A generative angular model of protein structure evolution

Michael Golden,*[,1] Eduardo García-Portugués,[2] Michael Sørensen[3], Kanti V. Mardia,[1,4] Thomas Hamelryck,[5] and Jotun Hein[1]

[1]Department of Statistics, University of Oxford
[2]Department of Statistics, Carlos III University of Madrid
[3]Department of Mathematical Sciences, University of Copenhagen
[4]Department of Mathematics, University of Leeds
[5]Bioinformatics Centre, Section for Computational and RNA Biology, Department of Biology and Image Section, Department of Computer Science, University of Copenhagen

***Corresponding author:** E-mail: golden@stats.ox.ac.uk
Associate Editor: Name Surname
Protein families were obtained from the HOMSTRAD database.

Abstract

Recently described stochastic models of protein evolution have demonstrated that the inclusion of structural information in addition to amino acid sequences leads to a more reliable estimation of evolutionary parameters. We present a generative, evolutionary model of protein structure and sequence that is valid on a local length scale. The model concerns the local dependencies between sequence and structure evolution in a pair of homologous proteins. The evolutionary trajectory between the two structures in the protein pair is treated as a random walk in dihedral angle space, which is modelled using a novel angular diffusion process on the two-dimensional torus. Coupling sequence and structure evolution in our model allows for modelling both "smooth" conformational changes and "catastrophic" conformational jumps, conditioned on the amino acid changes. The model has interpretable parameters and is comparatively more realistic than previous stochastic models, providing new insights into the relationship between sequence and structure evolution. For example, using the trained model we were able to identify an apparent sequence-structure evolutionary motif present in a large number of homologous protein pairs. The generative nature of our model enables us to evaluate its validity and its ability to simulate aspects of protein evolution conditioned on an amino acid sequence, a related amino acid sequence, a related structure or any combination thereof.

Key words: Evolution, protein structure, probabilistic model, directional statistics

Article

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 dependencies due to physical interactions. In their current formulation LFENMs do not distinguish between the differing chemical nature of different amino acids 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 (top of Figure 1).

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 in terms of a series of ϕ and ψ dihedral angles that defines the relationship between the planes in three-dimensional space. A benefit of this representation is that it bypasses the need for structural alignment, unlike in models on Cartesian coordinates which typically need to additionally superimpose the structures for comparison purposes (Herman *et al.*, 2014). Accordingly, having to account for 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 in order to model their evolution.

The evolution of dihedral angles in ETDBN is modelled using a novel stochastic diffusion process developed in García-Portugués *et al.* (2017). In addition to this, a coupling is introduced such that an amino acid change can lead to a *jump* in dihedral angles and a change in diffusion process, allowing 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 in order 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 much more naturally capture the underlying protein structure manifold.

Two or more homologous proteins will share a common ancestor, which leads to underlying tree-like dependencies. These dependencies manifest themselves most noticeably in the degree of amino acid sequence similarity between two 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.

Failing to account for evolutionary dependencies can lead to false conclusions (Felsenstein, 1985), whereas accounting for evolutionary dependencies allows information from homologous proteins to be incorporated in a principled manner. This can lead to more accurate inferences, such as the prediction of a protein structure from a homologous protein sequence and structure,

known as homology modelling (Arnold *et al.*, 2006). Stochastic models such as ETDBN are not expected to compete with homology modelling software such as SWISS-MODEL (Arnold *et al.*, 2006). However, they allow for estimation of evolutionary parameters and statements about uncertainty to be made in a statistically rigorous manner.

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.

Evolutionary model

Overview

ETDBN is a dynamic bayesian network model of local protein sequence and structure evolution along a pair of aligned homologous proteins p_a and p_b. ETDBN can be can be viewed as an HMM (see Figure 1). Each hidden node of the HMM, corresponding to an aligned position, adopts an evolutionary hidden state specifying a distribution over three different observations pairs: a pair of amino acid characters, a pair of dihedral angles and a pair of secondary structures classifications. A transition probability matrix

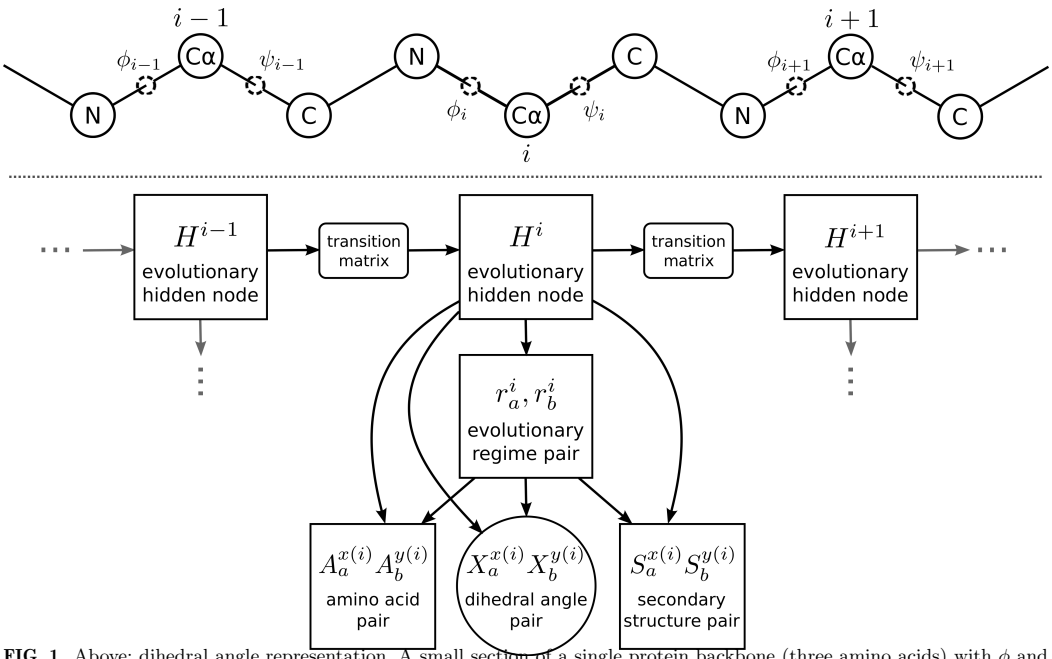

FIG. 1. Above: dihedral angle representation. A small section of a single protein backbone (three amino acids) with ϕ and ψ dihedral angles shown, together with C_α atoms which attach to the amino acid side-chains. Each amino acid side-chain determines the characteristic nature of each amino acid. Every amino acid position corresponds to a hidden node in the HMM below. Note that we only show a single protein, whereas the model considers a pair.
Below: depiction of HMM architecture of ETDBN where each H along the horizontal axis represents an evolutionary hidden node. The horizontal edges between evolutionary hidden nodes encode neighbouring dependencies between aligned sites. The arrows between the evolutionary hidden nodes and site-class pair nodes encode the conditional independence between the observation pair variables $A_a^{x(i)}, A_b^{y(i)}$ (amino acid site pair), $X_a^{x(i)} = \langle \phi_a^{x(i)}, \psi_a^{x(i)} \rangle, X_b^{y(i)} = \langle \phi_b^{y(i)}, \psi_b^{y(i)} \rangle$ (dihedral angle site pair) and $S_a^{x(i)}, S_b^{y(i)}$ (secondary structure class site pair). The circles represent continuous variables and the rectangles represent discrete variables.

specifies neighbouring dependencies between adjacent evolutionary states. For example, transitions along the alignment between hidden states encoding predominantly α-helix evolution would be expected to occur more frequently than transitions between a evolutionary hidden state encoding predominantly α-helix evolution and another encoding predominantly β-sheet evolution.

Ideally, the underlying hidden states would not just vary across the length of the alignment as captured by the HMM in the current model, but also evolve along the branches of the phylogenetic tree. This remains computationally intractable at present. Allowing the hidden states to evolve along the tree would allow capturing large structural changes, even induced by a single mutation. For now we model such events using a jump model (see below).

Partially in order to mitigate this, each hidden state specifies a distribution over a pair of site-classes at each aligned position. This gives rise

to the possibility of a 'jump event'. A jump event allows a large change in dihedral angle or secondary structure (e.g. helix to sheet) to occur at a given aligned position and also introduces a directional coupling between changes in amino acid that are informative of changes in dihedral angle or secondary structure conformation.

Observation types

The two proteins, p_a and p_b, in a homologous pair are associated with a pair of observation sequences O_a and O_b obtained from experimental data, respectively. An ith site observation pair, $O^i = (O_a^{x(i)}, O_b^{y(i)})$, is associated with every aligned site i in an alignment M_{ab} of p_a and p_b, where $M_{ab}^i \in \{\binom{x}{y}, \binom{x}{-}, \binom{-}{y}\}$ specifies the homology relationship at position i of the alignment (homologous, deletion with respect to p_a and insertion with respect to p_a, respectively.), i is taken to run from 1 to m, m is the length of the alignment M_{ab}, and $x \in \{1,\ldots,|p_a|\}$ and $y \in \{1,\ldots,|p_b|\}$ specify the indices of the positions in p_a and p_b, respectively. $|p_a|$ and $|p_a|$ specify the number of sites in p_a and p_b, respectively.

