structural models. In all cases, alignments are more accurate with structural information than under the sequence-only model, with a much narrower range of accuracy values. In many cases, the phylogenetic structural model also offers an additional improvement in alignment accuracy over the nonphylogenetic model. Simulated data results shown for ten realizations on an eight-taxon tree with $\sigma_k^2 = 0.7$ and $\varepsilon = 0.5$, with branch lengths multiplied by the multiplier indicated. Similar results were seen with the sum-of-pairs alignment accuracy metric (not shown).

The sequence-only model visits the most probable tree only 60.1% of the time, with 27.7% of the samples coming from a second topology (fig. 3). We also ran BAli-Phy (Suchard and Redelings 2006) on this data set, and the consensus tree yields a polytomy between 1lh1, 1hlb, and 2hbg, indicating even higher posterior tree uncertainty under the BAli-Phy sequence-only evolutionary model (supplementary fig. S4, Supplementary Material online).

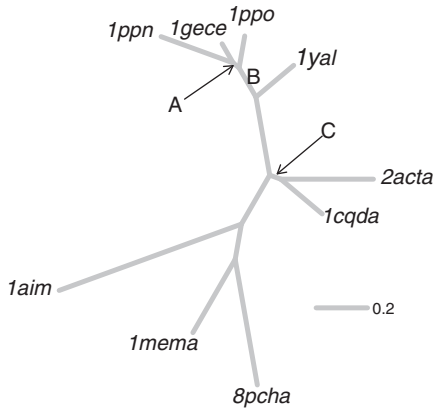
In contrast, under both structural model variants there is virtually no uncertainty in the topology, with more than 99% of the samples coming from the most probable topology, placing 2hbg (*G. dibranchiata* Hb) in between the other four structures. Acceptance for nearest-neighbor topology moves was 4% for sequence only, and less than 0.1% for the structural models, the latter reflecting the very low uncertainty in the topology when structural information is included.

These results clearly illustrate the ability of the joint sequence–structure model to concentrate the posterior around



seq-only	0.606	0.277
seq+struc (*non-phylo*)	0.997	0.000
seq+struc (*phylo*)	0.999	0.000

Fig. 3. The two most frequently sampled tree topologies (shown with equal branch lengths for each branch) for the 5-globin data set under the sequence-only model, with posterior probabilities shown under sequence-only and structural models. Posterior probabilities were computed using the program `trees-consensus`, written by Benjamin Redelings.



	A	B	C
seq-only	0.53	1.00	0.61
seq+struc (*non-phylo*)	0.81	0.97	0.96
seq+struc (*phylo*)	0.97	1.00	1.00

Fig. 4. For the cysteine proteinases, the consensus topology was the same under all model variants. The labeled edges correspond to splits with significant uncertainty under the sequence-only model (the other three splits had posterior probability 1.00 in all cases). The table below the figure shows the posterior probability of each of these labeled splits under the different model variants.

the most likely topology, indicating that additional information is contained within the structural portion of the model. This extra information can be incorporated with little additional computational cost in this case: The three model variants required the same number of iterations to achieve convergence, with the runtime of the structural models around 1.2–1.5 times that of the sequence-only model (see supplementary table S1, Supplementary Material online).

Similar results are observed with the larger cysteine proteinase data set (fig. 4). Again the structural consensus trees do not differ topologically from the sequence tree, and consensus branch lengths are very similar, but uncertain splits in the consensus tree are more highly resolved when structure is included. ASDSF = 0.000, 0.015, and 0.019 for sequence-only, nonphylogenetic (ε-only), and phylogenetic (ε and σ^2) structural models, respectively.

As discussed earlier, structural information can reduce topology uncertainty in at least three ways: By increasing alignment accuracy, by reducing alignment uncertainty, and by providing direct information regarding the topology and branch lengths. In the above cases, a decrease in topology uncertainty is also observed when the nonphylogenetic structural model is used, suggesting that alignment inaccuracy and/or uncertainty is a principal cause of topology uncertainty in these examples. Nevertheless, additional reductions in alignment and topology uncertainty are also seen from adding the phylogenetic drift component to the model (figs. 2 and 4).

Structural Information Reduces Tree Errors

For the simulated data sets where the true tree is known, we can also assess whether the structural model concentrates the tree posterior around the correct topology, using the Robinson–Foulds topology distance (Robinson and Foulds 1981). For trees with smaller branch lengths, the sequence-only and sequence + structure models performed similarly, with the structural model only slightly more accurate. However, when branch lengths are doubled, the structural information not only reduces uncertainty but also improves accuracy of the sampled topologies (fig. 5).

Structure Helps Select between Alternative Topologies

In cases where the majority of the tree is well resolved, the structural model often favors the same consensus tree as sequence. However, for trees with higher uncertainty, structure can also help to select between alternative hypotheses in regions that are difficult to resolve. Here, we illustrate this by analyzing a larger set of globins (table A1).

The known set of vertebrate globin types was expanded relatively recently with the discovery of two additional globins: neuroglobin (Ngb) (Burmester et al. 2000) and cytoglobin (Cygb) (Burmester et al. 2002). Ngb tends to occur in neurons and endocrine cells, while Cygbs appear in fibroblast-related cell types and have been observed to be present in all vertebrates. The function of both proteins is still somewhat unclear, although high levels of sequence conservation suggest a vital physiological function for Cygb (Hoffmann, Opazo, Storz 2012).

Since these discoveries, there has been a surge of interest in establishing the likely evolutionary history of the four vertebrate globin types: Hb, myoglobin (Mb), Ngb, and Cygb. All previous analyses have found Ngb to be the most distant outgroup, so we focus here on the order in which the other vertebrate globins split after diverging from the Ngbs.

Initial phylogenetic studies of Cygb using maximum-likelihood approaches suggested the topology (Ngb, (Hb, (Mb, Cygb))) (Burmester et al. 2002), although the support for this arrangement was found to be low. This topology may have initially appeared more plausible, because it requires O_2 transport to have evolved only once, along the branch to Hb. However, close homology was subsequently discovered between Cygb and the Hbs found in the jawless fishes known as cyclostomes (abbreviated as CycHbs).

