Population Genetics Based Phylogenetics Under Stabilizing Selection for an Optimal Amino Acid Sequence: A Nested Modeling Approach

JEREMY M. BEAULIEU[1,2,3], BRIAN C. O'MEARA[2,3], RUSSELL ZARETZKI[4], CEDRIC LANDER[2,3], JUANJUAN CHAI[2,5], AND MICHAEL A. GILCHRIST[2,3,*]

[1]Department of Biological Sciences, University of Arkansas, Fayetteville, AR 72701

[2]Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610

[3]National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996-3410

[4]Department of Business Analytics & Statistics, Knoxville, TN 37996-0532

[5]Current address: 50 Main St, Suite 1039, White Plains, NY 10606

[*]Corresponding author. E-mail: mikeg@utk.edu

Version dated: Wednesday 4th October, 2017

Abstract

We present a phylogenetic approach rooted in the field of population genetics that more realistically models the evolution of protein-coding DNA under the assumption of stabilizing selection for a gene specific, optimal amino acid sequence. In addition to being consistent with the fundamental principles of population genetics, our new set of models, which we collectively call SelAC (Selection on Amino acids and Codons), fit phylogenetic data much better than popular models, suggesting strong potential for more accurate inference of phylogenetic trees and branch lengths. SelAC also demonstrates that a large amount of biologically meaningful information is accessible when using a nested set of mechanistic models. For example, for each position SelAC provides a probabilistic estimate of any given amino acid being optimal. SelAC also assumes the strength of selection is proportional to the expression level of a gene and, therefore, provides gene specific estimates of protein synthesis rates. Finally, because SelAC's is a nested approach based on clearly stated biological assumptions, it can be expanded or simplified as needed.

Phylogenetic analysis now plays a critical role in most aspects of biology, particularly in the fields of ecology, evolution, paleontology, medicine, and conservation. While the scale and impact of phylogenetic studies has increased substantially over the past two decades, the realism of the mathematical models on which these analyses are based has changed relatively little by comparison. For example, the simplest but most popular models are nucleotide-based, which are naturally agnostic with regards to the different amino acid substitutions and their impact on gene function (e.g. F81, F84, HYK85, TN93, and GTR, see Yang (2014) for an overview).

Another set of models attempt to include a 'selection' term ω, but the link between ω and the key parameters found in standard population genetics models such as N_e, the distribution of fitness across genotype space, and mutation bias is far from clear. For instance, ω is generally interpreted as indicating whether a sequence is under 'purifying' ($\omega < 1$) or 'diversifying' ($\omega > 1$) selection. However, the actual behavior of the model is quite different. When $\omega < 1$ the model behaves as if the resident amino acid i at a given site is favored by selection since synonymous substitutions have a higher substitution rate than any possible non-synonymous substitutions. Paradoxically, this selection regime for the resident amino acid i persists *until* a substitution for another amino acid, j, occurs. As soon as amino acid j fixes, but not before, selection now favors amino acid j over all other amino acids, including i. This is now the opposite scenario to when i was the resident. Similarly, when $\omega > 1$, synonymous substitutions have a lower substitution rate than any possible non-synonymous substitutions the resident amino acid. In a parallel manner, this selection *against* the resident amino acid i persists until a substitution occurs at which point selection now *favors* the former resident amino acid i as well as the 18 others. Thus, the simplest and most consistent interpretation of ω is that it represents the rate at which the selection regime itself changes, and this change in selection perfectly coincides with the fixation of a new amino acid. As a result, ω based approaches only reasonably describe a

Does any one recall why we include mutation bias? These models usually include nt specific mutation rates.

subset of scenarios such as over/underdominance or frequency dependent selection (Hughes and Nei 1988; Nowak 2006). Because, as we show here, ω is well correlated with gene expression, its value is really an indicator of the strength of stabilizing selection on a coding sequence, rather than the 'nature' of that selection.

Given the continual growth in computational power available to researchers, it is now possible to utilize a more general set of population genetics based models for the purpose of phylogenetic analysis (e.g. Halpern and Bruno 1998; Robinson et al. 2003; Lartillot and Philippe 2004; Rodrigue and Lartillot 2014). One lesson from the field of population genetics is even when there are only a few fundamental evolutionary forces at play (mutation, drift, selection, and linkage effects), describing the evolutionary behavior of a system in which there are non-linear interactions between sites, such as epistasis, quickly becomes extremely challenging. The model formulation we evaluate here is a basic version of a more general cost-benefit model we've developed elsewhere (Gilchrist 2007; Gilchrist et al. 2009; Shah and Gilchrist 2011; Gilchrist et al. 2015). This basic version carefully avoids any non-linear interactions between evolutionary forces, resulting in simple additive effects between amino acid sites. This additivity between sites is critical to ensuring that calculation of our amino acid substitution matrix can be done in a site independent manner and, thus, dramatically reduce the computational cost of model fitting.

