Jian Xu

Modeling evolution at the protein level using an adjustable amino acid fitness model

Matthew W. Dimmic, David P. Mindell, and Richard A. Goldstein

Introduction

Today, majority of the phylogenetic analyses are performed using comparisons of DNA sequences between classifications. Several methods, such as maximum parsimony (MP) method[1] and maximum likelihood (ML) method[2], are available to do this task. But the MP method could repeatedly lead to wrong tree[3], although it is quick and exhaustive; and the ML method is computationally more intensive and requires and explicit model for the process of molecular evolution., although it does not suffer from the same biases as MP method. On the other hand, the site-specific substitution rate also reduces the accuracy of these methods.

Model

To overcome these problems, Dimmic et. al. presented an adjustable fitness model for amino acid site substitution. Based on the fact that amino acids have different biophysical characteristics and some will be more advantageous than others under certain conditions, they introduced a relative fitness $F(A_n)$ of each amino acid A_n. The values for the relative fitness are not assumed, but are adjustable parameters in the likelihood maximization scheme, which will be described later. Then they assumed that the probability Q_{ij} of substituting amino acid *i* with amino acid *j* is:

$$Q_{ij} = \begin{cases} v & \text{if } \Delta F_{ji} > 0 \\ v e^{\Delta F_{ji}} & \text{if } \Delta F_{ji} < 0 \end{cases}$$

where *v* is the average rate at which mutations occur and $\Delta F_{ji}=F(A_j)-F(A_i)$.

With this scheme, if the mutation is favorable then it is always accepted with a certain rate *v*, otherwise it is tolerated with a decaying exponential probability. The substitution matrix M at evolutionary time t is:

$$M_{ij}(t) = e^{tQ_{ij}}$$

At each site *s* in the amino acid sequence, the likelihood is represented by the sum of the probability of all possible paths to all possible ancestors, and can be written as $L_s = P_s(data \mid \theta, T)$, where "data" represents all possible paths and ancestors, ⌐ is a model-dependent parameter, and T is the evolutionary tree branch length. Therefore, the log-likelihood □ is the product of the likelihood of each site, equivalent to the sum of their logs:

$$\Omega = Log[\prod_s P_s(data \mid \theta, T)] = \sum_s Log(L_s)$$

In order to incorporate the site heterogeneity, they assumed *k* site classes, each class has its own fitness parameters $F_k(A_n)$ and mutation rate v_k. The concept of site class is a representation of different biophysical characters of amino acids. For instance, one site class might represent all sites where charge is important for the protein's function; then in this class, Glu would have a higher fitness than Ala. In their model, each site has a certain probability of being represented by each site class. This possibility is the likelihood function calculated using the parameters for that site class. Each site class, in turn, has a prior probability of representing any site $P(\cdot_k)$. After this site heterogeneity correction, the likelihood at each site is:

$$L_s = \sum_k P_s(data \mid \Theta_k, T) P(\Theta_k)$$

By adjusting the parameters to increase the likelihood, a maximum likelihood estimate for the parameters can be obtained.

Results

In order to test the utility of their amino acid fitness model, they compare their results with the mtREV model of Adachi and Hasegawa[4]. The mtRev model has 189 adjustable parameters and can be seen as the most general single-site class reversible model, thus it is a good basis for comparison. In both training sequences and test sequence, the 5-site-class fitness model performed better than mtREV model by exceeding nearly 900 log-likelihoods, while the 5-site-class model has 86 fewer adjustable parameters then the mtREV model.

In addition, they examined whether the site classes in the 5-site-class model had any correlation with know biophysical properties. They plotted the fitness parameter of each class against two amino acid characteristics: bulk and hydrophobicity. Most correlations with bulk were not strong, and site class #3 showed substantial negative correlation with amino acid hydrophobicity (Fig.1). The lack of correlation with biophysical properties among the site classes does not necessarily imply that the parameters have no physical meaning, but it imply that setting the parameters as simple functions of a few biophysical characteristics may not be adequate to capture the selective pressures at work on the protein. Or it may also imply that biophysical characteristics are "mixed" into the various site classes during the optimization scheme.

Conclusion

Dimmic et. al.'s adjustable fitness model accounts for site heterogeneity among substitution rates and among evolutionary constraints, and does not make any assumptions about which sites or characteristics of proteins are important to molecular evolution. This model has fewer adjustable parameters then the general reversible mtREV model, and outperforms mtREV in likelihood analysis on protein coding mitochondrial genes. In future, more comparisons need to be performed using other data sets to yield more convincing results. And applying statistical tests such as Monte Carlo simulation may be helpful to optimize the model.

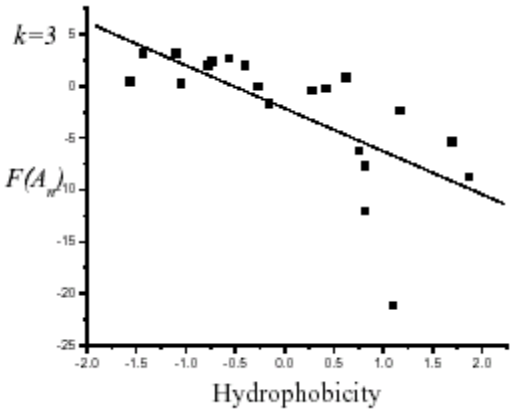

Fig.1. The fitness parameter of the third site class $F_3(A_n)$ versus hydrophobicity. The correlation R=-0.68, P<0.001.

---

[1] W. M. Fitch, On the problem of discovering the most parsimonious tree, *Am. Nat*. **111**, 223 (1997).
[2] L. L. Cavalli-Sforza and A. W. F. Edwards, Phylogenetic analysis: models and estimation procedures, Am. J. Hum. Genet. 19, 233 (1967).
[3] J. Felsenstein, Cases in which parsimony or compatibility methods will be positively misleading, Syst. Zool. 27, 401 (1978).
[4] J. Adachi and M. Hasegawa, "model of amino acid substitution in proteins encoded by mitochondrial DNA", J. Mol. Evol. 42(1), 138 (1999).