Reconstructing ancestral protein sequences and structures

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

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

March 14, 2019

Abstract

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

1 Introduction

2 Methods

2.1 Model

2.1.1 Model of a single protein

A single protein consisting of n amino acids:

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

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

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

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

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

The dihedral angle observations $(\omega_i, \phi_i, \psi_i)$ at a site, i, are each distributed according a univariate von Mises (vM) distribution with mean μ and concentration parameter κ conditional on the hidden state H_a^i and the amino acid S_a^i:

$$\begin{aligned} \omega_i &\sim \text{vM}\big(\mu_\omega(H_a^i, S_a^i),\, \kappa_\omega(H_a^i, S_a^i)\big) \\ \phi_i &\sim \text{vM}\big(\mu_\phi(H_a^i, S_a^i),\, \kappa_\phi(H_a^i, S_a^i)\big) \\ \psi_i &\sim \text{vM}\big(\mu_\psi(H_a^i, S_a^i),\, \kappa_\psi(H_a^i, S_a^i)\big) \end{aligned} \tag{1}$$

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

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

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

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

Note that the values of the three bond angles and three bond lengths are all largely invariant, implying that the values can be reasonably fixed. However, we opted to treat them as random variables so that the model gives a complete probabilistic description of the protein backbone structure.

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

Hidden Markov model The joint likelihood, $\Pi(a)$, of the structural observations, X_a, and hidden states, H_a conditioned on the amino acid observations, S_a, is given by a hidden Markov model as follows:

$$\begin{aligned} \Pi(a) = p(H_a, X_a \,|\, S_a, \theta) = \\ p\big(X_a^1 \,|\, H_a^1, S_a^1, \theta\big) p(H_a^1, \theta) \\ \times \prod_{i=2}^{n} p\big(X_a^i \,|\, H_a^i, S_a^i, \theta\big) p(H_a^i \,|\, H_a^{i-1}, \theta) \end{aligned} \tag{4}$$

where $p(H_a^i \,|\, H_a^{i-1}, \theta)$ is a transition probability matrix between hidden states intended to capture the amino acids (along with their associated structural values) that tend to occur next to one another in the

