A phylogenetic and population genetic model of amino acid substitution

May 10, 2013

1 Abstract

We introduce a new, mechanistic Markov model for studying the evolution of amino acid sequences within a phylogenetic context. Our model links together genotype, phenotype, and fitness of proteins, by calculating the fixation probability of a newly arisen mutant using a model of allele substitution from population genetics. As a result, our model explicitly includes the effects of mutation bias, genetic drift, and natural selection favoring an optimal amino acid sequence for a given protein. We assume that the strength of purifying selection is a function of the physiochemical differences between a given amino acid and the optimal amino acid for the site, how a protein's functionality declines with this distance, and the target expression level of the gene. Analysis of a multi-locus yeast data set using AIC shows that

our new model provides a substantially better fit to data than the standard empirical models and allows researchers to estimate biologically meaningful parameters such as the sensitivity of a protein's function to an amino acid substitution and the optimal amino acids at a given site. Further, because our model is based on explicit models of various biological processes, unlike empirical models it can easily be modified in the future to include other important biological phenomena such as selection on codon usage or test alternative hypotheses about the relationship between amino acid sequence and protein functionality.

Contents

2 Introduction

TO ADD: Importance of building accurate model for protein evolution.

In phylogenetics, models for evolution of protein-encoding sequences are usually formulated at three levels: mono-nucleotide level in DNA (e.g., see Felsenstein, 1981; Hasegawa et al., 1985; Jukes and Cantor, 1969; Kimura, 1980), codon level (see Goldman and Yang, 1994; Muse and Gaut, 1994; Yang and Nielsen, 2008; Yang et al., 1998), and amino acid (AA) level (see Kishino et al., 1990). The nucleotides, codons, or amino acids are assumed to evolve independently. DNA- and codon-based models use more data and are often the most powerful in terms of their ability to distinguish between closely related sequences. Of these two, codon-based models use all the information in DNA and know the product of amino acid in addition. On the other hand, AA-based models ignore synonymous differences between sequences by focusing not on the codons themselves but the amino acids they code for. Since synonymous codon usage is largely driven by mutation bias in low expression genes and selection on translational efficiency for high expression genes (see Gilchrist, 2007)[CHECK CITATIONS, add CUB citation], ignoring this aspect of the data has the advantage of reducing the noise in sequence data for low expression genes but at the cost of losing potentially useful information held in the high expression genes.

Based on how the substitution rates are formulated, models of amino acid substitution fall into two categories: empirical models and mechanistic models. In empirical models, the substitution rates are based solely on analysis of large quantities of sequence data complied from databases. Commonly used

models in this category include Dayhoff (Dayhoff et al., 1978), JTT (Jones et al., 1992), WAG (Whelan and Goldman, 2001), LG (Le and Gascuel, 2008) for nuclear proteins; mtREV, MTMAM (Adachi and Hasegawa, 1996; Yang et al., 1998) for mitochondrial proteins; and cpREV (Adachi et al., 2000) model for chloroplast proteins, etc. (see Cao et al., 1994; Gonnet et al., 1992; Henikoff and Henikoff, 1992) In contrast, mechanistic models are formulated based on the hypothesized biological processes thought to drive sequence evolution, such as mutation bias in DNA, translation of codons into amino acids, and natural selection.

The Goldman and Yang (1994) model, GY, is a codon level mechanistic model, which includes mutation and purifying selection for the aa of the ( ) of a lineage, transition vs. transversion bias. The strength of purifying selection is incorporated by multiplying the substitution rate by a factor $\exp(-d_{aa_i,aa_j}/V)$ where d_{aa_i,aa_j} is the physiochemical distance between amino acids aa_i and aa_j defined by Grantham (1974) (i.e. Grantham Distances) and V is a parameter representing the variability of the gene or its tendency to undergo non-synonymous substitution. (V scales sensitivity to d, as $V \to 0$ infinite selection, and as $V \to \infty$ there is no selection.) The model in common use is a simplified version of this model that ignores the effect of selection. Yang et al. (1998) implemented a few mechanistic models vague, are there variations of GY? on the codon level and found from analysis of mitochondrial genomes of 20 mammalian species that they fit the data better than empir-

ical models (based on -loglikelihood or AIC or something else?). One trait common to most phylogenetic models, whether empirical or mechanistic, and including the GY model, is time reversibility. In time-reversible models, the relative substation rate q_{ij} from state i to state j is assumed to satisfy the detailed balance condition $\pi_i q_{ij} = \pi_j q_{ji}$ for any $i \neq j$. While time reversibility provides substantial mathematical and computational advantages, it is difficult to interpret biologically. This, surprisingly, has been largely ignored by the phylogentics community. (CITATIONS of exceptions?)

For example, if amino acid i is favored by natural selection, then in the absence of mutation bias we expect the substitution rate from state j to i to be faster then the reverse. While mutation bias can alter this requirement when selection is weak, it can only do so when the assumption of time reversibility is violated. (JJ - is this true? substitution rates are different from exchange rates, even in time-reversible models. Under time-reversible model, the higher equilibrium state frequency indicates higher mutation rate to this state. In the discussion follows, consider the case where amino acid i has a higher equilibrium frequency than amino acid j, even if in the exchange rate matrix synonymous rates are bigger than non synonymous rates, it's still possible that the substation rates reflect the "optimality" of amino acids.) For example, consider a time-reversible codon substitution model where synonymous substitutions occur at a faster rate than non-synonymous substitutions. For any given state of the system, such a model implies that

the current amino acid is optimal since synonymous substitutions occur at a faster rate than non-synonymous ones. However, once a non-synonymous substitution has occurred (and as time goes to infinity it will), the time reversible aspects of the model now imply that the new state is the optimal state and the old state is sub-optimal. Thus, the only reasonable way to interpret such a time-reversible model is that the substitution matrix is actually describing the rate at which the optimal state switches at a given site and that once such a switch has occurred the system instantaneously shifts to the new state. If, in contrast, one were to assume the converse, that non-synonymous substitution occur at a faster rate than synonymous, then the interpretation of time-reversible models becomes even more problematic from a biological perspective. In such a scenario, not only is the optimal state constantly changing, the current state of any given site is always sub-optimal.