Each site observation, $O_a^{x(i)}$ and $O_b^{y(i)}$, contains amino acid and structural information corresponding to the two C_α atoms at aligned site i belonging to each of the two proteins. A site observation corresponding to a particular protein at aligned site i, $O_a^{x(i)}$, is comprised of three different data types associated with the C_α atom: an amino acid ($A_a^{x(i)}$; discrete, one of twenty canonical amino acids), ϕ and ψ dihedral angles ($X_a^{x(i)} = \langle \phi_a^{x(i)}, \psi_a^{x(i)} \rangle$; continuous, bivariate), and a secondary structure classification ($S_a^{x(i)}$; discrete, one of three classes: helix (H), sheet (S) or coil (C)). Therefore, $O_a^{x(i)} = (A_a^{x(i)}, X_a^{x(i)}, S_a^{x(i)})$ and $O_b^{y(i)} = (A_b^{y(i)}, X_b^{y(i)}, S_b^{y(i)})$.

Model structure

The sequence of hidden nodes in the HMM is written as $H = (H^1, H^2, \ldots, H^m)$. Each hidden node H^i in the HMM corresponds to a site observation pair, $O_a^{x(i)}$ and $O_b^{y(i)}$, at an aligned site i in the alignment M_{ab}. Initially we treat the alignment M_{ab} as given *a priori*, but later modify the HMM to marginalise out an unobserved alignment.

The model is parametrised by h hidden states. Every hidden node H^i corresponding to an aligned site i takes an integer value from 1 to q for the hidden state at node H^i. In turn, each hidden state specifies a distribution over a site-class pair: (r_a^i, r_b^i) as a function of evolutionary time. A site-class pair consists of two site-classes: r_a^i and r_b^i. Each of the two site-classes takes an integer value 1 or 2, *i.e.* $(r_a^i, r_b^i) \in \{(1,1),(1,2),(2,1),(2,2)\}$. We return to the specific role of the site-classes pairs in the next section.

The state of H^i together with the site-class pair, (r_a^i, r_b^i), and the evolutionary time separating proteins p_a and p_b, t_{ab}, specify a distribution over three conditionally independent stochastic processes describing each of the three types of site observation pairs: $A^i = (A_a^{x(i)}, A_b^{y(i)})$, $X^i = (X_a^{x(i)}, X_b^{y(i)})$ and $S^i = (S_a^{x(i)}, S_b^{y(i)})$. This conditional independence structure allows the

likelihood of a site observation pair at an aligned site i to be written as follows:

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

The assumption of conditional independence provides computational tractability, allowing us to avoid costly marginalisation when certain combinations of data are missing (e.g. amino acid sequences present, but secondary structures and dihedral angles missing).

Stochastic processes: modelling evolutionary dependencies

Each site-class couples together three time-reversible stochastic processes that separately describe the evolution of the three pairs of observation types, as in equation (1). Each site-class is intended to capture both physical and evolutionary features pertaining to sequence and structure. Parameters that correspond to a particular site-class are termed *site-class specific*, whereas parameters that are shared across all site-classes are termed *global*. The use of site-class specific parameters, such as site-class specific amino acid frequencies and dihedral angle diffusion parameters, as described in the next section, is intended to model site-specific physical-chemical properties (Halpern and Bruno, 1998; Koshi and Goldstein, 1998; Lartillot and Philippe, 2004).

Amino acid evolution As is typical with models of sequence evolution, amino acid evolution, $p(A_a^{x(i)},A_b^{y(i)}|H^i,r_a^i,r_b^i,t_{ab})$, is described by a Continuous-Time Markov Chain (CTMC). Each amino acid CTMC is parameterised 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 site-class specific set of 20 amino acid equilibrium frequencies $\Pi_r^h=\text{diag}\{\pi_1,\pi_2,\ldots,\pi_{20}\}$ (19 free parameters per site-class) and a site-class specific scaling factor Λ_r^h (1 free parameter per site-class). Together these parameters define a site-class specific time-reversible amino acid rate matrix $Q_r^h=\Lambda_r^h S\Pi_r^h$. The stationary distribution of Q_r^h is given by the amino acid equilibrium frequencies: Π_r^h.

Secondary structure evolution Secondary structure evolution, $p(S_a^{x(i)},S_b^{y(i)}|H^i,r_a^i,r_b^i,t_{ab})$, is also described by a CTMC. For the sake of simplicity we use only three discrete classes to describe secondary structure 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 site-class specific set of 3 secondary structure equilibrium frequencies $\Omega_r^h=\text{diag}\{\pi_1,\pi_2,\pi_3\}$. Together they define a site-class specific time-reversible secondary structure rate matrix $R_r^h=V\Omega_r^h$, with stationary distribution: Ω_r^h.

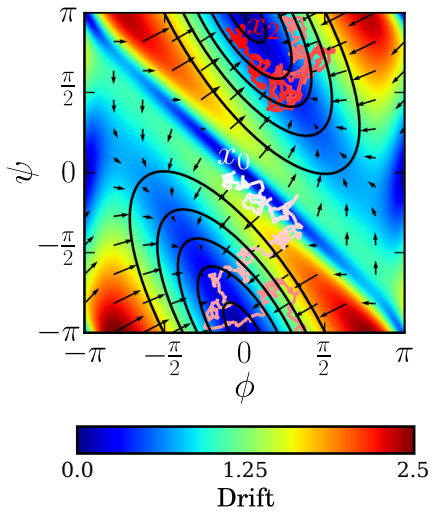

FIG. 2. Drift vector field for the WN diffusion with $A=(1,0.5;0.5,0.5)$, $\mu=(0,0)$ and $\Sigma=(1.5)^2 I$. The colour gradient represents the Euclidean norm of the drift. The contour lines represent the stationary distribution. An example trajectory starting at $x_0=(0,0)$ and ending at x_2, running in the time interval $[0,2]$ is depicted using a white to red colour gradient indicating the progression of time. The periodic nature of the diffusion can be seen by the wrapping of both the stationary diffusion and the trajectory at the boundaries of the square plane. The fact that stationary distribution is not aligned with the horizontal and vertical axes illustrates the dependence (given by α_3) between the ϕ and ψ dihedral angles.

Dihedral angle evolution Central to our model is evolutionary dependence between dihedral angles, $p(X_a^{x(i)},X_b^{y(i)}|H^i,r_a^i,r_b^i,t_{ab})$. Typically, the continuous-time evolution of the continuous-state random variables is modelled by a diffusive process such as the Ornstein-Uhlenbeck (OU) process, as in Challis and Schmidler (2012). However, an OU process is not appropriate for dihedral angles as they have a natural periodicity. For this reason, a bivariate diffusion that captures the periodic nature of dihedral angles, the *Wrapped Normal* (WN) *diffusion*, was specifically developed for this paper in García-Portugués *et al.* (2017).

Topologically, the WN diffusion (see Figure 2 for a pictorial example) can be thought of as the analogue of the OU process on the torus $\mathbb{T}^2=[-\pi,\pi)\times[-\pi,\pi)$. The WN diffusion arises as the wrapping on $\mathbb{T}^2$ of the following Euclidean diffusion:

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

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

$$w_k(\theta)=\frac{\phi_{\frac{1}{2}A^{-1}\Sigma}(\theta-\mu+2k\pi)}{\sum_{m\in\mathbb{Z}^2}\phi_{\frac{1}{2}A^{-1}\Sigma}(\theta-\mu+2m\pi)}, \tag{3}$$

$\theta\in\mathbb{T}^2$, is a probability density function (pdf) for $k\in\mathbb{Z}^2$. ϕ_Σ stands for the pdf of a bivariate Gaussian $\mathcal{N}(0,\Sigma)$. The pdf (3) weights the linear drifts of (2) such that they become smooth and periodic.

It is shown in García-Portugués *et al.* (2017) that the stationary distribution of the WN diffusion is a WN(μ,Σ), which has pdf:

$$p_{\mathrm{WN}}(\theta|\mu,\Sigma)=\sum_{k\in\mathbb{Z}^2}\phi_\Sigma(\theta-\mu+2k\pi). \tag{4}$$

Despite involving an infinite sum over $\mathbb{Z}^2$, taking just the first few terms of this sum provides a tractable and accurate approximation to the stationary density for most of the realistic parameter values.

Maximum Likelihood Estimation (MLE) for diffusions is based on the transition probability density (tpd), which only has a tractable

analytical form for very few specific processes. A highly tractable and accurate approximation to the tpd is given for the WN diffusion. This approximation results from weighting the tpd of the OU process in the same fashion as the linear drifts are weighted in (2), yielding the following multimodal pseudo-tpd:

$$\tilde{p}(\theta_2|\theta_1,A,\mu,\Sigma,t)=\sum_{m\in\mathbb{Z}^2} p_{\mathrm{WN}}(\theta_2|\mu_t^m,\Gamma_t)w_m(\theta_2), \tag{5}$$

with $\theta_1,\theta_2\in\mathbb{T}^2$, $\mu_t^m=\mu+e^{-tA}(\theta_1-\mu+2\pi m)$ and $\Gamma_t=\int_0^t e^{-sA}\Sigma e^{-sA^T}ds$. The pseudo-tpd provides a good approximation to the true tpd in key circumstances: *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 an OU process. Furthermore, it is shown in García-Portugués *et al.* (2017) that the pseudo-tpd has a lower Kullback-Leibler divergence with respect to the true tpd than the Euler and Shoji-Ozaki pseudo-tpds, for most typical scenarios and discretization times in the diffusion trajectory.

