A phylogenetic and population genetic model of amino acid substitution

January 12, 2013

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 rates based on time reversible substitution models for nucleotides; fixation probability obtained from population genetics theory. We assume there is an optimal amino acid at each position of a protein sequence. Selective restraints at amino acid level are characterized by the physiochemical distances from optimal amino acids and the Grantham sensitivity of protein fitness to the distances. Analysis of a set of yeast data set shows that the new model provides a better fit to data than the empirical models and reveals the variation 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 fall 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 pa-

rameter 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 good 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 protein. Proteins with better 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 the 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 this paper we characterize our amino acid-based model, which incorporates substitution rates of underlying coding nucleotides, the biological properties of amino acids, Grantham sensitivity of functionality to 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. We view the substitution as a 2-step process. First, mutations occur in the population according to certain rates. Second, the newly arisen amino acid mutant gets fixed with different probabilities depending on its fitness, which completes the substitution. Therefore, the rate matrix Q is obtained by multiplying the mutation rate matrix M and fixation probability matrix F. As usual the row sums of Q equal 0 and $P(t) = \exp(tQ)$, where $P_{ij}(t)$ is the probability that amino acid j replaces i after time t.

3.1 Methods

For all time reversible models, the substitution rate matrix is a product of a symmetric matrix S and the the base frequencies of different states: $Q = S\Pi$, where S is called the exchange rate matrix and Π is a diagonal matrix of the base frequencies. The mutation rate matrix for 20 amino acids are derived from the 4×4 exchange rate matrix for nucleotides. This reduces the number of rate parameters from 190 to 6 comparing to treating all the amino acid exchange rates as parameters.

We assume that mutations occur independently between nucleotides at the

same codon position. Therefore, mutations involving more than one position during time Δt will have probabilities on the order of Δt^2 and are, therefore, ignored. First, a 61×61 sense codon exchange rate matrix is obtained. The rate between codon i and j is equal to the rate between the pair of nucleotides if that is the only different pair in all 3 codon positions, and 0 otherwise.

Second, using the genetic code table, we group the codons that coding for the same amino acid to get the 20×20 exchange rate matrix S for all amino acids. Suppose the sets of codons for amino acids i and j are I and J correspondingly, i.e. $aa_u = i$ for $u \in I$, and $aa_v = j$ for $v \in J$. Combining synonymous codons that code for amino acid j into one state, we have $\pi_J = \sum_{v \in J} \pi_v$ as the equilibrium frequency of amino acid j. The exchange rate matrix for a reversible Markov process of amino acid mutation S has entries:

$$\mu_{IJ} = \sum_{u \in I} \sum_{v \in J} \pi_u \pi_v s_{uv} / (\pi_I \pi_J)$$

And $q_{IJ} = s_{IJ}\pi_J$ with $s_{IJ} = s_{JI}$ constitutes the mutation rate matrix M. For detailed derivation see Yang(MBE 1998). The mutation process is time reversible, i.e. $\pi_I \mu_{IJ} = \pi_J M_{JI}$ is satisfied for all $1 \leq I, J \leq 20$. We assume that for each amino acid its different genetic codes have the same frequency. Therefore, the mutation rates between amino acids only depend on the frequencies of amino acids.

Studies show that fixation of mutations between dissimilar amino acids is generally rare. Therefore, in addition to the effects of mutation bias our model also considers 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 are subjected to natural selection.

The functionality of an amino acid depends on its scaled physiochemical distance (Grantham, Science 1974) from the optimal amino acid, the sensitivity of the protein's functionality to the physicochemical distance and the protein production rate of the gene.

Among the properties of amino acid side chain that correlate with relative substitution rate, composition (c), polarity (p) and molecular volume (v) have strongest correlation. Composition is defined as the atomic weight ratio of non-carbon elements in end groups or rings to carbons in the side chain. The last two properties are from published data (add ref). The overall physicochemical difference between any two amino acids i and j combines the three properties: $d_{ij} = [\alpha(c_i - c_j)^2 + \beta(p_i - p_j)^2 + \gamma(v_i - v_j)^2]$ where α, β, γ are the corresponding weights for the 3 components. 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 for 20 amino acids, its weight α is defined as $(1/\bar{D}_c)^2 = 1.833$ where $\bar{D}_c = \sum[(c_i - c_j)^2]^{1/2}/\binom{20}{2}$. The weights for polarity and molecular volume obtained in the same way are $\beta = 0.1018$ and $\gamma = 0.000399$. These weights are used for all genes of any species. However, depending on the functions and environment of amino acids, it is reasonable to assume that the impact of different properties varies. (Example here) In our model, the weights α, β, γ are treated as estimable parameters rather than being fixed. The distances are scaled so that the mean pairwise distance is 1. Because of the scaling, only 2 of the 3 weights are free parameters. We fix the weight α to be 1.833 as in Grantham's weights, and estimate β, γ. The "Grantham sensitivity" is denoted by g.