While time-reversible models have played an important role in molecular phylogenetics for the last several decades (Tavaré, 1986) wrong citation, in order to model natural selection and mutation bias in a realistic manner the assumption of time reversibility must be relaxed. In this study we develop an amino-acid based model in which we assume that for each individual site i of a protein there is a corresponding optimal amino acid a'_i. The optimal state can be assigned or, as we demonstrate, estimated from the data itself. As with the GY model, we assume that the substitution rate between amino acids at a given site is a function of their Grantham distances from optimal amino

acids and, assume genes can vary in their sensitivities to such deviation from optimal amino acids. Here the sensitivity to amino acid changes is calculated using a cost-benefit framework we have developed previously for studying the evolution codon usage bias (Gilchrist, 2007; Gilchrist et al., 2009; Shah and Gilchrist, 2011). Furthermore, unlike most models in phylogenetics, we define the relative fitness of a phenotype, which is a given amino acid sequence, explicitly. We then use a model from population genetics to calculate the substitution rate between any two genotypes by explicitly taking into account the fitness differences between them as well as the effects of mutation bias and genetic drift. We use AIC to evaluate our model and alternative models by fitting them to the Rokas et al. (2003)'s data set of 106 genes sequenced from 8 different species of yeast. When fitting our model, we estimate the phylogenies of the yeast species, the Grantham sensitivity g of a gene (roughly comparable to $1/V$ in the GY model), as well as the optimal amino acid $\vec{a}'$ for each site within a coding sequence. We compare our model's fit to the Rokas data with other commonly used AA-based models using AIC criterion (Akaike, 1973, 1974, 1981). [LIST MODELS] Our model is similar to GY model in that grantham distance is used, but we include the effect of natural selection explicitly, and allow for different substitution matrices depending on which amino acid is optimal.

Our results show that even with our most parameter rich model in which we estimate the optimal amino acid at every site, thereby introducing tens

of thousands of additional parameters, our model still does a substantially better job fitting the Rokas dataset. So although the computational cost of our model is greater than most time reversible models, our ability to fit the phylogenetic data and extract biologically meaningful information is substantially greater than other models. Furthermore, because our approach explicitly links genotype to phenotype, phenotype to fitness, and fitness to fixation rate, the biological assumptions underlying our model are clearly stated and incorporation of additional biological factors, such as selection on codon usage bias, is much more straighforward.

3 Methods & Materials

In this study we use a series of continuous time Markov models with amino acid and gene specific parameterizations to describe the process of amino acid substitution for a given protein in a lineage.

Our approach, is applicable to any homologous protein-coding DNA sequence dataset where any gaps have been removed. In our model we assume that there is an optimal amino acid for any given site. The strength of natural selection for the optimal amino acid is a function of the gene's expression level, the physiochemical properties of a given amino acid, and the sensitivity of the gene's functionality to changes in these properties. It what can vary? can vary between genes and different sites in a gene. (in current implementation sensitivity does not vary within a gene) As a result, for each of 20

natural amino acids as optimal we have a 20×20 substation rate matrix with the same 20 amino acids as its states.

Conceptually the substitution process consists of two steps. First, a given amino acid i mutates to amino acid j, the rate of which depends on the mutation rates between nucleotides and the structure of the genetic code. Second, the newly arisen allele becomes fixed in the population with certain probability, which is based on the models from population genetics and includes the effects of natural selection and genetic drift. Therefore, when the optimal amino acid is k the substitution rate matrix can be written as $Q^{(k)}_{i,j} = M_{i,j}F^{(k)}_{i,j}$ for $i \neq j$, where M is the 20×20 mutation rate matrix between amino acids and F is the 20×20 matrix of fixation probabilities. Note that F, but not M, depends on optimal amino acid k for a given site. As usual the row sums of Q equal 0 and the matrix of substitution probabilities after time t is $P^{(k)} = \exp(Q^{(k)}t)$

Calculating the Mutation Rate Matrix M

We use a time reversible model for mutation between amino acids, i.e. $\pi_i M_{i,j} = \pi_j M_{j,i}$ where π_i is the equilibrium frequency of the state i. For all time reversible models, the substitution rate matrix can be written as $M = S\Pi$ where S is a symmetric matrix called the exchange rate matrix and Π is the diagonal matrix of base frequencies π_i's of the states. The model is formulated at the amino acid level, so the calculation of M involves 2 steps.

We begin with finding the mutation rate matrix between 61 sense codons. For simplicity we assume that the mutations occur independently between nucleotides within a codon, and denote the mutation rate matrix for nucleotides by M_ν. For codons that differ only by one nucleotide, the rate between codons is equal to the rate between the said pair of nucleotides. For all other codons, since the changes involving two or more nucleotides during time Δt have probabilities on the order of Δt^2, their mutation rates are set to 0. Therefore the 61×61 mutation rate matrix is rather sparse.

Second, from this codon level mutation rate matrix we can obtain 20×20 amino acid exchange rate matrix S_{aa} by grouping together the synonymous codons for each amino acid. For simplicity, the synonymous codon frequencies for any given amino acid are assumed to be the same. Following the approach developed by Yang (MBE 1998), S_{aa} for a reversible Markov process of amino acid mutation has entries:

$$(S_{aa})_{ij} = \frac{\sum_{u \in I} \sum_{v \in J} \pi_u \pi_v s_{uv}}{\pi_I \pi_J}$$

where i and j are two different amino acids, I and J are the corresponding sets of synonymous codons for i and j correspondingly, i.e. $c_u = i$ for $u \in I$ and $c_v = j$ for $v \in J$; $\pi_J = \sum_{v \in J} \pi_v$ is the equilibrium frequency of amino acid j by combining the frequencies of synonymous odors for it; and s_{uv} is the exchange rate between codons u and v. The matrix S_{aa} obtained in this way is symmetric, and we can find the mutation rate matrix by $(M)_{ij} = (S_{aa})_{ij} \pi_J$.

