



# Modeling Dependence in Evolutionary Inference for Proteins

Gary Larson<sup>1</sup>, Jeffrey L. Thorne<sup>2</sup>, and Scott Schmidler<sup>3</sup>(✉)

<sup>1</sup> Department of Statistical Science, Duke University, Durham, NC, USA

[gary.larson@stat.duke.edu](mailto:gary.larson@stat.duke.edu)

<sup>2</sup> Departments of Biological Sciences and Statistics, North Carolina State University, Raleigh, NC, USA

[thorne@statgen.ncsu.edu](mailto:thorne@statgen.ncsu.edu)

<sup>3</sup> Departments of Statistical Science and Computer Science, Duke University, Durham, NC, USA

[sschmid@duke.edu](mailto:sschmid@duke.edu)

**Abstract.** Protein structure alignment is a classic problem of computational biology, and is widely used to identify structural and functional similarity and to infer homology among proteins. Previously a statistical model for protein structural evolution has been introduced and shown to significantly improve phylogenetic inferences compared to approaches that utilize only amino acid sequence information. Here we extend this model to account for correlated evolutionary drift among neighboring amino acid positions, resulting in a spatio-temporal model of protein structure evolution. The result is a multivariate diffusion process convolved with a spatial birth-death process, which comes with little additional computational cost or analytical complexity compared to the site-independent model (SIM). We demonstrate that this extended, site-dependent model (SDM) yields a significant reduction of bias in estimated evolutionary distances and helps further improve phylogenetic tree reconstruction.

**Keywords:** Protein structure · Evolution · Dynamic programming  
Phylogeny · Diffusion process

## 1 Introduction

Protein alignment is an integral part of bioinformatic analyses and is a classic, widely studied problem in computational biology. Existing methods for aligning two or more proteins compare amino acid sequences and/or structures of the proteins, and encompass a variety of algorithms with different strengths and purposes. Such algorithms are a fundamental part of phylogenetic research in particular, where the degree and nature of evolutionary divergence between species is a quantity of interest. Alignment procedures that are widely used in studies of protein evolution are based only on the amino acid sequence and do not incorporate the tertiary (three-dimensional) structure of the proteins. Methods that

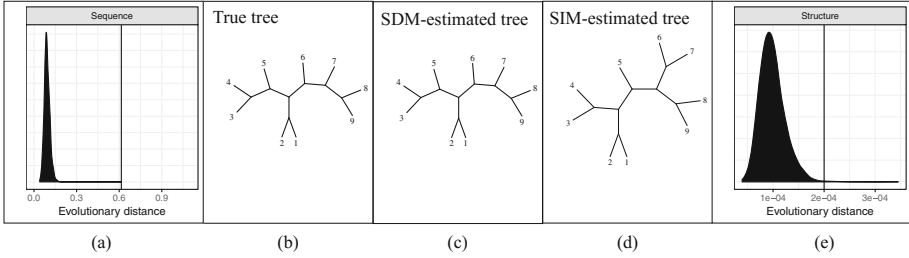
do incorporate tertiary structure, such as those mentioned in [1], do not account for the evolution over time of those structures. Recently Challis and Schmidler [2] introduced a stochastic evolutionary model of protein sequence and structure for this purpose; however, their approach, like the vast majority of alignment algorithms, assumes that “sites” (individual amino acid characters, or backbone atom coordinate triples) evolve independently of one other. This assumption is well-known to be violated since amino acid identities and spatial locations are highly dependent due to a combination of physico-chemical constraints and interactions, including bond lengths and excluded volume, hydrophobic and electrostatic attraction and repulsion, hydrogen bonding, and other cooperative effects in forming stable local and global protein structure. Nevertheless, alignment algorithms based on both sequence and structural information typically ignore the correlations induced by these interactions. Ignoring dependence is often justified by the computational intractability of site-dependent models [2,3]. In this paper we demonstrate that in structure-based alignment, as in sequence-based, ignoring site dependence systematically biases evolutionary inference. We present an expanded version of the Challis and Schmidler model which incorporates neighbor dependence without sacrificing computational tractability.

## 1.1 Motivation

Von Haeseler and Schöniger [4] examined the effect of site dependence on estimates of evolutionary distance between pairs of biological sequences. Using a model of whale mitochondrial DNA evolution whereby the sequence evolves as a collection of independent subsequences, each exhibiting Markovian dependence among its amino acids, the authors demonstrated the tendency to underestimate the true evolutionary distance between two sequences when using a site-independent model. Figure 1a replicates this effect using binary sequences from a nearest-neighbor site-dependent sequence model which does not assume independent subsequences, described in the Appendix A.2. When estimating the divergence time for these sequences under a site-independent version of the same model ( $b = 1$  for model in Appendix A.2), the posterior distribution (Fig. 1a) shows significant underestimation of the true value.

Despite a variety of efforts, no site-dependent sequence model has emerged as a widely applicable replacement for commonly used site-independent sequence models [5]. The primary hurdle to doing so is computational - adding realistic dependence generally prohibits the use of efficient alignment algorithms which rely on dynamic programming.

On the other hand, we demonstrate in Sect. 2 that the site-independent *structural* model (SIM) of [2] can be extended to a *site-dependent structural* model (SDM), incorporating site dependence while maintaining the same interpretability and mathematical and computational tractability as the SIM. Thus we can incorporate dependence into the evolutionary structural part of the model in a relatively straightforward way. Using data simulated from the SDM, we find a systematic underestimation effect for structural data due to the independent-site assumption, similar to that observed in sequences (Fig. 1e). The new SDM



**Fig. 1.** (a) Posterior distribution of evolutionary distance for sequences simulated under site-dependent model with  $b = 2$ ,  $t = 0.6$  (see Sect. A.2), when inference is performed under an assumption of site independence. Significant underestimation is seen relative to truth (vertical line). (b, c, d) This underestimation adversely affects phylogenetic reconstruction, as seen by comparing the true (b) and estimated trees under independent- (d) and dependent-site (c) models. (e) A similar effect is seen for 3D structures, with data simulated under the site-dependent model of Sect. 2.4.

can then be paired with a sequence evolution model to provide a site-dependent expansion of the joint sequence-structure model of Challis and Schmidler [2].

