A phylogenetic and population genetic model of amino acid substitution

November 7, 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), so 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. The rate matrix Q is obtained by multiplying the mutation rate matrix M and fixation probability matrix F. 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.

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 independently between nucleotides at the same codon position. 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 on the order of Δt^2 and are, therefore, ignored. The calculation follows two steps. First, a 61×61 (stop codons not included) codon mutation rate matrix M_{codon} is obtained; Second, using the genetic code table, we generate the 20×20 mutation rate matrix M for all amino acids. The mutation process is time reversible, i.e. $\pi_i M_{ij} = \pi_j M_{ji}$ is satisfied for all $1 \leq i, j \leq 20$.

In addition to the effects of mutation bias our model also includes the effects of natural selection on the amino acid sequence of a gene. We begin by assuming that there is an optimal amino acid for each position in a protein and non-optimal amino acids each position are subjected to natural selection. The strength of selection depends on the scaled physiochemical distance (Grantham, Science 1974) between the observed and optimal amino acids, the sensitivity of the protein's function to the physicochemical distance and the protein production rate of the gene.

Suppose a protein of length n has the optimal amino acid sequence $\hat{\mathbf{a}} = (\hat{a}_1, \hat{a}_2, \cdots \hat{a}_n)$, 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 physicochemical properties composition, polarity and molecular volume of the amino acid's side chain, and α, β, γ are the corresponding weights for the components. The values for physicochemical properties in the amino acid difference formula is given by Grantham (1974), each of them is weighted by dividing by the mean distance found with that property alone. In our model, the weights α, β, γ are treated as estimable parameters rather than being fixed.

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

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

In order to simplify notation, the parameters $\hat{\mathbf{a}}$ and $\mathbf{g}$ will be omitted from now on if there is no potential confusion.

Following Sella-Hirsh (Add reference), if there is a single mutant $\mathbf{a}_j$ from a diploid population with wild type $\mathbf{a}_i$, the fixation probability of it getting fixed is

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

where $f(\mathbf{a}_i)$ and $f(\mathbf{a}_j)$ are the fitnesses of genotypes $\mathbf{a}_i$ and $\mathbf{a}_j$. This formula is valid under the 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 a 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 for a given gene.

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 the S-H formula of the fixation probability, the determining value is f_i/f_j. Based on the definition of functionality in Equation 1, we have the following:

$$\frac{f(\mathbf{a}_i)}{f(\mathbf{a}_j)} = \prod_{k=1}^{n}\Big(\frac{f(\mathbf{a}_i^k)}{f(\mathbf{a}_j^k)}\Big)^{\frac{1}{n}} \tag{3}$$

i.e. the fitness ratio of the two genotypes 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, this fitness ratio simplifies to

$$\begin{aligned} \frac{f(\mathbf{a}_i)}{f(\mathbf{a}_j)} &= \Big(\frac{f(\mathbf{a}_i^k)}{f(\mathbf{a}_j^k)}\Big)^{\frac{1}{n}} \quad (4) \\ &= \exp\Big[-A\Big(\frac{1}{F(\mathbf{a}_i)} - \frac{1}{F(\mathbf{a}_j)}\Big)\Big] \\ &= \exp\Big[-\frac{C\Phi q g_k}{n}(d_k^{(i)} - d_k^{(j)})\Big] \quad (5) \end{aligned}$$

this quantity is only related to site k. From equation 5, 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.

For a single site, we have

$$\frac{f(a_i)}{f(a_j)} = \exp\Big(-C\Phi q g_k(d^{(i)} - d^{(j)})\Big) \quad (6)$$

From Equation 2 and 6, the fixation probability of a mutation at one site of a protein depends on the physicochemical distances d from the optimal amino acid at this site, sensitivity coefficient g of functionality to the distance d, and constants C, Φ, q, N_e.

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

$$u_{ij} = 2N_e\mu_{ij}\pi_{ij} \quad (7)$$

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

With the values for $(M_{nu}, g, \alpha, \beta, \gamma, C, \Phi, q, N_e)$, and the optimal amino acid a site, we can calculate the 20×20 instantaneous substitution rate matrix Q for the Markov process. Q is scaled by the frequencies of amino acids to satisfy $\sum_{i=1}^{20} \pi_i q_{ii} = -1$. Under this restraint, the length of a branch represents the expected number of substitutions on the branch. Since all sites are independent, we can calculate the likelihood of observing the sequence data at the tips of a phylogenetic tree T with given topology and branch lengths.

To calculate the likelihood values, the optimal amino acids need to be specified. We implement 3 approaches to do this. First, for each site (for each distinct data pattern at the tips), we calculate the likelihood values when each of 20 amino acids is optimal and choose the one that maximizes the likelihood. This method treats the optimal amino acids as parameters to be estimated in

Table 1: parameters in the model

s_{ij}	exchange rates between nucleotides
g	sensitivity coefficient of functionality to physicochemical distance
(α, β, γ)	weights for the 3 physicochemical properties in amino acid distance formula
C	cost of producing a protein
Φ	expression level
q	scaling factor
N_e	effective population size

the maximum likelihood computation. The number of parameters is increased by the number of distinct sequence patterns at the tips, which is often a big number. Second approach uses the majority rule, i.e. we pick the most frequent amino acid in the sequence as the optimal amino acid. When the sequences have evolved long enough to reach equilibrium, the optimal amino acid has the highest probability to be observed. If the evolving time is not long enough, or there are not enough substitutions during evolution process, the optimal amino acids estimated this way can be inaccurate. The third method is to assign weights to 20 amino acids being optimal. If the same weights are used for all sites, then the number of parameters added is 19 compared to hundreds or more in the first approach. The weights are expected to vary with the environment, function of proteins and other factors. Therefore an alternative is to use different weights for different genes (groups) in a protein sequence. It's apparent that the first method will give the best likelihood values but uses most parameters. The third method uses much fewer parameters. However, if the optimal amino acids vary a lot between different sites, the maximum likelihood values will decrease significantly. We'll compare the performance (including AIC scores) of different approaches in the Results section.

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 consistency

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.

4.2 Results on yeast data

The maximum likelihood approach using the new model is used on a real data set. This data set was used in Rokas (2003 Nature) to resolve incongruence in molecular phylogenies. This genome sequence data have been obtained for 7 *Saccharomyces* species (*S. cerevisiae, S. paradoxus, S. mikatae, S. kudriavzevii, S. bayanus, S.castellii* and *S. kluyveri*) as well as for the outgroup fungus *Candida albicans*. It includes 106 genes that are distributed throughout the *S. cerevisiae* genome on all 16 chromomosomes and comprise a total length of 42,342 amino acids (127,026 nucleotides), corresponding to roughly 1% of the genomic sequence and 2% of the predicted genes.

We first analyze the 106 gene sequences as a whole by concatenating them as 1 sequence. All the parameters are treated the same across all the sites except the optimal amino acids. The loglikelihood and AIC values are also compared with those under empirical models for amino acids from ProtTest (reference).

1. Treat the optimal amino acids as parameters to be estimated. The maximized loglikelihood is -236576.6, parameters are:
$g = 0.641372, \alpha = 1.83, \beta = 0.116, \gamma = 0.000577$
$Q = (3.854224, 19.926381, 6.221914, 4.096766, 8.051619, 1)$
and the branch lengths are
(0.09083797 0.12173561 0.07505142 0.19127314 0.24988114 0.07932873 0.06796140 0.27846822 1.93420199 1.50971356 0.69975745 1.95693652 1.79405040 3.34394967).