Probabilistic modelling of protein structure evolution

Michael Golden[1], Thomas Hamelryck[2], and Oliver Pybus[1]

[1]Department of Zoology, University of Oxford, UK

[2]Bioinformatics Centre, Section for Computational and RNA Biology, Department of Biology and Image Section, Department of Computer Science, University of Copenhagen

May 11, 2019

Abstract

We present a probabilistic model of protein evolution that captures several important features of protein sequence and local structure. The key feature being dependencies between neighbouring amino acid positions that are temporal in nature due to sequence mutations that occur during evolution. The model is trained on a large number of protein alignments and corresponding phylogenetic trees that represent the evolutionary history of the aligned proteins. This yields a model that acts as a rich prior distribution over protein evolution. Our model provides a complete probabilistic description of protein backbone structures using an angle and bond length representation. Structure evolution is modelled jointly with sequence, permitting ancestral structures and sequences to be reconstructed in a phylogenetically rigorous manner. Likewise, the model can perform homology modelling to predict the unknown local backbone structure of a known protein sequence using additional information from potentially large numbers of homologous proteins. The model is highly flexible with respect to input, implying that arbitrary combinations of protein sequences and structures can be used when performing various inference tasks. Our current model does not capture global features of protein structure that are necessary for accurate homology modelling or reconstruction of ancestral three-dimensional structures. However, it is ultimately expected to be combined with protein structure prediction models that account for such long-range dependencies, but that do not account for evolutionary information.

1 Introduction

Recently, several studies (Challis and Schmidler, 2012; Herman *et al.*, 2014) have proposed joint stochastic models of evolution which take into account simultaneous alignment of protein sequence and structure. These studies point out the limitations of earlier non-probabilistic methods, which often rely on heuristic procedures to infer parameters of interest. A major disadvantage of using heuristic procedures is that they typically fail to account for sources of uncertainty. For example, relying on a single fixed alignment, which is highly unlikely to be the *true* underlying alignment, may bias the inference of the posterior distribution over evolutionary trees.

We present a generative evolutionary model, ETDBN (Evolutionary Torus Dynamic Bayesian Network), for pairs of homologous proteins. ETDBN captures dependencies between sequence and structure evolution, accounts for alignment uncertainty, and models the local dependencies between aligned sites.

A key step in modelling protein structure evolution is selecting a suitable structural representation and corresponding evolutionary model. Early works by Gutin and Badretdinov (1994) and Grishin (1997) represented protein structure using three-dimensional Cartesian coordinates of protein backbone atoms and used diffusions processes to model the relationship between structural distance (measured using RMSD) and sequence similarity. More recent publications by Challis and Schmidler (2012) and Herman *et al.* (2014) likewise used the three-dimensional Cartesian coordinates of amino acid C_α atoms to represent protein structure and additionally used Ornstein–Uhlenbeck (OU) processes to construct Bayesian probabilistic models of protein structure evolution. These models emphasise estimation of evolutionary parameters such as the evolutionary time between species, tree topologies and alignment, and attempt to fully account for sources of uncertainty. For the sake of computational tractability, the aforementioned approaches treat the Cartesian coordinates associated with atoms as evolving independently of another. A non-probabilistic approach by Echave (2008) and Echave and Fernández (2010) referred to as the Linearly Forced Elastic Network Model (LFENM) treats protein structures as a collection of C_α atoms connected by spring forces. The major benefit of LFENMs is that they do not assume independence of atomic coordinates and take into account non-local structural dependencies. In their current formulation LFENMs do not capture the dependencies between amino acid sequence and structure and therefore do not account for the variable effect of sequence mutation on protein structure evolution.

Rather than using a Cartesian coordinate representation, our model, ETDBN, uses a dihedral angle representation motivated by the non-evolutionary TorusDBN model (Boomsma *et al.*, 2008, 2014). TorusDBN represents a single protein structure as a sequence of (ϕ, ψ) dihedral angle pairs, which are modelled using continuous bivariate angular distributions (Frellsen *et al.*, 2012). Likewise, ETDBN treats protein structure as a random walk in space, again making use of the ϕ and ψ dihedral angles (Figure 1 above).

The dihedral angle representation is informed by the chemical nature of peptide bonds. Each amino acid in a protein peptide chain is covalently bonded to the next via a peptide bond. Peptide bonds have a partial double bond nature that results in a planar configuration of atoms in space. This configuration allows the protein backbone structure to be largely described as a series of ϕ and ψ dihedral angles that defines the relationship between the planes in three-dimensional space. Because each amino acid (except for the N and C terminus amino acids) is associated with a (ϕ,ψ) dihedral angle pair, only a sequence alignment is necessary to compare structures. On the other hand, models that use Cartesian coordinates typically need to superimpose structures in addition to requiring a sequence alignment (Herman *et al.*, 2014). Such superimposition introduces an additional source of uncertainty. A further advantage of the dihedral angle representation is that there are fewer degrees of freedom per amino acid and therefore typically fewer parameters required to model their evolution.