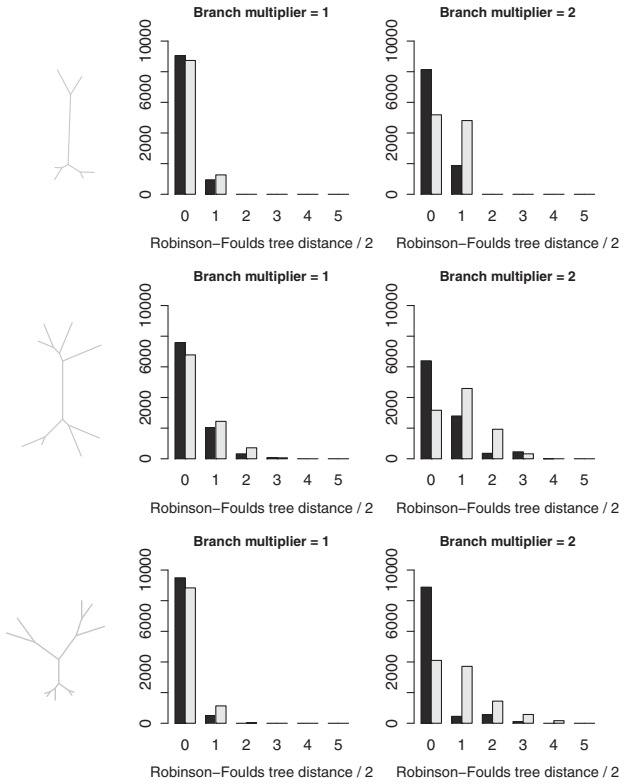


Fig. 5. Posterior distribution of topology errors relative to the true tree for simulated data, analyzed under the phylogenetic structural model (black) and the sequence-only model (gray), as branch lengths are doubled (left to right). The inclusion of structural information allows the tree to be accurately inferred even for large evolutionary distances, whereas the trees inferred by the sequence-only model become much less accurate. Frequencies shown for the trees on the left, with six (top), eight (middle), and ten (bottom) leaves, aggregated from ten independent samples from the model; the maximal half Robinson–Foulds distance for a tree with n leaves is $2(n-3)$, that is, 3, 5, and 7 for the three trees above.

Accounting for this relationship requires either double evolution of O_2 transport function or double loss of this functionality, as discussed by Hoffmann et al. (2010). Based on Bayesian phylogenetic analysis, the authors proposed the same phylogeny as Burmester et al. (2002), but with CycHb splitting from Cygb, yielding the topology (Ngb,(Hb,(Mb,(Cygb,CycHb)))), as shown in the top-left tree in figure 6. Under this scenario, oxygen transport functionality is proposed to have developed independently in the cyclostome Cygb, the ancestor of the current CycHb, with the orthologs of the Mb and Hb genes subsequently lost (Hoffmann et al. 2010, Hoffmann, Opazo, Storz 2012; Storz et al. 2013).

More recently, Hoffmann, Opazo, Hoogewijs, et al. (2012) conducted a Bayesian analysis on a larger data set including globins from plants and in this case reported a three-way split (Ngb,(Hb,Mb,(Cygb,CycHb))) (as shown in the bottom left tree in fig. 6, which contains a polytomy at the center). Using a similar data set including plant globins (without CycHb), Ebner et al. (2010) were also unable to resolve this three-way split, reporting the same polytomy.

Here, we compare the results obtained by Hoffmann et al. (2010) and Hoffmann, Opazo, Hoogewijs, et al. (2012) with those from our structural model, as well as the sequence-only indel model. To do so, we construct smaller versions of the two data sets, containing one or two representatives from each of the clades of interest (details in table A1). The first data set is the 8-globin set containing only Hb, Mb, Cygb, Ngb, and CycHb, and the second data set contains an additional four proteins, namely three plant globins and a recently crystallized bacterial globin known as Hell's gate, which has been observed to show high structural homology with human Ngb (Teh et al. 2011; Vázquez-Limón et al. 2012).

Although the original analyses of Hoffmann et al. (2010) and Hoffmann, Opazo, Hoogewijs, et al. (2012) used 68 and 110 sequences, respectively, we obtain the same consensus tree from just 8 and 12 sequences using our sequence-only statistical alignment model (fig. 6). However, as with the results of Hoffmann, Opazo, Hoogewijs, et al. (2012), the addition of the plant globins appears to destabilize the consensus tree, favoring other topologies in the posterior.

Specifically, our sequence-only model shifts from having 94% posterior probability on the split (Cygb, CycHb), Mb | Hb in the 8-globin case, to favoring this less than 50% of the time when the plant globins are added. In the 12-globin case, the sequence-only model visits the following three topologies between the clades of interest:

1) (Mb,((Cygb,CycHb),Hb))
2) ((Cygb,CycHb),(Mb,Hb))
3) (Hb,((Cygb,CycHb),Mb))

with relative frequency 2:1:1. The third topology is the same as the consensus topology on the 8-globin set.

As noted by Hoffmann, Opazo, Hoogewijs et al. (2012), globins are relatively short proteins and thus limited in the information that can be provided about evolutionary history. Hence, there is good reason to believe that more accurate inference can be obtained by including other sources of information such as structure.

Indeed, as shown in figure 7, the structural model favors topology 2 with almost 100% certainty regardless of whether the plants globins are added. This demonstrates that inference under the structural model is more robust to the choice of data set. Moreover, we can see that the sequence-only model is shifting to increasingly favor the structural tree as more sequences are included, illustrating the fact that structures can contain additional evolutionary information beyond what can be obtained from sequences alone.