The paper is organized as follows. We briefly review the site-independent structural diffusion model of [2], before describing the general form of a dependent structural diffusion model. Section 2 describes the details of incorporating dependence into the model, with computational tractability being the key constraint on the model’s form. Section 3 describes a reparameterization of the SDM necessary for analyzing the SDM’s effect on phylogenetic inference. Section 4 revisits the motivating example above and compares inferences and phylogenies from the expanded model on a number of real protein examples.

## 2 A Site-Dependent Structural Diffusion Model

Challis and Schmidler [2] introduced a stochastic model for protein structure evolution, extending a previously developed probabilistic framework for structural alignment of proteins [6, 7] into a model suitable for the study of molecular evolution. This work demonstrated the ability to significantly improve phylogenetic inference when structural information about the proteins is available [2, 3]. We briefly review the original Challis-Schmidler model before introducing our extended model incorporating site dependence. Throughout the paper, these structural models will be referred to as the SIM and SDM respectively.

### 2.1 Challis-Schmidler Model

Challis and Schmidler [2] model the diffusion of individual  $C_\alpha$  backbone positions in space, over time, via an Ornstein-Uhlenbeck (OU) process. Independence is assumed between each site along the backbone as well as between the  $(x, y, z)$

coordinates at each site, leading to the joint structure diffusion being modeled as a product of  $3n$  independent univariate OU processes:

$$dC_{ij}^{(t)} = \theta(\zeta_j - C_{ij}^{(t)})dt + \sigma dB \quad (1)$$

where  $C_{ij}^{(t)}$  denotes coordinate  $j \in \{x, y, z\}$  of  $\alpha$ -carbon  $i$  at time  $t$ . This setup admits tractable stationary and conditional distributions but, as noted by Challis and Schmidler, fails to account for known biophysical interactions which lead to strong observed dependence between sites, such as bond length constraints and the effect of excluded volume in the protein. Although a protein structure's coordinate frame is arbitrarily determined by the experiment, we assume the two structures in our pairwise analyses share a coordinate frame; thus for a pair of structures  $C^X, C^Y$ , we assume the coordinate frame of  $C^X$  and do not distinguish between  $C^Y$  and any rigid body rotation  $R$  and translation  $\eta$  thereof. We refer the reader to [2] for a detailed treatment of this issue, and for various other model details omitted here.

## 2.2 Dependence in a Multivariate Ornstein-Uhlenbeck Process

The independent site model (1) can be written as a multivariate diffusion in the form

$$d\mathcal{C} = -\Theta(\mathcal{C} - \zeta)dt + LdB_t \quad (2)$$

where  $\Theta$  and  $\Sigma = LL'$  are both assumed to be identity matrices. Here the  $3n \times 1$  vector  $\mathcal{C} = (\mathcal{C}_x, \mathcal{C}_y, \mathcal{C}_z)$  contains the backbone  $\alpha$ -carbon coordinates,  $\zeta$  is the  $3n \times 1$  long-term mean vector, and  $B_t$  represents  $3n$  independent univariate standard Brownian motion terms. Writing the model in this form makes clear that the assumption of site- (and coordinate-) independence can be relaxed by introduction of general  $\Theta$  and  $\Sigma$ , enabling a more expressive model. For convenience we factor  $\Theta = \Sigma_d \otimes \Theta_p$  and  $\Sigma = \Sigma_d \otimes \Sigma_p$  as Kronecker products, allowing coordinate dependence (subscript  $d$ ) and backbone site dependence (subscript  $p$ ) to be modeled separately.

For purposes of the current paper we set  $\Sigma_d = I_3$  allowing the  $x, y, z$  dimensions within an individual site to diffuse independently of each other. Observed data suggest that dependence between diffusion in the  $(x, y, z)$  dimensions is not strong: Table 1 shows average sample correlations between spatial dimensions for 549 structures comprised of a group of globins and a large group from the manually curated MALIDUP database [8], as well as sample lag-1 autocorrelations (i.e. correlations between consecutive backbone  $\alpha$ -carbons) within each spatial dimension. Although some proteins show weak to moderate correlation between spatial dimensions, the averages indicate the correlation is relatively weak compared to the strong autocorrelation along the backbone within a given spatial dimension. Consequently, we focus on incorporating dependence along the backbone rather than among spatial dimensions  $x, y, z$ .

Under the SDM then, the joint evolution of the  $3n$  scalar coordinates specifying all  $n$  backbone positions follows a multivariate OU process governed by

**Table 1.** Mean sample correlations between dimensions and mean lag-1 autocorrelations along dimensions for 71 globin and 478 MALIDUP protein structures.

	lag-1 autocorrelation			correlation		
	x	y	z	(x,y)	(x,z)	(y,z)
globins	0.95	0.95	0.95	-0.01	0.00	0.01
MALIDUP	0.93	0.93	0.93	0.01	0.02	-0.02

