A phylogenetic and population genetic model of amino acid substitution

October 11, 2012

1 Abstract

A new mechanistic model for the evolution of amino acid sequences is developed for studying the biological properties of proteins as well as phylogenetic estimation. Two steps are bridged together to form a Markov process to describe substitutions between amino acids: mutation is based on general time reversible models for underlying nucleotides; fixation is obtained using classical population genetics theory. Selective restraints at amino acid level are characterized by the physiochemical distances between amino acids and the Grantham sensitivity coefficient exerted on the distances. Analysis of a yeast data set shows that the new model provides a better fit to data than the empirical models and reveals the variance of Grantham sensitivities and optimal amino acids at different sites in proteins.

2 Introduction

Importance of building accurate model for protein evolution.

Known models of amino acid replacement can be divided into two categories: empirical models and mechanistic models. Models in the first category include Dayhoff, JTT, WAG, LG, etc. Yang et al. (1998) implemented a few mechanistic models at the level of codons and explicitly modeled the biological processes involved, including different mutation rates between nucleotides, translation of the codon triplet into an amino acid, and the acceptance or rejection of the amino acid due to selective pressure on the protein.

Mechanistic models for the evolution of protein-encoding sequences are on three levels: mono-nucleotide level in DNA sequences, codon level in DNA coding sequences and amino acid level in protein sequences. Models on the DNA level use the most information and are more powerful to distinguish closely related sequences such as those caused by synonymous substitutions which are

invisible at amino acid level. On the amino acid level, models can filter out some stochastic noise through the translation of DNA triplets to amino acids. Goldman and Yang (1994, MBE) constructed a codon-based model that uses the nucleotide-level information in DNA sequences and the amino-acid level information of synonymous and non synonymous nucleotide substitutions simultaneously. Their model incorporated transition / transversion bias, synonymous / nonsynonymous variation in a gene, and amino acid differences. The selective restraints at the amino acid level was accounted for by multiplying the substitution rate by a factor $\exp(d_{aa_i,aa_j}/V)$ where d_{aa_i,aa_j} is the distance between amino acids aa_i and aa_j given by Grantham (1974) and V is a parameter representing the variability of the gene or its tendency to undergo non synonymous substitution.

In Goldman and Yang's model, the Markov process is time reversible. In other words, the amino acids are equally as good in a protein and the substitution rates are only proportional to the frequencies of the amino acids. However, from population genetics, the selective restraints should be a function of the fitness of proteins. Proteins with higher fitnesses get fixed with higher probability than those with low fitnesses. Gilchrist (2007) showed that the fitness of a protein is a function of factors including protein production cost, gene expression level, and functionality of protein. A protein might have a sequence of "optimal" amino acids which give the protein best functionality, while other amino acids might also make the protein function but less well. Therefore, the functionality of a protein depends on what the optimal amino acids are and how far away the observed amino acids are from the optimal ones, as well as how sensitive the functionality is to the distance between amino acids.

In addition, the measure of difference between amino acids combines physiochemical properties that correlate best with protein residue substitution frequencies: composition, polarity and molecular volume. Grantham (1974) assigned weights to these three factors based on the average chemical distance given by the corresponding property alone. Take the composition for example, given the values for this property c_i's, the weight $\alpha = (1/\bar{D}_c)^2 = 1.833$ where $\bar{D}_c = \sum[(c_i - c_j)^2]^{1/2}/190$. Similarly the weights for polarity and molecular volume are 0.1018 and 0.000399. Since the values for the volume property is much bigger than the other 2 properties its weight is much smaller. We call the weights Grantham weights. It is reasonable to believe that in some genes one property might play a more important role while in some genes it might be another property. For example ? We present a new model that incorporates the above factors by including the Grantham weights α, β, γ and the sensitivity of functionality to distance from the optimal amino acid as parameters. We call the sensitivity coefficient "Grantham sensitivity" and denote is by g.

In this paper we characterize our amino acid-based model, which incorporates substitution rates of underlying coding nucleotides, the biological properties of amino acids, selection sensitivity of amino acid differences. We use

the model for maximum likelihood (m.l.) estimation of phylogenies and apply the model to Rokas's yeast data sets with 8 species. The results are compared with those under previous amino-acid models. We also investigate the evolution process of protein with different parameters.
(and use simulations and information from empirical data to find cases where populations of intermediate size may evolve faster than populations of large size.)