A further desirable property of the pseudo-tpd is that it obeys the time-reversibility equation, which in terms of $(X_a^{x(i)},X_b^{y(i)})$ is

$$\begin{aligned}&\tilde{p}(X_b^{y(i)}|X_a^{x(i)},A,\mu,\Sigma,t_{ab})p_{\mathrm{WN}}(X_a^{x(i)}|\mu,\tfrac{1}{2}A^{-1}\Sigma)\\&=\tilde{p}(X_a^{x(i)}|X_b^{y(i)};A,\mu,\Sigma,t_{ab})p_{\mathrm{WN}}(X_b^{y(i)}|\mu,\tfrac{1}{2}A^{-1}\Sigma).\end{aligned}$$

Indeed, the WN diffusion is the *unique* time-reversible diffusion with the stationary pdf (4), in the same way the OU is with respect to a Gaussian. Time-reversibility is an assumption of the overall model and many other models of sequence evolution. A benefit of time-reversibility in a pairwise model such as ETDBN is that one of the proteins in a pair may be arbitrarily chosen as the ancestor, thus avoiding a computationally expensive marginalisation of an unobserved ancestor.

The likelihood of a dihedral angle observation pair $(X_a^{x(i)},X_b^{y(i)})$, assuming that $X_a^{x(i)}$ is drawn from the stationary distribution, is given by:

$$\begin{aligned}&p(X_a^{x(i)},X_b^{y(i)}|H^i,r_a^i,r_b^i,t_{ab})\\&=p(X_a^{x(i)},X_b^{y(i)}|A,\mu,\Sigma,t_{ab})\\&\approx\tilde{p}(X_b^{y(i)}|X_a^{x(i)},A,\mu,\Sigma,t_{ab})p_{\mathrm{WN}}(X_a^{x(i)}|\mu,\tfrac{1}{2}A^{-1}\Sigma),\end{aligned} \tag{6}$$

A and Σ are constrained to yield a covariance matrix $A^{-1}\Sigma$. A parametrization that achieves this is $\Sigma=\mathrm{diag}(\sigma_1^2,\sigma_2^2)$ and $A=\left(\alpha_1,\frac{\sigma_1}{\sigma_2}\alpha_3;\frac{\sigma_2}{\sigma_1}\alpha_3,\alpha_2\right)$, $\alpha_1\alpha_2>\alpha_3^2$. α_1 and α_2 are the drift components for the ϕ and ψ dihedral angles, respectively. Dependence (correlation) between the dihedral angles is captured by α_3. A depiction of a WN diffusion with given drift and diffusion parameters is depicted in Figure 2.

Site-classes: constant evolution and jump events

We now turn to the meaning of the site-class pairs. Two modes of evolution are modelled: *constant evolution* and *jump events*. Constant evolution occurs when the site-class starting in protein p_a at aligned site i, r_a^i, is the same as the site-class

ending in protein p_b at aligned site i, r_b^i, *i.e.* $r_a^i = r_b^i$. Thus the distribution over observation pairs at a site is specified by a single site-class.

As already stated, a site-class specifies the parameters of the three conditionally independent stochastic processes describing amino acid, dihedral angle, and secondary structure evolution. A limitation of "constant evolution" is that the coupling between the three stochastic processes is somewhat weak. This in part stems from the time-reversibility of the stochastic processes – swapping the order of one of the three observation pairs at a homologous site, e.g. (Glycine, Proline) instead of (Proline, Glycine), does not alter the likelihood in equation (1). Alternatively restated from a generative perspective: a 'directional coupling' of an amino acid interchange does not inform the direction of change in dihedral angle or secondary structure. For example, replacing a glycine in an α-helix in one protein with a proline at the homologous position in a second protein would be expected to break the α-helix in the second protein and to strongly inform the plausible dihedral angle conformations in the second protein.

Ideally, we would consider a model in which the underlying site-classes were not fixed over the evolutionary trajectory separating the two proteins, as in the case of constant evolution as described above, but instead were able to 'evolve' in time. This would allow occasional switches in the underlying site-class at a particular homologous site, which would create a stronger dependency between amino acid, dihedral angle and secondary structure evolution, that furthermore captures the directional coupling we desire. Such an approach is considered computationally intractable due to the introduction of context-dependence when having to consider neighbouring dependencies amongst evolutionary trajectories at adjacent sites.

In order to approximate this 'ideal' model in a computationally efficient manner we introduce the notion of a *jump event*. A jump event occurs when $r_a^i \neq r_b^i$. Whereas constant evolution is intended to capture angular drift (changes in dihedral angles localised to a region of the Ramachandran plot), a jump event is intended to create a directional coupling between amino acid and structure evolution, and is also expected to capture angular shift (large changes in dihedral angles, possibly between distant regions of the Ramachandran plot).

The hidden state at node H^i, together with the evolutionary time t_{ab} separating proteins p_a and p_b, specifies a joint distribution over a site-class pair:

$$p(r_a^i, r_b^i | H^i, t_{ab}) = p(r_a^i | H^i, r_b^i, t_{ab}) p(r_b^i | H^i), \quad (7)$$

where

$$p(r_a^i|H^i,r_b^i,t_{ab}) = \begin{cases} e^{-\gamma_{H^i}t_{ab}} + \pi_{H^i,r_b^i}(1-e^{-\gamma_{H^i}t_{ab}}), & \text{if } r_a^i = r_b^i, \\ \pi_{H^i,r_b^i}(1-e^{-\gamma_{H^i}t_{ab}}), & \text{if } r_a^i \neq r_b^i, \end{cases}$$

and $p(r_a^i|H^i)=\pi_{H^i,r_a^i}$ and $p(r_b^i|H^i)=\pi_{H^i,r_b^i}$. π_{H^i,r_a^i} and π_{H^i,r_b^i} are model parameters specifying the probability of starting in site-class r_a^i or r_b^i, respectively, corresponding to the hidden state specified by node H^i. $\gamma_{H^i}>0$ is a model parameter specifying the jump rate corresponding to the hidden state specified by node H^i.

The site-class jump probabilities have been chosen so that time-reversibility holds, in other words:

$$p(r_a^i|H^i,r_b^i,t_{ab})p(r_b^i|H^i)=p(r_b^i|H^i,r_a^i,t_{ab})p(r_a^i|H^i).$$

The hidden state at node H^i, together with a site-class pair (r_a^i,r_b^i) and the evolutionary time t_{ab}, specifies the joint likelihood over site observation pairs:

$$p(O_a^{x(i)},O_b^{y(i)}|H^i,r_a^i,r_b^i,t_{ab}) = \begin{cases} p(O_a^{x(i)},O_b^{y(i)}|H^i,r_c^i,t_{ab}), & \text{if } r_a^i = r_b^i = r_c^i, \\ p(O_a^{x(i)}|H^i,r_a^i)p(O_b^{y(i)}|H^i,r_b^i), & \text{if } r_a^i \neq r_b^i. \end{cases} \quad (8)$$

In the case of constant evolution, evolution at aligned i is described in terms of the same site-class r_c^i. Evolution is considered constant because each observation type is drawn from a single stochastic process specified by H^i and r_c. Note that the strength of the evolutionary dependency within an observation pair is a function of the evolutionary time t_{ab}.

In the case of a jump event, the evolutionary processes are, after the evolutionary jump, restarted independently in the stationary distribution of the new site-class. Thus the site observations $O_a^{x(i)}$ and $O_b^{y(i)}$ are assumed to be drawn from the stationary distributions of two separate stochastic processes corresponding to site-classes r_a^i and r_b^i, respectively. This implies that, conditional on a jump, the likelihood of the observations is no longer dependent on t_{ab}. A jump event is an abstraction that captures the end-points of the evolutionary process, but ignores the potential evolutionary trajectory linking the two site observations. The advantage of abstracting the evolutionary trajectory is that there is no need to perform a computationally expensive marginalisation over all possible trajectories, as might be necessary in a model where the hidden states evolve along a tree. The likelihood of an observation pair is now simply a sum over the four possible site-class pairs:

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

where $R=\{(1,1),(1,2),(2,1),(2,2)\}$ is the set of four site-class pairs, $p(O^{x(i)},O^{y(i)}|H^i,r_a^i,r_b^i)$ is given by (8) and $p(r_a^i,r_b^i|H^i,t_{ab})$ is given by (7).

Identification of evolutionary motifs encoding jump events

In order to identify aligned sites having potential evolutionary motifs encoding jump events, a specific criterion was developed.

For a particular protein pair, inference was performed under the model conditioned on the amino acid sequence and dihedral angles for both proteins, (A_a, A_b, X_a, X_b). Homologous sites corresponding to a single hidden state and with evidence of a jump event ($r_a^i \neq r_b^i$) at posterior probability >0.90 were identified, that is, the i's such that $p(H^i, r_a^i \neq r_b^i | A_a, A_b, X_a, X_b) > 0.90$.

