A pairwise evolutionary model of protein sequence and structural conformation D6

May 12, 2016

Abstract

We give a generative evolutionary model of protein sequence and structural conformation. Recently, there have been stochastic models of structural evolution, which have shown that the inclusion of structural information leads to more reliable estimation of evolutionary parameters. We introduce a new pairwise evolutionary model which takes into account local dependencies between sequence and structural evolution. We treat each protein in a pair as random walk in space through the use of angular conformations. A coupling in our model is such that an amino acid change can lead to a jump in angle conformation and a change in diffusion process. This model is comparatively more realistic than previous stochastic models, since it allows improved understanding of the relationship between sequence and structure evolution. The generative nature of our model allows us to provide evidence of its validity.

Contents

List of figures

1. Diagram of model (JASSC) - done
2. Dihedral angle representation - done
3. Confirmatory Ramachandran plots: all angles, proline, glycine - partly done
4. A selection of interesting hidden states - to do
5. Homology modelling benchmarks (more combinations: s1+s2+angles1) - to do
6. Annotation of 3D protein with jumps - to do
7. Some plots depicting evolution / rates of evolution (for example: histograms or box and whisker plots of branch length) - to do
8. Diagram of alignment HMM - to do

General notes

1 Introduction

Recently it has been realised that one can make a joint stochastic model of evolution which takes into account simultaneous alignment of protein sequence and 3D structure (**? ?**]). These papers also point out limitations of earlier non-probabilistic methods, as well as the benefits of simultaneously taking protein sequence and 3D structure into account.

We present a pairwise model which takes into account local dependencies between sequence and structural evolution, providing more realism; this is a departure from **? ?**]. We treat each protein in a pair as a random walk in space, through the use of the phi and psi dihedral angle representations. This bypasses the need for structural alignment, unlike in . We use a stochastic diffusion process on phi and psi dihedral angles (See December abstract Eduardo), as well as the TKF92 model of insertion and deletion. We introduce a coupling in our model such that an amino acid change can lead to a jump in the phi and psi dihedral angles and a change in diffusion process.

Comparison with previous work NEED TO COMMENT ON 2014.

Both the 2012 (Challis and Schmidler) and 2014 (Herman and Schmidler) papers use the TKF92 model together with OU diffusions, but on atomic coordinates of C-alpha atoms rather than on dihedral angles, as in our model. Overall, we have various modelling virtues that provide more realism than the 2012 and 2014 papers. Our model has the potential for extension to the full phylogeny, rather than just pairs of proteins as we have considered in the current work.

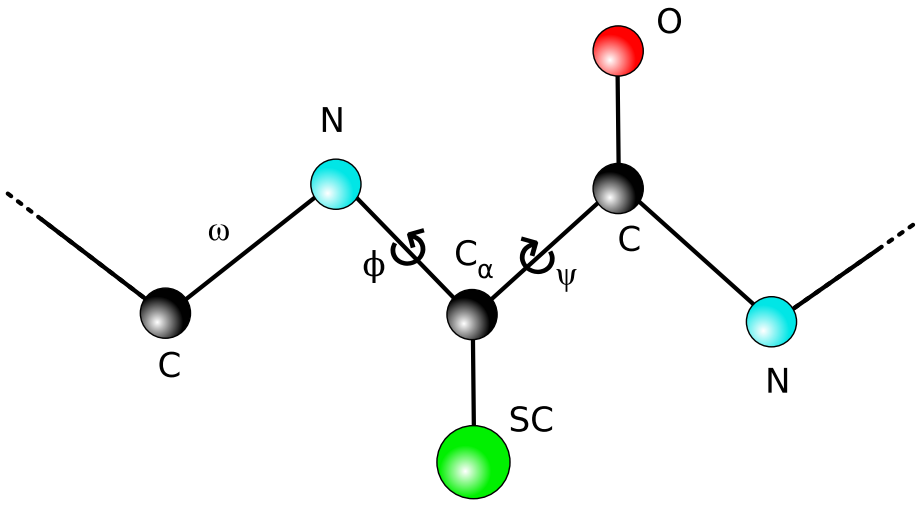

Figure 1: Dihedral angle representation. A very small section of the protein backbone is displayed in atomic detail: each amino acid contains four backbone atoms (N, C_α, C, O). The side chain is represented by one pseudo-atom (SC). The dihedral angles ϕ, ψ are shown. This gives a chemical representation of a typical amino acid.