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 vector is $\mathbf{g} = (g_1, g_2, \cdots, g_n)$. The distance vector $\mathbf{d} = (d_1, d_2, \cdots, d_n)$ represents the distance per amino acid basis from the optima. The functionality of a protein $\mathbf{a}$ with n amino acids is then 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.

As in Gilchrist (2007), the 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, Φ is a measure of gene expression, specifically protein production rate for a given gene, and $F(\mathbf{a})$ is the functionality. 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.

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

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

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 $s, \frac{1}{N}, Ns^2 \ll 1$ where s is the selelction advantage/disadvantage of the mutant over the wild type. (Add ref or delete)

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 amino acids at all sites. Therefore, when $\mathbf{a}_i$ and $\mathbf{a}_j$ only differ at position k, this fitness ratio simplifies to

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

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. Therefore we will only consider single site protein in the following.

For a single site, we have

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

From Equation 2 and 6, the fixation probability of a single mutant in a diploid population depends on the physicochemical distances d from the optimal amino acid, Grantham sensitivity coefficient g, and constants C, Φ, q, N_e.

Therefore, the instantaneous substitution rate q_{ij} from $\mathbf{a}_i$ to $\mathbf{a}_j$ is 2 times the product of effective population size N_e, mutation rate μ_{ij} from a_i to a_j and the fixation probability of a single mutant:

$$u_{ij} = 2N_e\mu_{ij}\pi_{ij} \tag{7}$$

Note that $\mu_{ij} = 0$ when more than 1 position differ in the genetic codes for a_i and a_j and that the mutation rate and fixation probability are both amino acid specific.

Given the values for $(M_{nu}, g, \alpha, \beta, \gamma, C, \Phi, q, N_e)$, the frequencies of different amino acids and the optimal amino acid at 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 along the branch. With the probabilities $P(t) = \exp(Qt)$ the likelihood for a given tree topology can be calculated following Felsenstein (1981). 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 by multiplying the likelihood values at all sites.

3.2 Identification of optimal amino acids

To calculate the likelihood values, the optimal amino acids need to be identified. We implement 3 approaches to identify the optimal amino acid at a certain site. First one is called "max rule". We calculate the likelihood values when each of 20 amino acids is optimal with all other parameters given and choose the one that maximizes the likelihood as optimal. This method treats the optimal

Table 1: parameters in the model

s_{ij}	exchange rates between nucleotides i and j
μ_{ij}	mutation rates between nucleotides i and j
π_{ij}	fixation probability of single mutant j with wild type i
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

amino acids as estimable parameters in the maximum likelihood computation. The number of parameters increases with the number of distinct sequence patterns at the tips, which often is a big number.

Second approach uses the "majority rule", i.e. the most frequent amino acid in the sequence is chosen as the optimal amino acid. If more than 1 amino acid has the same highest frequency, then one of them is picked randomly as optimal. If 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 short, or there are not enough substitutions during evolution process, the optimal amino acids estimated this way can be inaccurate.

The third method is "weighted rule". 20 amino acids are assigned weights of 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 or gene groups in a protein sequence.

Apparently the first method gives the best likelihood value but uses the most parameters. On the other hand, the third method uses much fewer parameters. However, if the optimal amino acids vary a lot between different sites, the likelihood values will decrease significantly. We'll compare the performance of different approaches in the Results section.

3.3 Identifiability of parameters

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. As mentioned earlier, for the weights used in the Grantham distance formula, α is fixed and β, γ are estimated. In addition, the effective population size is assumed to be fixed in this paper. Suppose the phylogenetic topology is given, we are estimating the following parameters: $g, \beta/\alpha$, γ/α, frequencies of amino acids, branch lengths, and the exchange rate matrix for nucleotides.