In a second filtering step, amino acid sequences and a single set of dihedral angles corresponding to one of the proteins were used (A_a, A_b, X_a or A_a, A_b, X_b) to infer the posterior probability, this time at a lower threshold: $p(H^i, r_a^i \neq r_b^i | A_a, A_b, X_a) > 0.50$ or $p(H^i, r_a^i \neq r_b^i | A_a, A_b, X_b) > 0.50$. This second criterion ensured that the evolutionary motif was identifiable under typical conditions where one has limited access to structural information (in this case a single protein structure in a pair). Only those aligned sites meeting both criteria were selected for downstream analysis.

Statistical alignment: modelling insertions and deletions

Protein sequences can not only undergo amino transitions due to underlying nucleotide mutations in the coding sequence, but also indel events. To account this, a modified pairwise TKF92 alignment HMM based on Miklós *et al.* (2004) was implemented. The TKF92 alignment HMM was augmented with the ETDBN's evolutionary hidden states in order to capture local sequence and structure evolutionary dependencies. Furthermore, it was modified such that neighbouring dependencies amongst hidden states at adjacent alignment sites were modelled. For more details see 'Statistical alignment' in the Supplementary Material.

Whilst it is possible to fix the alignment in advance by pre-aligning the sequences using one of the many available alignment methods (Katoh et al. (2002); Edgar (2004)) or using a curated alignment (such as from the HOMSTRAD database), doing so ignores alignment uncertainty.

Training and test datasets

A training dataset of 1200 protein pairs (2400 proteins; 417,870 site observation pairs) and a test dataset of 38 protein pairs (76 proteins; 14125 site observation pairs) were assembled from 1032 protein families in the HOMSTRAD database. For further details see 'Construction of test and training datasets' in the Supplementary Material.

Model training and selection

Maximum likelihood estimation of the model parameters, $\hat{\Psi}$, was done using Stochastic Expectation Maximisation (StEM, Gilks *et al.* (1995)).

For further details of the E- and M- steps of the StEM algorithm and for details about

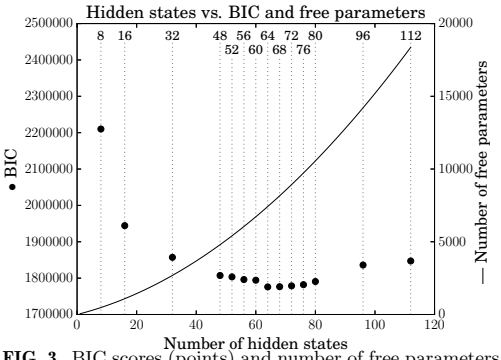

FIG. 3. BIC scores (points) and number of free parameters (curve) as a function of the number of hidden states across 14 models (indicated above the dotted vertical lines) trained using the 1200 protein pairs in the training dataset. A 64 hidden state model had the lowest BIC score. Each model represents the best of several attempts.

model selection please refer to 'Model training and selection' in the Supplementary Material.

Results and discussion

Selecting the number of hidden states

Fourteen models were trained (8, 16, 32, 48, 52, 56, 60, 64, 68, 72, 76, 80, 96 and 112 hidden state models). The 64 hidden state model was chosen as the best model, as it had the lowest Bayesian Information Criterion (BIC, Figure 3).

Stationary distributions over dihedral angles capture the empirical distribution

Figure 4 illustrates the sampled and empirical dihedral angle distributions. There is a good correspondence between dihedral angles sampled under the model (Figure 4, left) and the empirical distribution of dihedral angles in our training dataset (Figure 4, right) for all three cases illustrated (all amino acids, glycine only and proline only). The correspondence is not surprising given that ETDBN is effectively a mixture model with a large number of mixture components.

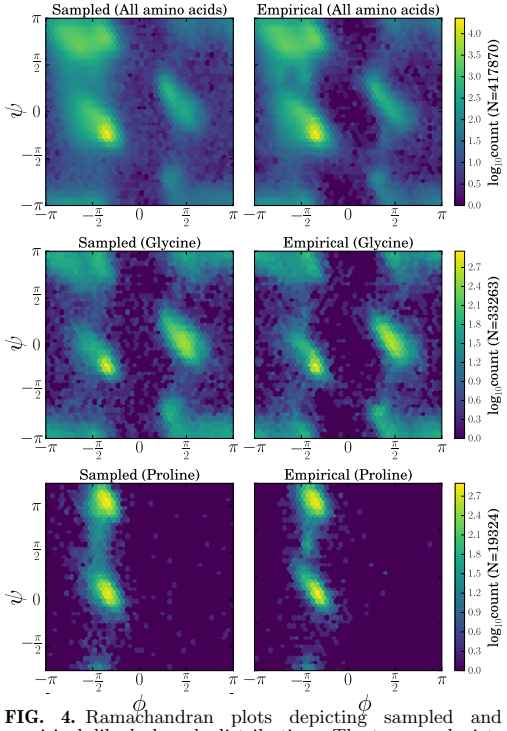

FIG. 4. Ramachandran plots depicting sampled and empirical dihedral angle distributions. The top row depicts the distributions for all amino acids, the middle for glycine only and the bottom for proline only. The leftmost plots show dihedral angles sampled under the jump model, whereas the rightmost plots show the empirical distributions of dihedral angles in the training dataset.

Estimates of evolutionary time from dihedral angles are consistent with estimates from sequence

Whilst ETDBN has a general scope with respect to applications (including acting as proposal distribution or as a building block in a homology modelling application), we envision the primary application being inference of evolutionary parameters.

Figure 5 compares evolutionary times estimated using only pairs of homologous amino acid sequences versus only pairs of homologous dihedral angles. As desired, the two estimates of evolutionary time for each protein pair are similar,

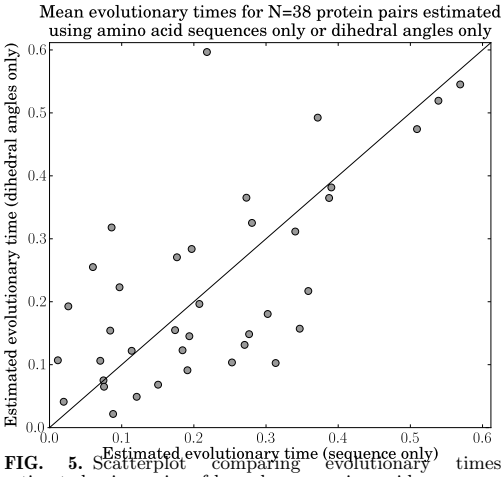

FIG. 5. Scatterplot comparing evolutionary times estimated using pairs of homologous amino acid sequences only versus pairs of homologous sets of dihedral angles only for $N=38$ proteins pairs in the test dataset. The x-coordinate of each point gives the estimated evolutionary time based only on the amino acid sequence, whereas the y-coordinate gives the estimated evolutionary time based only on the dihedral angles. The diagonal line represents $y=x$.

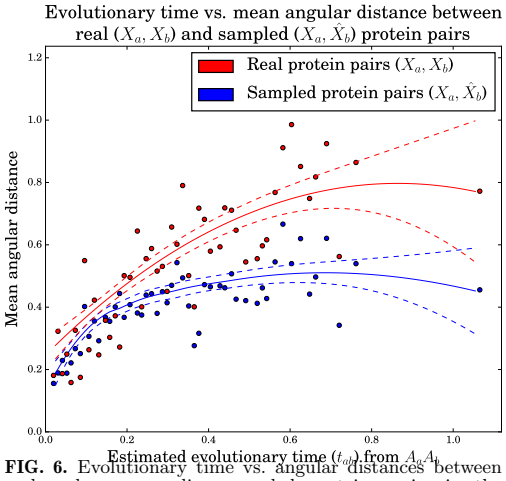

FIG. 6. Evolutionary time vs. angular distances between real and corresponding sampled proteins pairs in the training dataset at 50 representative evolutionary times. Mean angular distances (see 'Calculation of angular distances' in the Supplementary Material) between the real dihedral angles in a protein pair (red, X_a and X_b) and sampled dihedral angles in a sampled protein pair (blue, X_a and $\hat{X}_b$) were compared to test how well the sampled dihedral angles reproduced the real angular distances. The dihedral angles ($\hat{X}_b$) of each sampled protein pair were sampled by conditioning on both amino acid sequences and the homologous dihedral angles (A_a, A_b, X_a), and the estimated evolutionary time ($\hat{t}_{ab}$) for the real protein pair. The regression curves were obtained by a quadratic LOcally-weighted regrESSion (LOESS), with smoothing parameter chosen by leave-one-out cross-validation. The 95% confidence intervals for the mean assume error normality.

as can be seen by the proximity of the points to the identity line.

A paired t-test gave a p-value of 0.578, thus failing to reject the null hypothesis that there is no difference between branch lengths estimated using sequence only vs. angles only. This indicates that there is sufficient evolutionary information in the dihedral angles to estimate the evolutionary times and that the model is consistent in its estimates, lacking a significant tendency to under-estimate or over-estimate the evolutionary times when either sequence or dihedral angles are used.

Interestingly, the variance in the sampled evolutionary times is higher when dihedral angles only are used, as compared to sequence only (see Supplementary Figure 1).