Both structural models favor (CycHb, Cygb) as the first split from the root. It should be emphasized that in the nonphylogenetic (ε-only) structural model, only the alignment is directly informed by structural information (rather than evolutionary distance), which reiterates the fact that the alignment can have a large impact on the resulting phylogenetic inference. When phylogenetic structural drift is also included in the model, the posterior probability of (CycHb,Cygb) diverging before the Mb–Hb split increases further (from 0.72 to 1.00), demonstrating that the phylogenetic structural drift model does indeed allow for additional structural information to be used in estimating tree topologies.

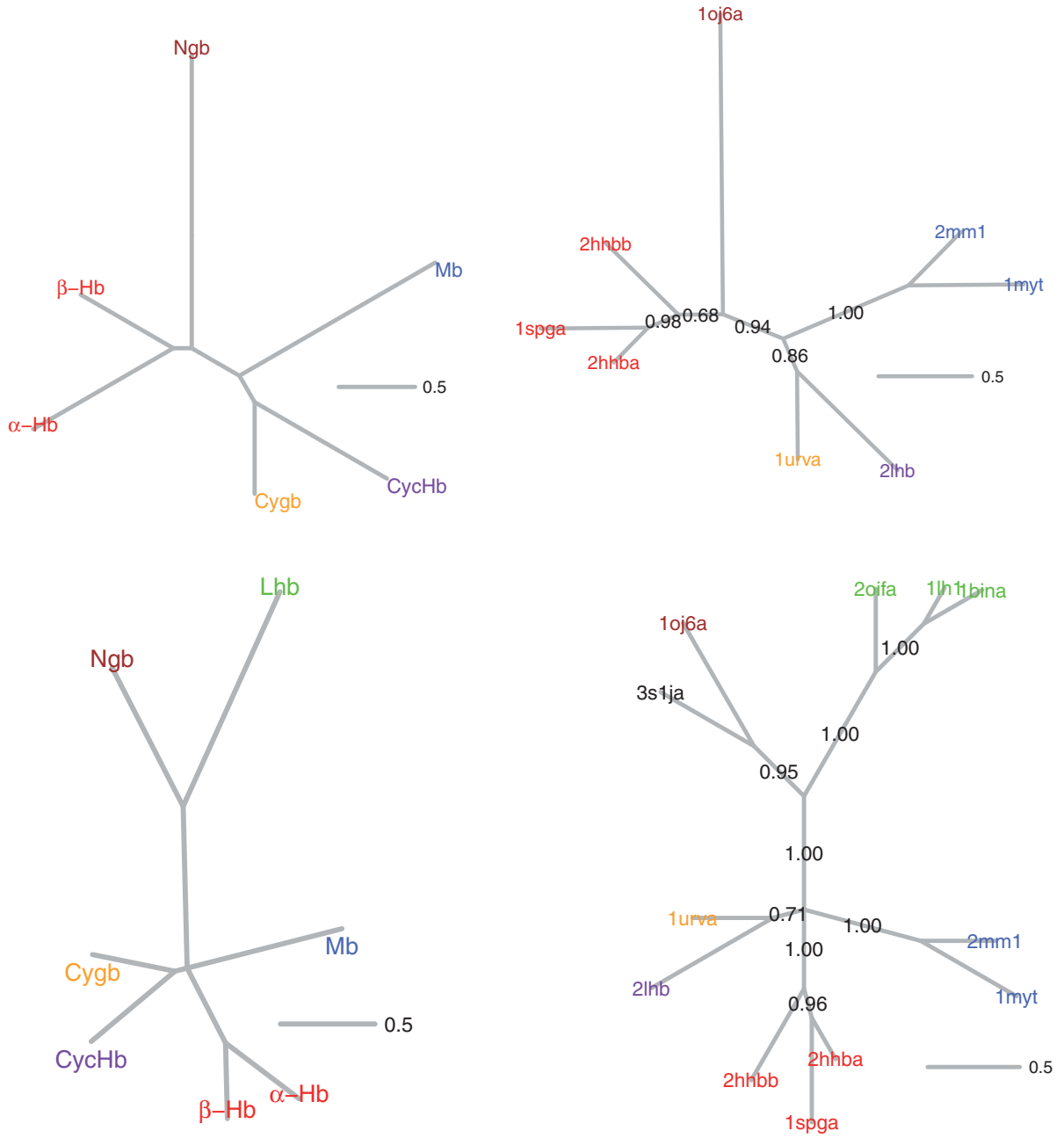


Fig. 6. Consensus trees for globin data sets, taken from Hoffmann et al. (2010) and Hoffmann, Opazo, Hoogewijs, et al. (2012) (top left and bottom left, respectively), and inferred using the sequence-only evolutionary model of Miklós et al. (2008) (top right and bottom right, ASDSF = 0.011, 0.008, respectively). The bottom row features an augmented data set containing plant globins, as well as a bacterial globin in our analysis. In both cases, we obtain the same consensus tree as Hoffmann et al., including the four-way polytomy in the 12-globin case.

Phylogenetic Structural Model Improves Fit

As shown by the results in the previous sections, structural information is able to reduce topology uncertainty, concentrating the topology distribution around the posterior mode, as well as offering improvements in alignment accuracy. These improvements are often greater with the phylogenetic structural drift model than with the nonphylogenetic (ε-only) model.

To measure whether the phylogenetic model also achieves a better model fit to the data, we make use of the deviance information criterion (DIC) (Spiegelhalter et al. 2002), given by $\mathrm{DIC} = \mathbb{E}[D] + P_V$, where $D = -2\log L$ is the deviance and $P_V = Var[D]/2$ is a measure of the effective number of parameters in the model (Gelman et al. 2003). Smaller values of DIC indicate a better model fit. The DIC measure is particularly suited to analyzing the output of MCMC inference in hierarchical models when Bayes factors are not easily available (Spiegelhalter et al. 2002). It should be noted that the effective number of parameters includes a contribution from the alignment and the tree, such that lower posterior uncertainty in these parameters will reduce the effective dimensionality of the model.

