Reconstructing ancestral protein sequences and structures

Michael Golden[1], Jotun Hein[2], Thomas Hamelryck[3], and Oliver Pybus[1]

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

March 4, 2019

Abstract

We present a probabilistic model of protein evolution that captures several important features of protein sequence and 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 represents the complete backbone structure of each protein using an angle and bond length representation that realistically captures local features of protein structure. Such local structure evolution is modelled jointly with sequence, permitting ancestral structures 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 the information that the user can choose to provide, implying that arbitrary combinations of protein sequences and structures can be used when performing various inference tasks. The current model does not capture long-range dependencies that are necessary for accurate homology modelling and reconstruction of ancestral three-dimensional structures. However, it is ultimately expected to be combined with state-of-the-art machine-learning models of protein structure that account for these long-range structural dependencies, but that do not account for evolutionary information that can strongly inform predictions.

1 Introduction

2 Methods

2.1 Model

2.1.1 Hidden Markov model

The hidden Markov model is given by: Π

2.1.2 Rate matrix

$$M_{ab} = \begin{cases} \sqrt{\frac{\Pi(b)}{\Pi(a)}} Q_{a_i b_i} \pi^{h_i^b}_{aa_i^b} & \text{single amino acid difference at site } i. \\ \sqrt{\frac{\Pi(b)}{\Pi(a)}} R_{a_i b_i} \pi^{h_i^b}_{aa_i^b} & \text{single hidden state difference at site } i. \\ 0 & \text{both hidden state and amino acid differences at site } i. \\ 0 & \text{differences at two or more sites.} \\ -\sum\limits_{c \neq a} M_{ac} & a = b \end{cases} \tag{1}$$

2.1.3 Stationary distribution

The stationary distribution is given by $p(a)$.

3 Results

4 Software availability

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

5 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

Knudsen, B. and Hein, J. 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6): 446–454.

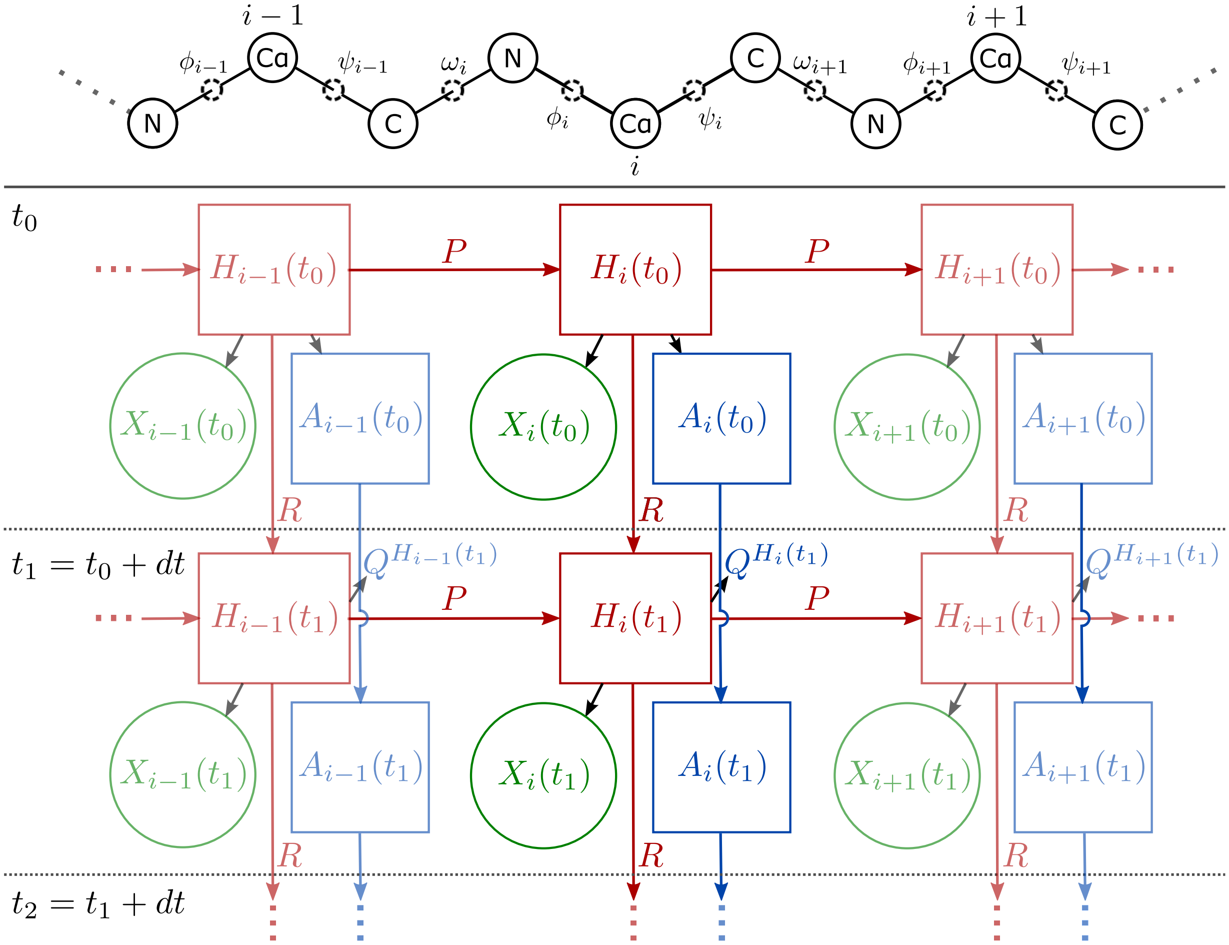

Figure 1: Graphical description of the model architecture.

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