A novel stochastic diffusion process developed in García-Portugués *et al.* (2017) was used to model the evolution of dihedral angles. Additionally, a coupling was introduced such that an amino acid change can lead to a *jump* in dihedral angles and a change in diffusion process. This allows the model to capture changes in amino acid that are directionally coupled with changes in dihedral angle or secondary structure. As in Challis and Schmidler (2012) and Herman *et al.* (2014), the insertion and deletion (indel) evolutionary process is also modelled to account for alignment uncertainty (Thorne *et al.*, 1992).

The Ornstein–Uhlenbeck processes used in Challis and Schmidler (2012) and Herman *et al.* (2014) ignore bond lengths and treat C_α atoms as evolving independently for the sake of computationally tractability. Furthermore, the OU process makes Gaussian assumptions. From a generative perspective these properties will lead to evolved proteins with C_α atoms that are unnaturally dispersed in space. Bond lengths are also ignored in ETDBN, but can be plausibly fixed, or modelled. As a result, it is expected that the use of angular diffusions will more naturally capture the underlying local protein structure manifold; notwithstanding, global structural constraints are not imposed which may also lead to simulated proteins that lack certain global properties of protein structure, such as protein compactness.

Tree-like dependencies typically arise when considering two or more homologous proteins related via a common ancestor. These dependencies manifest themselves most noticeably in the degree of amino acid sequence similarity between the homologous proteins. The strength of these dependencies is assumed to be a result of two major factors: the time since the common ancestor and the rate of evolution.

Most models of structural evolution ignore dependencies amongst sites because of the increased computational demand and model complexity associated with such models. These dependencies are expected to influence patterns of evolution, specifically patterns of amino acid substitution. The current model deals with local dependencies only – dependencies that are expected to arise due to interactions between neighbouring amino acids, for example, between amino acids in an α-helix. ETDBN does not account for global dependencies – dependencies that result in the globular nature of proteins (Boomsma *et al.*, 2008). In ETDBN, we attempt to model local dependencies only by using a Hidden Markov Model (HMM) to capture dependencies amongst neighbouring aligned positions. HMMs such as PASSML (Liò *et al.*, 1998) have been successfully used to predict protein secondary structure from aligned sequences; however, these models typically have the disadvantage that they assume a canonical secondary structure shared amongst all the sequences being analysed. This restricts analysis to closely related sequences where conservation of secondary structure is a reasonable assumption. ETDBN does not assume a canonical secondary structure, but instead uses a phylogenetic HMM approach (similar to Siepel and Haussler (2004)) that assumes dependencies between evolutionary processes at neighbouring aligned positions.

Parameters of ETDBN were estimated using 1200 homologous protein pairs from the HOMSTRAD database (Mizuguchi *et al.*, 1998). The resulting model provides a realistic prior distribution over proteins and protein structure evolution in comparison to previous stochastic models. Doing so enables biological insights into the relationship between sequence and structure evolution, such as patterns of amino acid change that are informative of patterns of structural change (Grishin, 2001). It was with these features in mind that ETDBN was developed.