3 Model

Our model works for homologous protein-coding sequence without gaps or with gaps removed. We use a continuous time Markov process to model substitutions among the amino acids within a protein-coding sequence. The states of the Markov process are the 20 natural amino acids (nonnatural amino acids can be easily added), and we use a 20×20 rate matrix $Q = (Q_{ij})$ where Q_{ij} represents the instantaneous rate that amino acid i will be substituted by amino acid j. As usual the row sum of (Q_{ij}) equals 0 and $P(t) = \exp(tQ)$, where $P_{ij}(t)$ is the probability that amino acid j replaces i after time t. Rate matrix Q is obtained by multiplying the mutation rate matrix M and fixation probability matrix F.

Based on the 4×4 general time reversible (GTR) mutation rate matrix M_{nu} for nucleotides the mutation rates μ_{ij} among 20 amino acids are calculated. We assume that mutations occur on the codon level at the three codon positions independently. Therefore, more than one nucleotide substitutions are not allowed to occur instantaneously as mutations involving more than one position during time Δt will have probabilities Δt^2 and should be ignored. The calculation follows two steps. First, a 61×61 (stopping codons not included) codon mutation rate matrix M_{codon} is obtained; Second, according to the codon translation table, we can find the mutation rate matrix M (20×20) for all amino acids. The mutation process is time reversible, i.e. $\pi_i M_{ij} = \pi_j M_{ji}$ is satisfied for all $1 \leq i, j \leq 20$.

Next we consider the selection. Assume that there is an optimal amino acid for each position in a protein. Any non-optimal amino acid at a position is subjected to selection, the strength of which depends on the physiochemical difference (Grantham, Science 1974) between the observed and optimal amino acids and magnitude of the Grantham sensitivity.

Suppose a protein of length n has a sequence of optimal amino acids $\hat{\mathbf{a}} = (\hat{a}_1, \hat{a}_2, \cdots \hat{a}_n)$, the observed sequence of amino acids is $\mathbf{a} = (a_1, a_2, \cdots, a_n)$, the Grantham sensitivity coefficient is g_k and let $\mathbf{g} = (g_1, g_2, \cdots, g_n)$. The overall physiochemical difference (distance) between amino acids i and j consist of 3 components: $d_{ij} = [\alpha(c_i - c_j)^2 + \beta(p_i - p_j)^2 + \gamma(v_i - v_j)^2]$, where c, p and v represent values of properties composition, polarity and molecular volume,

and α, β, γ are the corresponding weights for the components. The values for properties in the amino acid difference formula is given by Grantham (1974), he presented a way to compute the weights for the 3 components. In our model, the weights are treated as parameters rather than being fixed.

Given the distance vector $\mathbf{d} = (d_1, d_2, \cdots, d_n)$ from the optimal protein and the Grantham sensitivity $\mathbf{g}$, the functionality of a protein $\mathbf{a}$ with n amino acids is defined as

$$F(\mathbf{a}|\hat{\mathbf{a}}, \mathbf{g}) = \frac{n}{\sum_{k=1}^{n} (1 + d_k g_k)} \tag{1}$$

The condition $\hat{\mathbf{a}}, \mathbf{g}$ will be omitted from now on if there is no potential confusion.

If there is a single mutant $\mathbf{a}_j$ from a diploid population with wild type $\mathbf{a}_i$, the fixation probability is

$$\pi_{ij} = \pi(\mathbf{a}_i \to \mathbf{a}_j) = \frac{1 - f(\mathbf{a}_i)/f(\mathbf{a}_j)}{1 - (f(\mathbf{a}_i)/f(\mathbf{a}_j))^{2N_e}} = \frac{1 - f_i/f_j}{1 - (f_i/f_j)^{2N_e}} \tag{2}$$

according to Sella-Hirsh (Add reference) where $f(\mathbf{a}_i)$ and $f(\mathbf{a}_j)$ are the fitnesses of $\mathbf{a}_i$ and $\mathbf{a}_j$. It is an approximation to the canonical formula

$$\pi(\mathbf{a}_i \to \mathbf{a}_j, p) = \frac{1 - e^{-2N_e ps}}{1 - e^{-2N_e s}} \tag{3}$$

where p is the initial frequency of the mutant, and $s = (f_j - f_i)/f_i$ is the selection advantage of $\mathbf{a}_j$ comparing to $\mathbf{a}_i$ (note here s is different from the selection strength defined above on the distance from optimal protein). When there is a single mutant in the population, i.e. $p = 1/(2N_e)$, the formula simplifies to $(1 - e^{-s})/(1 - e^{-2N_e s})$. Both S-H and the canonical formulae are valid under the same condition: $s, \frac{1}{N}, Ns^2 \ll 1$.