2 Methods

2.0.1 A pairwise evolutionary model

Our pairwise evolutionary model is an HMM (Figure 2) designed to model local protein sequence and structure evolution along a pair of aligned homologous proteins, p_a and p_b in two different species a and b, respectively.

An ith observation pair, $O^i = (O^i_a, O^i_b)$, is associated with every aligned site i in an alignment M of p_a and p_b. Where we take i to run from 1 to m, where m is the length of the alignment M. Each observation, O^i_a and O^i_b, contains amino acid sequence and structural information corresponding to the two Cα atoms at aligned site i associated with proteins p_a and p_b, respectively. An observation corresponding to a particular protein at aligned site i, O^i_a, is comprised of three different data types associated with the Cα atom: an amino acid (A^i_a, discrete, one of twenty canonical amino acids), phi and psi dihedral angles ($X^i_a = \langle \phi^i_a, \psi^i_a \rangle$; continuous, bivariate), and it's secondary structure classifcation (S^i_a; discrete, one of three classes: helix, sheet or coil). That is $O^i_a = (A^i_a, X^i_a, S^i_a)$ and $O^i_b = (A^i_b, X^i_b, S^i_b)$.

Each hidden node H^i in the HMM corresponds to an aligned site i in the alignment M. Initially we treat the alignment M as given apriori, but later on modify the HMM to allow an unknown alignment.

The model is parametrised by q hidden states. Every hidden node, H^i, corresponding to an aligned site i takes an integer value from 1 to q for the

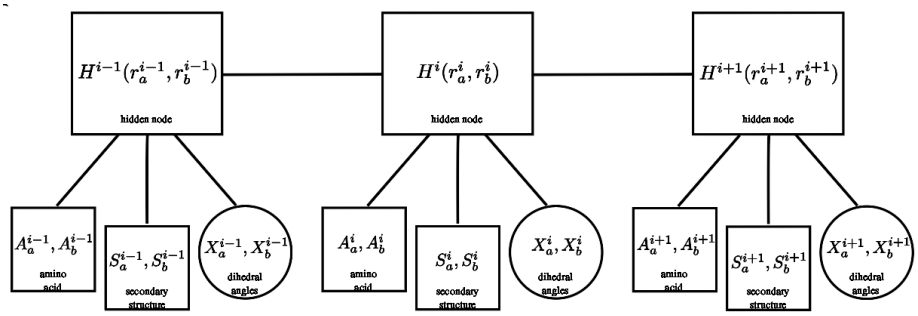

Figure 2: The graph represents the architect of TorusDBN where each H along the horizontal axis represents a node, and the edges with arrows encode the conditional independencies between the observed variables A, S and X for given $H(r_a, r_b)$ where A = amino acid (20 possible values), S = secondary structure (helix, β–sheet or loop), and X = angle pair (bivariate von Mises distribution). The rectangles represent discrete variables, and the circles represent continuous variables..

hidden state at node H^i. Furthermore, every hidden state couples together an evolutionary regime pair (r_a^i, r_b^i), written as $H^i(r_a^i, r_b^i)$. Each evolutionary regime takes on integer values 1 or 2, i.e. $(r_a^i, r_b^i) \in \{1, 2\}^2$. Therefore, there are $4q$ possible combinations of hidden state and evolutionary regime pairs at every aligned site i. We will return the role of the evolutionary regime pairs later.

From this point forward we may sometimes write $H^i(r_a^i, r_b^i)$ as H^i when the meaning is clear. We write $H = (H^1, H^2, \ldots, H^m)$, and assume that, conditional on $H^i(r_a^i, r_b^i)$ and branch length t_{ab}, the corresponding random variables A^i, X^i and S^i are independent.

$$\begin{aligned} p(O^i|H^i(r_a^i, r_b^i), t_{ab}) &= \overbrace{p(A^i|H^i(r_a^i, r_b^i), t_{ab})}^{\text{amino acid sequence evolution (CTMC)}} \\ &\times \overbrace{p(X^i|H^i(r_a^i, r_b^i), t_{ab})}^{\text{dihedral angle evolution (wNOU diffusion)}} \\ &\times \overbrace{p(S^i|H^i(r_a^i, r_b^i), t_{ab})}^{\text{secondary structure evolution (CTMC)}} \end{aligned} \tag{1}$$