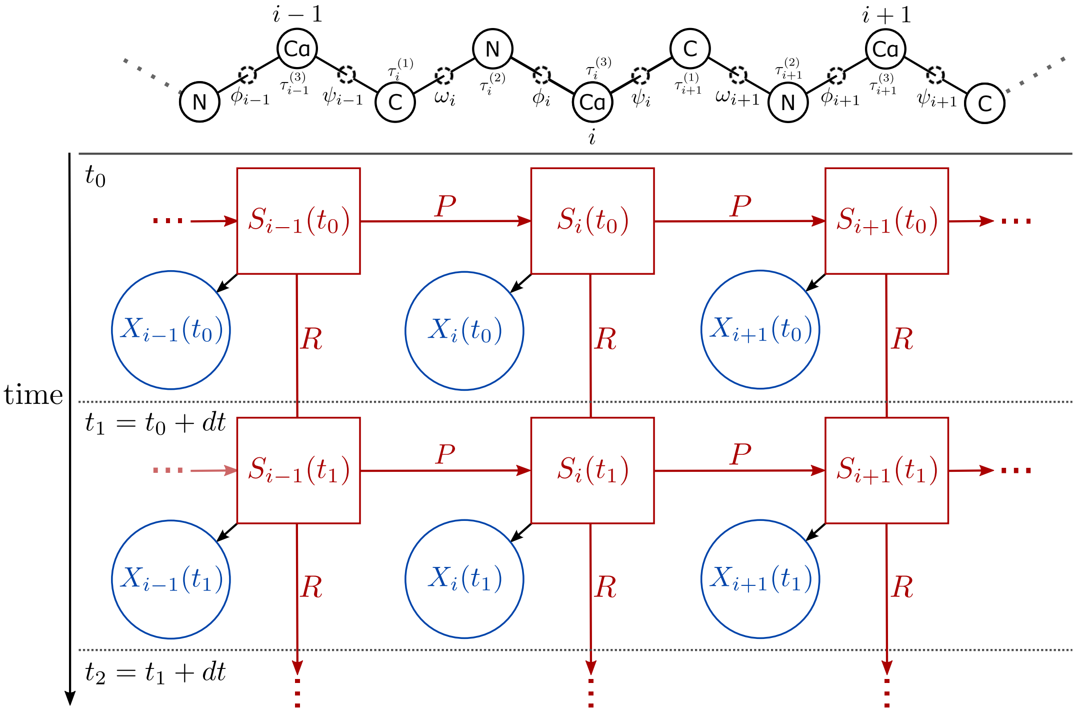

Figure 1: Above: a depiction of a protein backbone (three amino acids long) with the ω, ϕ and ψ dihedral angles and the three additional bond angles ($\tau_i^{(1)}$, $\tau_i^{(2)}$, $\tau_i^{(3)}$) shown. Bond lengths are implicit. Bond angles and bond lengths are not to scale. Also shown are C_α atoms which attach to the amino acid side-chains. Each amino acid side-chain determines the characteristic nature of each amino acid. Every amino acid position corresponds to a hidden node in the model below.
Below: Graphical depiction of the model architecture showing three amino acid positions ($i-1$, i, and $i+1$) at two time instants (t_0 and t_1) along a single branch of a phylogenetic tree. Note that $S_i(t) = \big(H_i(t), A_i(t)\big)$.

2 Methods

2.1 Model

2.1.1 Model of a single protein

A single protein consisting of n amino acids:

$$\begin{aligned} P_a &= (H_a, A_a, X_a) \\ &= \langle (H_a^1, A_a^1, X_a^1), \ldots, (H_a^n, A_a^n, X_a^n) \rangle \end{aligned}$$

is a sequence of aligned sites where each site i is associated with a discrete-valued hidden state, H_a^i (taking on one of h possible values), a discrete-valued amino acid observation, A_a^i (representing one of the twenty possible amino acids), and a corresponding vector of continuous-valued structural observations X_a^i representing the backbone structure of the protein.

Structural observations The set of structural observations, X_a^i, at a particular site, i, consists of nine continuous-valued variables: three dihedral angles (ϕ_i, ψ_i, ω_i), three additional bond angles ($\tau_i^{(1)} = \widehat{C\alpha_{i-1}, C_{i-1}, N_i}$, $\tau_i^{(2)} = \widehat{C_{i-1}, N_i, C\alpha_i}$, $\tau_i^{(3)} = \widehat{N_i, C\alpha_i, C_i}$), and three bond lengths ($b_i^{(1)} = \overrightarrow{C_{i-1}, N_i}$, $b_i^{(2)} = \overrightarrow{N_i, C\alpha_i}$, $b_i^{(3)} = \overrightarrow{C\alpha_i, C_i}$).

Note that ϕ_1, $\tau_1^{(1)}$, and $\tau_1^{(2)}$, are undefined at the first position in the peptide backbone of an unaligned protein. Similarly, ψ_n and ω_n are undefined for the last position, n, in each unaligned protein.

Given the structural observations, X_a, it is possible to exactly reconstruct the three-dimensional coordinates of each atom in a protein's backbone (Parsons *et al.*, 2005).

The ω_i dihedral angle (which determines the cis/trans conformation) at each site i is assumed to be distributed according a univariate von Mises (vM) distribution with mean μ_ω and concentration parameter κ_ω conditional on the hidden state H_a^i and the amino acid A_a^i:

$$\omega_i \quad \sim \quad \text{vM}\big(\mu_\omega(H_a^i, A_a^i),\, \kappa_\omega(H_a^i, A_a^i)\big). \tag{1}$$

The ϕ_i and ψ_i dihedral angles are assumed to be drawn from a bivariate von Mises (bvM) distribution with mean vector $\mu_{\phi,\psi} = \langle \mu_\phi, \mu_\psi \rangle$ and covariance parameters $\kappa_{\phi,\psi} = \langle \kappa_1, \kappa_2, \kappa_3 \rangle$:

$$(\phi_i, \psi_i) \quad \sim \quad \text{bvM}\big(\mu_{\phi,\psi}(H_a^i, A_a^i),\, \kappa_{\phi,\psi}(H_a^i, A_a^i)\big), \tag{2}$$

where κ_1 is the variance associated with ϕ, κ_2 is the variance associated with ψ, and κ_3 is the correlation between ϕ and ψ,

The three additional bond angles ($\tau_i^{(1)}$, $\tau_i^{(2)}$, $\tau_i^{(3)}$) are each distributed according a univariate (vM) distribution conditional on the hidden state H_a^i only:

$$\begin{array}{rcl} \tau_i^{(1)} & \sim & \text{vM}\big(\mu_{\tau^{(1)}}(H_a^i),\, \kappa_{\tau^{(1)}}(H_a^i)\big) \\ \tau_i^{(2)} & \sim & \text{vM}\big(\mu_{\tau^{(2)}}(H_a^i),\, \kappa_{\tau^{(2)}}(H_a^i)\big) \\ \tau_i^{(3)} & \sim & \text{vM}\big(\mu_{\tau^{(3)}}(H_a^i),\, \kappa_{\tau^{(3)}}(H_a^i)\big). \end{array} \tag{3}$$

