Towards a Null Model for Detecting Epistasis in Protein Multiple Sequence Alignments

Grey Wilburn
MIT 8.592 Final Project

(Dated: May 18, 2018)

Multiple sequence alignments (MSAs) of related biological sequences allow for the study of the evolution of biological molecules over time. In recent work [1], Konate et al claim that differing patterns of amino acid conservation in enzyme MSAs across different lineages is evidence of the effect of epistasis in protein evolution. Here, I use simple evolutionary simulations to determine whether a simple, non-epistatic model of evolution can account for heterogenous patterns of site conservation in related proteins across different regions of phylogeny. I find that by using site-independent profile models, simulations produce similar patterns of site conservation across distant lineages compared to [1].

I. INTRODUCTION

Epistasis describes the dependence of the effect of a mutation on the state of the genome. As an example of epistasis, consider independent sequences of the same gene with mutations A and B with fitness S_A and S_B relative to the wild type, respectively. Epistasis is said to exist of the fitness S_{AB} of the double mutant AB is not equal to the sum of S_A and S_B. The role of intragenic epistasis in protein evolution is an open question subject to much recent debate. For example, Breen et al recently argued that epistasis is a critical factor in protein evolution [2], while McCanlish et al refuted this claim while examining similar data [3].

To determine the role of epistasis in protein evolution, Konate et al analyze MSAs of enzyme sequences with conserved molecular function [1]. Among their many findings is that in deep MSAs of a bacterial enzyme, dihydrofolate reductase FoIA, and an *E. coli* translation initiation factor, InfA, the location of conserved residues varies heavily between lineages. For example, they note that the set of conserved sites ($\geq 90\%$ conserved) between two lineages with a divergence time of around 2 billion years only overlap by 10-20%. They conclude that epistasis is a major force in protein evolution, with not all combinations of amino acids allowing for the conservation of molecular function.

I seek to answer the following question: *Can the observation that a heterogenous set of sites are conserved across different lineages of related proteins be described by an evolutionary model that does not allow for epistasis?* I believe that other factors such as phylogeny may lead to the results seen by Konate et al. To do so, I have performed simply evolutionary simulations using the globin family of proteins.

II. METHODS

Null models for the analysis of coevolution of residues in biological sequences have been used in structural biology before. Covariation between sites has been used to predict the three dimensional structure of proteins and RNA from primary sequence data. Lapedes et al performed evolutionary simulations to examine the effects of phylogeny on covariation of sites in protein MSAs [5]. Rivas et al performed similar simulations for structural RNA MSAs [6].

I have attempted to adapt these methods towards creating a null model for identifying the signal of epistasis in protein MSAs. I have simulated sequence mutation down a binary phylogenetic tree using three separate simple models of evolution:

- A site-independent, non site-specific model based on the Jukes-Cantor model of nucleotide evolution
- A site-independent, site-specific model using profile models, coupled with Markov Chain Monte Carlo
- A site-specific Potts model that accounts for the coevolution of sites in a protein, coupled with Markov Chain Monte Carlo.

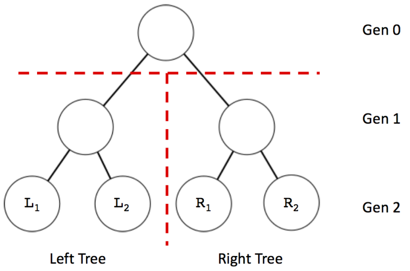
The first two models do not allow for epistasis between sites in the protein and can be considered candidate null models. The Potts model allows for pairwise epistasis.

A. Construction of Phylogenetic Tree

I evolve sequences down a binary tree with 11 levels ("generations"). The evolutionary time is fixed for each sequence and each generation, but varies between the three evolutionary models. After a fixed time with evolution, each sequence at each node in the tree is duplicated (a "speciation" event), and its sequence is passed to two daughter nodes.