Let us denote the transition probability of the $(i-1)$th node to the ith node by $P(H^i|H^{i-1})$ and the initial probability at the 1st node by $p(H^1)$. It can be seen that the joint distribution of (O, H) conditioned on the branch length t_{ab} seperating proteins p_a and p_b is given by:

$$p(O, H|t_{ab}) = p(H^1)p(O^1|H^1, t_{ab})\sum_{i=2}^{m} p(H^i|H^{i-1})p(O^i|H^i, t_{ab})$$

where

$$p(O^i|H^i,t_{ab}) = \sum_{(r_a^i,r_b^i)\in\{1,2\}^2} p(O^i|H^i(r_a^i,r_b^i),t_{ab})p(r_a^i,r_b^i|H^i,t_{ab})$$

and $p(r_a^i,r_b^i|H^i,t_{ab})$ is the evolutionary regime pair probability.
Hence

$$p(O|t_{ab}) = \sum_H p(O,H|t_{ab})$$

Where the sum is taken over all H which are combinations of m integers from $(1,2,\ldots,q)$. Thus we have as many as q^m terms in the last expression.

We now turn to the meaning of the evolutionary regimes. Two modes of evolution are modelled: "Constant evolution" and "Jump evolution". Constant evolution occurs when the evolutionary regime starting in protein a at aligned site i, r_a^i, is the same as the evolutionary regime ending in protein b at aligned site i, r_b^i, i.e. $r_a^i = r_b^i$. Conversely, jump evolution occurs when the evolutionary regime starting in protein a, r_a^i, differs from the evolutionary regime ending in protein b, r_b^i, at aligned site i, i.e. $r_a^i \neq r_b^i$. Constant evolution in our model, is intended to capture structural drift (changes in dihedral angles restricted to a region of the Ramachandran plot), whereas jump evolution allows us to capture structural shift (large changes in dihedral angles, possibly between distant regions of the Ramachandran plot).

The hidden state at node H^i, together with the branch length, t_{ab}, separating proteins p_a and p_b specifies a joint distribution over the evolutionary regime pair probability. The evolutionary regime pair probabilities have been chose as follows:

$$p(r_a^i,r_b^i|H^i,t_{ab}) = \begin{cases} \pi(H^i,r_a^i)[e^{-\alpha(H^i)t_{ab}} + \pi(H^i,r_a^i)(1-e^{-\alpha(H^i)t_{ab}})] & \text{if } r_a^i = r_b^i \\ & \text{i.e. constant evolution} \\ \pi(H^i,r_a^i)\pi(H^i,r_b^i)(1-e^{-\alpha(H^i)t_{ab}}) & \text{if } r_a^i \neq r_b^i \\ & \text{i.e. jump evolution} \end{cases}$$

Where $\pi(H^i,r_a^i)$ and $\pi(H^i,r_b^i)$ are model parameters specifying the probability of starting in regime r_a^i or r_b^i conditional on the hidden state at node H^i. Similarly, $\alpha(H^i)$ is a parameter of the model specifying the jump rate corresponding to the hidden state at node H^i.

Note that the regime pair jump probabilities have been chosen such that time reversibility holds, i.e. $p(r_a^i,r_b^i|H^i,t_{ab}) = p(r_b^i,r_a^i|H^i,t_{ab})$.

Each regime pair, (r_a^i,r_b^i), and branch length t_{ab}, specifies a joint distribution over observation pairs: $p(O_a^i,O_b^i|r_a^i,r_b^i,t_{ab})$.

The joint likelihood of the observations O_a^i and O_b^i at position i, conditional on regime pair, (r_a^i,r_b^i) , and branch length t_{ab} separating the two protein is:

$$p(O_a^i, O_b^i | H^i(r_a^i, r_b^i), t_{ab}) = \begin{cases} p(O_a^i, O_b^i | H^i, r_\#^i, t_{ab}) & \text{if } r_a^i = r_b^i, \text{ where } r_\#^i = r_a^i = r_b^i \\ & \text{Constant evolution, likelihood given by the} \\ & \text{evolutionary model specified by evolutionary regime } r_\#^i \\ p(O_a^i | H^i, r_a^i) p(O_b^i | H^i, r_b^i) & \text{if } r_a^i \neq r_b^i \\ & \text{Jump evolution, likelihood given by the stationary} \\ & \text{distributions of regimes } r_a^i \text{ and } r_b^i, \text{ respectively.} \end{cases}$$

In the case of constant evolution, evolution at aligned i is described in terms of the same evolutionary regime $r_\#^i = r_a^i = r_b^i$ and is therefore considered "constant". In the case of jump evolution, the observations O_a^i and O_b^i are assumed to be drawn from the stationary distributions corresponding to evolutionary regimes r_a^i and r_b^i, respectively. The assumption that they are drawn from the stationary distribution conditioned on the occurrence of a jump allows us to ignore the branch length t_{ab}. Furthermore, it is no longer necessary to consider the unobserved evolutionary trajectory linking the two observations and hence there is no need to marginalise over all possible trajectories, as in a model with continuous Markovian switching between evolution regimes. This allows us to model structural drift in a computationally tractable manner. Our jump model can be justified by noting that in the case where $t_{ab} \to \infty$, O_a^i and O_b^i, are conditionally independent, the evolutionary regimes and the behaviour of the model tends to that of a non-evolutionary TorusDBN model on each protein.