The three bond lengths ($b_i^{(1)} > 0$, $b_i^{(2)} > 0$, $b_i^{(3)} > 0$) are distributed according a truncated Multivariate Normal (MVN) with mean vector, μ, of length 3 and a 3×3 covariance matrix, Σ, conditional on the hidden state H_a^i:

$$(b_i^{(1)}, b_i^{(2)}, b_i^{(3)}) \sim \text{MVN}\big(\mu(H_a^i),\, \Sigma(H_a^i)\big). \tag{4}$$

Note that the parameters in (3) and (4) are no longer conditional upon the amino acid observations A_a^i. This was done to reduce the number of model parameters, as the values of the three bond angles and three bond lengths are all largely invariant, implying that the values can be reasonably fixed. Despite this, we still opted to treat them as random variables so that the model gives a complete probabilistic description of a protein backbone structure.

Site likelihood The likelihood of a structural observation, $p\big(X_a^i \,|\, H_a^i, A_a^i, \hat{\theta}\big)$, at site i conditional on the hidden state, H_a^i, and amino acid, S_a^i, is given by a product of the densities in Equations 1, 2, 3, and 4.

Hidden Markov model A sequence of structural observations representing a single protein backbone structure is modelled using a Hidden Markov Model (HMM). Hidden states in the HMM are primarily intended encode the angle and bond lengths distribu-

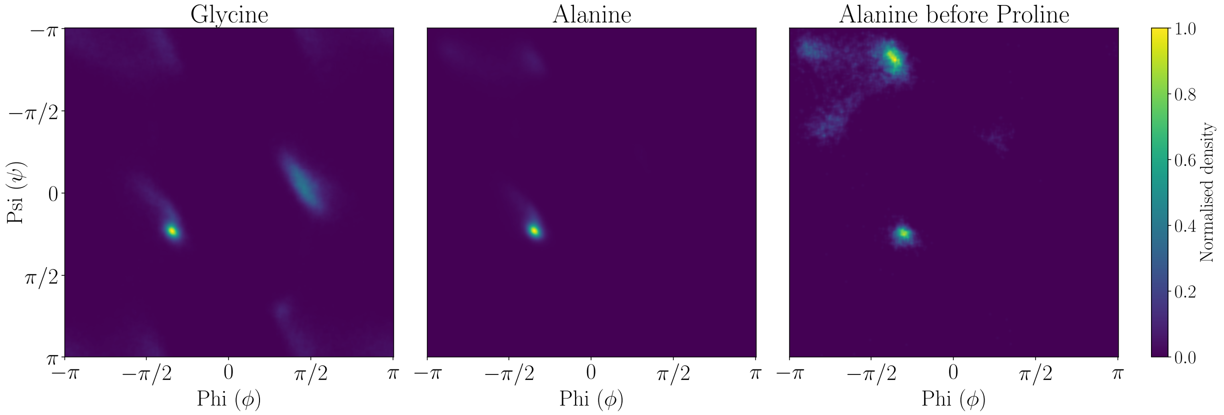

Figure 2: Ramachandran plots depicting empirical distributions of ϕ and ψ angle combinations under three different amino acid contexts. Glycine has the smallest amino acid side-chain and is therefore the least constrained. Alanine has a larger side-chain than glycine, constraining most ϕ and ψ angle to lie near a single peak. When alanine precedes proline in the peptide backbone, the ϕ and ψ angle combinations previously favoured become sterically hindered MG: is this the correct terminology?, favouring angle combinations at multiple peaks.

tions and their association with the different amino acids as specified in Equations 1-4. However, the HMM is also critically to capture neighbouring dependencies, such as steric effects on dihedral angle conformations (Figure 2). These neighbouring dependencies are captured using a $h \times h$ transition probability matrix $P = p(H_a^i \,|\, H_a^{i-1}, \hat{\theta})$.

We let $\Pi(a)$ denote the joint likelihood of a sequence of hidden states (H_a) and structural observations (X_a) conditional on a sequence of amino acids (A_a) defined as follows:

$$\begin{aligned} \Pi(a) = p(H_a, A_a, X_a \,|\, \hat{\theta}) &= \\ p\big(A_a^1, X_a^1 \,|\, H_a^1, \hat{\theta}\big) p(H_a^1 \,|\, \hat{\theta}) & \\ \times \prod_{i=2}^{n} p\big(A_a^i, X_a^i \,|\, H_a^i, \hat{\theta}\big) p(H_a^i \,|\, H_a^{i-1}, \hat{\theta}), & \end{aligned} \tag{5}$$

where $p(H_a^1 \,|\, \hat{\theta})$ is the initial probability of starting in state H_a^1 at the first site. Whilst $\Pi(a)$ describes a single protein, it is used in the next section to ‘weight’ an evolutionary model such that protein evolutionary trajectories are visited with probability proportional to $\Pi \times \pi$, where π is weighting corresponding to the amino acid sequences.