$3n \times 3n$  matrix-valued parameters  $\Theta$  and  $\Sigma$ . This model introduces site dependence while preserving the analytical tractability of the conditional and limiting distributions of the process, important properties for phylogenetic inference. Under the diffusion process defined by the stochastic differential equation in (2), the joint distribution of  $\mathcal{C}^{(t)}$  (the full coordinate set at time  $t$ ) conditional on  $\mathcal{C}^{(s)}$  is multivariate normal:

$$P(\mathcal{C}^{(t)}|\mathcal{C}^{(s)}) \sim N\left(e^{-\Theta\tau}\mathcal{C}^{(s)} + (I - e^{-\Theta\tau})\zeta, \Sigma_\tau\right) \quad (3)$$

with  $\tau$  denoting the time difference ( $t - s$ ) and with conditional covariance  $\Sigma_\tau$  given by

$$vec(\Sigma_\tau) = (\Theta \oplus \Theta)^{-1} \left( I - e^{-(\Theta \oplus \Theta)\tau} \right) vec(\Sigma) \quad (4)$$

where  $vec()$  is the linear operator converting a matrix into a column vector. Letting  $\tau \rightarrow \infty$  in the conditional mean and covariance gives the stationary distribution

$$P(\mathcal{C}) \sim N(\zeta, \Sigma_\infty) \quad (5)$$

where the stationary covariance  $\Sigma_\infty$  is expressed as

$$vec(\Sigma_\infty) = (\Theta \oplus \Theta)^{-1} vec(\Sigma). \quad (6)$$

Although these closed-form solutions exist for general  $\Sigma_p, \Theta_p$ , they are in general not computationally tractable when convolved with the indel process of the evolutionary model from [2] (i.e. the Links model of [9]) because the conditional independence required for dynamic programming is not preserved. To maintain computational tractability in phylogenetic applications, we require forms of  $\Theta_p$  and  $\Sigma_p$  for which both the conditional and stationary distributions of the multivariate OU exhibit certain conditional independencies, as described in the next section.

### 2.3 Computational Tractability in Phylogenetic Models

Common uses of evolutionary models, in phylogenetic or homology detection contexts, require the ability to optimize or average over the set of possible alignments. In a Bayesian or maximum likelihood context, the alignment must be

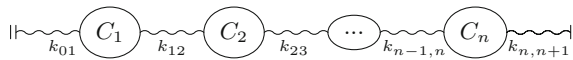
inferred simultaneously with the other parameters. Because of the (exponentially large) size of the alignment space, algorithmic efficiency considerations in these calculations play a key role. In particular, calculating the joint likelihood  $p(X, Y)$  of two structures  $X$  and  $Y$  marginalized over all possible alignments  $\mathcal{M}$  is possible in site-independent models by use of dynamic programming (the so-called forward algorithm for pair hidden Markov models (HMMs); see [10]). These algorithms depend on conditional independence properties of the (marginal) likelihood of the backbone coordinates at a single backbone site given all previous backbone sites:

$$P(C_{ij}^X, C_{ij}^Y \mid C_{1j}^X, C_{2j}^X, \dots, C_{(i-1)j}^X, C_{1j}^Y, C_{2j}^Y, \dots, C_{(i-1)j}^Y) = P(C_{ij}^X, C_{ij}^Y) \quad (7)$$

with  $X$  and  $Y$  denoting ancestor and descendant structures respectively. Models with long-range dependence among sites, including the dependent diffusion model (2) with general  $\Theta, \Sigma = LL'$ , do not exhibit these conditional independence relationships and therefore prohibit the recursive decomposition which forms the basis of efficient dynamic programming calculations. Since an evolutionary model without efficient alignment algorithms is far too expensive to use in the context of phylogenetic tree inference, we desire a model that incorporates site dependence while still preserving sufficient conditional independence structure to permit use of a forward-type algorithm.

## 2.4 Constructing a Dependent Structural Diffusion Model

A natural approach to introducing limited neighbor dependence into the diffusion model is to consider the backbone sites' coordinates as a series of nodes with forces acting upon each pair of neighboring sites, for example as in a ball and spring model. Figure 2 shows a general ball and spring model with spring constants  $k_{ij}$ . This model corresponds to a probability distribution for the equilibrium positions of the backbones coordinates which has precision matrix  $\Sigma^{-1} = (b_{ij})$  where  $b_{ij} = b_{ji}$ ,  $b_{ii} = k_{i-1,i} + k_{i,i+1}$  and  $b_{ij} = 0$  for  $|i - j| > 1$ .



**Fig. 2.** General ball and spring model for  $n$  backbone positions.

The corresponding Gaussian model with neighbor dependence is a spatial first-order auto-regressive process, denoted AR(1). However, setting the spring matrix equal to an AR(1) precision matrix gives a set of equations for the spring constants  $k_{ij}$  with no solution. We therefore instead approach the problem of incorporating dependence by starting with a general  $\Theta$  and  $\Sigma$  and determining what specific forms will correspond to an AR(1) process along the backbone.

We used symbolic algebra software to assist in solving for general matrices  $\Theta_p$  and symmetric, positive definite  $\Sigma_p$  such that the constraints  $\Lambda_\tau(i, j) = \Lambda_\infty(i, j) = 0 \quad \forall i, j : |i - j| > 1$  are satisfied for conditional and stationary

precision matrices  $A_\tau, A_\infty$ . Solutions to low-dimensional problems allowed us to identify the general form for a single pair of suitable  $\Theta_p, \Sigma_p$ . For five backbone positions this nearest-neighbor SDM takes the form:

$$\Theta_p = \theta \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \rho & 1 - \rho^2 & 0 & 0 & 0 \\ \rho^2 & -\rho^3 & 1 & 0 & 0 \\ \rho^3 & -\rho^4 & 0 & 1 & 0 \\ \rho^4 & -\rho^5 & 0 & 0 & 1 \end{pmatrix} \quad \Sigma_p = \sigma^2 \begin{pmatrix} 1 & a\rho & a\rho^2 & a\rho^3 & a\rho^4 \\ a\rho & 1 & \rho & \rho^2 & \rho^3 \\ a\rho^2 & \rho & 1 & \rho & \rho^2 \\ a\rho^3 & \rho^2 & \rho & 1 & \rho \\ a\rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \quad (8)$$

where  $a = (3 - \rho^2)/2$ . The conditional and stationary distributions given by (3) and (5) have tri-diagonal precision matrices. Thus dynamic programming is preserved, albeit with some modification to the standard pair HMM recursion formulas required as described in Sect. 2.5.