The relationship between evolutionary time and angular distance is adequately modelled

We investigated the relationship between evolutionary time and angular distance between real protein pairs and protein pairs where the dihedral angles of p_b (X_b) were treated as missing and hence sampled (Figure 6).

As expected, for both real and sampled pairs, angular distance tends to increase as a function evolutionary time. For larger evolutionary times a plateau begins to emerge, which is expected as the maximum possible theoretical angular distance is $\sqrt{8} \approx 2.828$.

When the evolutionary time is exactly zero ($t_{ab}=0$) under our model, the angular distance between sampled dihedral angles is exactly zero (not shown in Figure 6), however, this is not expected to be the case for real protein pairs when the two sequences are identical (due to the inherently flexible nature of proteins, different experimental conditions, experimental noise, etc.). It is therefore not surprising that the regression curve for the real protein pairs does not pass through zero.

For small evolutionary times (<0.2) the curves for the real and sampled protein pairs show a good correspondence, however, for larger evolutionary times the model tends to under-estimate angular distances. This may reflect the fact that the tpd of the WN diffusion specified is localised around its mean, even when the evolutionary time is large, therefore dihedral angles distant from this mean are unlikely to be sampled. To a certain extent this is mitigated by the jump model, which occasionally allows for large changes in dihedral angle, but may still be somewhat limited in its flexibility, as jumps can only occur between two site-classes. The majority of protein pairs in our training dataset represent smaller evolutionary times (81.7% of evolutionary times are smaller than 0.4) and therefore protein pairs with larger evolutionary times and their associated jumps are under-represented in our dataset, which may also explain the under-estimation.

An additional possibility is that ETDBN does not attempt to model global dependencies. Echave and Fernández (2010) use a LFENM model (which does take into account global dependencies) provide evidence showing that the majority of structural changes are due to collective global deformations rather than local deformations. A local model such as ETDBN, by definition, does not take into account global dependencies and therefore does not fully account for their contribution to structural divergence.

Evaluation of the model

The conditional independence structure in (1) enables computationally efficient sampling from the model under different combinations of observed or missing data. For example, ETDBN can be used to sample (*i.e.* predict) the dihedral angles of a protein from its corresponding amino acid sequence, a homologous amino acid sequence, a homologous set of dihedral angles, the corresponding secondary structure, a homologous secondary structure or any combination of them.

Predictive accuracy was measured using 38 homologous protein pairs in the test dataset. For every protein pair (p_a,p_b), the dihedral angles of p_b in each pair were treated as missing, and these missing dihedral angles were sampled under the model given a particular combination of observation types. The average angular distance between the sampled and known dihedral angles was used as the measure of predictive accuracy.

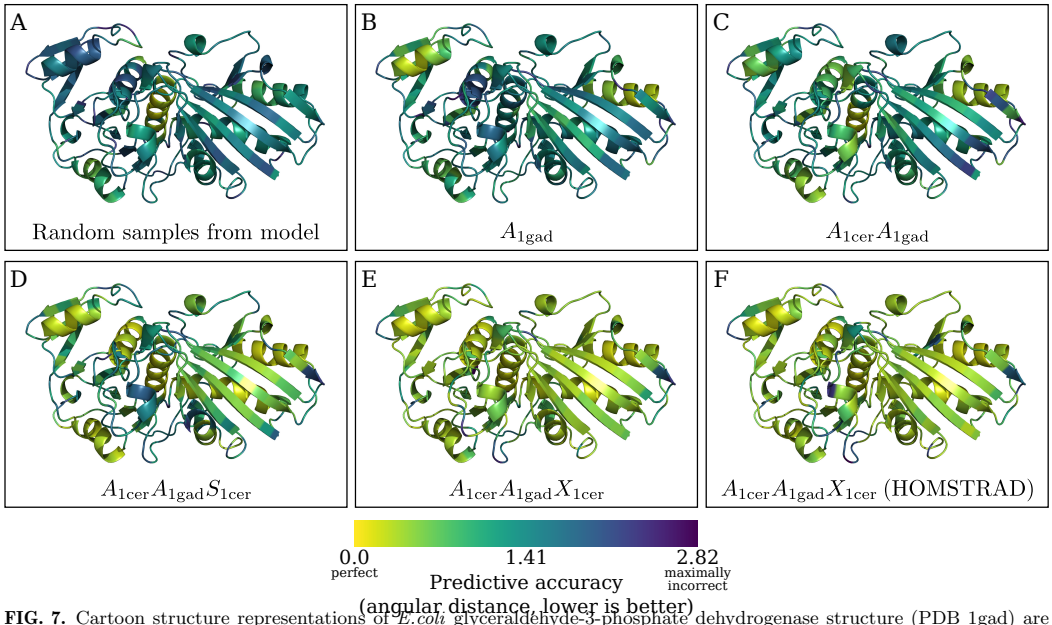

FIG. 7. Cartoon structure representations of *E.coli* glyceraldehyde-3-phosphate dehydrogenase structure (PDB 1gad) are depicted in each panel, overlaid with predictive accuracy when using different combinations of observed data to predict missing dihedral angles in 1gad. *Thermus aquaticus* glyceraldehyde-3-phosphate dehydrogenase (PDB 1cer) was used as a homolog for the purposes of prediction. Predictive accuracy is indicated using a colour gradient depicting the mean angular distance between the true dihedral angle (X^i_{1gad}) and the predicted (sampled) dihedral angles ($\hat{X}^i_{1gad}$) at each amino acid position. The label at the bottom of each panel indicates the data combination used. In A, no data was used for prediction. In B, only the amino acid sequence corresponding to 1gad (A_{1gad}) was used. In C, the amino acid sequence of 1gad (A_{1gad}) and the amino acid sequence of the homologous protein (A_{1cer}) were used. In D, both amino acids sequences (A_{1cer} and A_{1gad}) and the secondary structure of the homologous protein (S_{1cer}) were used. In E, both the amino acid sequences (A_{1cer} and A_{1gad}) and the dihedral angles of the homologous protein (X_{1cer}) were used. Finally, in panel F the same combination of observations was used as in E, but the alignment was treated as known *a priori*.

Figure 7 gives an example of predictive accuracy under different combinations of observations types overlaid on a cartoon structure of the protein structure being predicted, whereas Figure 8 provides a representative view of predictive accuracy across 10 different protein pairs in the test dataset for different combinations of observations types. We highlight some of the key patterns identified in Figures 7 and 8 as follows.

Combination 1 refers to random sampling from the model, implying no data observations were conditioned on besides the respective lengths of proteins p_a and p_b. The average angular distance between the true and predicted dihedral angles was 1.6. Random sampling acts as a baseline for predictive accuracy. It is apparent from Figure 7 that the model has a propensity to predict right-handed α-helices, which is the most populated region in the Ramachandran plot.

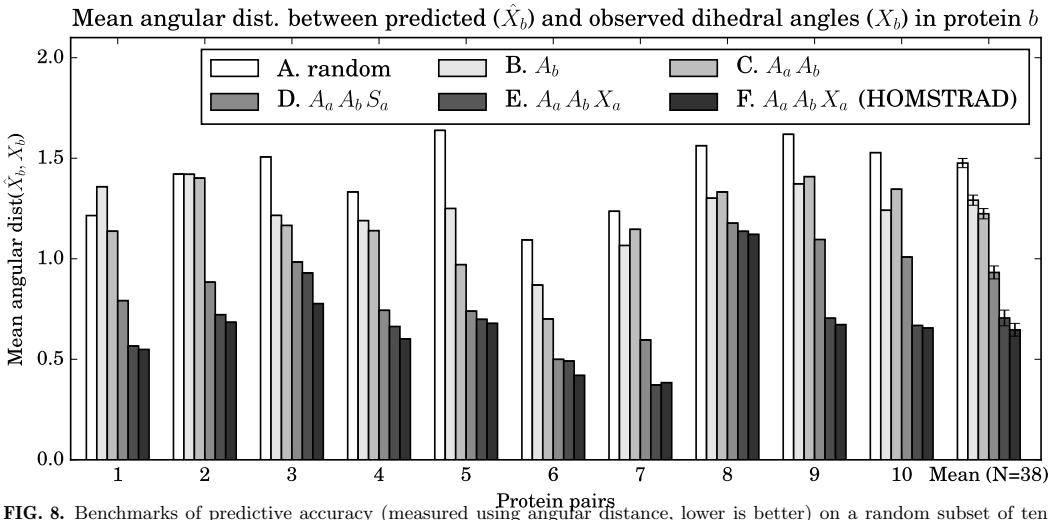

FIG. 8. Benchmarks of predictive accuracy (measured using angular distance, lower is better) on a random subset of ten protein pairs in the test dataset, giving a representative view of predictive accuracy under six different combinations of observations. The dihedral angles X_b of protein b were treated as missing and were sampled under the model, whereas p_a was a homologous protein used for the purposes of prediction. See the legend of Figure 7 for a description of each combination (A-F). The final set of bars, denoted 'Mean (N=38)', are the mean values for the entire test dataset of N=38 protein pairs. The error bars are the standard errors.

Under combination 2, only the amino acid sequence corresponding to p_b is observed. As expected in Figures 7 and 8 there is an increase in predictive accuracy with the addition of the amino acid sequence relative to combination 1.