4 Results

4.1 Results on Rokas et al.'s data on yeast

We analyzed data previously studied by Rokas (2003 Nature). 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 comprises a total length of 42,342 amino acids. Rokas et.al analyzed this data set to investigate the conflict of gene trees. We use the tree topology that

is supported by the concatenated genome sequence, which is also supported by the majority of the genes. Since the new model is not time reversible the tree is rooted at the out group *C.alb*.

4.1.1 maximum likelihood estimation

First, the 106 gene sequences are concatenated as 1 whole sequence with 42,342 amino acids. We use ProtTest to find maximum log likelihood values under empirical models and compare their AIC values. We also find the maximum log likelihood values under our new model, with all 3 approaches to treat the optimal amino acids. In all the analyses, tree branch lengths are optimized while the topology is not.

The log likelihood values and the AIC values are compared in Table 2. Under the empirical models, the substitution rates are fixed instead of being optimized. I denotes that proportion of invariable sites is estimated in the model, G means that Gamma distributed rate variation across all sites is included in the model. In models with F, amino acid frequencies are treated as free parameters and estimated by the observed frequencies in the sequence data. Otherwise, the amino acid frequencies are the equilibrium (stationary) frequencies under the substitution rate matrix.

Under the new model, amino acid frequencies are treated as 19 free parameters and estimated from the observed frequencies in the sequence data. In addition, 5 free parameters for exchange rates between nucleotides, 14 branch lengths for the 8-species phylogeney, Grantham sensitivity g, 2 free parameters for the weights in the physicochemical distance formula β and γ are optimized

in the maximum likelihood analyses.

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).

Table 2: Log-likelihood values and parameter estimates under empirical models and new model for the amino acid sequences of length 42,342

Model	ΔAIC	l	Tree length	Parameters
New+maj	0.00	-257790.10	10.43	41
New+max	48576.60	-239736.40	11.75	42,383
New+weights	123709.40	-319625.80	2.74	60
LG+I+G+F	81803.98	-298699.09		34
LG+G+F	81801.98	-298699.09		33

Note: The last two models in the table are the best models picked out by ProtTest.(What happens when the weights of amino acids being optimal are gene specific and other parameters are fixed across genes? Total number of parameters is 40,369. Better case senario, we estimate different optimal weights and other parameters genewise, this should give a better likelihood value in total compared to only optimal weights are gene specific. In this better case, the total loglikelihood value is -311186. Even with this loglikelihood value and number of parameters 40,369, the ΔAIC value is 187447.8, which means it performs worse than the third model in the table. The real loglikelihood value under this model is -317214.68; it gives a larger AIC value.)

If the optimal amino acids are not counted as free parameters being estimated, the majority approach gives the best AIC value. ΔAIC value for the

best empirical model LG + I + G + F is about 50,000 units, which indicates a substantially better fit under the new model. One thing need to point out is that under the maximizing approach for identifying optimal amino acids at each position in the sequence, the number of parameters is much bigger. However, the improvement of log likelihood is so big that this model still performs much better than the best empirical model, with AIC value 40,000 units smaller. With the weighted approach, i.e. across all sites, every amino acid has the same probability of being optimal, the increase in the log likelihood outweighs the reduction in the number of parameters. This also indicates that the optimal amino acids vary a lot across the sites.

4.1.2 Parameter variation between genes

We also analyzed Rokas et.al.'s data gene by gene. The best tree branch lengths and the exchange rate matrix for the concatenated genomic sequence are used in the analysis. Grantham sensitivity and the weights in the physicochemical distance formula are optimized for each gene. Under both approaches (max and maj) of obtaining the optimal amino acids, the estimates for all 3 parameters are on the similar scale. As expected, the weights for physicochemical properties are also similar to the ones that Grantham used. Figure 2 showed the correlation between β and γ. Linear regression suggests strong linear relationship between the 2 parameters, especially under the maximizing rule where R^2 is very close to 1.

Since the variation of Grantham weights across genes is small, we set β and γ across all genes to be the same and optimized g for each gene to get the maximum likelihood. We then did optimization on the common physicochemical weights β and γ. The ML estimates are $\beta = 0.1244$ and $\beta = 0.0005917$,

comparing to Grantham's weights $\beta = 0.1018$ and $\gamma = 0.000399$.