Calculating the Fixation Probability Matrix F

While the mutation rate matrix M accounts for the effects of the structure of the genetic code and variation in mutation rates on the generation of new alleles, F describes the probabilities of any such mutation going to fixation.

Modeling the relationship between amino acid sequence and protein function is a complex and challenging problem [CITATION]. No general techniques that accurately and reliably predict a protein's structure, much less function, currently exist. However, empirical data suggests that the effect of an amino acid on a protein's function depends largely on its physiochemical properties. Therefore, we assume that for each site i of the protein there is an optimal amino acid k_i. If a protein consists solely of the optimal amino acids at all sites then it is defined to have 100% functionality. Non-optimal amino acids reduce functionality and are, therefore subjected to purifying selection. In our model, the strength of purifying selection is a function of the physiochemical differences between the non-optimal amino acid and the optimal amino acid, a functional sensitivity term, and the expression level of the gene, specifically its average protein production rate ϕ.

Linking Amino Acid Sequence to Protein Functionality: The relative functionality of a given amino acid sequence $\vec{a} = (a_1, a_2, \cdots, a_n)$ is a function of the differences between its physiochemical properties and those of the optimal amino acid sequence $\vec{a}' = (a'_1, a'_2, \cdots a'_n)$ where n is the protein's

length. Grantham (1974) developed a physiochemical distance metric based on the composition (c), polarity (p) and molecular volume (v) of an amino acid's side chain. Composition is defined as the atomic weight ratio of the on-carbon elements in ending group or rings to carbons in the side chain while side chain polarity and molecular volume are well established (CITATION). Numerous studies (CITATION) have since shown that there is a strong negative correlation between the substitution rates between amino acids and the differences in the three physiochemical properties. Following Grantham (1974) we define the Grantham distance $d(a_i, a_j)$ between amino acids a_i and a_j as a function of their weighted distances in c, p and v physiochemical space. More precisely, $d_{ij} = [\alpha(c(a_i)-c(a_j))^2+\beta(p(a_i)-p(a_j))^2+\gamma(v(a_i)-v(a_j))^2]^{1/2}$ where α, β, γ are the corresponding weights for the 3 components. Other properties, distance measures and scalings could also be used.

Grantham weighted each property by dividing them by the mean distance found with it alone in the formula, afterwards the distances are scaled so that the mean distance between all possible amino acid pairs is 100. For example, given the values for property c of 20 amino acids, its weight α is defined as $(1/\bar{D}_c)^2 = 1.833$ where $\bar{D}_c = \sum_{i=1}^{20}\sum_{j=1}^{20}[(c\,(a_i) - c\,(a_j))^2]^{1/2}/\binom{20}{2}$. Correspondingly, the weights for polarity and molecular volume are $\beta = 0.1018$ and $\gamma = 0.000399$. Subsequent studies using these Grantham distances have used these same weights across different genes and taxa. Although these weights have thus been shown to be useful, one might expect these weights

to vary between different proteins or taxa. For example, changes in polarity might have a bigger effect on the functionality of a transmembrane protein than changes in composition, while in cytosolic enzyme, the opposite could hold. Consequently, in our model, the weights β, γ are treated as estimable parameters rather than being fixed.

With the distance between amino acids defined as above, we define the relative functionality of a protein $\vec{a}$ with optimal sequence $\vec{a}'$ as follows:

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

where g is a gene specific Grantham sensitivity coefficient which quantifies the sensitivity of a given protein's function to the deviation of physiochemical properties from the optimal sequence, and d_k is the Grantham distance between the given and optimal amino acids at the k^{th}position. Note that the optimal amino acid sequence has a relative functionality of 1. In order to simplify the notation we will drop $\vec{a}'$ and g when there is no potential ambiguity.

Defining Protein Fitness: Following our previous work relating protein production cost, relative functionality, and energy expenditure to fitness [CITATION] Gilchrist (2007); Shah and Gilchrist (2011), we define a cost-benefit function $\eta(\vec{a}|\vec{a}')$ as the expected cost, in ATPs of producing one unit or protein function, i.e. the equivalent of one optimal protein sequence. Here we assume that the cost of protein translation is simply proportional to the

length of the protein produced. Thus,

$$\eta\left(\vec{a}|\right) = \frac{C(n)}{F\left(\vec{a}|\vec{a}', \vec{g}\right)}, \tag{2}$$

where $C(n)$ is the cost of producing a protein of length n. Based on the basic biology of protein translation, we define $C(n) = a_1 + a_2(n-1)$ where $a_1 =$ 4 ATPs is the cost in ATPs of ribosome assembly on an mRNA transcript and $a_2 = 4$ ATPs is the cost in ATPs of tRNA charging and the moving the ribosome forward one codon.

For a given gene, we assume that there is a mean target functionality production rate, i.e. the average rate at which the organism requires the production of the functionality encoded by that gene. Thus, if $\eta(\vec{a})$ represents the cost of producing one unit of functionality and the organism, on average, needs to produce that functionality at rate ϕ, then $\eta(\vec{a}) \times \phi$ represents the rate at which the organism must spend energy to meet the functionality requirement provided by a given gene. Letting q represent the proportional gain in fitness for each ATP saved per unit time, we can define the relative fitness of a protein $\vec{a}_i$ as

$$f(\vec{a}_i) = f_i \propto \exp\{-\frac{\phi q C(n)}{F\left(\vec{a}_i|\vec{a}'\right)}\}.$$

Clearly, fitness f_i is an increasing function of functionality F and the strength of selection on F increases with protein length n and expression level ϕ. Note that n, ϕ, and $\vec{a}'$ vary between loci, $\vec{a}$ varies between alleles for a given locus, but q is the same for all genes.