2.1.2 Evolutionary model

Thus far we have only considered a single protein. In this section we outline how multiple phylogenetically related proteins are modelled evolutionarily. Following Choi *et al.* (2008) we construct a rate matrix that represents changes between two sequences, a and b, instead of character states, such as amino acids, as is typical of substitution models. Furthermore, each sequence position combines a hidden state (h_i), an amino acid (aa_i), and a set of structural observations (x_i), into a joint character state (h_i^s, aa_i^s, x_i^s), where s refers to a particular sequence. The rate matrix is given as follows:

$$R_{ab} = \begin{cases} \sqrt{\frac{\Pi(b)}{\Pi(a)}} U_{a_i b_i} \pi^{h_i^b}_{aa_i^b} & \text{Single amino acid difference at site } i. \\ \sqrt{\frac{\Pi(b)}{\Pi(a)}} V_{a_i b_i} \pi^{h_i^b}_{aa_i^b} & \text{Single hidden state difference at site } i. \\ 0 & \text{Both hidden state and amino acid differences at site } i. \\ 0 & \text{Differences at two or more sites.} \\ -\sum\limits_{k \neq a} R_{ak} & a = b \end{cases} \tag{6}$$

where i is the position that differs between sequence a and b, U is a symmetric 20×20 amino acid exchangeability matrix, and V is a symmetric $h \times h$ hidden state exchangeability matrix. The term, $\frac{\Pi(b)}{\Pi(a)}$, weights each hidden state or amino acid change, such that sequences are visited with probability given by it's probability under the hidden Markov model multiplied by the amino acid sequence probability. This gives an evolutionary model on amino acid sequences and protein structures that accounts for neighbouring dependencies between adjacent sites and introduces temporal evolutionary dependencies between proteins.

The evolutionary dependencies between structures is introduced via the hidden states, thus avoiding having to directly implement an evolutionary process on structure, which is cumbersome given the continuous nature of the structural observations. We have previously developed a continuous diffusion process on angles for modelling protein dihedral angles, (García-Portugués *et al.*, 2018; Golden *et al.*, 2017), however, protein backbone angles and bond lengths do not evolve in a continuous fashion, rather they are expected to 'jump' when changes occur. The hidden states capture this jump behaviour.

Note that proteins P_a and P_b referred to in the ratio $\frac{\Pi(b)}{\Pi(a)}$ in Equation 6 always differ at exactly one site, implying that at most three terms in Equation 5 need to be considered when computing the ratio.

Additionally note, although the summation in Equation 6 appears to involve an exponential number of terms, most terms are equal to zero, except those that differ from P_a at one position. Furthermore, an amino acid transition and a hidden state transition are not permitted to occur simulateneously, further reducing the number of terms that are summed ($19n$ amino acid terms plus $(h-1)n$ hidden state terms).

Stationary probability of proteins Following Choi *et al.* (2008), and by construction, the stationary probability of a protein a is given by:

$$p(P_a|\hat{\theta}) = \frac{\Pi(a) \prod\limits_i \pi^{h_i^a}_{aa_i^a}}{\sum_k \Pi(k) \prod\limits_i \pi^{h_i^k}_{aa_i^k}} \tag{7}$$

Time-reversibility Since U and V in (6) are symmetric matrices, i.e. $U_{a_i b_i} = U_{b_i a_i}$ and $V_{a_i b_i} = V_{b_i a_i}$, time-reversibility of the model holds, in other words:

$$p(P_a|\hat{\theta}) M_{ab} = p(P_b|\hat{\theta}) M_{ba}. \tag{8}$$

Time-reversibility implies that at any rooting of the tree can be used if the equilibrium probabilities are taken to be the initial probabilities (Felsenstein, 1981), which is indeed the case for our model.

Dataset likelihood The likelihood of a given dataset $\mathcal{D}_d$ of proteins related by a tree $\mathcal{T}_d$ consisting of a set of branch paths $\mathcal{B}_d$ is given as follows:

$$\begin{aligned} p(\mathcal{D}_d|\mathcal{T}_d, \mathcal{B}_d, \hat{\theta}) &= \\ p(P_{root}) \prod_{b \in \mathcal{B}_d} p\big(X_b(t_{end}) \,|\, H_b(t_{end}), A_b(t_{end}), \hat{\theta}\big) \\ \times \big[e^{R_{b_n}(t_{end}-t_n)} \prod_{k=1}^{n} e^{-R_{b_k b_k}(t_k - t_{k-1})} R_{b_{k-1} b_k}\big]. \end{aligned} \tag{9}$$

The first term, $p(P_{root})$, is the probability of the protein, P_{root}, at the root of the tree. The outermost product is a product over branches in $\mathcal{B}_d$, where the first term is the likelihood of any structural observations at the tip of each branch. The terms in square brackets represent the likelihood of the the hidden states and amino acids along a branch paths, as specified the rate matrix R. The first term in square parentheses is the probability that no events occur after the last event in a given branch path, whereas the second term is the probability of the events in a branch path and the waiting times between them.