Notice in Figure 1 and 3 there is an outlier for the estimates of Grantham weights with the maximizing rule, which corresponds to gene 63; this also happens to gene 48 with estimates of sensitivity values. There are several possible explanations. One is discrepancy between the gene trees. We use the same tree (topology and branch lengths) for all the genes, however, sequences in some gene might support a different phylogeny. Another possibility is that these genes have very different structures from other genes so that the importance of the physicochemical properties is very different. (Look at the gene trees and see if they have different topologies from the one that is used in the ML analysis)

4.1.3 Confidence of estimates of optimal amino acids

To get the confidence level of the estimates for optimal amino acid at each site with the maximizing approach, we found the smallest set of amino acids being optimal that cover more than 95% of the total likelihood. In Rokas's data there are about 9000 different state patterns at the 8 species. For each of the 9000+ sites, the likelihood values achieved by assuming each amino acid as optimal is ordered decreasingly, therefore the first is the likelihood under the max optimal amino acid. Then the next amino acid is included in the optimal set of amino acids until the total likelihood exceeds 95% of the total likelihood. Figure 4 shows the histogram of numbers of optimal amino acids in the set. The mean value for all 9000+ patterns is 5.855, and mode is 6. The case where there are more than 10 amino acids in the set rarely happened. Figure 5 showed the density of percentages of total likelihood value covered by the optimal amino acid found with max rule only. Mean percentage is 0.4749 and the peak of the

density distribution is between 0.3 and 0.4. (How confident are we now??)

Figure 1: The figures in the first row are the estimates of β and γ using the maximizing rule, in the second row are the estimates of β and γ using the majority rule. Notice the difference in the range of y-axis

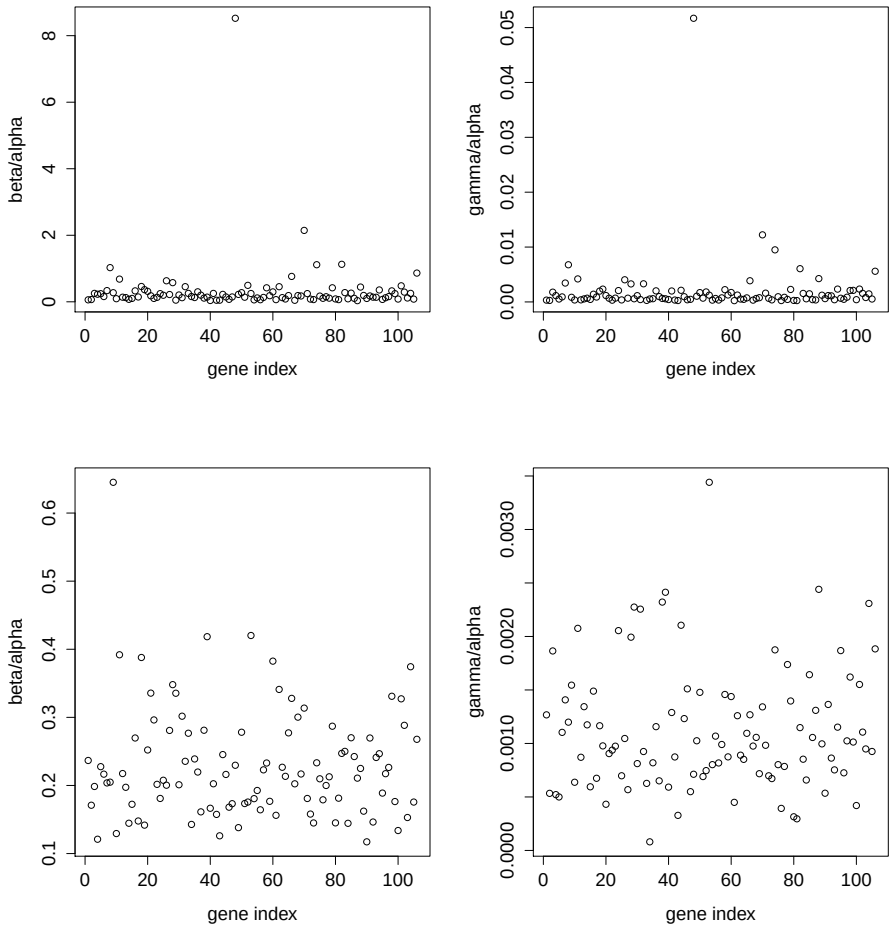

Figure 2: Correlation between β and γ. Blue line is the maximizing rule, with $R^2 = 0.9899$, Red line is the majority rule with $R^2 = 0.3258$