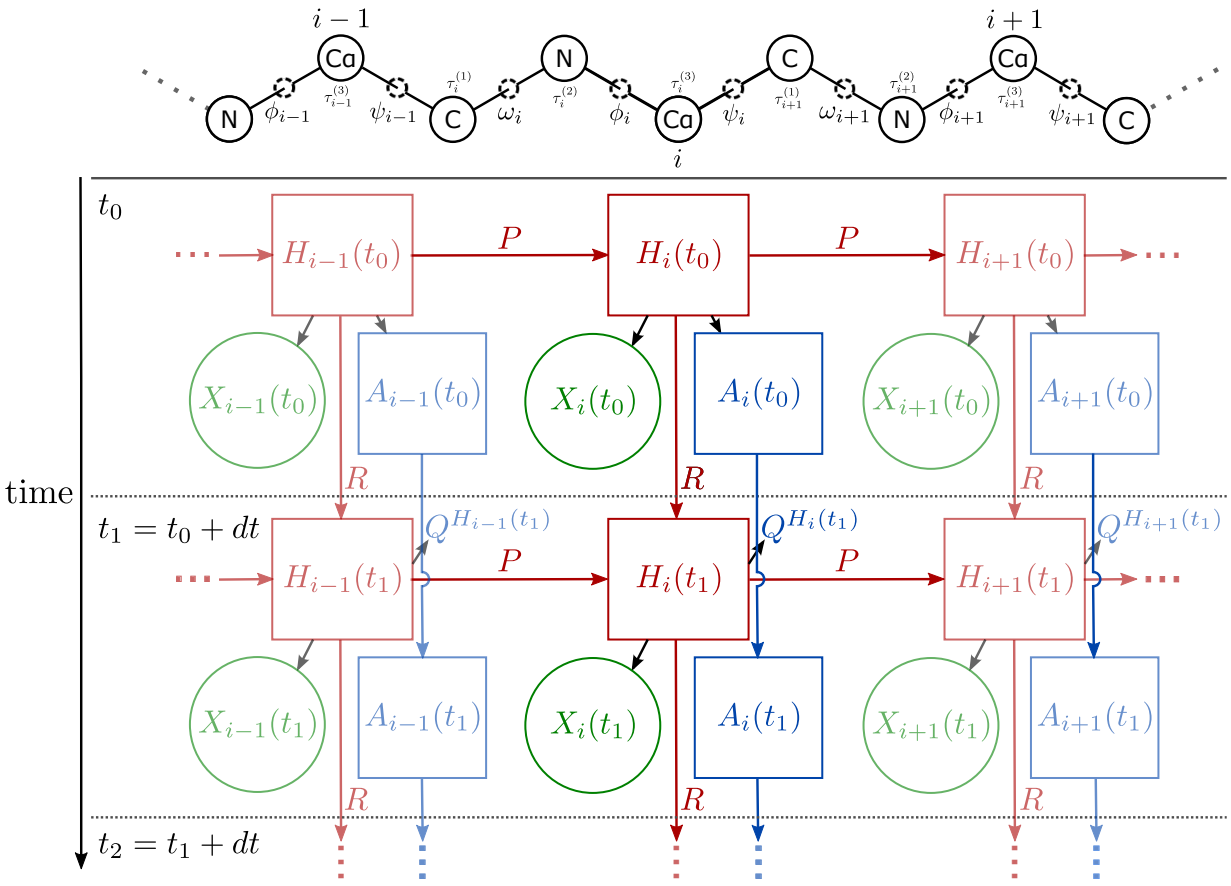

Figure 1: Above: a depiction of a protein backbone (three amino acids long) with the ω, ϕ and ψ dihedral angles and the three additional bond angles ($\tau_i^{(1)}$, $\tau_i^{(2)}$, $\tau_i^{(3)}$) shown. Bond angles and bond lengths are not to scale. Also shown are C_α atoms which attach to the amino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

peptide backbone.

2.1.2 Evolutionary model

$$M_{ab} = \begin{cases} \sqrt{\frac{\Pi(b)}{\Pi(a)}} S_{a_i b_i} \pi_{a a_i^b}^{h_i^b} & \text{Single amino acid difference at site } i. \\ \sqrt{\frac{\Pi(b)}{\Pi(a)}} R_{a_i b_i} \pi_{a a_i^b}^{h_i^b} & \text{Single hidden state difference at site } i. \\ 0 & \text{Both hidden state and amino acid differences at site } i. \\ 0 & \text{Differences at two or more sites.} \\ -\sum_{c \neq a} M_{ac} & a = b \end{cases} \tag{5}$$

where S is a symmetric 20×20 amino acid exchangeability matrix, and R is a symmetric $h \times h$ hidden state exchangeability matrix.

Note that proteins P_a and P_b referred to in the ratio $\frac{\Pi(b)}{\Pi(a)}$ in Equation 5 always differ at exactly one site, implying that at most three terms in Equation 4 need to be considered when computing the ratio.

Additionally note, although the summation in Equation 5 appears to involve an exponential number of terms, most terms are equal to zero, except those that differ from P_a at one position. Furthermore, an amino acid transition and a hidden state transition are not permitted to occur simulatenously, further reducing the number of terms that are summed ($19n$ amino acid terms, and $(h-1)n$ hidden state terms).

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

$$p(P_a|\theta) = \frac{\Pi(a) \prod_i \pi_{a a_i^a}^{h_i^a}}{\sum_k \Pi(k) \prod_i \pi_{a a_i^k}^{h_i^k}} \tag{6}$$

Time-reversibility Since S and R are symmetric matrices, i.e. $S_{a_i b_i} = S_{b_i a_i}$ and $R_{a_i b_i} = R_{b_i a_i}$, it can easily be shown that time reversibility holds, in other words:

$$p(P_a|\theta) M_{ab} = p(P_b|\theta) M_{ba} \tag{7}$$

Dataset likelihood

$$p(\mathcal{D}_d|\mathcal{B}_d, \Upsilon_d, \theta) = \prod_{b \in \mathcal{B}_d} p\big(X_b^i \,|\, H_b^i, S_b^i, \theta\big) \times \big[e^{q_{last} t_b} \prod_{i \to j} q_{ii} e^{-q_{ii}(t_b - t_b)} q_{ij}/q_{ii}\big] \tag{8}$$

2.2 Model training

2.2.1 Datasets

2.2.2 Parameter estimation

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

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

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

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

Table 1: Ancestral sequence reconstruction benchmarks

Dataset	Our model	LG2008	BEAST	ASR
Influenza	0.896	0.876	0.88	0.87
HIV	0.896	0.876	0.88	0.87

2.3 Inference

2.3.1 Gibbs sampling

3 Results and Discussion

3.1 Benchmarks of ancestral sequence reconstruction

4 Conclusions

5 Software availability

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

6 Acknowledgements

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

References

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

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

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