Some might consider "jump evolution" to be fully non-evolutionary given that the observations are drawn from the stationary distributions of the respective evolutionary regimes, however, the probability of a jump is informed by evolutionary information such as the branch t_{ab} and any observations, such as the amino acids corresponding to each of the two proteins.

Time reversibility The three stochastic processes in Equation 1 are assumed to be time-reversible. This, together with assumption of time-reversibility in jumping between evolutionary regimes, ensures overall time-reversibility in Equation **??**. This allows to treat the phylogenetic tree relating proteins $p_a$and p_b as unrooted, implying we can arbitrarily pick p_a or p_b as a root of the phylogenetic tree (CITE). This avoids the need to marginalise over the common ancestor protein of proteins p_a and p_b and branch times. Note that whilst time-reversibility of the evolutionary processes at each site holds, this is different from time-reversibility of the HMM. The HMM is specified by a non-reversible transition probability matrix, implying different probabilities going in opposite directions along a pair of proteins (N to C terminus vs. C to N terminus).

2.0.2 Stochastic processes

Each evolutionary regime couples together three time-reversible Markovian stochastic processes that each separately describe the evolution of the three pairs of observation types, as in Equation 1. Because each evolutionary regime is intended

to capture different features of sequence and structural evolution, a different set of parameters is used to describe each. Parameters that correspond to a particular evolutionary regime are termed 'regime-specific', whereas parameters that are shared across all evolutionary regimes are termed 'global'.

Amino acid evolution Amino acid evolution is described by a Continous-Time Markov Chains (CTMC), which is the standard approach for modelling amino acid substitution processes.

Amino acid evolution in our model is parameterized in the following way: the exchangeability of amino acids is described by a 20×20 symmetric global exchangeability matrix S (190 free parameters; Whelan and Goldman (2001)), a regime-specific set of 20 amino acid equilibrium frequencies $\Pi_r^h = \text{diag}\{\pi_1, \pi_2, \ldots, \pi_{20}\}$ (19 free parameters per evolutionary regime), and a regime-specific scaling factor Λ_r^h (1 free parameter per evolutionary regime). Together these define a regime-specific amino acid rate matrix $Q_r^h = \Lambda_r^h S \Pi_r^h$. Due to the symmetry of S, Q_r^h is guaranteed to have a stationary distribution corresponding to the set of 20 amino acid equilibrium frequencies: Π_r^h .

Secondary structure evolution Secondary structure evolution is likewise described by a CTMC. In our model we use three classes to describe secondary structure class at each position: helix (H), sheet (S) and random coil (C).

The exchangeability of secondary structure classes at a position is described by a 3×3 symmetric global exchangeability matrix V and a regime-specific set of 3 secondary structure equilibrium frequencies $\Xi_r^h = \text{diag}\{\pi_1, \pi_2, \pi_3\}$. Together these define a regime-specific secondary structure rate matrix $R_r^h = V \Xi_r^h$, with stationary distribution: Ξ_r^h.

Dihedral angle evolution Unlike the previous stochastic processes and in order to account or full-atomic detail, the evolution of dihedrals is modelled by a continuous-time and *continuous-state* Markovian process, i.e. a diffusion. Constructing a *toroidal diffusion* \citep{Garcia-Portuges2016} with the required properties of time-reversibility, ergodicity, parsimonity and tractability presents several a non-trivial challenges. The last point is specially challenging since the maximum likelihood estimation depends on the transition probability density (tpd), which almost always has an intractable analytical form and is only known implicitly as the solution to the Fokker-Planck equation. This is particullarly true when the drift and diffusion coefficients are periodic - hence highly non-linear - as it happens with toroidal diffusions. Albeit the unavailability of the tpd can be bypassed by different approaches (pseudo-likelihoods, numerical solution of the Fokker-Planck equation, etc), these methods either make severe compromises in the accuracy of the approximation to the tpd (mainly aimed for $t \to 0$) or in its computational expediency (particularly on multivariate diffusions). Both problems result in a major drag down for the traaining and performance of the model, specially for a large dataset.