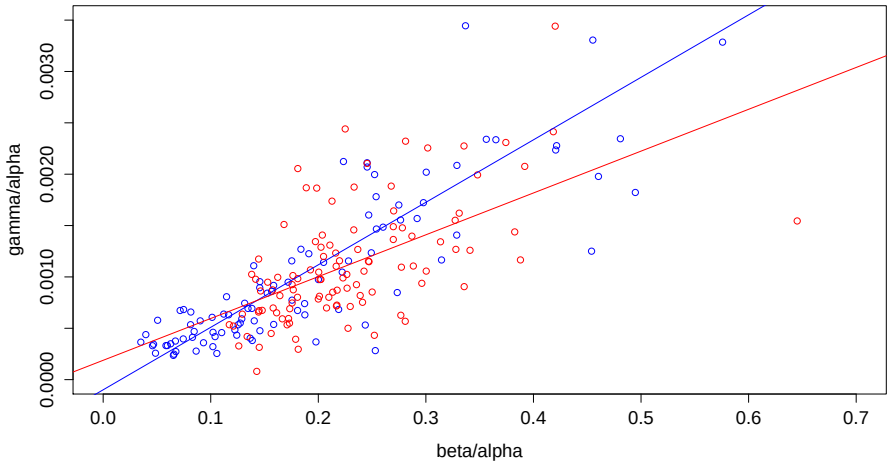

Figure 3: Plots of Grantham sensitivity across all the genes. On the left, optimal amino acids are obtained using maximizing rule; on the right majority rule

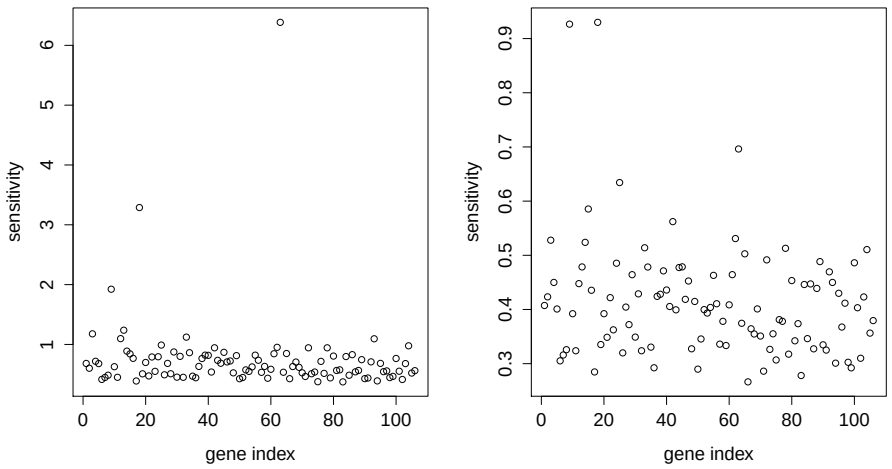

Figure 4: Histogram of the number of optimal amino acids together to cover at least 95% of the total likelihood attained by all possible optimal amino acids.

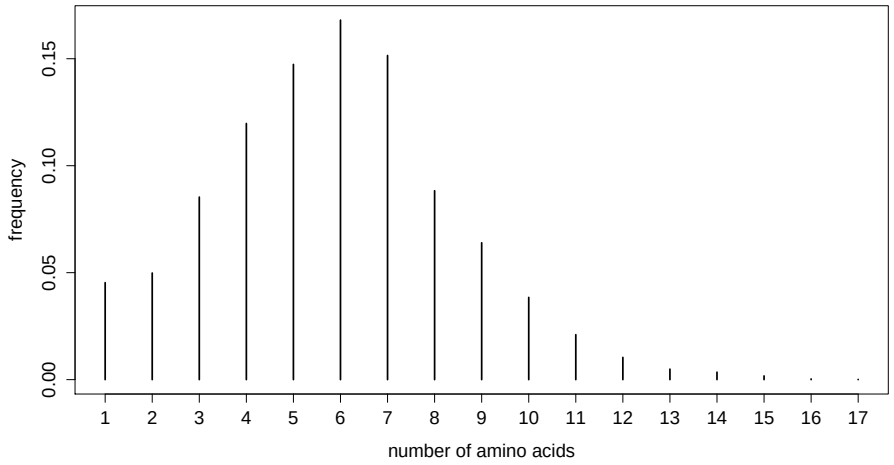

Figure 5: Density plot of percentages of the likelihood achieved by the optimal amino acid found by max rule.