To compare sequences across two different regions of phylogeny, I break the tree at the first generation, yielding a "left subtree" and a "right subtree". I then compare sequences and sites across subtrees. For each evolutionary model, the branch length is set so that the average pairwise identity between an 11th generation sequence in the left subtree and an 11th generation sequence in the right subtree is approximately 25 %, which is roughly

FIG. 1: Diagram of the construction of the out-of-clade datasets. The binary tree is split between the two daughter nodes of the root. The nodes of the last generation in the left subtree are then compared to the nodes in the last generation of the right tree.



the percent identity Konate et al observe for enzyme orthologs with a divergence time of approximately two billion years [1]. I have independently simulated 100 such trees of nodes containing 96-residue sequences using each of the three evolutionary models.

B. Evolutionary Models

1. *Jukes-Cantor Style Model*

For the site independent, non site-specific model, I have modeled protein evolution as a continuous time Markov chain with the assumptions of equal substitution rates across sites and equal amino acid frequencies. This is an extension of the Jukes-Cantor model for nucleotide evolution. Under the Jukes-Cantor model [4], the nucleotide probability of a specific site in a nucleic acid is governed by a set of coupled equations.

$$\begin{pmatrix} p_A(t) \\ p_G(t) \\ p_T(t) \\ p_C(t) \end{pmatrix} = \mathbf{Q}(t) \begin{pmatrix} p_A(0) \\ p_G(0) \\ p_T(0) \\ p_C(0) \end{pmatrix}$$

The operator $\mathbf{Q}(t)$ is

$$\mathbf{Q}(t) = \frac{1}{4} \begin{pmatrix} 1+3e^{-\mu t} & 1-e^{-\mu t} & 1-e^{-\mu t} & 1-e^{-\mu t} \\ 1-e^{-\mu t} & 1+3e^{-\mu t} & 1-e^{-\mu t} & 1-e^{-\mu t} \\ 1-e^{-\mu t} & 1-e^{-\mu t} & 1+3e^{-\mu t} & 1-e^{-\mu t} \\ 1-e^{-\mu t} & 1-e^{-\mu t} & 1-e^{-\mu t} & 1+3e^{-\mu t} \end{pmatrix}$$

The mutation rate μ is uniform across sites and nucleotides. This model can be generalized to a protein alphabet of 20 amino acids. The coupled equations can be solved [7], and the probability of a mutation in a time t under such a model is given by:

$$P_{mut} = \frac{19}{20}\left(1 - e^{-\mu t}\right)$$

At each speciation event, I randomly determine whether or not each site has mutated given the above probability, branch length, and mutation rate (set to 0.01 inverse branch units for convenience). If a mutation is allowed, I randomly select one of 19 amino acids to be passed on to the daughter sequences.

2. *Profile Model*

Profile models have been used to align protein sequences since the 1980s [8]. Under a profile model, the probability of an aligned sequence $\vec{x}$ of length L belonging to a protein family is given by

$$p(\vec{x}) = \prod_{i=1}^{L} p_i(x_i)$$

where $p_i(x_i)$ is the *emission* probability of observing residue x_i at site i [7]. Unlike the Jukes-Cantor model, profile models are site specific: the set of probabilities for a site i are generally unique for each site. However, individual sites are modeled to mutate independently, which does not allow for epistasis.

To evolve sequences using a profile model, I use Markov Chain Monte Carlo. At each discrete time step, I follow the Metropolis algorithm. I randomly mutate one residue at one site i from x_i to x'_i and then calculate the ratio of probabilities before and after the mutation.

$$\alpha = \frac{p(\vec{x})}{p(\vec{x}')} = \frac{p_i(x'_i)}{p_i(x_i)}$$

If $\alpha \geq 1$, I always accept the mutation. If $\alpha < 1$, I accept the mutation with probability α. During speciation events, I copy sequences to the daughter nodes and evolve them independently.