2.2 Model parameters

The number of model parameters is a function of h, the number of hidden states:

vM for ω :	$20 \times 2 \times h$
bvM for ϕ, ψ :	$20 \times 5 \times h$
vMs for τ_1, τ_2, τ_3 :	$3 \times 2 \times h$
MVN for b_1, b_2, b_3	$3 \times 8 \times h$
AA freqs.	$(20 - 1) \times h$
AA exchangeability matrix	190
Hidden transition prob. matrix	$h^2 - h$
Hidden rate matrix	$(h^2 - h)/2$

2.3 Inference

2.3.1 Branch path inference: a phylogeny

Inference for a given dataset $\mathcal{D}_d$ consists of sampling the set of branch paths, $\mathcal{B}_d$, conditional on the tree topology and branch lengths, $\mathcal{T}_d$, and model parameters $\hat{\theta}$:

$$\mathcal{B}_d \sim p(\mathcal{B}_d | \mathcal{D}_d, \mathcal{T}_d, \hat{\theta}). \tag{10}$$

To sample this distribution, for each site i, Felsenstein's algorithm was used to calculate likelihoods in a forward pass up the tree, followed by a backwards sampling pass down the tree to propose new hidden node states at the tip of each branch. The hidden state rate matrices used were conditional on the amino acid branch paths at site i and hidden state branch paths at site $i - 1$ and $i + 1$.

Conditional upon proposed internal node states, modified rejection sampling was used to sample hidden state branch paths using the parent branch's proposed hidden node tip state as the start state and the current branches proposed hidden node tip state as the end state.

The proposed branch paths for site i were then accepted or rejected using the Metropolis-Hastings ratio together with proposal ratio.

An analogous algorithm was used for sampling the amino acid branch paths, where the amino acid rate matrices used were conditional on the hidden states branch paths at each site i.

2.3.2 Inference: a single protein

Inference for a dataset consisting of a single protein P_a is much simpler, given that the protein is assumed to be drawn from the stationary distribution of the model, which is given by (7)

$$H_a \sim p(H_a | A_a, \hat{\theta}). \tag{11}$$

This distribution can be sampled exactly using the forward-filtering backward-sampling algorithm for HMMs (Frühwirth-Schnatter, 1994) in $\mathcal{O}(h^2|P_a|)$ computational time. Note that the amino acid sequence, A_a, is typically observed However, regardless of which combinations of A_a and X_a are observed it remains possible to use the forward-filtering backward-sampling algorithm to efficiently sample H_a.

2.3.3 Backbone structure inference

Branch path inference gives the distribution of $S_b(t) = \big(H_b(t), A_b\big)(t))$ at every point in time t along a branch b. Conditioned on $H_b(t)$ and $A_b(t)$ the angles and bond lengths, $X_b(t)$, comprising the backbone structure can be trivially sampled using Equations 1-4. The posterior marginal $p(X_b(t)|\mathcal{D}_b, \mathcal{T}_b)$ is therefore obtained by first sampling $H_b(t)$ and $A_b(t)$:

$$\big(H_b(t), A_b(t)\big) \sim p(H_b(t), A_b(t) | \mathcal{D}_b, \mathcal{T}_b), \tag{12}$$

followed by sampling $X_b(t)$ conditional on $H_b(t), A_b(t)$:

$$X_b(t) \sim p(X_b(t) | H_b(t), A_b(t)). \tag{13}$$

Table 1: Composition of training datasets

Category	Number in category	Amino acid observations	Structural observations
One sequence (no structure)	310	84,806	0
Two sequences (no structures)	208	138,892	0
Three or more sequences (no structures)	118	390,913	0
One structure (one corresponding sequence)	4,565	1,259,635	1,259,342
One structure (two or more sequences)	192	147,679	58,887
Two structures (two or more sequences)	81	72,331	50,982
Three or more structures (three or more sequences)	18	44,273	20,745
Total	5,492	2,138,529	1,389,956

2.4 Model training

2.4.1 Datasets

2.4.2 Model estimation

Stochastic EM (StEM, Gilks *et al.* (1995)) was used to train the model. StEM is a stochastic version of the well known Expectation-Maximization algorithm (Gilks *et al.*, 1995). Its distinguishing feature is that the E-step consists of filling in the values of the latent variables using sampling. Only a single value is sampled. StEM is attractive due to its computational efficiency and its tendency to avoid getting stuck in local minima (Gilks *et al.*, 1995).

Sampling was used in the E-step to sample branch paths and times. In other words, at iteration k for each dataset, d, consisting of an aligned set of proteins, $\mathcal{D}_d$, and a corresponding set of branch paths, $\mathcal{B}_d$, we draw samples, from the following joint-distribution:

$$Z_d^{(k)} \sim p(\mathcal{B}_d | \mathcal{D}_d, \Psi^{(k)}).$$

In the M-step the samples from the previous E-step, were used to update the hidden node parameters ($\hat{\Psi}$) using efficient sufficient statistics (ESSs).