Under combination 3, we add in the amino acid sequence of a homologous protein (p_a). In all ten cases there is an improvement in predictive accuracy. The improvement in predictive accuracy is reasonable, as knowledge of the sequence evolutionary trajectory is expected to encode information about structure evolution and hence will inform the dihedral angle conformational possibilities.

Under combination 4, in addition to the two amino acid sequences we treat the homologous secondary structure as observed. This results in a substantial improvement in predictive accuracy as one would expect. Knowledge of the amino acid sequence and a homologous secondary structure strongly informs regions of the Ramachandran plot that are likely to be occupied.

Under combination 5 (which we consider the *canonical* combination – the standard homology modelling scenario), we treat both amino acid sequences as observed, as well as the dihedral angles of the homologous protein (p_a) – in all cases the predictive accuracy improves over combination 4. This is anticipated as the homologous dihedral angles are expected to be the best proxy for missing dihedral angles and are therefore expected to be more informative than secondary structure alone. Note that the availability of a homologous amino acid sequence pair here and in combination 4 is consequential as it informs the evolutionary time t_{ab} parameter,

which will typically constrain the distribution over dihedral angles and reduce the associated uncertainty.

Finally, in combination 6, the same data observations as in combination 5 are used, except the alignment is treated as given *a priori* (by the HOMSTRAD alignment) rather than as unobserved. The HOMSTRAD alignment is based on a structural and sequence alignment of p_a and p_b and therefore is expected to encode a higher degree of homology and structural information than combination 5 (where the alignment is treated as unobserved and therefore a marginalisation over alignments is performed). On average, there is a slight improvement in predictive accuracy when fixing the alignment, albeit the magnitude of improvement is not substantial. This demonstrates the accuracy of the alignment HMM.

The alignment HMM accounts for alignment uncertainty in a principled manner, which is particularly useful when an appropriate alignment is unavailable. However, it should be noted that inference scales $\mathcal{O}(|p_a||p_b|h^2)$ when treating the alignment as unobserved, where $|p_a|$ and $|p_b|$ are the lengths of p_a and p_b, respectively. Inference scales $\mathcal{O}(mh^2)$ when the alignment is fixed *a priori*, where m is the length of alignment M_{ab} and is typically much smaller than $|p_a||p_b|$.

It should be emphasised that we do not expect ETDBN to compete with structure prediction packages such as Rosetta (Rohl *et al.*, 2004) or homology modelling software such as Arnold *et al.* (2006) in terms of predictive accuracy. Our current model is a local model of structure evolution – it is not even expected capture fundamental constraints such as the radius of gyration of a protein or other global features typical of proteins.

Evolutionary hidden states reveal a common evolutionary motif

One benefit of ETDBN is that the 64 evolutionary hidden states learned during the training phase are interpretable. We give an example of a hidden state encoding a jump event that was subsequently found to represent an *evolutionary motif* present in a large number of protein pairs in our test and training datasets.

Evolutionary hidden state 3 (Figure 9) was selected from the 64 hidden states as an example of a hidden state encoding a jump event and capturing angular shift (a large change in dihedral angle). A notable feature of this hidden state is that the change in dihedral angles between site-classes r_1 and r_2 is associated with specific amino acid changes. In site-class r_1 the amino acid frequencies are relatively spread out amongst a number of amino acids, whereas in site-class r_2 the frequencies are particularly concentrated in favour of glycine (Gly) and asparagine (Asp), with glycine being significantly more probable in site-class r_2 than r_1. This suggests that, conditioned on hidden state 3, an exchange between a glycine to another amino acid is likely indicative of a jump and hence a corresponding change in dihedral

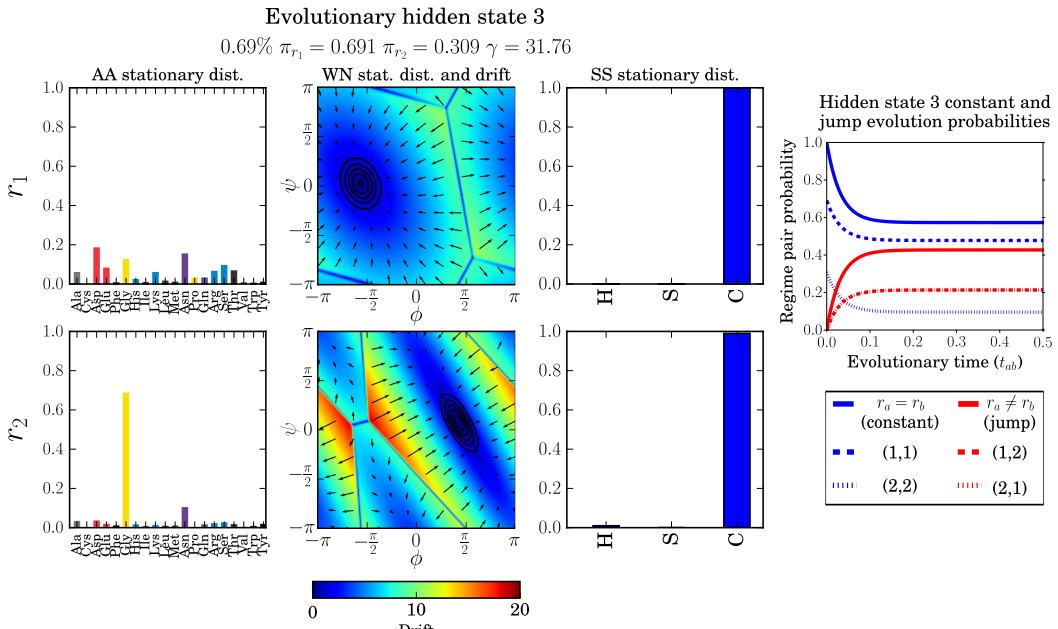

FIG. 9. Depiction of evolutionary hidden state 3. This hidden state was sampled at 0.69% of sites (the average was 1.56%). The equilibrium frequencies of r_1 and r_2 were $\pi_1 = 0.691$ and $\pi_2 = 0.309$, respectively. The jump rate was $\gamma = 31.76$. The corresponding site-class pair probabilities are depicted to the right as a function of evolutionary time. Note that the dashed red lines depicting the probabilities for (1,2) and (2,1) superimpose exactly, because the probabilities are equal – this holds for the jump probabilities of all hidden states as it is required for time-reversibility. In the main figure, the two rows depict the parameters encoded by the two site-classes, respectively. Columns 1 and 3 depict the parameters governing the the amino acid and secondary structure stochastic processes, respectively. The secondary structure classes correspond to H=helix, S=sheet and C=coil. Column 2 depicts the WN diffusions. The stationary distributions of the WN diffusions are shown using black contour lines, the direction of the drifts are indicated by the arrows and the magnitude of the drifts at each position indicated using the colour gradient.

angle. This is consistent with what we find in a subsequent analysis of evolutionary motifs. This particular jump occurs in coil regions.

Having selected hidden state 3, positions in 238 protein pairs were analysed for evidence of the corresponding evolutionary motif. 38 protein pairs in the test dataset and a further 200 from the training dataset were analysed using the criteria described in the Methods section. Using the first criterion, 84 protein sites in 59 protein pairs corresponding to $H^i = 3$ (evolutionary hidden state 3) were identified. Of the 84 protein sites, 34 protein sites met the second criterion.

We give an example of a homologous protein pair illustrating the identified evolutionary motif. Two histidine-containing phosphocarriers, 1pch (*M. capricolum*) and 1poh (*E. coli*), were identified as having the evolutionary motif (Figure 10) at homologous site E39/G39.

Most positions in the homologous pair have low posterior jump probabilities (≈ 0.0), with the exception of positions N38/N38 and E39/G39, which both have high posterior jump probabilities (≈ 1.0). The exchange between a glutamate (at position 39 in 1poh) and a glycine (at position 39 in 1pch) appears to be responsible for

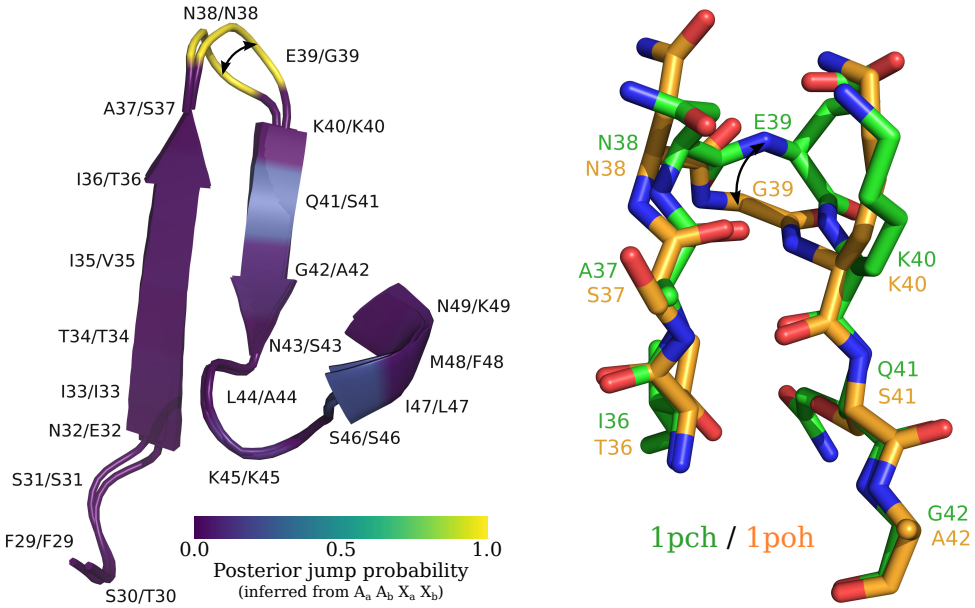

FIG. 10. Depiction of two histidine-containing phosphocarriers, PDB 1pch and 1poh, superimposed. On the left is a cartoon representation of the two proteins corresponding to regions F29-K49 and F29-A42, respectively, with posterior jump probabilities at each position overlaid. On the right is a ball-and-stick representation giving atomic detail for a smaller region (I36-G42 and T36-A42, respectively). The exchange between a glutamate (E39 in 1poh) and a glycine (G39 in 1pch) is associated with a large change in dihedral angle as indicated by the curved arrows.