A suitable diffusion satisfying the above conditions was specifically developed for this purpose in \cite{Garcia-Portugues2016}, where a toroidal analogue of the celebrated Ornstein-Uhlenbeck (OU) process was considered. The diffusion is driven by the (bivariate) Wrapped Normal (WN) \citep{Mardia2000} distribution, denoted by $\mathrm{WN}(\mu, \Sigma)$ and having pdf

$$f_{\mathrm{WN}}(\theta; \mu, \Sigma) = \sum_{k \in \mathbb{Z}^2} \phi_\Sigma(\theta - \mu + 2k\pi), \quad \theta \in \mathbb{T}^2,$$

with ϕ_Σ standing for the pdf of a bivariate Gaussian with zero mean and covariance Σ. Jointly with the bivariate von Mises (see Chapter 6 in \cite{Hamelryck2012} for a review on bivariate von Mises models), the WN can be considered as a toroidal analogue of the Gaussian distribution. The WN toroidal diffusion $\{\Theta_t\} \subset \mathbb{T}^2$ arises as the wrapping $\Theta_t = (X_t + \pi) \mod 2\pi - \pi$ of the Euclidean diffusion

$$dX_t = \sum_{k \in \mathbb{Z}^2} A(\mu - X_t - 2k\pi) w_k(X_t) dt + \Sigma^{\frac{1}{2}} dW_t \tag{2}$$

where W_t is the Wiener process, A is the drift matrix, $\mu \in \mathbb{T}^2$ is the stationary mean, Σ is the diffusion matrix and

$$w_k(x) = \frac{\phi_{\frac{1}{2}A^{-1}\Sigma}(x - \mu + 2m\pi)}{\sum_{k \in \mathbb{Z}^2} \phi_{\frac{1}{2}A^{-1}\Sigma}(x - \mu + 2k\pi)}, \quad k \in \mathbb{Z}^2,$$

represents the distribution over the winding numbers of $\{X_t\}$. Intuitively, the drift of (2) can be regarded as a smoothed weighting of piecewise linear drifts that are binded to achieve periodicity. The WN process is the *characteristic diffusion*, in the sense of \cite{Kent1978}, of the WN, likewise the OU process is the characteristic diffusion of the Gaussian distribution. It is ergodic in $\mathbb{T}^2$ with stationary distribution $\mathrm{WN}\big(\mu, \frac{1}{2}A^{-1}\Sigma\big)$ and is time-reversible. Indeed, it is the *unique* diffusion satisfying both properties. Other toroidal diffusions with given stationary distributions can be constructed following a similar argument, we refer to \cite{Garcia-Portugues2016} for further considerations.

No analytical solution for the tpd of the WN diffusion is known. However, the peculiar form of the drift of (2) allows to make a connection with the (bivariate) OU process

$$dX_t = A(\mu - X_t) dt + \Sigma^{\frac{1}{2}} dW_t$$

which is analytically tractable. We can conjecture that an approximation to the tpd follows from weighting the exact tpd of \eqref{eq:ou} in the same fashion as the linear drifts are weighted in \eqref{eq:wnd}. If we wrap this the resulting density, we get

$$\tilde{p_t}(\theta \,|\, \theta_0; A, \mu, \Sigma) = \sum_{m \in \mathbb{Z}^2} f_{\mathrm{WN}}(\theta; \mu_t^m, \Gamma_t) w_m(\theta_0),$$

with

$$\mu^{\boldsymbol{t}} = \mu + e^{-tA}(\theta_0 - \mu + 2\pi k), \quad \Gamma_t = \int_0^t e^{-sA}\Sigma e^{-sA^T}\, ds.$$

