Reconstruction of ancestral protein sequences using autoregressive generative models

Pierre Barrat-Charlaix[*] and Andrea Pagnani

DISAT, Politecnico di Torino, Italy

(Dated:)

Abstract

abstract

A. Introduction

B. Results

1. *Autoregressive model of sequence evolution*

Models of evolution commonly used in phylogenetics rely on the assumptions that sequence positions evolve independently and that evolution at each position i follows a continuous time Markov chain (CTMC) parametrized by a substitution rate matrix $\mathbf{Q}^i$. Matrix $\mathbf{Q}^i$ is of dimensions $q \times q$ where $q = 4$ for DNA, 20 for amino acids or 64 for codon models. The probability of observing a change from state a to state b during evolutionary time t is then given by $P(b|a,t) = \left(e^{t\mathbf{Q}^i}\right)_{ab}$.

If the model is time-reversible, it is a general property of CTMCs that the substitution rate matrix can be written as

$$\mathbf{Q} = \mathbf{H} \cdot \mathbf{\Pi} = \mathbf{H} \cdot \begin{pmatrix} \pi_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \pi_q \end{pmatrix}, \tag{1}$$

where $\mathbf{H}$ is symmetric and $\mathbf{\Pi}$ is diagonal with entries that sum to 1 [1]. The two matrices have simple interpretations. On the first hand, $\mathbf{\Pi}$ fixes the long-term equilibrium frequencies, that is $P(b|a,t) \xrightarrow[t\to\infty]{} \pi_b$. On the other, $\mathbf{H}$ influences the dynamics of the Markov chain but does not change the equilibrium distribution. Most commonly, $\mathbf{\Pi}$ is considered to be independent of the sequence position i, while $\mathbf{H}$ can be multiplied by position-dependent rates in order to model the different variability of different sites [2–4].

In order to incorporate constraints coming from a protein's structure and function into the evolutionary model, we develop a family-specific model of protein sequence evolution based on the the autoregressive generative model ArDCA [5]. ArDCA models the diversity

[*] Correspondance to: PBC, DISATemail@univ.com

of sequences in a protein family using a set of learned conditional probabilities. In practice, the model assigns a probability to any sequence $\mathbf{a} = \{a_1, \ldots, a_L\}$ of L amino acids:

$$P^{AR}(\mathbf{a}) = \prod_{i=1}^{L} p_i(a_i|a_{<i}), \tag{2}$$

where the product runs over positions in the sequence and $a_{<i} = a_1, \ldots, a_{i-1}$ represents all amino acids before position i. Functions p_i represent the probability according to the model to observe state a_i in position i, given that the previous amino acids were $a_1, \ldots, a_{i-1}$. Their precise functional form is given in the methods section. They are learned using the aligned sequences of members of the family. In actual implementations, the order in which the product in Eq. 2 is performed is not the natural $(1, \ldots, L)$ but rather an order where positions are sorted by increasing variability. This does not significantly effect the model we present below, and we keep the notation of Eq. 2 for simplicity.

It has been shown in [5] that the generative capacities of ArDCA are comparable to that of state of the art models such as bmDCA [6]. This means that a set of sequences sampled from the probability in Eq. 2 is statistically hard to distinguish from the natural sequences used in training or, in other words, that the model can be used to sample new artificial homologs of a protein family. Generative capacities of a protein model come from its ability to represent epistasis, that is the relation between the effect of a mutation and sequence context in which it occurs. Here, epistasis is modeled through the conditional probabilities p_i: the distribution of amino acids at position i depends on the states of the previous positions $1, \ldots i-1$.

We take advantage of the autoregressive architecture to define the following evolution model. Given two amino acid sequences $\mathbf{a}$ and $\mathbf{b}$, we propose

$$P(\mathbf{b}|\mathbf{a}, t) = \prod_{i=1}^{L} q_i(b_i|a_i, b_{<i}, t), \tag{3}$$

where the conditional propagator q_i is defined as

$$q_i(b_i|a_i, b_{<i}, t) = \left(e^{t \cdot Q^i(b_{<i})}\right)_{a_i, b_i}, \quad Q^i(b_{<i}) = \mathbf{H} \cdot \begin{pmatrix} p_i(1|b_{<i}) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & p_i(q|b_{<i}) \end{pmatrix}. \tag{4}$$

According to these equations, evolution for each position i follows a standard CTMC. However, we use the decomposition of Eq. 1 to set the equilibrium frequency at i to $p_i(b|b_{<i})$. In other words, we consider that position i evolves in the context of $b_1, \dots, b_{i-1}$, and that its dynamics are constrained by its long term frequency given by the autoregressive model. An important consequence of this choice is that our evolution model will converge at long times to the generative distribution P^{AR}:

$$q_i(b_i|a_i, b_{<i}, t) \xrightarrow[t\to\infty]{} p_i(b_i|b_{<i}), \quad P(\mathbf{b}|\mathbf{a}, t) \xrightarrow[t\to\infty]{} P^{AR}(\mathbf{b}). \tag{5}$$

We argue here that such a property is essential to build a realistic protein sequence evolution model, particularly when considering evolution over a relatively long time frame. Note that to converge to a generative distribution, accurate modeling of epistasis is required. Using site-specific frequencies would not be sufficient, as the effect of mutations in a protein sequence typically depends on the context [7]. The technique proposed here allows us to represent epistasis through the context dependent probabilities p_i, while still considering each sequence position one at a time.