Similar computer algebra experiments were used to demonstrate that no such solutions exist for any diffusion of the form (2) where  $\Sigma_p = I$ . With  $\Theta = I_3 \otimes \Theta_p$  and  $\Sigma = I_3 \otimes \Sigma_p$ , (3-6) give the marginal or conditional distributions for matched positions.

## 2.5 Dynamic Programming

The recursive equations used for the pair hidden Markov model underlying the SIM [10] require several modifications in order to be used with the SDM. These modifications are specific to the form of  $\Theta$  and  $\Sigma = LL'$  chosen for the structural diffusion parameters. The primary reason for the changes is that the backbone coordinate emission probabilities in the SIM are independent of neighboring sites, whereas in the SDM the emission probabilities depend on neighboring sites. The details of the changes required to the dynamic programming algorithm are given in Appendix A.1.

## 2.6 Bayesian Inference for the Site-Dependent Model

Under the new site-dependent model specified by (2, 8), the joint distribution  $p(X, Y | \mathcal{M})$  of backbone coordinates for ancestor  $X$  and descendant  $Y$  given any alignment  $\mathcal{M}$  can be expressed

$$p(X, Y | \mathcal{M}) = \prod_{m \in M} p(X_{[m]}, Y_{[m]} | m, \mathcal{N}_m) \prod_{d \in D} p(X_{[d]}, Y_{[d]} | d, \mathcal{N}_d) \prod_{i \in I} p(X_{[i]}, Y_{[i]} | i, \mathcal{N}_i) \quad (9)$$

where  $M, D$ , and  $I$  respectively are the sets of matched, deleted, and inserted sites in  $\mathcal{M}$ .  $X_{[m]}$  denotes the backbone coordinates of the positions of  $X$  aligned in  $m \in M$ , and  $\mathcal{N}_i$  is the set of backbone positions neighboring position  $i$ . In other words  $p(X, Y | \mathcal{M})$  can be expressed in a decomposed form, each factor of which is either the joint density for a contiguous block of matches given its neighbors or the density of an insertion or deletion distribution for a particular site given its neighbors.

Bayesian inference based on this joint distribution (and that including indels) uses priors and sampling techniques detailed in [2] with trivial additions to accommodate priors and sampling for the model's dependence parameter  $\rho$ .

### 3 Joint Sequence-Structure Model for Phylogenetic Inference

Phylogenetic inference involves constructing a phylogenetic tree using estimates of the evolutionary distance between proteins, or equivalently models of the time-dependent evolution. Traditionally this is done using site-independent sequence evolution models parameterized by a matrix  $Q$  of relative substitution rates, defining a likelihood over the time  $\tau$  over which evolution occurs. The joint sequence-structure evolution model introduced by [2] multiplies this likelihood by one derived similarly from the time-dependent structure diffusion process (SIM) given by (1), allowing both structural and sequence differences to inform the estimation of divergence time  $\tau$ .

#### 3.1 Amino Acid Sequence Model

The sequence portion of our joint sequence and structure model is identical to that used in [2], where the joint likelihood for the two sequences  $S^X, S^Y$  and an alignment  $\mathcal{M}$  between them is given by

$$\begin{aligned} p(S^X, S^Y, \mathcal{M} | \lambda, \mu, \tau, Q) &= P(S^X, S^Y | \mathcal{M}, \tau, Q) P(\mathcal{M} | \lambda, \mu, \tau) \\ &= P(S_M^Y | S_M^X, \tau, Q) P(S_M^Y | \pi) \times P(S^X | \pi) P(\mathcal{M} | \lambda, \mu, \tau) \end{aligned} \quad (10)$$

where  $S_M^X, S_M^Y$  denote the matched (aligned) positions of the amino acid sequences  $S^X$  and  $S^Y$ ,  $S_M^Y$  the unmatched positions of  $S^Y$ ,  $Q$  the substitution rate matrix, and  $\pi$  the equilibrium distribution of amino acid labels. The probabilities  $P(S_M^Y | S_M^X, \tau, Q)$  are given by a product of independent substitution probabilities at each site via the transition probability matrix  $e^{Q\tau}$ .  $P(S_M^Y | \pi)$  and  $P(S^X | \pi)$  are given by the equilibrium distribution  $\pi$ , and we refer the reader to [2] for a discussion of the Links indel model which specifies  $P(\mathcal{M} | \lambda, \mu, \tau)$ .

#### 3.2 Site-Dependent Random Effect Model

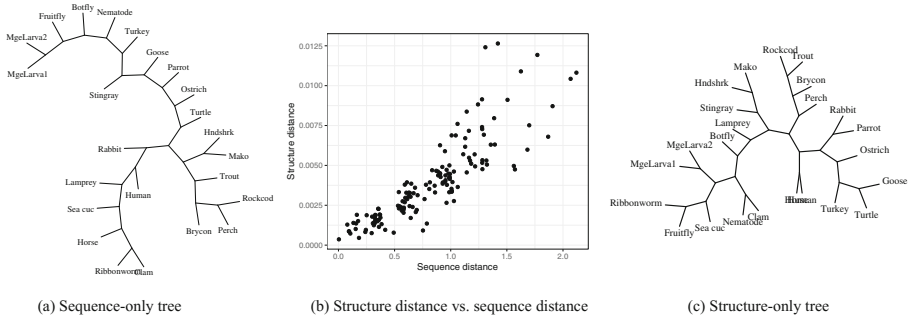
In a sequence evolution model (10), only the product  $Q\tau$  is identifiable - one cannot simultaneously estimate absolute rates and  $\tau$  itself. As a result, it is standard to scale the substitution rate matrix  $Q$  to a single expected substitution per unit time [11]. As a result, the time  $\tau$  is interpreted as the expected number of substitutions per site, which can be estimated from sequences. The structural model exhibits a similar identifiability issue: in pairwise estimation with a structure-only model, with neither rate  $\theta$  nor time  $\tau$  fixed, only the structural distance  $\theta\tau$  would be identifiable. In the Challis-Schmidler model this was not thought to



be a concern, since when the joint model is used  $\tau$  becomes determined by the sequence information, making  $\theta$  identifiable as well.