The density \eqref{eq:ap} is indeed the conditional density of $(X_t + \pi) \bmod 2\pi - \pi$ and, albeit is not the true tpd, it provides a good approximation to it in key circunstances: *i)* $t \to 0$, since it collapses in the Dirac's delta; *ii)* $t \to \infty$, since it converges to the stationary distribution; *iii)* high concentration, since the WN diffusion becomes a OU process after a suitable centering. CHECK[Point *ii)* is specially relevant in our context as it implies that the tpd is going to be correctly approximated when there is littledependence/no changes]. Besides *ii)* and *iii)*, a further key advantage of \eqref{eq:ap} over pseudo-likelihoods is that is able to capture the multimodality of the tpd and has the correct stationary distribution. In addition, as the true tpd, \eqref{eq:ap} satisfies the time-reversibility equation, namely $\tilde{p}_t(\theta_2 \,|\, \theta_1; A, \mu, \Sigma) f_{\mathrm{WN}}(\theta_1; \mu, \frac{1}{2}A^{-1}\Sigma) = \tilde{p}_t(\theta_1 \,|\, \theta_2; A, \mu, \Sigma) f_{\mathrm{WN}}(\theta_2; \mu, \frac{1}{2}A^{-1}\Sigma)$. The approximation is shown to have lower Kullback-Leibler divergences with respect to the true tpd than toroidal pseudo-likelihoods, for most of the usual scenarios and discretization times in the diffusion trajectory \citep{Garcia-Portugues2016}.

The drift and diffusion matrices need to satisfy that $A^{-1}\Sigma$ is a covariance matrix in order to have a non-degenerate stationary WN. Different parametrizations are possible and we opted for a compromise between flexibility and simplicity: $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$, σ_1, $\sigma_2 > 0$ and

$$A = \begin{pmatrix} \alpha_1 & \frac{\sigma_1}{\sigma_2}\alpha_3 \\ \frac{\sigma_2}{\sigma_1}\alpha_3 & \alpha_2 \end{pmatrix}, \quad \alpha_1,\, \alpha_2 > 0, \alpha_1\alpha_2 > \alpha_3^2.$$

With this formulation, the correlation between components is only captured at the drift level by α_3 and this is reflected in the stationary covariance matrix:

$$\frac{1}{2}A^{-1}\Sigma = \frac{1}{2(\alpha_1\alpha_2 - \alpha_3^2)} \begin{pmatrix} \alpha_2\sigma_1^2 & -\alpha_3\sigma_1\sigma_2 \\ -\alpha_3\sigma_1\sigma_2 & \alpha_1\sigma_2^2 \end{pmatrix}.$$

The computation of \eqref{eq:ptilde} in practise involves computing e^{-tA} and Γ_t, which we can work out explicitly. First, in virtue of \cite{Bernstein1993}, the exponential matrix of any 2×2 matrix$A$ has the explicit expression $e^{tA} = a(t)I + b(t)A$ with $a(t) = e^{rt}(\cosh(qt) - r\frac{\sinh(qt)}{q})$, $b(t) = e^{rt}\frac{\sinh(qt)}{q}$, $r = \frac{\mathrm{tr}(A)}{2}$, $q = \sqrt{|\det(A - rI)|}$ and I the identity matrix. Second, since $A^{-1}\Sigma$ is constrained to be symmetric, the integral in Γ_t can be computed analytically. Combining this two facts, we have

$$\Gamma_t = \tfrac{1}{2}A^{-1}(I - e^{-2At})\Sigma = s(t)\tfrac{1}{2}A^{-1}\Sigma + i(t)\Sigma,$$

with $s(t) = 1 - a(-2t)$, $i(t) = -\frac{1}{2}b(-2t)$. This gives a neat linear combination of the stationary and infinitesimal covariance matrices, particularly convenient for evaluating the approximating tpd at different times as many

computations can be reused. The maximum likelihood estimates from a sample $\{(\phi_1, \psi_1, \phi_2, \psi_2, t)_i\}_{i=1}^n$ are given by

$$(\hat{\alpha_1}, \hat{\alpha_2}, \hat{\alpha_3}, \hat{\mu_1}, \hat{\mu_2}, \hat{\sigma_1}, \hat{\sigma_2})_{\text{MLE}} = \arg\max_{A,\mu,\Sigma} (\sum_{i=1}^{n} \log f_{\text{WN}}((\phi_1, \psi_1)_i; \mu, \tfrac{1}{2}A^{-1}\Sigma) + \sum_{i=1}^{n} \log \tilde{p}_{t_i}((\phi_2, \psi_2)_i \mid (\phi_1, \psi_1)_i; A, \mu$$

and can be obtained by CITE NLOPT METHOD USED. COMPLETE MORE THIS PART.