As presented in table 1, despite increasing the actual number of parameters, the addition of phylogenetic drift

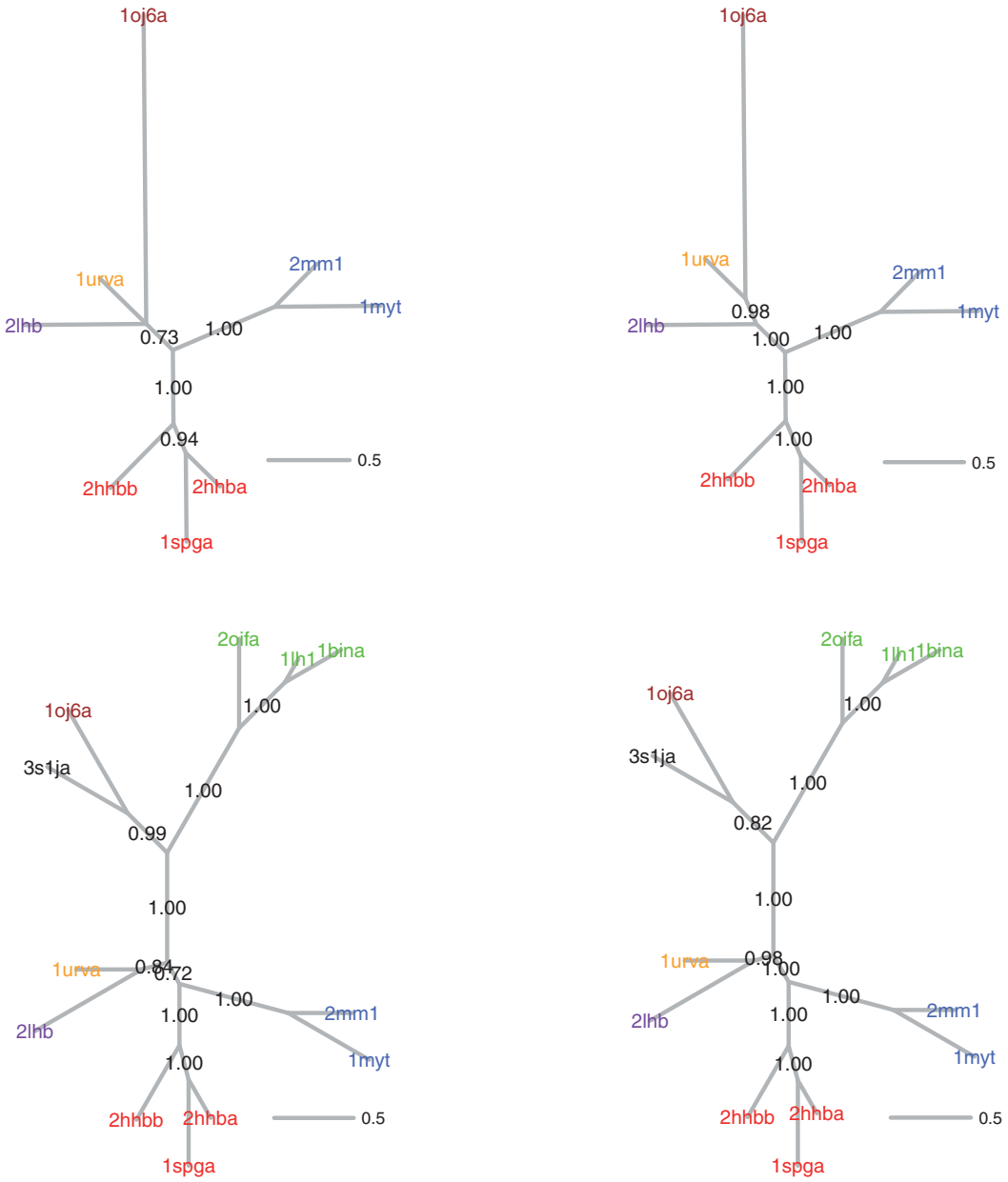


Fig. 7. The structurally derived trees have very low uncertainty, and the order of the splits of interest is unchanged by the inclusion of additional sequences. Consensus trees derived under the nonphylogenetic (ε-only) structural model (top left and bottom left, ASDSF = 0.010, 0.026, respectively) and the phylogenetic structural drift model (top right and bottom right, ASDSF = 0.002, 0.016, respectively).

rates for each branch reduces the overall uncertainty associated with the model, hence decreasing the effective number of parameters, P_V, and resulting in a substantial improvement in model fit, as measured by the DIC.

With complete shrinkage ($\gamma = 1$), the model retains only a single global σ_g^2, decrease the effective number of parameters (on the 5-globin set this results in a reduction in average P_V from 148 to 140). However, the model fit generally suffers as a result (average DIC increases from 13,640 to 13,700 on the 5-globin data set) and tends to result in trees with very different branch lengths from those obtained with sequence-only data. In contrast, the heterogeneous diffusivity model ($\gamma < 1$) results in a better model fit and estimates branch lengths similar to those in the sequence-only trees. This suggests that branch-specific drift rates are indeed needed to explain the heterogeneity in the data. We examine this in more detail in the sequel to this work (in preparation).

Parameter Inference

In addition to alignments and phylogenies, the model also provides the ability to estimate several scalar parameters of interest in the evolutionary process, such as indel rates and structural diffusivity coefficients.

On simulated data, the structural parameters are recovered to a high degree of accuracy, lying within the 95% highest posterior density interval in all cases, with the posterior median usually very close to the true value (see supplementary figs. S6–S8, Supplementary Material online). Importantly, we are able to clearly resolve the different

Table 1. Effective Number of Parameters, P_V, and Model Fit as Measured by DIC for Structural Models with and without a Phylogenetic Drift Component.

	8-Globins		12-Globins		Cys Proteinase	
	Nonphylogenetic	Phylogenetic	Nonphylogenetic	Phylogenetic	Nonphylogenetic	Phylogenetic
P_V	150	140	258	229	226	213
DIC	16,759	15,959	25,110	23,743	18,739	17,075

NOTE.—Results averaged over four independent repetitions for each data set.