the shift in dihedral angle. This exchange corresponds to a significant jump in dihedral angle: $\langle \phi_{1\text{poh,E39}}, \psi_{1\text{pch,E39}} \rangle = \langle -1.63, -0.06 \rangle \rightarrow \langle \phi_{1\text{poh,G39}}, \psi_{1\text{pch,G39}} \rangle = \langle 1.40, 0.22 \rangle$. The angular distance between the two dihedral angles is 2.01. This is consistent with the amino acid frequency parameters specified by the two site-classes for hidden state 3 (Figure 9).

Site-class r_1 indicates that a number of amino acids (alanine, aspartic acid, glycine, histidine, lysine, asparagine, proline, glutamine, arginine, serine and theorine) other than glutamate plausibly coincide with the particular dihedral angle conformation specified by site-class r_1. The involvement of glycine in a jump is not surprising as it is a small and flexible amino acid, whereas the role of asparagine is less clear. In our analysis of 238 protein pairs we found that of the seven positions meeting the criteria for hidden state 3 and involving an exchange with asparagine (Asn), four were an exchange between an asparagine and a glycine, whereas the remaining three were between asparagine and one of lysine, histidine or serine.

Using dihedral angles for alignment

A valuable feature of our model is its ability to account for alignment uncertainty by summing over possible pairwise alignments using the TKF92 model as a prior distribution over indel histories, whilst simultaneously taking

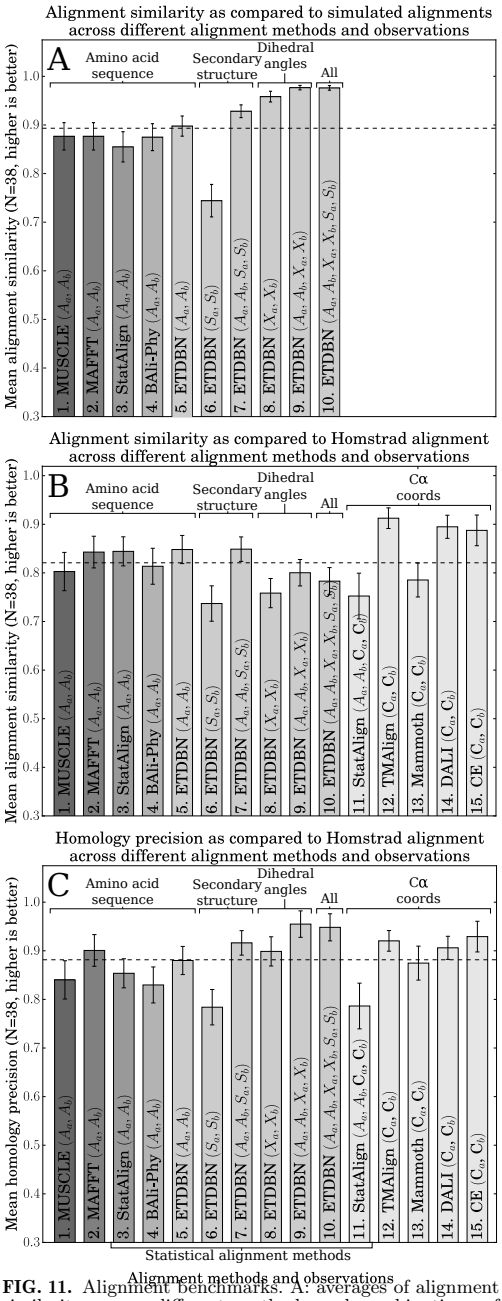

FIG. 11. Alignment benchmarks. A: averages of alignment similarity across different methods and combinations of data observations, where the observations were simulated from the ETDBN model conditioned on a single known alignment. B: averages of alignment similarity, where the 38 HOMSTRAD alignments in the test dataset were taken as the true alignments. C: averages of precision in predicting homologous site pairs in 38 sequence pairs from the test dataset, where homologous site pairs in the HOMSTRAD alignments were taken to be the true homologous site pairs.

into account neighbouring dependencies amongst aligned sites. Doing so results in a sample of alignments rather than a single alignment. Nevertheless, a single Maximum A Posteriori (MAP) pairwise alignment may be obtained from the alignment samples and used for downstream analysis.

ETDBN and several other alignment methods (namely StatAlign, BAli-Phy, MUSCLE and MAFFT) were used to infer pairwise alignments from simulated and real data under various combinations of data observations, for example: an amino acid sequence pair (A_a,A_b), a secondary structure sequence pair (S_a,S_b), a dihedral angle sequence pair (X_a,X_b) and combinations thereof.

In the first set of benchmarks (Figure 11A), pairs of proteins were simulated from the ETDBN model conditioned on 38 different pairwise alignments and corresponding evolutionary times. This resulted in a set of 38 simulated pairwise alignments together with corresponding observations, implying that the true underlying alignments were known for each of the simulated protein pairs. ETDBN and a number other alignment methods were used to infer pairwise alignments for each. The alignment similarity metric (Schwartz *et al.*, 2005) was used to measure the similarity between the inferred alignments and the true alignments, where higher similarity indicates better predictions. It was found that, when using the simulated amino acid sequences alone, ETDBN (11A.5) outperformed all four

other methods tested (11A.1 MUSCLE, 11A.2 MAFFT, 11A.3 StatAlign, 11A.4 BAli-Phy). However, the greater performance of ETDBN compared to other methods can not be considered a fair comparison, as the data were simulated under the ETDBN model.

More revealing in Figure 11A was the alignment similarity under ETDBN when using different combinations of simulated data observations. It was found that secondary structure alone (11A.6) performed the worst, which is unsurprising given that only three states were available to align the proteins. The second worst in terms of alignment similarity was amino acid sequences alone (11A.5), followed by amino acid sequences and secondary structures (11A.7). Interestingly, using dihedral angles only (11A.8) outperformed both 11A.5 (sequences only) and 11A.6 (secondary structures only). Finally, using amino acid sequence together with dihedral angles (11A.9) or all three data types combined (11A.10) outperformed all other combinations. This illustrates that, at least under simulation conditions, increasing the number of data observations results in better alignment accuracy.

Following that, the various alignment methods were benchmarked against 38 pairwise alignments consisting of real sequence and structure observations in the test dataset. These pairwise alignments were obtained from the HOMSTRAD alignments. The sequence identity of these pairwise alignments ranged from 10% to 93%, with an average sequence identity of 39%. In addition to methods $1-10$ in Figure 11A, five structural alignment methods were also used: 11. StatAlign (Herman *et al.*, 2014), 12. TMAlign (Zhang and Skolnick, 2005), 13. Mammoth (Ortiz *et al.*, 2002), 14. Dali (Holm and Rosenström, 2010) and 15. CE (Shindyalov and Bourne, 1998). These methods were not used in 11A due to the lack of an appropriate model for simulating the evolution of three-dimensional protein structures.

When benchmarking the MAP estimated alignments against the HOMSTRAD alignments (Figure 11B), using real sequences alone for inference (A_a, A_b), ETDBN (11B.5) had a similar degree of accuracy when compared to several other sequence-based methods (11B.1 StatAlign, 11B.2 BaliPhy, 11B.3 MUSCLE and 11B.4 MAFFT). This demonstrates that ETDBN has performance comparable to that of other commonly-used sequence alignment methods.

Using (S_a, S_b) alone, ETDBN (11B.6) had substantially lower alignment similarity compared to sequence only, which was expected given that a similar result was obtained for the simulated data (11A.6). However, when including the real sequences (11B.7) the predictive accuracy was once again comparable to sequence only inferences (11B.1–11B.5).

When using (X_a, X_b) alone (11B.8), the alignment similarity was found to be somewhat worse than the sequence only cases. Furthermore, when introducing the sequences (11.9) and

secondary structures (11B.10) in addition to the dihedral angles, the similarity remained worse than the sequence only methods (11B.1–11B.5), despite the additional information. These results are in contrast to the results we obtained for simulated data (11A.8–11A.10).