However this means that disagreement between the structural evolution model and sequence evolution model regarding the divergence time  $\tau$  will be resolved by compensation in the estimate of  $\theta$ . Because we do not currently have a computationally tractable site-dependent sequence evolution model, we do not wish the information in the structural SDM to be overridden by the site-independent sequence model, which we know to be susceptible to underestimation. We address this by introducing a distinct sequence time  $Q\tau = \tau_q$  and structural time  $\tau_s$  related by a stochastic model. This differs from the approach of [2, 3], which assumed a common time shared by both structural and sequence components of the likelihood.

The importance of distinguishing these two quantities is highlighted by the plot in Fig. 3, where we estimated divergence time separately using the sequence-only model of (10) and the independent structure-only model (see e.g. [2]) for a set of globins. There is a strong, arguably linear relationship between the structure-only evolutionary distance  $\theta\tau$  and the sequence-only evolutionary distance  $\tau$ , but the relationship between them is clearly noisy. Forcing the two models to share a common parameter ignores the different amounts of information and uncertainty provided about the evolutionary distance by sequence and structural data. The sequence-only and structure-only phylogenetic trees are shown as well, where we see the implications for tree topology.



**Fig. 3.** Pairwise sequence-only distance ( $\tau_q$ ) and structure-only distance ( $\theta\tau_s$ ) estimates from a set of 24 globin proteins under the SIM. The estimates are plotted against each other in panel (b) with the respective phylogenetic tree estimates (via neighbor-joining) in panels (a) and (c). In panel (b), we excluded pairs whose sequence distances could not be reliably estimated due to high sequence divergence.

Instead, we introduce a random effect model defining a stochastic linear relationship between sequence and structure distances:

$$(\theta\tau_s) = \beta\tau_q + \epsilon \quad \text{where } \epsilon \sim N(0, \omega^2). \quad (11)$$

Here  $\tau_s, \tau_q$  are the structural and sequence divergence times respectively. A simple linear regression gives  $\hat{\beta} = 0.005$  and an estimate for  $\omega$ . Under this formulation, the sequence model is now given by

$$\begin{aligned} p(S^X, S^Y, \mathcal{M} | \lambda, \mu, \tau_q, Q) &= P(S^X, S^Y | \mathcal{M}, \tau_q, Q) P(\mathcal{M} | \lambda, \mu, \tau_q) \\ &= P(S_M^Y | S_M^X, \tau_q, Q) P(S_M^Y | \pi) \times P(S^X | \pi) P(\mathcal{M} | \lambda, \mu, \tau_q) \end{aligned} \quad (12)$$

and the PDE governing the structural diffusion is

$$dC = -\Theta(C - \zeta) dt_s + L d\mathbf{B}_{t^{(s)}}. \quad (13)$$

To ensure the structure distance variable  $\tau_s$  is on a similar scale to  $\tau_q$ , in each pairwise estimation under this model we fix  $\theta$  at its posterior mean under the SIM. Hereafter we refer to this joint sequence and structure model with random effect as the SDMre.

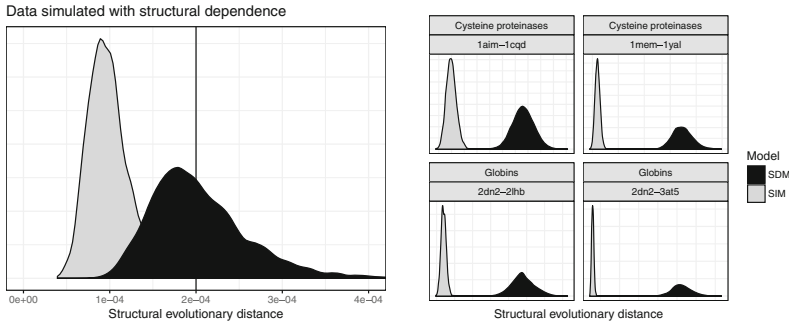
## 4 Results

All inferences were performed on the Duke Computer Cluster (DCC), a heterogeneous network of shared computing nodes; a typical node CPU is an Intel Xeon 2.6 GHz. Average runtimes for the SIM range from 20-60 iterations per second depending primarily on the length of the proteins, while SDM computations are roughly an order of magnitude slower than the SIM. All model parameters were sampled via random-walk Metropolis Hastings, augmented with a library sampling step for rotation parameter  $R$  as described in [2].

### 4.1 Improved Estimation of Evolutionary Distances

We first revisit the example of underestimation in the SIM, shown in Fig. 1(e). The left panel of Fig. 4 shows the posteriors from both the site-independent and site-dependent models. We see again that the SIM underestimates the true evolutionary distance, while the SDM corrects for this.

While this is not surprising on data simulated from the SDM, similar results are observed on real data for which the ‘true’ distance is unknown. The four plots at right in Fig. 4 compare the SIM and SDM posterior distributions for structural distance  $\theta\tau$  between two pairs of cysteine proteinases from [3] (top row) and two pairs of globins (human-turtle and human-lamprey, bottom row). In each pairwise estimation, the SIM is significantly underestimating structural distance relative to the SDM. This result is consistently observed across the other pairs of globins and cysteine proteinase pairs from [2, 3] (results omitted for brevity). In each case the SDM posterior is somewhat more diffuse, presumably due to the lower effective sample size in the structural information induced by dependence in the structural model. Although the ‘true’ distances for these pairs cannot be known, these results strongly suggest that including site dependence in the structural model can significantly reduce systematic bias in the estimated evolutionary distances.



**Fig. 4.** Estimation of evolutionary distance using SIM (light) and SDM (dark), for (a) simulated data with known true distance, and (b) real data from two cysteine proteinase pairs (b, top row) and two globins (b, bottom row). In all cases the SIM estimate is significantly lower than the SDM estimate, strongly suggesting systematic underestimation under the SIM assumption. Simulation parameters:  $\sigma^2 = 1$ ,  $\theta = 0.002$ ,  $t = 0.1$ ,  $\rho = 0.95$ .