2.1 Accounting for alignment uncertainty

Protein sequences can not only undergo point mutation events, but also also insertion and deletion events. We describe a modified pairwise TKF92 alignment HMM that models both local sequence/structure evolution and sequence alignment. Whilst, it is possible to fix the alignment in advance by pre-aligning the sequences using one of the many available optimisation-based alignment methods (CITE CITE), doing so ignores alignment uncertainty. An alignment can be thought of as a statement about homology, such that when amino acid positions are aligned in order to indicate homology they are considered to have evolved solely via mutation along the evolutionary trajectory linking them and not via insertion or deletion. As the sequence of insertion and deletions is rarely observed in practice, it is difficult to make statements about the true underlying alignment (homology relationships), especially when the sequences being compared are distantly related and/or the rate of evolution is high. The TKF92 model [CITE CITE] suitably provides a stochastic process describing the evolution of insertions and deletions. [DESCRIBE PROCESS IN MORE DETAIL]

For the pairwise case, the TKF92 model can be represented as an HMM - we use the TKF92 HMM described in [CITE] with the modification that each emitted pair of characters is drawn from one of H evolutionary hidden states (Figure 3). This HMM formulation allows one to integrate over all possible pairwise alignments in $\mathcal{O}(nm)$ time, where n and m are the respective lengths of the two sequences. The introduction of evolutionary hidden states increases this complexity to $\mathcal{O}(nmH)$, where H is the number of evolutionary hidden states. The probability of transitioning from one evolutionary hidden state from one alignment position to the next is given by the probability transition matrix $p(h^i, h^{i+1})$, which is intended to encode the local sequence/structure evolution as in the TorusDBN model.

The likelihood of a sequence pair is given by:

$$p(O|t_{ab}, \Theta, \Omega) = \sum_{A,H} p(O|A, H, t_{ab}, \Theta) p(A, H|t_{ab}, \Omega) \tag{3}$$

Where Ω is the alignment and hidden state transition parameters, t_{ab} is the time separating the two sequences a and b, and Θ corresponds to the evolution-

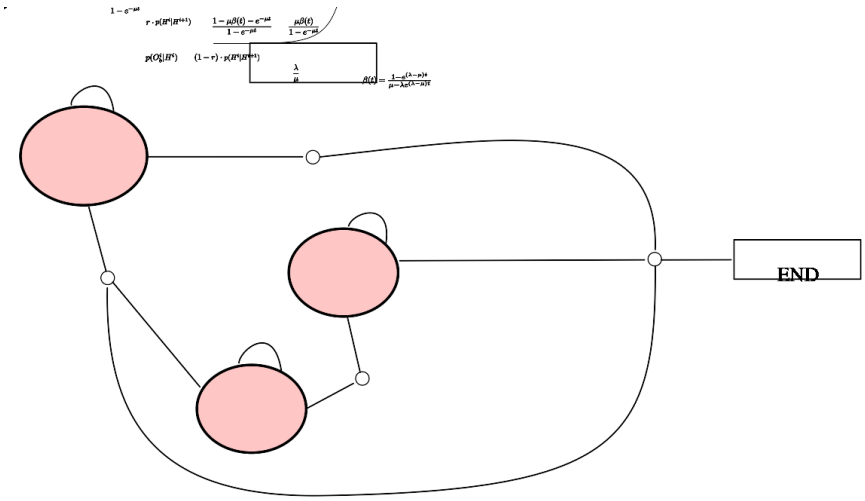

Figure 3: Diagrammatic representation of TKF92 alignment HMM modified to incorporate evolutionary hidden states and corresponding transitions.

ary parameters. The sum over alignments and hidden states ($\sum_{A,H}$) can be computed using the forward algorithm. The likelihood given by $p(O|A,H,t_{ab},\Theta)$ in Equation 3 can be decomposed into a product of alignment site likelihoods when conditioning on a particular alignment and set of hidden states:

$$p(O|A,H,t_{ab},\Theta) = \prod_{i \in A,H} p(O_a^i, O_b^i | H^i, t_{ab}, \Theta)$$

$$p(O_a^i, O_b^i | h^i, t_{ab}) = \begin{cases} p(O_a^i, O_b^i | H^i, t_{ab}) & \text{are matched are homologous (homologous)} \\ p(O_a^i | H^i) & \text{if the observation in sequence } a\text{is the result of an indel.} \\ p(O_b^i | H^i) & \text{if the observation in sequence } b\text{is the result of an indel.} \end{cases}$$

Note that in the case of insertion or deletion the probability of observing a single observation is no longer dependent on the time parameter as it is assumed that the observations are drawn from the stationary distribution.