3. *Potts Model*

Potts models are a generalization of the Ising model in which a site can have q possible spins. 1-D Potts models were first applied to modeling protein sequence families by Lapedes et al [5]. Under a Potts model, the probability that an aligned sequence $\vec{x}$ belongs to a protein family is given by

$$p(\vec{x}) = \frac{1}{Z} e^{U(\vec{x})}$$

where Z is the partition function and $U(\vec{x})$ is the pseudo-energy.

$$U(\vec{x}) = \sum_{i=1}^{L} h_i(x_i) + \sum_{i=1}^{L} \sum_{j=i+1}^{L} e_{ij}(x_i, x_j)$$

Potts models are not site independent, modeling the co-evolution of residues across sites. Therefore, pairwise epistasis can occur under a Potts model simulation.

FIG. 2: Histogram of pairwise sequence identities of all possible pairs of 11th generation sequences across the left subtree and the right subtree. For each evolutionary, the aggregate results of 100 independent simulations is shown.

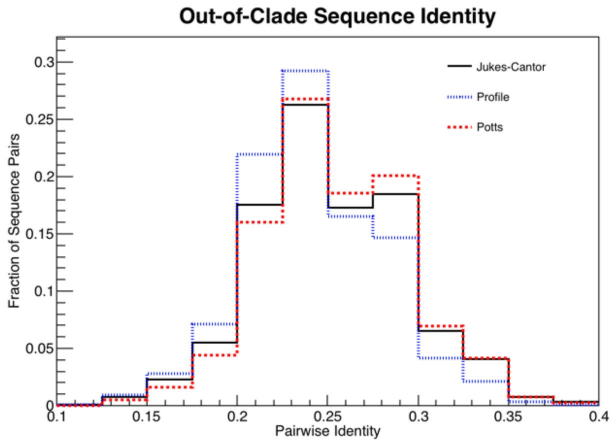
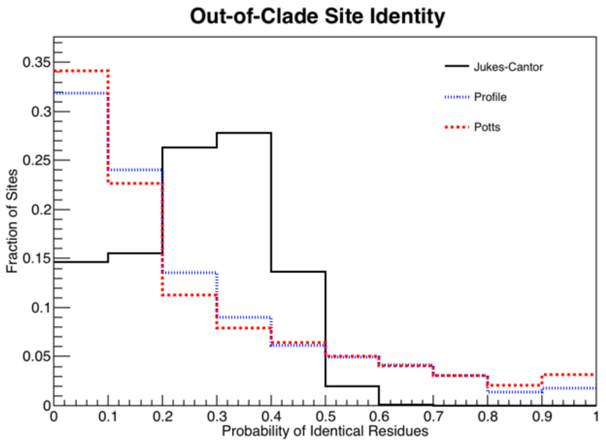


FIG. 3: Distribution of the probabilities that site pairs are identical across pairs of sequences in the left subtree and the right subtree. For each evolutionary model, the aggregate of 100 independent simulations is shown.



As with profile models, I use Markov Chain Monte Carlo to evolve sequences down the binary tree. At each step, I randomly mutate one site i and calculate α.

$$\alpha = e^{\Delta U\left(x_i, x_i'\right)}$$

C. Data

As a model protein family, I use the globin family (PFam id PF00042 [9]) Globins are involved in metal binding in oxygen transport, and they are widely distributed in eukarya, archaea, and bacteria. Globins have a well conserved tertiary structure [10]. In PFam, over 4000 globins are aligned over 96 conserved match positions. Therefore, I use a sequence length L=96 in all my simulations.

To infer the profile model parameters, I construct a profile hidden Markov model from the PFam full alignment using HMMER on the full PF00042 alignment [11]. To infer the Potts model parameters, I use the Gremlin structure prediction software package and the full PFam alignment [12]. For both methods, I construct the root of the tree after running the Metropolis algorithm for 1000 iterations. For comparison, the branch length between generations is 22 iterations for the profile model simulation and 40 iterations for the Potts model simulation.

III. RESULTS

For each tree, I calculated the pairwise identity of all possible pairs of 11th generation sequences across the left and right subtrees. By adjusting the generation time for the Jukes-Cantor simulation and the number of iteration times between generations for the profile and Potts model MCMC simulations, I was able to achieve a mean out-of-clade pairwise identity distribution that averaged around 25% for each model (see Figure 2). This is comparable to the pairwise identity observed by Konate et al in FoIA homologs with a divergence time of approximately 2 billion years (See Figure 1 in [1]).