As in Gilchrist 2007, fitness of a protein is related to its functionality in the following way:

$$f(\mathbf{a}) \propto \exp\{-\frac{C\Phi q}{F(\mathbf{a})}\}$$

where C is the expected cost of producing a single complete protein, q is the scaling constant (seconds per ATP) determining the relationship between the rate of ATP usage and fitness f, and Φ is a measure of gene expression, specifically protein production rate (protein per second).

Combining $C\Phi q$ as one constant A, we have $f(\mathbf{a}) \propto \exp\{-\frac{A}{F(\mathbf{a})}\}$. Clearly, protein fitness is an increasing function of functionality.

In either S-H formula or the canonical formula of the fixation probability, the determining value is f_i/f_j. From the definition of functionality in Equation 1, we have the following:

$$\frac{f(\mathbf{a}_i)}{f(\mathbf{a}_j)} = \prod_{k=1}^{n} \Big(\frac{f(\mathbf{a}_i^k)}{f(\mathbf{a}_j^k)}\Big)^{\frac{1}{n}} \tag{4}$$

i.e. the fitness ratio of the whole protein is the geometric mean of the fitness ratios between the two proteins for all sites. Therefore, when $\mathbf{a}_i$ and $\mathbf{a}_j$ only differ at position k, it becomes

$$\frac{f(\mathbf{a}_i)}{f(\mathbf{a}_j)} = \Big(\frac{f(\mathbf{a}_i^k)}{f(\mathbf{a}_j^k)}\Big)^{\frac{1}{n}}$$

and

$$\begin{aligned} \frac{f(\mathbf{a}_i)}{f(\mathbf{a}_j)} &= \exp\Big[-A\Big(\frac{1}{F(\mathbf{a}_i)} - \frac{1}{F(\mathbf{a}_j)}\Big)\Big] \\ &= \exp\Big[-\frac{C\Phi q g_k}{n}(d_k^i - d_k^j)\Big] \end{aligned} \tag{5}$$

this quantity is only related to site k. It is easy to see that all the sites are independent in the sense that if there are more than 1 site that differ, the ratio is simply a product of ratios at all sites.

From above the fixation probability depends on Grantham sensitivity coefficient, the difference between amino acids, therefore the weights for all components, and constants C, Φ, q, N_e.

Instantaneous substitution rate u_{ij} from $\mathbf{a}_i$ to $\mathbf{a}_j$ is the product of effective population size, mutation rate and fixation rate:

$$u_{ij} = 2N_e \mu_{ij} \pi_{ij} \tag{6}$$

where μ_{ij} is the mutation rate from $\mathbf{a}_i$ to $\mathbf{a}_j$, and $\mu_{ij} = 0$ when more than 1 position differ in the codons that code for $\mathbf{a}_i$ and $\mathbf{a}_j$. Note that the mutation and fixation are both at amino acid level.

With the values for $(M_{nu}, g, \alpha, \beta, \gamma, C, \Phi, q, N_e)$, we can find the 20×20 instantaneous substitution rate matrix Q and have the Markov process set up. Then we can calculate the likelihood of observing the sequence data at the tips of a phylogenetic tree T with given topology and branch lengths, therefore find the maximum likelihood estimates for parameters.

Identifiability — Since C, Φ, q and g are multiplied together as a composite parameter, we fix the values of C, Φ, q and search for MLE for g. For the weights of 3 components in the amino acid difference formula, if they are multiplied by a same constant, the likelihood will not be affected. So we will fix α and look

for the MLEs for β, γ since only the relative ratios are identifiable. In addition, the effective population size is assumed to be fixed in this paper. Suppose the phylogenetic topology is given, the parameters that we are estimating: $s, \beta/\alpha$, γ/α, branch lengths, and the GTR rate matrix for nucleotides.

4 Results

4.1 Model accuracy

To assess the model accuracy, we first simulate data using different parameter values, find the MLEs for the parameters from the simulated data, and then investigate the accuracy of the estimates by looking at the mean squared error and confidence intervals.

We did simulation with number of sites 100, 300, 500, 700, 1000, $s = 0.1$, 0.3, 0.5, 0.7, 0.9, 1, under 14 different trees, including 2 trees for number of tips 4, 8, 10, 12, 14, 16, 18. Simulation is done with different branch lengths: one with all branches of length 10 (expected 10 transitions on each branch) and the other one with shorter (more realistic) lengths randomly chose between 0 and 1.

4.2 Results on yeast data