2.5 Calculation of angular distances

For benchmarking and comparison purposes, the angular cosine distance was used to measure distances between pairs of dihedral angles, $\langle \phi_a, \psi_a \rangle$ and $\langle \phi_b, \psi_b \rangle$. It is defined as follows (Downs and Mardia, 2002):

$$\begin{aligned} & d(\langle \phi_a, \psi_a \rangle, \langle \phi_b, \psi_b \rangle) \\ &= \sqrt{4 - 2\cos(\phi_a - \phi_b) - 2\cos(\psi_a - \psi_b)}. \end{aligned} \tag{14}$$

The maximum possible distance is $\sqrt{8} \approx 2.828$. Angular distance defined in this way has the property that when $\phi_a - \phi_b \approx 0$ and $\psi_a - \psi_b \approx 0$ are near zero it may be approximated by the Euclidean distance – using the small angle approximation for cosine ($\cos\theta \approx 1 - \frac{\theta^2}{2}$ when θ is near zero):

$$\begin{aligned} & d(\langle \phi_a, \psi_a \rangle, \langle \phi_b, \psi_b \rangle) \\ &\approx \sqrt{4 - 2(1 - (\phi_a - \phi_b)^2/2) - 2(1 - (\psi_a - \psi_b)^2/2)} \\ &= \sqrt{(\phi_a - \phi_b)^2 + (\psi_a - \psi_b)^2}. \end{aligned}$$

3 Results and Discussion

3.1 Benchmarks of ancestral sequence reconstruction

4 Conclusions

5 Software availability

Julia code (compatible with Windows and Linux) is available at: https://github.com/michaelgoldendev/protein-evolution