Following the model of allele fixation presented by Sella and Hirsh (2005), the fixation probability of a newly introduced mutant allele $\vec{a}_j$ in a Fisher-Wright population with the resident allele $\vec{a}_i$ and an effective size of N_e is,

$$fix_{ij} = fix\left(\vec{a}_i \to \vec{a}_j\right) = \frac{1 - \left(f\left(\vec{a}_i\right)/f\left(\vec{a}_j\right)\right)^{\alpha}}{1 - \left(f\left(\vec{a}_i\right)/f\left(\vec{a}_j\right)\right)^{2N_e}}. \tag{3}$$

(need a letter name for fixation probability) where $\alpha = 1$ for a diploid population and 2 for a haploid population. Here we focus on diploid populations. This formula for the fixation of a mutant allele is valid under weak mutation assumptions, i.e. $s, \frac{1}{N_e}, Ns^2 \ll 1$ where $s = 1 - \frac{f(\vec{a}_i)}{f(\vec{a}_j)}$. Alternative fixation calculations, such as the canonical forms derived by Fisher (1930); Moran et al. (1962); Wright (1931) or Kimura (1962) could also be used.

The fixation probability described by Equation 3 depends on the ratio of the resident and mutant alleles fitnesses, f_i/f_j. Using the definitions of protein translation cost $C(n)$ and functionality $F\left(\vec{a}|\vec{a}'\right)$ in Equation 1, we get

$$\frac{f\left(\vec{a}_i\right)}{f\left(\vec{a}_j\right)} = \prod_{k=1}^{n}\left(\frac{f\left(\vec{a}_i^{\,k}\right)}{f\left(\vec{a}_j^{\,k}\right)}\right)^{\frac{1}{n}}. \tag{4}$$

Thus under the assumptions of our model, the fitness ratio of the resident and mutant genotypes $\vec{a}_i$ and $\vec{a}_j$ is the geometric mean of the fitness ratios between the two amino acids at all sites. When $\vec{a}_i$ and $\vec{a}_j$ only differ at a

single amino acid position k, the fitness ratio f_i/f_j simplifies to

$$\begin{aligned}
\frac{f\left(\vec{a}_i\right)}{f\left(\vec{a}_j\right)} &= \left(\frac{f\left(\vec{a}_i^k\right)}{f\left(\vec{a}_j^k\right)}\right)^{\frac{1}{n}} && (5)\\
&= \exp\left[-q\phi\left(\frac{C(n)}{F\left(\vec{a}_i\right)}-\frac{C(n)}{F\left(\vec{a}_j\right)}\right)\right]\\
&= \exp\left[-q\phi\frac{C(n)}{n}\left(d_k^{(i)}-d_k^{(j)}\right)g\right] && (6)\\
&= \exp\left[-q\phi a_2\left(d_k^{(i)}-d_k^{(j)}\right)g\right] && (7)
\end{aligned}$$

where, again, a_2 is the cost of each elongation step during protein translation. Equation (7) shows that in our model even though the $F(\vec{a})$ is a non-linear function of the entire sequence, the fitness ratio of two alleles depends only on the site that differs. Thus, within a given gene each amino acid site evolves independently of the other. This equation also indicates that the strength of selection on these distance differences increases with the value q, the cost of an elongation step a_2, gene's expression level ϕ, and the sensitivity of protein function to its deviation from the optimal sequence g.

Substituting Equation (7) into Equation (3) gives,

$$fix_{ij} = \frac{1-\exp\left[-q\phi a_2\left(d_k^{(i)}-d_k^{(j)}\right)g\right]}{1-\exp\left[-q\phi a_2\left(d_k^{(i)}-d_k^{(j)}\right)g2N_e\right]}. \qquad (8)$$

Equation (3) shows that the fixation probability of an allele is a function of the fitness ratio f_i/f_j as well as effective population size N_e.

Therefore, connecting mutation and fixation steps together, the instantaneous substitution rate q_{ij} from $\vec{a}_i$ to $\vec{a}_j$ is equal to the rate at which a

mutant is introduced, times fixation probability of the mutant fix_{ij}:

$$q_{ij} = 2N_e fix_{ij}\pi_{ij}. \tag{9}$$

Given the values for $M_\nu, g, \alpha, \beta, \gamma, C, \phi, q, N_e$, the frequencies of different amino acids Π and the optimal amino acid k at a site, we can calculate the 20×20 instantaneous substitution rate matrix $Q^{(k)}$ 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 scaling, the length of a branch represents the expected number of substitutions along the branch. With the probabilities $P(t) = \exp(Qt)$ the likelihood at 1 site 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.

3.1 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 of 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 N_e is assumed to be fixed across all lineages in this paper. Suppose the phylogenetic topology is given, we are estimating the following parameters: $g, \beta/\alpha, \gamma/\alpha$, frequencies of

Table 1: parameters in the model

s_{ij}	exchange rates between nucleotides i and j
fix_{ij}	mutation rates from nucleotides i to j
π_{ij}	fixation probability of single mutant j from i
q_{ij}	substitution rate from amino acid i to amino acid j
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 Π, branch lengths, the exchange rate matrix M_ν for nucleotides, and the optimal protein sequence for the given gene.

3.2 Identification of optimal amino acids

To calculate the likelihood values, the optimal amino acids need to be identified. We implemented 3 approaches to identify the optimal amino acid at a certain site. First one is called "max rule". The likelihood values are calculated 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 amino acids as estimable parameters in the maximum likelihood estimation. 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, there will not be enough substitutions, the optimal amino acids estimated this way can be inaccurate.

The third method is "weighted rule". 20 amino acids are assigned weights (probabilities) of being optimal. If the same set of 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.

3.3 Inference of ancestral states

Unlike empirical models (and other models?) where the exchange rate matrix is fixed and ancestral state frequencies are usually either equilibrium empirical frequencies from observed data, the exchange rate matrix under our model depends on several parameters. For every site, once the optimal amino acid is chosen, there are several options for choosing the ancestral state frequencies when calculating likelihood values.

- EmpRoot: The empirical frequencies from all observed data can be used like all other empirical models.

- EqmRoot: The equilibrium frequencies. Since the substitution rate matrix depends on optimal amino acid while other parameters are fixed, different sites can have different ancestral state frequencies if their optimal amino acids are different.

- MaxRoot: The ancestral state can be specified as the one that provides the maximum likelihood value.