2.2 Data selection

2.3 Model training

2.3.1 E-step

2.3.2 Sequence-specific parameters

Priors

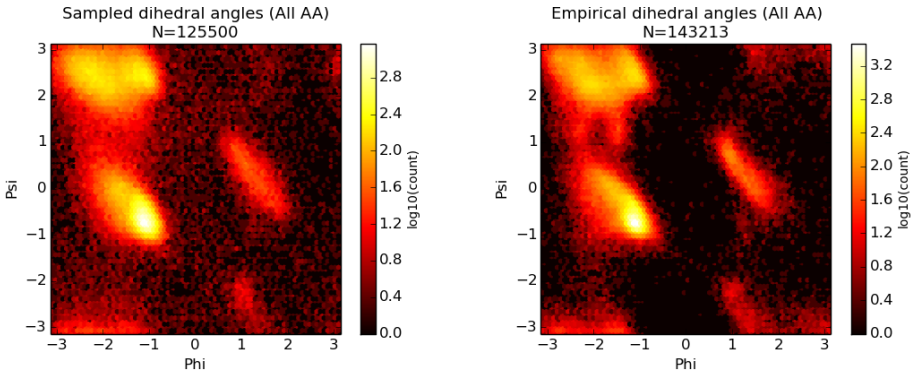

Figure 4:

Marginalisation using MCMC

2.3.3 M-step

2.3.4 Paralleisation

2.4 Posterior inference

2.5 Benchmarks

2.5.1 Angular distance

Distances between angle pairs, $\langle \phi_a, \psi_a \rangle$ and $\langle \phi_b, \psi_b \rangle$, are calculated using the cosine distance (CITE MARDIA, DOWNS?? 2002):

$$d(\langle \phi_a, \psi_a \rangle, \langle \phi_b, \psi_b \rangle) = \sqrt{4 - 2\cos(\phi_a - \phi_b) - 2\cos(\psi_a - \psi_b)}$$

This can be justified by noting that when $\phi_a - \phi_b \approx 0$ and $\psi_a - \psi_b \approx 0$ are small, this distance 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):

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

3 Results

4 Discussion

For reasons of tractability we have chosen a similar HMM framework for our pairwise evolutionary model, although more general frameworks such as Factorial HMMs, that allow additional hidden layers, or Continous Time Bayesian

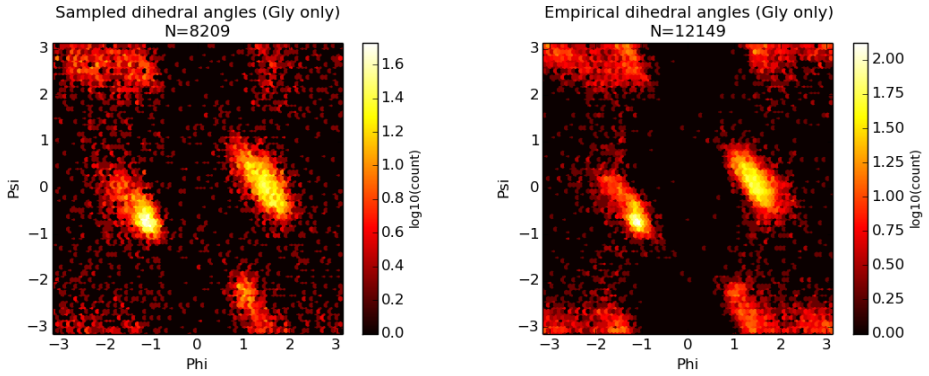

Figure 5:

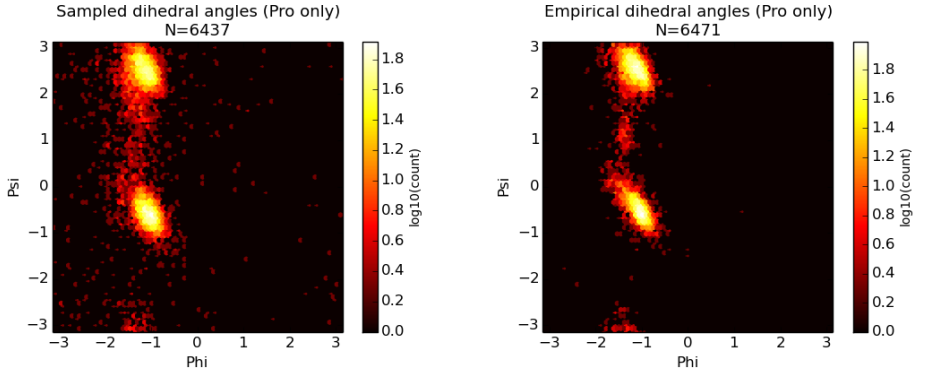

Figure 6:

Networks, that allow hidden states that evolve between protein as well as capturing dependencies along the proteins, could also be considered.

References