*Non-neighbor dependence:* Proteins exhibit significant non-neighbor dependencies due to shared environments and physico-chemical interactions between amino acids that are distant in sequence but proximal in space. Simulations were run using general (non-banded) covariance matrices to simulate structural evolution with long-range correlations, with the SDM then used to estimate evolutionary distance. The results (omitted for brevity) are very similar to the left panel of Fig. 4: the SIM noticeably underestimates the true structural distance while the SDM accurately estimates it. This indicates the robustness of the nearest-neighbor approximation, required for efficient computation, to more general dependency patterns.

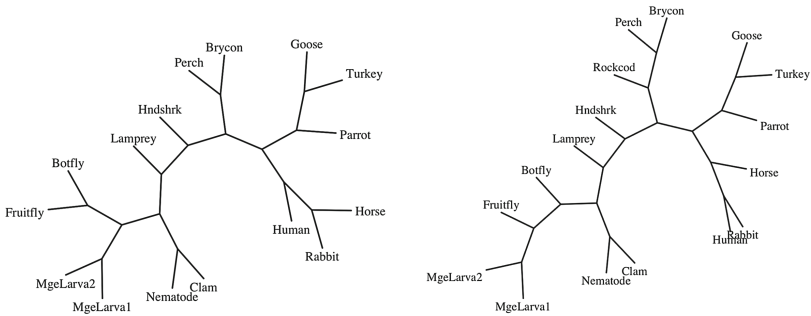
## 4.2 Effect on Phylogeny of Ignoring Structural Dependence in Globin Structures

Errors in estimation of pairwise evolutionary distances have the potential to undermine phylogenetic inference as well. To explore this, we compare phylogenetic trees reconstructed via neighbor-joining for a group of 16 globins using the SIM versus that obtained under the SDMre of Sect. 3. In each case, the respective model was used to estimate the pairwise distances for all pairs of proteins, and the resulting pairwise distance matrix was used to produce a neighbor-joining tree with the PHYLIP and Drawtree software [12]. Differences observed in these trees can be expected to also appear in trees if the SDM were used to replace the SIM component of the fully Bayesian joint sequence-structure tree estimation [3].

The phylogenetic trees estimated using posterior mean evolutionary distances are shown in Fig. 5. The SIM and SDMre trees are very similar, and neither matches the accepted NCBI taxonomy exactly. However, the SDMre tree improves upon the SIM tree in that botfly and fruitfly are now placed together

in a single clade with no other species, as in the NCBI taxonomy. This example demonstrates that phylogeny estimation can be adversely affected by ignoring structural dependence, even for proteins with high structure similarity such as these globins.

The SIM and SDMre models leading to the trees in Fig. 5 differ in two ways: incorporation of dependence in the diffusion, and incorporation of the random-effect relation between the sequence and structure time parameters. For comparison, we also ran the SIM with the random-effect incorporated, but without dependence in the diffusion model. This SIMre does not correctly group botfly and fruitfly, indicating that it is the site dependence which leads to the improved tree topology. For comparison, the sequence-only tree is also shown (for a superset of globins) in panel (a) of Fig. 3; it is highly inaccurate due to many pairs with highly divergent sequences. Without the structural component of the model included, these divergent sequences yield highly uncertain distance estimates which significantly destabilize the tree.



**Fig. 5.** The SDMre tree (left) improves upon the SIM tree (right) by grouping the botfly and fruitfly in their own clade, matching the accepted NCBI taxonomy.

## 5 Discussion

The site-dependent structural evolution model described here allows a significant improvement in model realism while retaining the computational tractability necessary for use in phylogenetic inference. As shown, the incorporation of dependence into the model significantly reduces bias in the estimates of evolutionary distance, and can have a resulting stabilizing effect on phylogenetic tree reconstruction. These results suggest a need for continued research on computationally efficient site-dependent *sequence* evolution models, which can be expected to further improve inference in these problems. This is because our current combined sequence-structure model pairs the site-dependent structural model with a site-*independent* sequence model, which likely still retains some downward bias on the estimated evolutionary distance due to the independence assumption in the sequence side of the model.

A natural next step will be to incorporate the site-dependent structural model presented here into the fully Bayesian simultaneous alignment and phylogeny reconstruction model of [3], which currently uses the site-independent structural model. This extension would be straightforward and may improve inference of multiple sequence alignments in addition to improving inference of phylogenetic trees.

**Acknowledgments.** This work was partially supported by NSF grant DMS-1407622 and NIH grant R01-GM090201 (S.C.S.). Jeffrey L. Thorne was supported by NIH grant GM118508. Gary Larson was partially supported by NSF training grant DMS-1045153 (S.C.S.).

## A Appendix

### A.1 Modified Dynamic Programming for a Pair HMM with Dependence

In the SDM, the dynamic programming equations' coordinate emission probabilities for each site will now involve preceding positions' coordinates. Because these probabilities are specified by distributions conditional on an alignment, we must know the form of the joint distribution  $p(X, Y|\mathcal{M})$  given any alignment  $\mathcal{M}$ .