- OpaaRoot: The ancestral state can be chosen as the same one as the optimal amino acid. However, it seems paradoxical to assume that the evolution starts from the optimal amino acid and changes to worse states.

Again, MaxRoot always gives the best likelihood values while treats the ancestral state as estimable parameters at each site. The AIC values of different approaches will be compared in the Results section.

4 Results

We analyzed three sets of nucleotide data: yeast, insects, and mammals.

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

We analyzed data previously studied by Rokas et al. (2003). 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 chromosomes and comprises a total length of 42,342 amino acids. Rokas et al. (2003) analyzed this data set to investigate the conflict of gene trees. We use the tree topology that is supported by the concatenated genome sequence, and the majority of the genes as found by Rokas et al. (2003). Since the new model is not time reversible the tree is rooted with 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 (Darriba et al., 2011; Guindon and Gascuel, 2003) 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 different approaches to determine the optimal amino acids and ancestral state frequencies for each site. In all the analyses, tree branch lengths are optimized while the topology is fixed.

The log-likelihood values and the AIC values are compared in Table 2.

Under the new model, amino acid frequencies are the observed frequencies in the sequences therefore counted as 19 free parameters. In addition, there are 5 free parameters for exchange rates between nucleotides with the rate between G and T fixed as 1, Grantham sensitivity g, 2 free parameters for the weights in the physicochemical distance formula β and γ. These 27 parameters are estimated for all data sets. Depending on the number of branch lengths, and number of different site patterns at the tips, the total number of parameters vary between data sets.

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

Model	ΔAIC	l	Parameters
New+max+MaxRoot	0.00	-133891.20	18,297
New+max+OpaaRoot	103816.00	-194927.20	9,169
New+max+EqmRoot	168993.00	-227515.70	9,169
New+max+EmpRoot	193432.40	-239735.40	9,169
New+maj+EmpRoot	211285.80	-257790.10	41
New+weights+EmpRoot	334974.40	-319615.40	60
LG+I+G+F	293089.78	-298699.09	34
LG+G+F	293094.66	-298702.53	33

Table 2: Comparison between empirical models and new model for the yeast sequence data with 42,342 amino acids and 9,128 different site patterns.

are gene specific and other parameters are fixed across genes? Total number of parameters is 40,369. Better case scenario, 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 -311,186. Even with this loglikelihood value and number of parameters 40,369, the ΔAIC value is 187,447.8, which means it performs worse than the third model in the table.The real loglikelihood value under this model is -317,214.68; it gives a larger AIC value.)

From Table 2 the best model according to AIC values is the new model

with max rule for both optimal amino acid and the ancestral state. Even though the number of parameters is much bigger than the empirical models, the improvement of likelihood is so large that this model still outperforms the best empirical model, with AIC value 293,089 units smaller.

4.1.2 Parameter variation between genes

We also analyzed the yeast data gene by gene. The estimates for Grantham sensitivity and weights are on the similar scale across all 106 genes in the data. As expected, the weights for physicochemical properties are also similar to what Grantham proposed. Figure 1 showed the correlation between β and γ. Linear regression suggests strong linear relationship between the 2 parameters. Note that α value is fixed for all sites as 1.833, the results indicate that the ratios between weights for the 3 components in the distance formula do not vary a lot.

There are several possible explanations for the variation of g values between genes. One, these genes have different structures, which caused the different degrees of sensitivity to the distance from the optimal amino acids. For example, hydrophobic cores of proteins can be efficiently repacked with different hydrophobic sequences. All polar amino acids can form hydrogen bonds whose thermodynamic energy varies sharply with distance and angle, providing a rationale for the greater variability of the fitness of polar amino

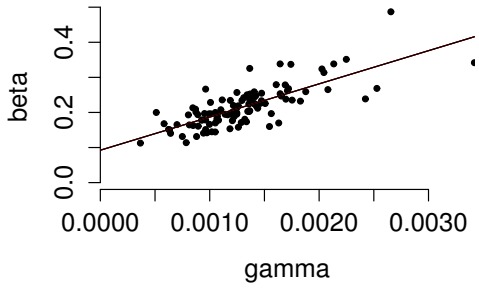

Figure 1: Correlation between β and γ, $R^2 = 0.5844$.

acids. Two, we use the same tree topology for all genes. However, sequences in some gene might better support a different tree topology, therefore causes other parameter estimates to be inaccurate.

4.1.3 Relationship between sensitivity g and expression level ϕ

It is believed (and found) [citation] that genes with higher expression level are under stronger selection against non-optimal product. In our model, the selection strength is reflected through the Grantham sensitivity g to the difference from the optimal sequence. When the sensitivity is bigger, the selection is stronger. Therefore there should be a positive relationship between g and ϕ. Since we only have estimates of $g\phi$ from the model (since we estimate g with ϕ fixed), ϕ values for the genes in the Rokas's data come from Gilchrist (2007). From figure 2 it seems like that the results from our

model give opposite conclusion, i.e. selection is negatively correlated with gene expression level. However, the fact that the estimates of ϕ values are much noisier than those of $g\phi$ could make the plot deceptive, and simple linear regression cannot give correct estimate of the slope either. In fact, we conducted some simulation to verify this. Suppose that $\log(\phi)$ follows a random normal distribution across all genes, with relatively big standard deviation; and that $\log(g) = a + b\log(\phi)$ is a linear function of ϕ. Then we simulated values of $\log(g \cdot \phi)$ with smaller standard deviation. We then look at the relationship between $\log(g \cdot \phi/\phi)$ and $\log(\phi)$. From the right plot in figure 2 we can see that even when the simulation is done under the assumption $b = 2$, which represents positive linear relationship, with big noise in ϕ the result indicates the opposite relationship. Therefore, without more accurate estimates of ϕ values, it is not possible to draw useful conclusion on the relationship between the 2 parameters. And we cannot say that the result from our model contradicts the expectation.