Next, by once again examining all possible pairs of 11th generation sequences across subtrees in each tree, I create a histogram of the probability of site identity across sequences in the two clades under all three models (see Figure 3). This is analogous to Figure 4 in [1].

The Jukes-Cantor simulation, which is independent and non-site specific, does not have any sites conserved in over 60% of independent lineages, which is contrary to the data found by Konate et al in the FoIA alignments. Therefore, I conclude a site-independent, non-site specific model is not effectively modeling the protein evolution observed in [1].

The profile model and Potts Model simulations yield very similar distributions; only 5% of sites are identical in greater than 90% of out-of-clade sequence pairs. In comparison, only 10% of sites in the FoIA alignment are identical across independent lineages. The fact that these two distributions mirror similar data for enzymes analyzed in [1] indicates that both profile models and Potts models accurately model real protein evolution, and that the data Konate et al attribute to epistasis can likely be explained by a null evolutionary model in which epistasis is excluded.

IV. DISCUSSION

In [1], the relatively low fraction of conserved sites across distant FoIA homologues is used to proposed the possible role of epistasis in protein evolution. However, I am able to produce similar results in evolutionary simulations of the globin family that do not account for epistasis using profile models and Markov Chain Monte Carlo. In addition, simulations using Potts models and MCMC do not yield a significantly larger fraction of highly conserved sites across clades compared to profile model simulations. Therefore, the claim that low conservation of sites across distant phylogenies likely implies epistasis needs to be examined further in tandem with a null hypothesis that such results could be the result of site-independent evolution and phylogenetic effects. Evolution under a profile model can be used in such a capacity.

My simulations are simple and could be improved in future work. For instance, the number of branching events is fixed in my simulations. It would be interesting to modify the number of generations to see how site conservation was affected across sites, varying the topology of the phylogenetic tree. Also, I only model pairwise epistasis, while more complicated models of protein evolution certainly can be simulated. However, this project demonstrates that purported evidence for epistasis in protein evolution must be evaluated rigorously, as even simple simulations suggest such evidence could be the result of other evolutionary processes.

[1] M. Konate, G. Plata, J Park, H Wang, and D. Vitkup, *bioRxiv, USA* (2017) (preprint)

[2] M. Breen, C. Kemena, P Vlasov, C. Notredame and F Kondrashov, *Nature (USA)* **490**, *535-538* (2013)

[3] M. McCandlish, E. Rajon, P. Shah, Y. Ding, and J. Plotkin, *Nature (USA)* **497**, *E1-E3* (2013)

[4] T. Jukes and C. Cantor, in *Evolution of Protein Molecules* (Academic Press, New York, 1969), pp. 21-132.

[5] A. Lapedes, B Giraud, L.C. Liu, and G. Stormo. *Statistics in Molecular Biology (USA)* **33**, *236-256* (1999)

[6] E. Rivas, J. Clements, and S. Eddy. *Nature Methods (USA)* **14**, *45-48* (2017)

[7] R. Durbin, S. Eddy. A. Krogh, and G. Mitchison. *Biological Sequence Analysis.* (Cambridge Press, Cambridge, UK, 1998)

[8] M. Gribskov, A. McLachlan, and D. Eisenberg. *Proceedings of the National Academy of Sciences (USA)*, **84**, *4355-4358* (1987)

[9] R. Eberhardt et al. *Nucleic Acids Research (USA)* **44**, *D279-D285* (2016)

[10] R. Hardison. *Cold Spring Harbor Perspectives in Medicine (USA)* **2** *:a011627* (2012)

[11] S. Eddy. *PLoS Computational Biology (USA)* **7**, *e1002195* (2011)

[12] H. Kamisetty, S. Ovchinnikov, and D. Baker. *Proceedings of the National Academy of Sciences (USA* **39**, *15674-15679* (2013)