In our model, as in [2], a pair HMM is used to model the distribution of pairwise alignments between two proteins. As described in [10], the use of a pair HMM allows one to calculate the probability of two protein structures marginalized over all possible alignments between the two structures. This is accomplished via dynamic programming by using the well-known forward algorithm to recursively calculate values of  $f^k(i, j)$  (i.e., the total probability of all partial alignments through position  $(i, j)$  in the ancestor  $(i)$  and descendant  $(j)$  that end in state  $k \in \{\text{Match, Delete, Insert}\}$ ). The forward equations typically used for this purpose are presented in [10] as:

$$f^M(i, j) = p_{X_i, Y_j} \cdot (a_{MM}f^M(i-1, j-1) + a_{DM}f^D(i-1, j-1) + a_{IM}f^I(i-1, j-1)) \quad (14)$$

$$f^D(i, j) = p_{X_i} \cdot (a_{MD}f^M(i-1, j) + a_{DD}f^D(i-1, j) + a_{ID}f^I(i-1, j)) \quad (15)$$

$$f^I(i, j) = p_{Y_j} \cdot (a_{MI}f^M(i, j-1) + a_{DI}f^D(i, j-1) + a_{II}f^I(i, j-1)) \quad (16)$$

where  $p_{X_i, Y_j}, p_{X_i}, p_{Y_j}$  are the three emission probabilities for (respectively): a matched pair  $X_i, Y_j$ , a deletion  $X_i$ , and an insertion  $Y_j$ . Terms of the form  $a_{JK}$  give the probability of transition from state  $J$  to state  $K$  in the pair HMM. The emission probability terms  $p_{X_i, Y_j}, p_{X_i}$  and  $p_{Y_j}$  involve only the sites denoted and are independent of neighboring sites<sup>1</sup>. The SDM emission probabilities are not independent of other sites, so the forward equations must be modified.

<sup>1</sup> For a detailed explanation of the standard forward equation terms we refer the reader to the pair HMM material in [10].

To illustrate the set of changes needed, we focus only on the Match equation (14); analogous changes are required for the other two recursive equations. Equation (14) gives the total probability of all alignments up to position  $(i, j)$  which end with a Match at position  $(i, j)$ . The three terms on the right hand side arise because a path through the pair HMM could arrive at a Match at  $(i, j)$  from one of three previous states in the path: either a Match, Delete, or Insert at  $(i - 1, j - 1)$ . The term  $p_{X_i, Y_j}$  is a single factor on the right hand side, indicating that the Match emission probability at  $(i, j)$  is the same regardless of the previous state in the path. In our case, the Match emission probability at  $(i, j)$  depends on the previous state in the path. Accordingly, the first step in modifying the equation for our purposes is to define unique emission probabilities that depend on the previous state in the path through the pair HMM. We write the site-dependent version of (14) as

$$\begin{aligned} f^M(i, j) &= (\bar{p}_{X_i, Y_j}^M) \cdot a_{MM} f^M(i - 1, j - 1) \\ &+ (\bar{p}_{X_i, Y_j}^D) \cdot a_{DM} f^D(i - 1, j - 1) \\ &+ (\bar{p}_{X_i, Y_j}^I) \cdot a_{IM} f^I(i - 1, j - 1) \end{aligned} \quad (17)$$

where the superscripts on  $\bar{p}$  terms indicate the previous state before the Match at  $X_i, Y_j$ . The modified equations for  $f^D(i, j)$  and  $f^I(i, j)$  are analogous. Any of the emission distributions  $\bar{p}$  can be derived by first writing down the joint distribution for the appropriate backbone positions given an alignment (see Sect. 2.6) and then conditioning on that multivariate normal distribution as needed.

When determining the emission distributions, obvious edge cases must be dealt with. In addition, note that the emission distribution for a matched pair given a previous Match ( $\bar{p}_{X_i, Y_j}^M$ ) depends on where in the alignment the emitted matched pair occurs. In other words, calculation of  $\bar{p}_{X_i, Y_j}^M$  should take into account two possibilities: one, that the state prior to the previous Match was also a Match, or two, that it was an Insertion or Deletion. This can be verified by considering the joint distribution for 3 consecutive matched pairs and noting that the distribution of the 2nd matched pair conditional on previous positions is different than the distribution of the 3rd matched pair conditional on previous positions. This characteristic arises due to the specific forms chosen for the OU process'  $\Theta$  and  $\Sigma$  in our site-dependent model. Thus, the term  $\bar{p}_{X_i, Y_j}^M$  in (17) will itself be calculated as a sum over possible states preceding the prior state:

$$\begin{aligned} \bar{p}_{X_i, Y_j}^M &= \bar{p}_{X_i, Y_j}^{(M)_1} [f^D(i - 2, j - 2) \cdot a_{DM} \cdot \bar{p}_{X_{i-1}, Y_{j-1}}^D \\ &+ f^I(i - 2, j - 2) \cdot a_{IM} \cdot \bar{p}_{X_{i-1}, Y_{j-1}}^I] \cdot a_{MM} \\ &+ \bar{p}_{X_i, Y_j}^{(M)_2} [f^M(i - 2, j - 2) \cdot a_{MM} \cdot \bar{p}_{X_{i-1}, Y_{j-1}}^M] \cdot a_{MM}. \end{aligned} \quad (18)$$

The presence of the recursive term  $\bar{p}_{X_{i-1}, Y_{j-1}}^M$  in the equation above requires that an additional dynamic programming matrix be tracked. There are no other emission probabilities which depend on more than one previous hidden state of the pair HMM.

**Derivation of Emission Probabilities.** Suppose  $\mathcal{M}_p$  is a known partial alignment of all matches, aligning  $n$  positions  $X_i$  through  $X_{i+n-1}$  to positions  $Y_j$  through  $Y_{j+n-1}$  with no indels. The joint distribution of these backbone coordinates  $p(X_{i,i+n-1}, Y_{j,j+n-1} | \mathcal{M}_p)$  has a block covariance matrix:

$$p(X_{i,i+n-1}, Y_{j,j+n-1} | \mathcal{M}_p) \sim N \left( \mathbf{0}, \begin{pmatrix} \Sigma_{n \times n} & R^T \\ R & \Sigma_{n \times n} \end{pmatrix} \right) \quad (19)$$

where  $\Sigma_{n \times n}$  is equal to the stationary OU solution obtained using (8) and  $R$  is  $n \times n$ , equal to:

$$R = \frac{\sigma^2 e^{-\theta\tau}}{2\theta} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho k & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 k & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} k & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}$$

with  $k = \frac{1 - (1 - \rho^2)e^{\theta\rho^2\tau}}{\rho^2}$ . The emission probability for an Insertion  $Y_j$  or Deletion  $X_i$  at a particular site given its previous neighbor has an AR(1) form:

$$p(X_i | X_{i-1}, Y_j) \equiv p(X_i | X_{i-1}) \sim N(\rho X_{i-1}, \sigma^2(1 - \rho^2)) \quad (20)$$

$$p(Y_j | X_i, Y_{j-1}) \equiv p(Y_j | Y_{j-1}) \sim N(\rho Y_{j-1}, \sigma^2(1 - \rho^2)). \quad (21)$$