Table 2. Comparison of Posterior Quantiles for Global Structural Parameters on Three Data Sets under the Phylogenetic and Nonphylogenetic Variants of the Model.

		8-Globins		12-Globins		Cys Proteinase	
		Nonphylogenetic	Phylogenetic	Nonphylogenetic	Phylogenetic	Nonphylogenetic	Phylogenetic
$\hat{\epsilon}$	5%	3.23	0.762	5.16	1.54	1.03	0.239
	50%	3.53	0.902	5.78	1.76	1.09	0.275
	95%	3.81	1.046	6.37	1.99	1.14	0.310
	GR	1.02	1.00	1.49	1.00	1.01	1.04
$\hat{\sigma_g^2}$	5%	0	0.085	0	0.112	0	0.032
	50%	0	0.192	0	0.232	0	0.049
	95%	0	0.336	0	0.386	0	0.069
	GR	—	1.00	—	1.00	—	1.01

NOTE.—Results averaged over four repetitions from independent starting points. Gelman-Rubin potential scale reduction factors (GR) are shown below each column, and effective sample sizes are shown in supplementary table S2, Supplementary Material online. In the cysteine proteinase case, most of the variability is explained by baseline variance rather than evolutionary drift, although drift coefficients are significantly higher in certain regions of the tree (not shown). Underlined values correspond to the median in each case.

contributions from ε and σ even without repeated observations at the leaves.

Table 2 presents posterior quantiles for ε and σ_g^2 (the global diffusivity) on two globin data sets (with 8 and 12 taxons), and the cysteine proteinase data set, under the nonphylogenetic (ε-only) and phylogenetic structural models. The phylogenetic drift model estimates $\sigma_g^2 > 0$ even with ε in the model, indicating that there is always a time-dependent component to the structural variation. ε is a multiplicative scale factor (in units of Å^2) for the site-specific variance parameters, which in our case are proportional to normalized *B*-factors. Hence, $\epsilon = 1$ signifies that an atom with *B*-factor equal to the mean has baseline variance equal to 1 Å^2. The parameter σ_g^2 has units of Å^2 per substitution per site. For example, from the 12-globin set, we expect phylogenetic drift to lead to an increase in mean square deviation of approximately 0.23 Å^2 per substitution per site (see table 2), although there are also noticeable heterogeneities in drift rates across the tree.

In all cases, Gelman-Rubin potential scale reduction factors were very close to 1, except for the nonphylogenetic (ε-only) model on the 12-globin data set, since a single ε parameter struggles to explain the variability in this data set, leading to slow convergence. In the cysteine proteinase case, although the global σ_g^2 is estimated to be very low (around 0.05), some branch-specific diffusivity coefficients are estimated to be substantially higher, hence there is still a substantial improvement in model fit using the phylogenetic structural drift model in this case (table 1).

Table 3 also summarizes the posterior distributions of the TKF92 parameters with and without (phylogenetic) structural information. Increasing the data set from 8 to 12, sequences reduce the uncertainty associated with the parameter estimates in all cases, but a similar reduction in uncertainty in the alignment length and *r* is also observed when structural information is included. Alignments are typically slightly longer with the structural model, and the indel rate parameters, λ and μ, are estimated slightly higher. Both of these increases are likely due to the fact that the sequence-only model can have a tendency to over-align in loop regions, whereas the joint sequence–structure model favors more indels in such regions, due to the high structural divergence. This shows that the estimation of these parameters can also be affected by alignment uncertainty; hence, the inclusion of structural information also has the potential to improve estimates of insertion and deletion rates by improving alignment accuracy.

Discussion

The main achievement of this work is the development of a tractable probabilistic model for joint evolution of sequences and structures on a phylogenetic tree. Our results demonstrate that inclusion of structural information reduces posterior uncertainty over alignments and topologies, improves alignment accuracy, and reduces the number of tree errors, allowing for more reliable inference over larger evolutionary distances. The structural model is also more robust to the particular data set chosen for analysis, whereas sequence-only models can be highly sensitive to this choice.

Using this approach, we are able to provide structural insights into the evolutionary history of the globin family, whereas sequence-only methods encounter high uncertainty and sensitivity to choice of data set, making it difficult to confidently characterize deep splits in the tree.

Structural information can reduce topology uncertainty both by reducing alignment uncertainty and by adding

Table 3. Posterior Quantiles for Alignment Lengths (L) and TKF92 Indel Model Parameters for Globin Data Sets, Aggregated from Four Independent MCMC Chains in Each Case.

		8-Globins		12-Globins	
		Sequence-Only	Sequence + Structure	Sequence-Only	Sequence + Structure
L	5%	167	177	174	184
	50%	173	182	184	188
	95%	183	186	194	194
	GR	1.00	1.06	1.01	1.02
r	5%	0.669	0.700	0.644	0.681
	50%	0.787	0.796	0.742	0.761
	95%	0.887	0.880	0.833	0.832
	GR	1.00	1.04	1.00	1.02
λ	5%	0.021	0.035	0.028	0.045
	50%	0.049	0.071	0.050	0.073
	95%	0.092	0.121	0.079	0.109
	GR	1.00	1.00	1.00	1.00
μ	5%	0.021	0.037	0.029	0.047
	50%	0.053	0.077	0.053	0.080
	95%	0.103	0.137	0.087	0.123
	GR	1.00	1.00	1.00	1.00

NOTE.—All runs used a burn-in of 10 m iterations, followed by a sampling period of 20 m (sequence-only) and 40 m (sequence + phylogenetic structural drift), with samples for all parameters recorded every 200 iterations; hence, 100,000 samples were taken for the sequence-only runs and 200,000 for the structural variants. Gelman-Rubin potential scale reduction factors (GR) are shown in each column, and average effective sample sizes are shown in supplementary table S3, Supplementary Material online. Underlined values correspond to the median in each case.

additionally information regarding divergence times for estimating topology and branch lengths. We observe that in some cases, a large decrease in topology uncertainty can be obtained even with a nonphylogenetic structural model (the ε-only model), which affects the tree only via the alignment. This suggests that alignment inaccuracy and/or uncertainty can be a major cause of topology uncertainty and further highlights the benefits of approaching alignment and topology inference in a joint framework, as we have done here.