4.1.4 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 site patterns at the 8 species. For each of the 9000+ sites, the likelihood values achieved by assuming each amino acid

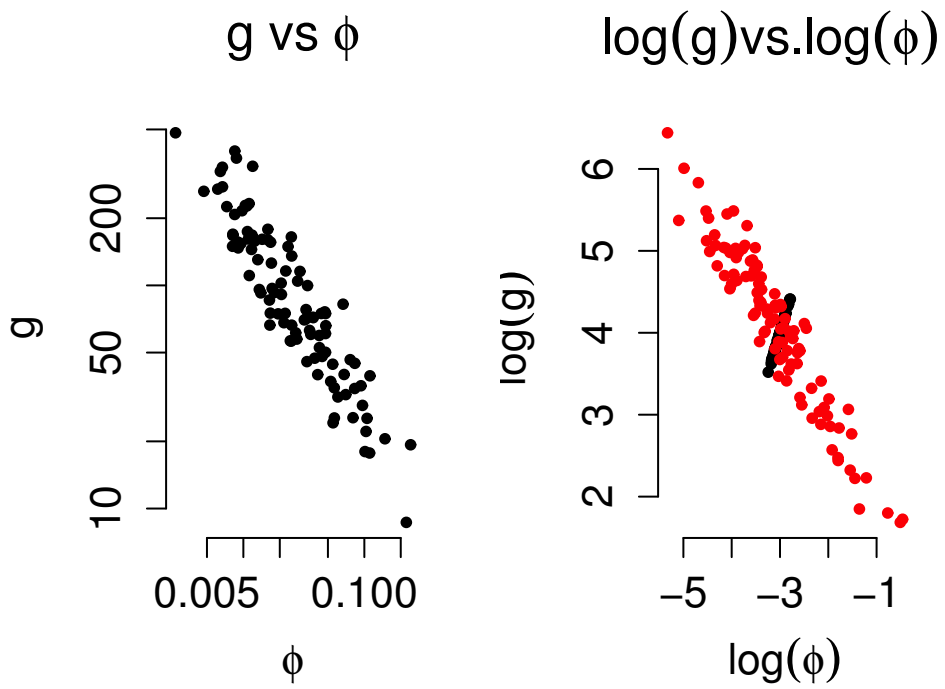

Figure 2: Left: Plot of sensitivity g against SEMPPR ϕ values for 106 genes in yeast data on log scale. Right: Plot of simulated data, black dots are from true values of ϕ and red dots are ϕ values with noise.

as optimal is ordered decreasingly, therefore the likelihood under the max optimal amino acid is ranked the first. Then the next amino acid is included in the optimal set of amino acids until the total likelihood exceeds 95% of the total likelihood. In the following figures, the equilibrium frequencies are used for the ancestral states. Figure 3 shows the histogram of numbers of optimal amino acids in the set. The mode for all 9000+ patterns is 6 amino

acids. The case where there are more than 10 amino acids in the set rarely happened. Figure 4 showed the density of percentages of total likelihood value covered by the optimal amino acid found with max rule only. Mean percentage is 0.4265 and the peak of the density distribution is between 0.3 and 0.4. (Can we say anything based on these?)

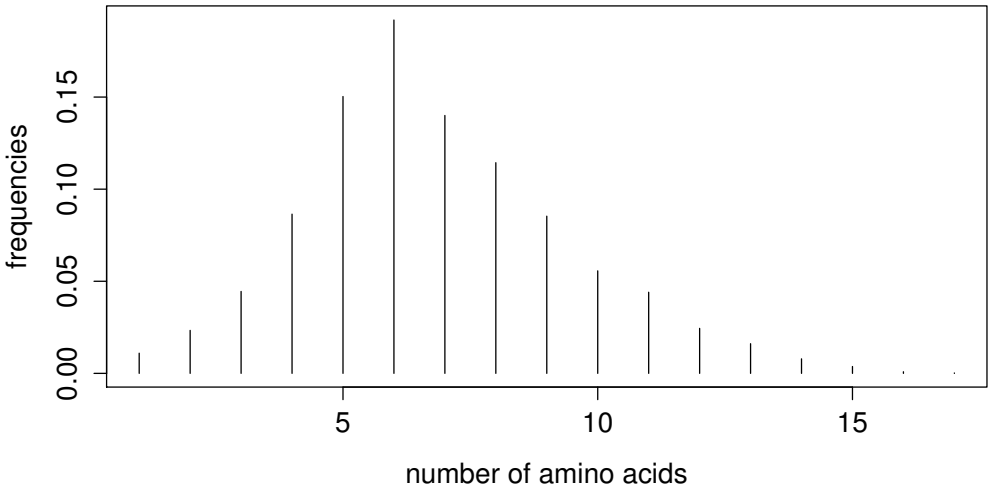

Figure 3: Histogram of the number of optimal amino acids together to cover at least 95% of the total likelihood attained by all possible optimal amino acids.

4.1.5 Simulated data vs. observed data

We analyzed model adequacies of new model and the best empirical model in the following way. For a given data set, we compare the new model with

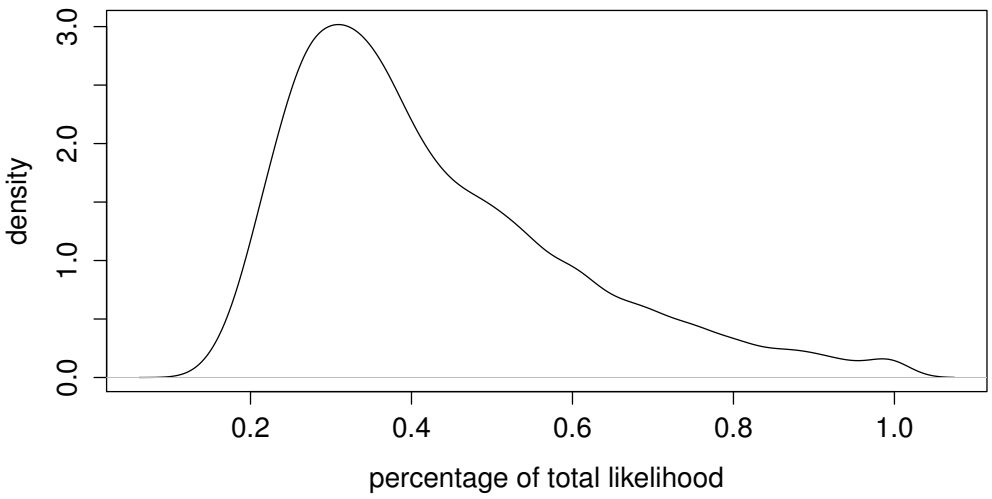

Figure 4: Density plot of percentages of the likelihood achieved by the optimal amino acid found by max rule.