Interestingly, we note that the model in Eq. 3 is not time reversible, although context dependent site propagators in Eq. 4 are reversible. We show in the SM that this is mainly an artifact of the autoregressive nature of the model coupled with epistasis. Using non-time reversible evolutionary models is uncommon in the field, but this is mainly due to practical considerations and there are no fundamental reasons for evolution itself to be reversible [8]. In practice, this means that algorithms using this model have to be adapted accordingly.

We underline that this approach has important differences with standard models of evolution used in phylogenenetics. In phylogenetic reconstruction, the tree and the sequence evolution model are usually inferred at the same time and from the same data. The number of parameters of the evolution model is then kept low to reduce the risk of overfitting, for instance by using site specific rates to account for variable and conserved sites. Methods that introduce more complex models such as site specific frequencies do so by jointly inferring the parameters and the tree, leading to relatively complex algorithms [9, 10].

Here instead, parameters of the generative model in Eq. 2 are learned from a protein family, *i.e.* a set of diverged homologous protein sequences. While it is true that these sequences share a common evolutionary history and cannot be considered as independent samples, common learning procedures only account for this in a very crude way [5, 11]. Despite this, it appears that the generative properties of such models are not strongly affected by ignoring the phylogeny [12, 13]. This allows us to proceed in two steps: first construct the model from data while ignoring phylogeny, and then only use it for phylogenetic inference tasks.

An advantage of this approach is that once the model of Eq. 2 is inferred, the propagator in Eq. 3 comes "for free" as no additional parameters are required. Importantly, our model does not use site specific substitution rates. Indeed, it has been shown that these can be seen as emergent properties when using more complex evolution models such as the one presented here [14]. However, a disadvantage is that the technique is only applicable to a given protein family at a time, and requires the existence of an appropriate training set for the model.

2. Ancestral sequence reconstruction

C. Discussion

D. Methods

- proper presentation of ardca (sequence ordering, functional form of cond probs)
- how I run iqtree

[1] Ziheng Yang. *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press, Oxford, New York, October 2006. ISBN 978-0-19-856702-8.

[2] Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3):306–314, September 1994. ISSN 1432-1432. doi:10.1007/BF00160154.

[3] Alexandros Stamatakis. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, 30(9):1312–1313, May 2014. ISSN 1367-4811. doi:10.1093/bioinformatics/btu033.

[4] Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, January 2015. ISSN 0737-4038. doi: 10.1093/molbev/msu300.

[5] Jeanne Trinquier, Guido Uguzzoni, Andrea Pagnani, Francesco Zamponi, and Martin Weigt. Efficient generative modeling of protein sequences using simple autoregressive models. *Nature Communications*, 12(1):5800, October 2021. ISSN 2041-1723. doi:10.1038/s41467-021-25756-4.

[6] Francisco McGee, Sandro Hauri, Quentin Novinger, Slobodan Vucetic, Ronald M. Levy, Vincenzo Carnevale, and Allan Haldane. The generative capacity of probabilistic protein sequence models. *Nature Communications*, 12(1):6302, November 2021. ISSN 2041-1723. doi:10.1038/s41467-021-26529-9.

[7] Michael Socolich, Steve W. Lockless, William P. Russ, Heather Lee, Kevin H. Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058): 512–518, September 2005. ISSN 1476-4687. doi:10.1038/nature03991.

[8] Felsenstein, Joseph. *Inferring Phylogenies*. Oxford university press edition, September 2003. ISBN 978-0-87893-177-4.

[9] A L Halpern and W J Bruno. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15(7):910–917, July 1998. ISSN 0737-4038. doi:10.1093/oxfordjournals.molbev.a025995.

[10] Vadim Puller, Pavel Sagulenko, and Richard A Neher. Efficient inference, potential, and limitations of site-specific substitution models. *Virus Evolution*, 6(2), August 2020. ISSN

2057-1577. doi:10.1093/ve/veaa066.

[11] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Remi Monasson, and Martin Weigt. Inverse Statistical Physics of Protein Sequences: A Key Issues Review. *Reports on Progress in Physics*, 81(3):032601, March 2018. ISSN 0034-4885, 1361-6633. doi:10.1088/1361-6633/aa9965.

[12] Adam J. Hockenberry and Claus O. Wilke. Phylogenetic Weighting Does Little to Improve the Accuracy of Evolutionary Coupling Analyses. *Entropy*, 21(10):1000, October 2019. ISSN 1099-4300. doi:10.3390/e21101000.

[13] Edwin Rodriguez Horta and Martin Weigt. On the effect of phylogenetic correlations in coevolution-based contact prediction in proteins. *PLoS computational biology*, 17(5):e1008957, May 2021. ISSN 1553-7358. doi:10.1371/journal.pcbi.1008957.

[14] Jose Alberto de la Paz, Charisse M. Nartey, Monisha Yuvaraj, and Faruck Morcos. Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proceedings of the National Academy of Sciences*, page 201913071, March 2020. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1913071117.

Supplementary Material: Reconstruction of ancestral protein sequences using autoregressive generative models

Appendix A: Autoregressive evolution model

- Show the simplified expression for H=1
- discuss irreversibility, with example