Future Work

As discussed, several modeling assumptions are made to ensure tractability of likelihood computations. These are likely to be reasonable for modeling local fluctuations around a particular fold but may be less appropriate for modeling larger deviations. In particular, the assumption of independence between sites under the structural model becomes questionable when considering large displacements of secondary structure or other structural motifs. We are currently exploring extensions to allow for dependency between sites, although this is computationally very demanding, just as it is for sequence-based models.

The current model requires experimental structural data for all sequences included in the analysis. This is somewhat restrictive, and we are also developing extensions to allow analyses when only a subset of the sequences has structural data available. A number of other extensions to the model could be considered, including using mixture models in the diffusion process to increase flexibility of the model and potentially locate differing rates of evolution along the sequences, for example, to identify structural features that are under strong selection.

Another modification that may improve model fit would be to allow the priors for each σ_k^2 to depend on the rate of the parent branch, as discussed by Thorne et al. (1998) and Arisbrosou and Yang (2002), to account for the fact that evolutionary rates are likely to diverge as a function of time. From a biophysical perspective, this may reflect the fact that the σ^2 parameters are related to the ability of a structure to accommodate sequence mutations, and this property is likely to be inherited to some extent from the parent structure.

Currently, the model uses the magnitude of the crystallographic B-factor to estimate the expected standard deviation for each atom. In the cases we have examined, this relationship appears to hold very well (see e.g., supplementary fig. S5, Supplementary Material online), but there may be cases where anisotropy and the presence of multiple conformers could lead to noticeable deviations from the expected behavior (DePristo et al. 2004). By instead using the B-factor information to specify a prior distribution for each ϵ_{ki}, it would be possible to allow the data to override the B-factors where appropriate, although a larger number of structures may be needed to carry out parameter estimation in such a model.

Finally, as mentioned earlier, the structural model presented here is independent of the particular choice of indel model. By combining structural drift with other stochastic models of insertion and deletion, for example, the recently developed Poisson indel model (Bouchard-Côté and Jordan 2013), which allows for analytical marginalization of indel histories as a result of some simplifying model assumptions, it may be possible to increase the size of data sets that can be analyzed using this type of joint approach.

Availability

We have implemented the joint sequence–structure model as a plugin for the StatAlign software package (Novák et al. 2008), which can be downloaded, along with example data sets, from http://statalign.github.io/ (last accessed June 22, 2014).

Supplementary Material

Supplementary tables S1–S3, figures S1–S8, and sections S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Jeff Thorne, Willie Taylor, and Serge Vinogradov for helpful discussions and comments. This work was supported by the EPSRC to J.L.H., BBSRC to Á.N., and National Institutes of Health grant R01GM090201 to S.C.S.

Appendix

Table A1. The 5-, 8-, and 12-Globin Data Sets, Grouped According to Observed Clades.

Structure	Protein	Organism	Resolution	*R* Value	Length[a]
2oif	NsGb	*H. vulgare* (barley)	1.80	20.2	153
1bin	Lhb	*Glycine max* (soybean)	2.20	19.8	143
1lh1[b]	Lhb	*L. luteus* (lupin bean)	2.00	27.3	153
1oj6[c]	Ngb	*H. sapiens* (human)	1.95	17.8	147
3s1j	HGbI	*M. infernorum* (thermophile)	1.80	21.0	131
1urv[c]	Cygb	*H. sapiens* (human)	2.00	22.2	154
2lhb[b,c]	CycHb	*P. marinus* (lamprey)	2.00	14.2	149
1myt[b,c]	Mb	*T. albacares* (tuna)	1.74	17.7	146
2mm1[c]	Mb	*H. sapiens* (human)	2.80	15.8	153
1spga[c]	α-Hb	*L. xanthrus* (spot croaker)	1.95	19.1	143
2hhba[c]	α-Hb	*H. sapiens* (human)	1.74	16.0	141
2hhbb[c]	β-Hb	*H. sapiens* (human)	1.74	16.0	146
2hbg[b]	Hb	*G. dibranchiata* (bloodworm)	1.50	12.7	147
1hlb[b]	Hb	*C. aurenicola* (sea cucumber)	2.50	15.0	157

NOTE.—NsGb, nonsymbiotic plant globin; Lhb, leghemoglobin; HGbI, bacterial Hell's gate globin I; CycHb, cyclostome Hb. All structures are present in the 12-globin set, except 2hbg and 1hlb.

[a]Length shown for the portion present in the PDB file.

[b]Structures present in the 5-globin data set.

[c]Structures present in the 8-globin data set.

Table A2. The Cysteine Proteinase Data Set.

Structure	Protein	Organism	Resolution	*R* Value	Length[a]
1aim	Cruzain	*T. cruzi* (trypanosome)	2.00	18.8	216
8pcha	Cathepsin H	*S. scrofa* (wild boar)	2.10	NA	221
1mema	Cathepsin K	*H. sapiens* (human)	1.80	18.3	216
2acta	Actinidin	*A. chinensis* (kiwi fruit)	1.70	16.5	219
1cqda	Proteinase II	*Z. officinale* (ginger)	2.10	21.3	217
1yal	Chymopapain	*C. papaya*	1.70	19.2	217
1ppn	Monoclinic papain	*C. papaya*	1.60	16.0	213
1gece	Glycyl endopeptidase	*C. papaya*	2.10	19.6	217
1ppo	Protease omega	*C. papaya*	1.80	15.5	217

NOTE.—Average pairwise identity using the HOMSTRAD alignment is 42%.

[a]Length shown for the portion present in the PDB file.