The joint distribution  $p(X, Y | \mathcal{M})$  can be specified by combining these insertion and deletion distributions with the distribution for contiguous matches in (19). Then, the nine dynamic programming emission distributions can be verified using standard techniques for conditioning multivariate normal distributions.

## A.2 Dependent Binary Sequence Model

Let  $\sigma$  represent a length  $n$  binary sequence. The space of all  $2^n$  possible sequences is  $\Omega = \{\sigma_1, \sigma_2, \dots, \sigma_{2^n}\}$ . A given sequence  $\sigma_i$  consists of  $n-1$  pairs of neighboring labels. To characterize members of  $\Omega$ , let  $k_i$  denote the number of neighbor pairs in  $\sigma_i$  with identical labels ( $k$  for “keeps” the same label from one position to the next), and let  $c_i$  denote the number of neighbor pairs in  $\sigma_i$  with different labels ( $c$  for “changes”). Now define  $\lambda_{\sigma_i} := k_i - c_i$ . We can refer to  $\lambda_\sigma$  as a degree of dependence: for sequences with  $\lambda_\sigma > 0$ , more than half the neighboring label pairs will have the same label and overall the sequence labels will appear non-randomly distributed along the sequence. If  $\lambda_{\sigma_i} < 0$ , the sequence will look more like a uniform distribution of labels.

To construct a simple model for site-dependent binary sequence evolution, we construct a Markov chain on the state space of binary sequences such that the transitions are site-dependent. We first specify a set of (identical) transition rates  $\{a_i\}$  and a corresponding probability jump matrix  $P$  having entries  $P_{ij}$ . The generator  $Q$  for the corresponding Markov chain has entries  $Q_{ij} = a_i P_{ij}$ . In defining  $P$ , we follow the convention that multiple substitutions cannot occur

simultaneously, so that the  $(i, j)$  entry of  $Q$  and  $P$  will be 0 if the configurations  $\sigma_i, \sigma_j$  differ at more than one position. To induce dependence into such a model, we set  $Q_{ij} = b^{\lambda_{\sigma_j} - \lambda_{\sigma_i}} / Z_i$  with  $b \geq 1$  an adjustable parameter controlling the strength of neighbor dependence ( $b = 1$  represents neighbor independence) and  $Z_i$  a normalizing constant for the row such that the off-diagonal row elements sum to 1. Suppose the Markov chain is currently in state  $i$ . After an exponential waiting time elapses (given by rate  $a_i$ ), the Markov chain is more likely to transition to states  $j$  having larger  $\lambda_{\sigma_j} - \lambda_{\sigma_i}$  than to states  $j$  having smaller  $\lambda_{\sigma_j} - \lambda_{\sigma_i}$ . In other words, in this model a binary sequence is more likely to evolve into a sequence with a more contiguous blocks of identical labels than into a sequence where the sequence labels are uniformly distributed along the sequence length.

## References

1. Wang, S., Ma, J., Peng, J., Xu, J.: Protein structure alignment beyond spatial proximity. *Sci. Rep.* **3**, 1448 (2013). <https://doi.org/10.1038/srep01448>
2. Challis, C.J., Schmidler, S.C.: A stochastic evolutionary model for protein structure alignment and phylogeny. *Mol. Biol. Evol.* **29**(11), 3575–3587 (2012). <https://doi.org/10.1093/molbev/mss167>
3. Herman, J.L., Challis, C.J., Novák, A., Hein, J., Schmidler, S.C.: Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. *Mol. Biol. Evol.* **31**(9), 2251–2266 (2014). <https://doi.org/10.1093/molbev/msu184>
4. von Haeseler, A., Schöniger, M.: Evolution of DNA or amino acid sequences with dependent sites. *J. Comput. Biol.* **5**(1), 149–163 (1998). <https://doi.org/10.1089/cmb.1998.5.149>
5. Arenas, M.: Trends in substitution models of molecular evolution. *Front. Genet.* **6**, 319 (2015). <https://doi.org/10.3389/fgene.2015.00319>
6. Schmidler, S.C.: *Bayesian Statistics*, vol. 8. Oxford University Press, New York (2006)
7. Wang, R., Schmidler, S.C.: Bayesian multiple protein structure alignment. In: Sharan, R. (ed.) *RECOMB 2014*. LNCS, vol. 8394, pp. 326–339. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-05269-4\\_27](https://doi.org/10.1007/978-3-319-05269-4_27)
8. Cheng, H., Kim, B.H., Grishin, N.V.: MALIDUP: a database of manually constructed structure alignments for duplicated domain pairs. *Proteins* **70**(4), 1162–1166 (2008). <https://doi.org/10.1002/prot.21783>
9. Thorne, J.L., Kishino, H., Felsenstein, J.: An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33**(2), 114–124 (1991). <https://doi.org/10.1007/BF02193625>
10. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University, Cambridge (1998). <https://doi.org/10.1110/ps.8.3.695>
11. Kosiol, C., Goldman, N.: Different versions of the Dayhoff rate matrix. *Mol. Biol. Evol.* **22**(2), 193–199 (2005). <https://doi.org/10.1093/molbev/msi005>
12. Felsenstein, J.: *Phylip - phylogeny inference package (version 3.2)*. Cladistics (1989)