max rule for optimal amino acids and ancestral states, (Other methods for ancestral state frequencies give similar results) and the best empirical model. First, a single taxon and the branch connecting to it is pruned from the phylogenetic tree and its sequence is deleted in the observed data. The models are fitted to the resulted smaller data set with remaining taxa, and maximum likelihood estimates for all parameters are estimated. Then the deleted taxon is put back. With the parameters just found on the smaller tree, the location and the length of the extant branch are estimated with maximum likelihood. With all the parameters and branch lengths, we found the probabilities of observing different sequences at the start of the grafted branch. After this is

done with both models in consideration, we have the probabilities from which the starting sequence at the grafting point are drawn, and the length of the grafted extant branch. Then we simulate sequences on this new branch under both models, record the change of functionalities on the evolution path, and compare the simulated sequences at the tip with the observed sequence.

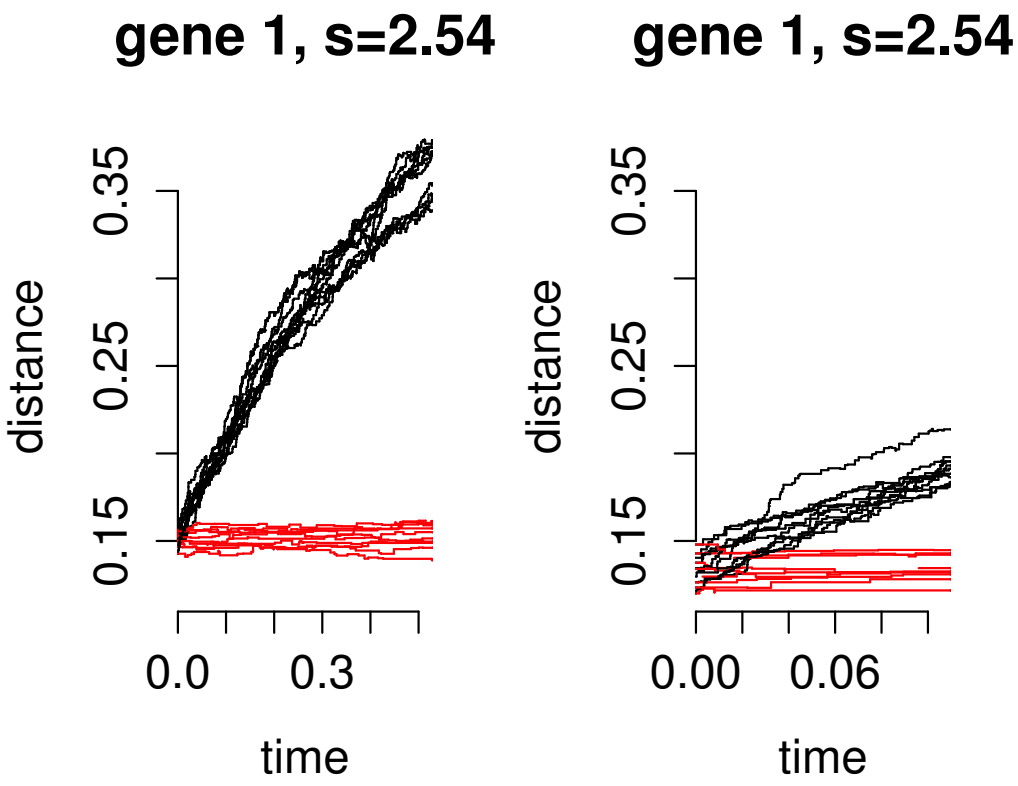

Figure 5: Analysis of model adequacy, simulations on one branch. Left: parameter values under new model; Right: parameter values under the best empirical model.

On the left, each dot represents the proportion of amino acids that differ between the simulated and the observed sequence for a given gene. Our new model performed much better than the standard WAG model in matching sequences, especially for genes under high selection that are shown in brighter dots.

For repeated simulations under our new model (red) and empirical model (black) starting from the ancestral sequence estimated under the new model (left) and empirical model (right) for a single gene (gene1 in yeast data in this case), Figure 5 plotted the distance between the simulated sequences along the evolution paths and the observed sequence.

The ancestral sequences estimated from both new and empirical models have similar distance from the observed sequence. However, the estimated length of new branch is much bigger under the new model. If evolved under new model, the distance does not change much at the end of the branch. On the other hand there is steady increase in the distance from observed sequence if the sequences evolve under the empirical model.

No matter how the ancestral sequences are obtained, the new model presents a better match to the observed data. This realistic behavior shows that the new model is more adequate than the empirical models for Rokas's data.

4.2 Result on insect data and mammal data

We also analyzed the model performance on the insect data used in McKenna and Farrell (2010) and mammal data used in Zhou et al. (2012). DNA sequence data on insect comprised of approximately 13 kb of aligned data from 7 single-copy nuclear protein-coding genes: elongation factor-1α, alanyl-tRNA synthetase (AATS), carbamoylphosphate synthase domain (CAD), 6-phosphogluconate dehydrogenase (PGD), sans fille (SNF), triosephosphate isomerase (TPI), and RNA polymerase II (RNA Pol II). The taxon sample was comprised of 34 insects, including 32 exemplars representing all orders of holometabolous insects, and two hemimetabolous insect outgroups. Mammal data comprised of 97 orthologs, 46,152 bp, for 15 taxa, representing all laurasiatherian orders.