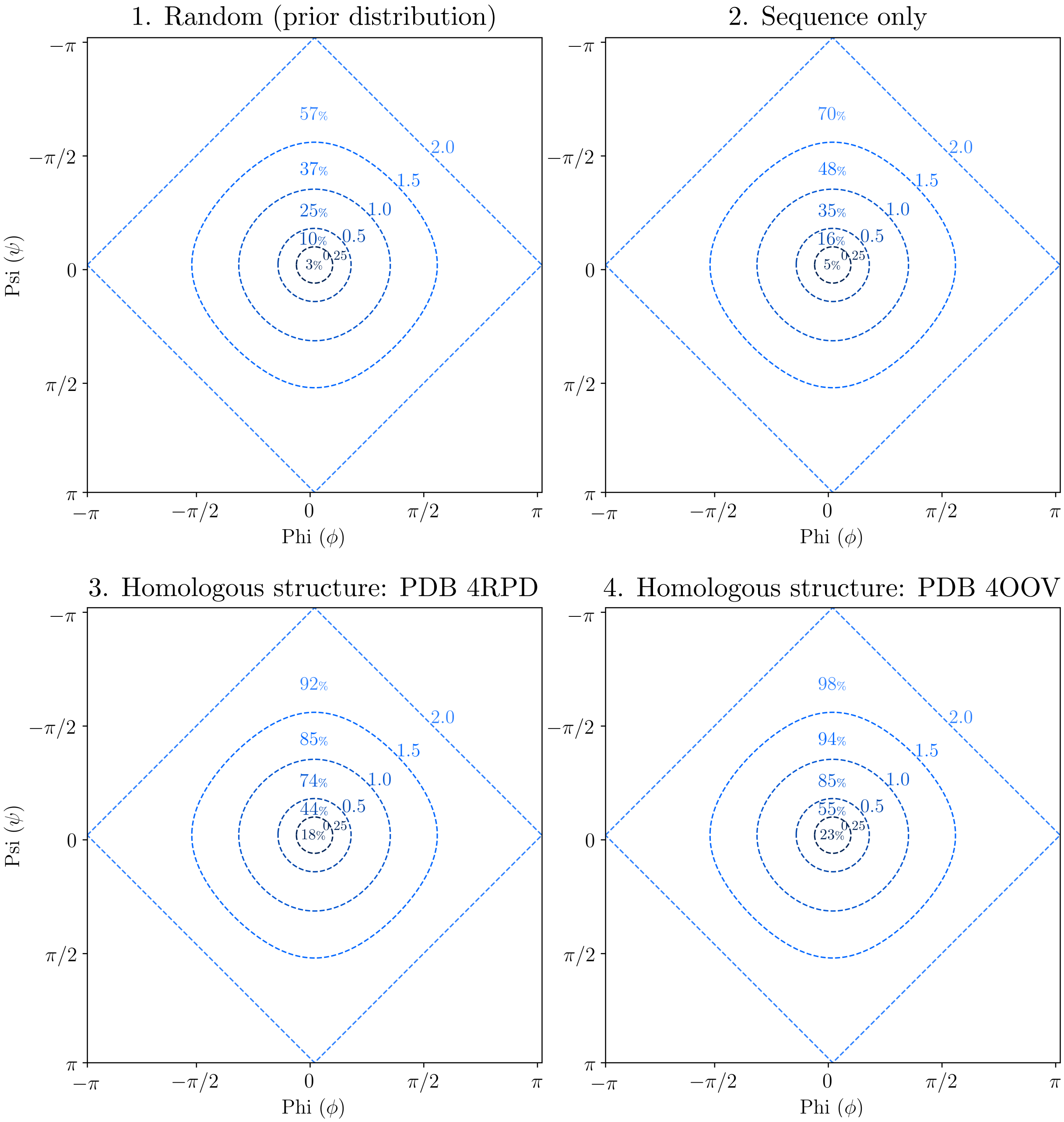

Figure 3: Example target diagrams The angular cosine distance (14) was used to compare our model's predictions of phi-psi angles in PDB 5KON under four increasingly informative observation conditions (1-4). The plots show the percentage of phi-psi angles correctly predicted within five different cosine angular distance radii (0.25, 0.5, 1.0, 1.5, and 2.0; dashed blue lines). If the centre of the plot is taken to be location of each of the true phi-psi angles in PDB 5KON, the objective is to the maximise the percentage of angles predicted within the innermost radius.

6 Acknowledgements

MG is supported by the ERC under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 614725-PATHPHYLODYN.

References

Boomsma, W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A., and Hamelryck, T. 2008. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105(26): 8932–8937.

Boomsma, W., Tian, P., Frellsen, J., Ferkinghoff-Borg, J., Hamelryck, T., Lindorff-Larsen, K., and Vendruscolo, M. 2014. Equilibrium simulations of proteins using molecular fragment replacement and NMR chemical shifts. *Proceedings of the National Academy of Sciences*, 111(38): 13852–13857.

Challis, C. J. and Schmidler, S. C. 2012. A stochastic evolutionary model for protein structure alignment and phylogeny. *Molecular biology and evolution*, 29(11): 3575–3587.

Choi, S. C., Redelings, B. D., and Thorne, J. L. 2008. Basing population genetic inferences and models of molecular evolution upon desired stationary distributions of DNA or protein sequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512): 3931–3939.

Downs, T. D. and Mardia, K. 2002. Circular regression. *Biometrika*, 89(3): 683–698.

Echave, J. 2008. Evolutionary divergence of protein structure: the linearly forced elastic network model. *Chemical physics letters*, 457(4): 413–416.

Echave, J. and Fernández, F. M. 2010. A perturbative view of protein structural variation. *Proteins: Structure, Function, and Bioinformatics*, 78(1): 173–180.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6): 368–376.

Frellsen, J., Mardia, K. V., Borg, M., Ferkinghoff-Borg, J., and Hamelryck, T. 2012. Towards a general probabilistic model of protein structure: the reference ratio method. In *Bayesian methods in structural bioinformatics*, pages 125–134. Springer.

Frühwirth-Schnatter, S. 1994. Data augmentation and dynamic linear models. *Journal of time series analysis*, 15(2): 183–202.

García-Portugués, E., Sørensen, M., Mardia, K. V., and Hamelryck, T. 2017. Langevin diffusions on the torus: estimation and applications. *Statistics and Computing*.

García-Portugués, E., Golden, M., Sørensen, M., Mardia, K. V., Hamelryck, T., and Hein, J. 2018. Toroidal diffusions and protein structure inference. In C. Ley and T. Verdebout, editors, *Applied Directional Statistics: Modern Methods and Case Studies*. Chapman and Hall/CRC.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. 1995. *Markov chain Monte Carlo in practice*. CRC press.

Golden, M., García-Portugués, E., Sørensen, M., Mardia, K. V., Hamelryck, T., and Hein, J. 2017. A generative angular model of protein structure evolution. *Molecular Biology and Evolution*, 34: msx137.

Grishin, N. V. 1997. Estimation of evolutionary distances from protein spatial structures. *Journal of molecular evolution*, 45(4): 359–369.

Grishin, N. V. 2001. Fold change in evolution of protein structures. *Journal of structural biology*, 134(2): 167–185.

Gutin, A. M. and Badretdinov, A. Y. 1994. Evolution of protein 3D structures as diffusion in multidimensional conformational space. *Journal of molecular evolution*, 39(2): 206–209.

Herman, J. L., Challis, C. J., Novák, Á., Hein, J., and Schmidler, S. C. 2014. Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. *Molecular biology and evolution*, 31(9): 2251–2266.

Liò, P., Goldman, N., Thorne, J. L., and Jones, D. T. 1998. PASSML: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics*, 14(8): 726–733.

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

Parsons, J., Holmes, J. B., Rojas, J. M., Tsai, J., and Strauss, C. E. 2005. Practical conversion from torsion space to cartesian space for in silico protein synthesis. *Journal of computational chemistry*, 26(10): 1063–1068.

Siepel, A. and Haussler, D. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology*, 11(2-3): 413–428.

Thorne, J. L., Kishino, H., and Felsenstein, J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. *Journal of molecular evolution*, 34(1): 3–16.