# 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]Department of Computer Science, University of Copenhagen, Denmark

March 3, 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 mutations that occur throughout the course of evolution. The model is trained on a large numbers of protein alignments and corresponding phylogenetic tree topologies that represent the evolutionary history of the aligned proteins. This yields a model that acts as 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 modeled 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 structures of proteins using information from large numbers of related proteins. The model is highly flexible with respect to the information that the user provides. Implying that arbitary combinations of protein sequences and structures can be used when performing various inference tasks.

# 1 Introduction

The primary role of nucleic acid molecules, such as DNA (deoxyribonucleic acid) and RNA (ribonucleic acid), is encoding genetic information for storage and transfer. However, both types of molecules can form structures with additional functions (Mattick, 2003). DNA is ordinarily thought of as a double-stranded molecule forming the now iconic double helical configuration (Watson and Crick, 1953), although many viral genomes consist entirely of single-stranded DNA (ssDNA) or single-stranded RNA (ssRNA) molecules. Such single-stranded nucleic acid molecules are far less constrained than double-stranded nucleic acid molecules in the variety of structures that they can form. For example, the Rev response element (RRE) within the single-stranded HIV RNA genome plays a crucial role in the regulation of HIV virion expression by binding the HIV Rev protein to facilitate the transfer of HIV genomes from the nucleus to the cytoplasm where translation and virion packaging occur (Heaphy *et al.*, 1990; Daugherty *et al.*, 2010).

### 1.0.1 The KH99 grammar

The KH99 SCFG (Knudsen and Hein, 1999) was chosen as a prior over secondary structures. The rules and associated probabilities are given as follows:

$$\mathcal{G}_{\text{KH99}} = \begin{array}{ccccccc} S & \rightarrow & \bullet & \text{or} & LS & \text{or} & (F) \\ & & 0.118 & & 0.869 & & 0.014 \\ & & & & & & \\ L & \rightarrow & \bullet & \text{or} & (F) & & \\ & & 0.895 & & 0.105 & & \\ & & & & & & \\ F & \rightarrow & (F) & \text{or} & LS & & \\ & & 0.788 & & 0.212 & & \end{array} \tag{1}$$

Note that $S$ is the start symbol.

# 2 Software availability

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

# 3 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

Daugherty, M. D., Liu, B., and Frankel, A. D. 2010. Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. *Nature structural & molecular biology*, 17(11): 1337–1342.

Heaphy, S., Dingwall, C., Ernberg, I., Gait, M. J., Green, S. M., Kern, J., Lowe, A. D., Singh, M., and Skinner, M. A. 1990. HIV-1 regulator of virion expression (Rev) protein binds to an RNA stem-loop structure located within the Rev response element region. *Cell*, 60(4): 685–693.

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

Mattick, J. S. 2003. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 25(10): 930–939.

Watson, J. D. and Crick, F. H. 1953. Molecular structure of nucleic acids. *Nature*, 171(4356): 737–738.

Table 1: SHAPE structure ranking. Top 10 of 86 non-overlapping HIV NL4-3 substructures ranked from highest to lowest z-score based on the estimated degrees of coevolution within an alignment of HIV-1 subtype B sequences. Where the HIV NL4-3 SHAPE-MaP secondary structure was used as the canonical structure.

| Rank | Alignment position | NL4-3 position | Length | Name | Median degree of coevolution | z-score |
|---|---|---|---|---|---|---|
| 1 | 8233 - 8582 | 7249 - 7595 | 350 | Rev Response element (RRE) | 5.38 | 5.02 |
| 2 | 2608 - 2943 | 1991 - 2326 | 336 | Longest continuous helix | 5.17 | 2.92 |
| 3 | 10155 - 10383 | 8982 - 9170 | 229 | 3' Untranslated region (3'UTR) | 5.27 | 2.69 |
| 4 | 588 - 838 | 105 - 344 | 251 | 5' Untranslated region (5'UTR) | 5.65 | 2.61 |
| 5 | 9570 - 9584 | 8440 - 8454 | 15 | | 5.91 | 2.29 |
| 6 | 860 - 979 | 366 - 485 | 120 | 5' Untranslated region (5'UTR) | 5.54 | 2.28 |
| 7 | 1710 - 1845 | 1177 - 1312 | 136 | | 5.17 | 2.28 |
| 8 | 2115 - 2301 | 1561 - 1711 | 187 | Gag-pol frameshift | 5.31 | 2.21 |
| 9 | 1479 - 1490 | 946 - 957 | 12 | | 5.85 | 2.04 |
| 10 | 3886 - 3907 | 3269 - 3290 | 22 | | 5.80 | 2.01 |