Data	Model	ΔAIC	l	Parameters
Insects	New+max+MaxRoot	0.00	-23,276.19	3,565
	New+max+OpaaRoot	4,758.80	-27,391.60	1,829
	New+max+EqmRoot	4,765.70	-27,395.05	1,829
	LG+I+G+F	5,941.30	-29,725.85	86
	LG+G+F	5,956.23	-29,752.31	85
	New+maj	7,615.78	-30,556.09	93
	New+max+EmpRoot	8,245.66	-29,135.03	1,829
	New+opw	15,206.78	-34,282.59	112
Mammal	New+max+MaxRoot	0.00	-40,574.92	6,679
	New+max+OpaaRoot	25,248.82	-56,511.33	3,367
	New+max+EqmRoot	41,606.42	-64,690.13	3,367
	New+max+EmpRoot	72,443.58	-80,108.71	3,367
	New+maj			
	HIVb+I+G+F	99,387.24	-96,899.54	48
	JTT+I+G+F	99,566.15	-96,950.90	48
	New+opw	107,394.76	-100,877.30	74

Table 3: Comparison of model performance for insect and mammal data

References

Jun Adachi and Masami Hasegawa. Model of amino acid substitution in proteins encoded by mitochondrial dna. *Journal of Molecular Evolution*,

42(4):459–468, 1996.

Jun Adachi, Peter J Waddell, William Martin, and Masami Hasegawa. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast dna. *Journal of Molecular Evolution*, 50(4):348–358, 2000.

Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory, 2 nd, Tsahkadsor, Armenian SSR*, pages 267–281, 1973.

Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

Hirotugu Akaike. Likelihood of a model and information criteria. *Journal of econometrics*, 16(1):3–14, 1981.

Ying Cao, Jun Adachi, Axel Janke, Svante Pääbo, and Masami Hasegawa. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *Journal of Molecular Evolution*, 39(5):519–527, 1994.

Diego Darriba, Guillermo L. Taboada, Ramón Doallo, and David Posada. Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–1165, 2011. doi: 10.1093/bioinformatics/

btr088. URL http://bioinformatics.oxfordjournals.org/content/27/8/1164.abstract.

MO Dayhoff, RM Schwartz, and BC Orcutt. A model of evolutionary change in proteins. *In Atlas of Protein Sequences and Structure*, 5:345–352, 1978.

Joseph Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981. ISSN 0022-2844. doi: 10.1007/BF01734359. URL http://dx.doi.org/10.1007/BF01734359.

RA Fisher. *The Theory of Natural Selection*. Oxford University Press, London, 1930.

M. A. Gilchrist. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Mol Biol Evol*, 24(11):2362–72, 2007.

M A Gilchrist, P Shah, and R Zaretzki. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics*, 183(4):1493–1505, Dec 2009. doi: 10.1534/genetics.109.108209. URL http://www.hubmed.org/fulltext.cgi?uids=19822731.

N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Mol Biol Evol*, 11(5):725–36, 1994.

Gaston H Gonnet, Mark A Cohen, Steven A Benner, et al. Exhaustive matching of the entire protein sequence database. *Science*, 256(5062): 1443–1445, 1992.

R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864, 1974. doi: 10.1126/science.185.4154.862. URL `http://www.sciencemag.org/content/185/4154/862.abstract`.

Stéphane Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, 2003. doi: 10.1080/10635150390235520. URL `http://sysbio.oxfordjournals.org/content/52/5/696.abstract`.

Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, 22:160–174, 1985. ISSN 0022-2844. doi: 10.1007/BF02101694. URL `http://dx.doi.org/10.1007/BF02101694`.

Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22): 10915–10919, 1992.

David T Jones, William R Taylor, and Janet M Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences: CABIOS*, 8(3):275–282, 1992.

TH Jukes and CR Cantor. Evolution of protein molecules. *Manmmalian Protein Metabolism*, III:21–132, 1969.

Motoo Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713, 1962.

Motoo Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980. ISSN 0022-2844. doi: 10.1007/BF01731581. URL `http://dx.doi.org/10.1007/BF01731581`.

Hirohisa Kishino, Takashi Miyata, and Masami Hasegawa. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, 31(2):151–160, 1990.

Si Quang Le and Olivier Gascuel. An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7):1307–1320, 2008.

Duane D. McKenna and Brian D. Farrell. 9-genes reinforce the phylogeny of holometabola and yield alternate views on the phylogenetic placement of strepsiptera. *PLoS ONE*, 5(7):e11887, 07 2010. doi: 10.1371/journal.pone.0011887. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0011887`.

Patrick Alfred Pierce Moran et al. The statistical processes of evolutionary theory. *The statistical processes of evolutionary theory.*, 1962.

S V Muse and B S Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, 1994. URL `http://mbe.oxfordjournals.org/content/11/5/715.abstract`.

A Rokas, B L Williams, N King, and S B Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804, Oct 2003. doi: 10.1038/nature02053. URL `http://www.hubmed.org/fulltext.cgi?uids=14574403`.

G Sella and A E Hirsh. The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci U S A*, 102(27):9541–9546, Jul 2005. doi: 10.1073/pnas.0501865102. URL `http://www.hubmed.org/fulltext.cgi?uids=15980155`.

P Shah and M A Gilchrist. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proc Natl Acad Sci U S A*, 108(25):10231–10236, Jun 2011. doi: 10.1073/pnas.1016719108. URL `http://www.hubmed.org/fulltext.cgi?uids=21646514`.

Simon Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lect. Math. Life Sci*, 17:57–86, 1986.

Simon Whelan and Nick Goldman. A general empirical model of protein evo-

lution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5):691–699, 2001.

Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.

Z Yang and R Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*, 25(3):568–579, Mar 2008. doi: 10.1093/molbev/msm284. URL `http://www.hubmed.org/fulltext.cgi?uids=18178545`.

Z. Yang, R. Nielsen, and M. Hasegawa. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol*, 15 (12):1600–11, 1998.

Xuming Zhou, Shixia Xu, Junxiao Xu, Bingyao Chen, Kaiya Zhou, and Guang Yang. Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the laurasiatherian mammals. *Systematic Biology*, 61(1):150–164, 2012. doi: 10.1093/sysbio/syr089. URL `http://sysbio.oxfordjournals.org/content/61/1/150.abstract`.