Table A3. Mathematical Notation Used in the Article.

Model component	Symbol	Domain	Meaning
Data structures	Υ	Binary tree	Phylogenetic tree
	$\mathcal{M}, \tilde{\mathcal{M}}$	*eqn (1)*	Observed (+ ancestral) alignments
	$\mathcal{S}, \tilde{\mathcal{S}}$	$\mathbf{S}^{(k)} \in \chi^{L_k}$	Observed (+ ancestral) sequences
	$\mathcal{C}, \tilde{\mathcal{C}}$	$\mathbf{C}^{(k)} \in \mathbb{R}^{3\times L_k}$	Observed (+ ancestral) coordinates
TKF92 indel model (Λ)	μ	$(0, \mu)$	Insertion rate
	λ	(λ, ∞)	Deletion rate
	r	(0, 1)	Geometric rate for indel length
Structural model (Θ)	τ	$(0, \infty)$	Average structural radius of gyration
	ϵ_i	$(0, \infty)$	Baseline structural variance for the ith column
	θ_k	$(0, \infty)$	Rate structure loses memory along the kth branch
	σ_k^2	$(0, \infty)$	Structural diffusivity of the kth branch
	σ_g^2	$(0, \infty)$	Mean structural diffusivity
	ν	$(0, \infty)$	Variance of σ_k^2 parameters (on log scale)
	γ	[0,1]	Proportion of σ_k^2 parameters fixed at σ_g^2
Substitution model (Φ)	–	Unspecified	Substitution model parameters

NOTE.—χ denotes the set of characters (usually amino acids) utilized by the substitution model.

References

Aris-Brosou S, Yang Z. 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Syst Biol.* 51(5):703–714.

Blackburne BP, Whelan S. 2013. Class of multiple sequence alignment algorithm affects genomic analysis. *Mol Biol Evol.* 30(3):642–653.

Bouchard-Côté A, Jordan MI. 2013. Evolutionary inference via the Poisson indel process. *Proc Natl Acad Sci U S A.* 110(4):1160–1166.

Bujnicki JM. 2000. Phylogeny of the restriction endonuclease-like superfamily inferred from comparison of protein structures. *J Mol Evol.* 50(1):39–44.

Burmester T, Ebner B, Weich B, Hankeln T. 2002. Cytoglobin: a novel globin type ubiquitously expressed invertebrate tissues. *Mol Biol Evol.* 19(4):416–421.

Burmester T, Weich B, Reinhardt S, Hankeln T. 2000. A vertebrate globin expressed in the brain. *Nature* 407(6803):520–523.

Challis CJ, Schmidler SC. 2012. A stochastic evolutionary model for protein structure alignment and phylogeny. *Mol Biol Evol.* 29(11):3575–3587.

Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL. 2007. Quantifying the impact of protein tertiary structure on molecular evolution. *Mol Biol Evol.* 24(8):1769–1782.

Chothia C, Lesk AM. 1986. The relationship between the divergence of sequence and structure in proteins. *EMBO J.* 5(4):823–826.

Cruickshank DWJ. 1960. The required precision of intensity measurements for single-crystal analysis. *Acta Crystallogr.* 13(10):774–777.

Cruickshank DWJ. 1999. Remarks about protein structure precision. *Acta Crystallogr D Biol Crystallogr.* 55(3):583–601.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure, Vol. 5. Washington (DC): National Biomedical Research Foundation. p. 345–352.

DePristo MA, de Bakker PI, Blundell TL. 2004. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure* 12(5):831–838.

Dessimoz C, Gil M. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biol.* 11(4):R37.

Drummond A, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 7(1):214.

Dutheil JY, Galtier N, Romiguier J, Douzery EJ, Ranwez V, Boussau B. 2012. Efficient selection of branch-specific models of sequence evolution. *Mol Biol Evol.* 29(7):1861–1874.

Ebner B, Panopoulou G, Vinogradov S, Kiger L, Marden M, Burmester T, Hankeln T. 2010. The globin gene family of the cephalochordate amphioxus: implications for chordate globin evolution. *BMC Evol Biol.* 10(1):370.

Eidhammer I, Jonassen I, Taylor WR. 2000. Structure comparison and structure patterns. *J Comput Biol.* 7:685–716.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17(6):368–376.

Garau G, Di Guilmi AM, Hall BG. 2005. Structure-based phylogeny of the metallo-lactamases. *Antimicrob Agents Chemother.* 49(7):2778–2784.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2003. Bayesian data analysis. 2nd ed. Boca Raton (FL): Chapman & Hall/CRC.

Goodall CR, Mardia KV. 1993. Multivariate aspects of shape theory. *Ann Stat.* 21(2):848–866.

Green PJ, Mardia KV. 2006. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika* 93(2):235–254.

Green PJ, Mardia KV, Nyirongo VB, Ruffieux Y. 2010. Bayesian modelling for matching and alignment of biomolecules. In:. In: O'Hagan A, West M, editors. The Oxford handbook of applied Bayesian analysis. Oxford: Oxford University Press. p. 27–50.

Grishin NV. 1997. Estimation of evolutionary distances from protein spatial structures. *J Mol Evol.* 45:359–369.

Groussin M, Boussau B, Gouy M. 2013. A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. *Syst Biol.* 62(4):523–538.

Gutin AM, Badretdinov AY. 1994. Evolution of protein 3D structures as diffusion in multidimensional conformational space. *J Mol Evol.* 39: 206–209.

Hansen TF, Martins EP. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution* 50(4):1404–1417.

Hasegawa H, Holm L. 2009. Advances and pitfalls of protein structural alignment. *Curr Opin Struct Biol.* 19(3):341–348.

Hoffmann FG, Opazo JC, Hoogewijs D, Hankeln T, Ebner B, Vinogradov SN, Bailly X, Storz JF. 2012. Evolution of the globin gene family in deuterostomes: lineage-specific patterns of diversification and attrition. *Mol Biol Evol.* 29(7):1735–1745.