The non-statistical structural alignment methods (11B.12–11B.15) faired the best, likely because they use a criteria similar to that used to align the HOMSTRAD alignments. When interpreting these results it is important to note the HOMSTRAD alignments should not be considered the true underlying alignments and may even be strongly biased. For example, they may favour the closest structural superimposition of structures or the most parsimonious alignments, with the fewest number of indels. In the evolutionary modelling context our goal is to distinguish between homologous sites (sites that have evolved via mutation alone) and indels. In practice, it is extremely difficult to obtain the true underlying alignment (sets of homologies and indels), because it would require an experiment where every indel event since the common ancestor is observed, a seeming impossible task outside of simulation or laboratory conditions.

After further investigation, the trend of lower alignment similarity seen in Figure 11B.6–11B.10 when using ETDBN with structural observations compared to sequence only or non-statistical structural alignment methods was found to reverse (11C.6–11C.10) upon calculating the precision of predicting homologous sites (the fraction of sites which were predicted as homologous and were correctly predicted as such). Therefore when only dihedral angle observations are used, ETDBN underpredicts the number of homologous sites, however, when a homologous site is predicted, it is correctly predicted more often than when using only amino acid sequences. In particular, ETDBN predicted fewer homologous sites with coiled secondary structure compared to homologous sites with helical or sheet secondary structure. This pattern of results may be in part due to the WN diffusion used to model evolution of dihedral angles. The WN diffusion is suitable for modelling angular drift (small changes in angles localised around a region of the Ramachandran plot) but does not sufficiently capture angular shift (large changes in angles between regions of the Ramachandran plot, which are more likely in coiled regions) due to stationarity. As noted before, the jump model is an abstraction intended to capture the end-points of evolution by allowing a jump between two regions of the Ramachandran plot, abstracting a potential intermediate evolutionary trajectory for the sake of computational tractability. Note that the jump model accurately captures the common cases where a single mutation induces a large conformational shift.

Concluding remarks

The main achievement of this work is a computationally tractable, generative and

interpretable probabilistic model of protein sequence and structure evolution on a local scale.

Previous stochastic models of protein sequence and structure evolution emphasised estimation of evolutionary parameters (Challis and Schmidler, 2012; Herman *et al.*, 2014). ETDBN is somewhat of a departure from these previous models, but is likewise capable of estimating evolutionary parameters. We show that estimates of evolutionary times inferred under ETDBN are consistent regardless of whether amino acid sequence or dihedral angle observations are used. In addition, the relationship between evolutionary time and angular distance in real proteins is adequately recapitulated in protein pairs sampled under the model, albeit the angular distance is under-estimated for larger evolutionary times, which might be explained by the limited flexibility of the jump model and the lack of taking into account global dependencies.

Like previous models, ETDBN is capable of dealing with alignment uncertainty by marginalising over indel histories; it predicts pairwise MAP consensus alignments with accuracy similar to that of score-based and statistical alignment methods.

The generative nature of ETDBN allows us to demonstrate that the underlying empirical distributions over dihedral angles (depicted using Ramachandran plots) are captured and that the model is capable of predicting missing observations, such as dihedral angles, from a variety of different data types. For example, an amino acid sequence, a homologous amino acid sequence, a homologous secondary structure, a homologous set of dihedral angles or any combination thereof.

Due to its local nature, ETDBN does not constitute a homology modelling method in itself. Rather, it can be used as a building block, much like fragment libraries model local structure in protein structure prediction methods. ETDBN places the homology modelling problem on a statistical footing, enabling a number of approaches to later be used, such as multi-level modelling, *i.e.* combining fine-grained distributions (for example, distributions over dihedral angles, such as ETDBN) and coarse-grained distributions (for example, distributions describing the global properties of proteins, such as compactness). In particular, a method referred to as 'the Reference Ratio method' can be used to combine fine-grained and coarse-grained distributions in a statistically principled manner (see Frellsen *et al.* (2012); Hamelryck *et al.* (2010)). However, we have shown that the current model can already be used for the inference of evolutionary parameters in its present form.

In addition to multi-level modelling, probabilistic models such as ETDBN allow one to account for and to make statements about uncertainty (e.g. with respect to evolutionary time, alignment, etc.) in a rigorous manner. In principle, ETDBN, like TorusDBN (Boomsma

et al., 2008), could be used as a proposal distribution. In other words, ETDBN could be used to sample protein structures (possibly conditioned on various data observations) in a computationally efficient manner, such that the resulting samples are expected to be located in regions of high probability density with respect to the true underlying distribution.

A final key feature of our evolutionary model is its interpretable nature. This interpretability enables the identification of potential evolutionary motifs – common patterns of sequence-structure evolution. We identify one such evolutionary motif in 34 different homologous protein pairs. A major direction for future research is the further identification of such evolutionary motifs. Understanding these evolutionary motifs, may *i*) improve homology modelling predictions; *ii*) provide more accurate estimates of evolutionary parameters; and *iii*) produce better models of protein evolution that more realistically capture evolutionary trajectories through sequence and structure space, which may help identify functionally relevant positions that are potential drug targets.

Future challenges

Pairwise to phylogeny

For reasons of computational tractability the implemented model is pairwise, but it is theoretically possible to generalise it to a phylogeny, such as in Herman *et al.* (2014). In practice, for three or more sequences on a phylogeny it is necessary to marginalise out the unobserved ancestral protein states in order to compute likelihoods. Felsenstein's algorithm can be used to marginalise over discrete ancestral states, such as amino acids in a computationally efficient manner. However, we do not know whether a similar efficient algorithm exists for marginalising the continuous ancestral dihedral angle states under the WN diffusion, thereby necessitating a more expensive MCMC algorithm. A possibly greater computational hindrance to considering a phylogeny is the alignment problem, which scales $\mathcal{O}(l_1 \times l_2 \times \ldots \times l_N)$, where l_i is the length of sequence i and N is the number of sequences, although MCMC approaches are possible (Herman *et al.*, 2014).

Context-dependence

Although we believe our model provides a substantial improvement over current stochastic models of sequence and structural evolution, there is still scope for improvement. The WN diffusions used to model dihedral angle evolution adequately capture angular drift (small local changes in dihedral angle), but are less capable of capturing angular shift (large changes in dihedral angle). This is to a considerable extent mitigated by the introduction of jump events, as discussed before. A more realistic model would model the entire evolutionary trajectory, allowing an an arbitrary number of switches between site-classes together with neighbouring dependencies amongst adjacent sites along the evolutionary

trajectory. Similar context-dependent models are typically computationally expensive and require sophisticated inference procedures (Robinson *et al.*, 2003; Yu and Thorne, 2006).

Software availability

Julia code (tested on both Windows and Linux platforms) is available at: `http://www.computingforbiology.org/software/etdbn`

Supplementary material

Supplementary material is available at Molecular Biology and Evolution online: `http://www.mbe.oxfordjournals.org/`

Acknowledgements

The authors acknowledge funding from the University of Copenhagen 2016 Excellence Programme for Interdisciplinary Research (UCPH2016-DSIN). The second author acknowledges support from project MTM2016-76969-P from the Spanish Ministry of Economy and Competitiveness and ERDF. Authors acknowledge valuable comments from three referees that led to substantial improvements in the manuscript, as well as initial discussions with Christian Havn, Ian Lim and Mathias Cronjäger.

References

Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2): 195–201.

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.

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. 1985. Phylogenies and the comparative method. *American Naturalist*, 125(1): 1–15.

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.

García-Portugués, E., Sørensen, M., Mardia, K. V., and Hamelryck, T. 2017. Langevin diffusions on the torus: estimation and applications. *arXiv:1704.XXXXX*.

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

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.

Halpern, A. L. and Bruno, W. J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7): 910–917.

Hamelryck, T., Borg, M., Paluszewski, M., Paulsen, J., Frellsen, J., Andreetta, C., Boomsma, W., Bottaro, S., and Ferkinghoff-Borg, J. 2010. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PloS one*, 5(11): e13714.

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.

Holm, L. and Rosenström, P. 2010. Dali server: conservation mapping in 3D. *Nucleic acids research*, page gkq366.

Koshi, J. M. and Goldstein, R. A. 1998. Models of natural mutations including site heterogeneity.

Lartillot, N. and Philippe, H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6): 1095–1109.

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.

Miklós, I., Lunter, G., and Holmes, I. 2004. A long indel model for evolutionary sequence alignment. *Molecular Biology and Evolution*, 21(3): 529–540.

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.

Ortiz, A. R., Strauss, C. E., and Olmea, O. 2002. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Science*, 11(11): 2606–2621.

Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., and Thorne, J. L. 2003. Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution*, 20(10): 1692–1704.

Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. 2004. Protein structure prediction using Rosetta. *Methods in enzymology*, 383: 66–93.

Schwartz, A. S., Myers, E. W., and Pachter, L. 2005. Alignment metric accuracy. *arXiv preprint q-bio/0510052*.

Shindyalov, I. N. and Bourne, P. E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein engineering*, 11(9): 739–747.

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.

Whelan, S. and Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5): 691–699.

Yu, J. and Thorne, J. L. 2006. Dependence among sites in RNA evolution. *Molecular biology and evolution*, 23(8): 1525–1537.

Zhang, Y. and Skolnick, J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, 33(7): 2302–2309.