Hoffmann FG, Opazo JC, Storz JF. 2010. Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates. *Proc Natl Acad Sci U S A.* 107(32):14274–14279.

Hoffmann FG, Opazo JC, Storz JF. 2012. Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. *Mol Biol Evol.* 29(1):303–312.

Holder MT, Lewis PO, Swofford DL, Larget B. 2005. Hastings ratio of the LOCAL proposal used in Bayesian phylogenetics. *Syst Biol.* 54(6):961–965.

Huelsenbeck JP, Ronquist F. 2001. MrBayes: Bayesian inference in phylogenetic trees. *Bioinformatics* 17(8):754–755.

Illergård K, Ardell DH, Elofsson A. 2009. Structure is three to ten times more conserved than sequence: a study of structural response in protein cores. *Proteins* 77(3):499–508.

Johnson MS, Sali A, Blundell TL. 1990. Phylogenetic relationships from three-dimensional protein structures. In: Methods in enzymology. Vol. 183. Waltham (MA): Academic Press. p. 670–690.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16(2):111–120.

Kleinman CL, Rodrigue N, Lartillot N, Philippe H. 2010. Statistical potentials for improved structurally constrained evolutionary models. *Mol Biol Evol.* 27(7):1546–1560.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. *Mol Biol Evol.* 29(2):457–472.

Lake JA. 1991. The order of sequence alignment can bias the selection of tree topology. *Mol Biol Evol.* 8(3):378–385.

Larget B, Simon D. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol Biol Evol.* 16(6):750.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A.* 102(30):10557–10562.

Lundin D, Poole AM, Sjoberg BM, Hogbom M. 2012. Use of structural phylogenetic networks for classification of the ferritin-like superfamily. *J Biol Chem.* 287(24):20565–20575.

Lunter G, Drummond AJ, Miklós I, Hein J. 2005. Statistical alignment: recent progress, new applications, and challenges. In: Statistical methods in molecular evolution, statistics for biology and health. New York: Springer. p. 375–405.

Lunter G, Miklós I, Drummond AJ, Jensen JL, Hein J. 2003. Bayesian phylogenetic inference under a statistical insertion-deletion model. In: Benson G, Page R, editors. Algorithms in bioinformatics. Vol. 2812. Lecture Notes in Computer Science. Berlin (Germany): Springer. p. 228–44.

Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. 2008. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.* 18(2):298–309.

Lunter GA, Miklós I, Drummond A, Jensen HL, Hein JL. 2005. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* 6:83.

Miklós I, Novák A, Dombai B, Hein J. 2008. How reliably can we predict the reliability of protein structure predictions? *BMC Bioinformatics* 9:137.
Mizuguchi K, Deane CM, Blundell TL, Overington JP. 1998. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* 7(11):2469–2471.
Morrison DA, Ellis JT. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Mol Biol Evol.* 14(4):428–441.
Novák A, Miklós I, Lyngsø R, Hein J. 2008. StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics* 24(20):2403–2404.
Panchenko AR, Wolf YI, Panchenko LA, Madej T. 2005. Evolutionary plasticity of protein families: coupling between sequence and structure variation. *Proteins* 61(3):535–544.
Redelings BD, Suchard MA. 2005. Joint Bayesian estimation of alignment and phylogeny. *Syst Biol.* 54:401–418.
Robinson D, Foulds L. 1981. Comparison of phylogenetic trees. *Math Biosci.* 53(12):131–147.
Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol.* 20(10):1692–1704.
Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* 347(2):207–217.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572–1574.
Rueda M, Ferrer-Costa C, Meyer T, Prez A, Camps J, Hospital A, Gelp JL, Orozco M. 2007. A consensus view of protein dynamics. *Proc Natl Acad Sci U S A.* 104(3):796–801.
Satija R, Novák A, Miklós I, Lyngsø R, Hein J. 2009. BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. *BMC Evol Biol.* 9(1):217.
Schmidler SC. 2006. Fast Bayesian shape matching using geometric algorithms (with discussion). In: Bernardo JM, Bayarri S, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M, editors. Bayesian statistics. Vol. 8. Oxford: Oxford University Press. p. 471–490.
Schneider TR. 2000. Objective comparison of protein structures: error-scaled difference distance matrices. *Acta Crystallogr D Biol Crystallogr.* 56(6):714–721.
Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. 2002. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B Stat Methodol.* 64(4):583–639.
Storz JF, Opazo JC, Hoffmann FG. 2013. Gene duplication, genome duplication, and the functional diversification of vertebrate globins. *Mol Phylogenet Evol.* 66(2):469–478.
Suchard MA, Redelings BD. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* 22(16):2047–2048.
Teh AH, Saito JA, Baharuddin A, Tuckerman JR, Newhouse JS, Kanbe M, Newhouse EI, Rahim RA, Favier F, Didierjean C, et al. 2011. Hells Gate globin I: an acid and thermostable bacterial hemoglobin resembling mammalian neuroglobin. *FEBS Lett.* 585(20):3250–3258.
Thorne J, Kishino H, Felsenstein J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol.* 33(2):114–124.
Thorne J, Kishino H, Felsenstein J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. *J Mol Evol.* 34(1):3–16.
Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol.* 15(12):1647–1657.
Vázquez-Limón C, Hoogewijs D, Vinogradov SN, Arredondo-Peter R. 2012. The evolution of land plant hemoglobins. *Plant Sci.* 191–192:71–81.
Wang R, Schmidler SC. 2014. Bayesian multiple protein structure alignment. In: Sharan R, editor. Research in Computational Molecular Biology—Lecture Notes in Computer Science. Vol. 8394. Springer-Verlag. p. 326-339.
Westesson O, Lunter G, Paten B, Holmes I. 2012. Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. *PLoS One* 7(4):e34572.
Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. *Science* 319:473–476.
Wood TC, Pearson WR. 1999. Evolution of protein sequences and structures. *J Mol Biol.* 291(4):977–995.