This additivity between sites also means our model could be generalized further and simply posed as a more generic, non-mechanistic, additive model. While often useful in the early stages of a field's development, given the maturity of the field of phylogenetics, we believe such model generalization is now counterproductive. The misinterpretation of GY94's ω we discuss above is a case in point. Another example, which we touch upon in the Discussion, is the natural emergence of epistsis between sites when site independent selection on both the codon usage and the amino acid usage occur. While this epistasis may be negligible under certain conditions, identifying such conditions is impossible without considering the mechanisms of selection.

Materials & Method

We model the substitution process as a classic Wright-Fisher process which includes the forces of mutation, selection, and drift (Fisher 1930; Kimura 1962; Wright 1969; Iwasa 1988; Berg and Lässig 2003; Sella and Hirsh 2005; McCandlish and Stoltzfus 2014). For simplicity, we ignore linkage effects and, as a result of this and other assumptions, our method behaves in a site independent manner. Our approach, which we call SelAC (Selection on Amino acids and Codons), is developed in the same vein as previous phylogenetic applications of the Wright-Fisher process (e.g. Muse and Gaut 1994; Halpern and Bruno 1998; Yang and Nielsen 2008; Rodrigue et al. 2005; Koshi and Goldstein 1997; Koshi et al. 1999; Dimmic et al. 2000; Thorne et al. 2012; Lartillot and Philippe 2004; Rodrigue and Lartillot 2014). Similar to Lartillot's work (Lartillot and Philippe 2004; Rodrigue and Lartillot 2014), we assume there is a finite set of rate matrices describing the substitution process and that each position within a protein is assigned to a particular rate matrix category. Unlike this previous work, we assume *a priori* there are 20 different families of rate matrices, one family for when a given amino acid is favored at a site. As a result, SelAC allows us to quantitatively evaluate the support for a particular amino acid being favored at a particular position within the protein encoded by a particular gene.

Because SelAC requires twenty families of 61 × 61 matrices, the number of parameters needed to implement SelAC would, without further assumptions, be extremely large (i.e. on the order of 73,000 parameters). To reduce the number of parameters needed, while still maintaining a high degree of biological realism, we construct our gene and amino acid specific substitution matrices using a submodel nested within our substitution model, similar to approaches in Gilchrist (2007); Shah and Gilchrist (2011); Gilchrist et al. (2015).

Could you confirm this 73k value correct? I get it from

One advantage of a nested modeling framework is that it requires only a handful of genome-wide parameters such as nucleotide specific mutation rates (scaled by effective population size N_e), side chain physicochemical weighting parameters, and a shape parameter describing the distribution of site sensitivities. In addition to these genome-wide parameters, SelAC requires a gene g specific expression parameter ψ_g which describes the average rate at which the protein's functionality is produced by the organism. (For notational simplicity, we will ignore the gene specific indicator $_g$, unless explicitly needed.) Currently, ψ is fixed across the phylogeny, though relaxing this assumption is a goal of future work. The gene specific parameter ψ is multiplied by additional model terms to make a composite term ψ' which scales the strength and efficacy of selection for the optimal amino acid sequence relative to drift (see). In terms of the functionality of the protein encoded, we assume that for any given gene there exists an optimal amino acid sequence $\vec{a}^*$ and that, by definition, is a complete, error free peptide consisting of $\vec{a}^*$ and provides one unit of the gene's functionality. We also assume that natural selection favors genotypes that are able to synthesize their proteome more efficiently than their competitors and that each savings of an high energy phosphate bond per unit time leads to a constant proportional gain in fitness A_0. SelAC also requires the specification (as part of parameter optimization) of an optimal amino acid at each position or site within a coding sequence which, in turn, makes it the largest category of parameters we estimate. Because we use a submodel to derive our substitution matrices, SelAC requires the estimation of a fraction of the parameters required when compared to approaches where the substitution rates are allowed to vary independently (Halpern and Bruno 1998; Lartillot and Philippe 2004; Rodrigue and Lartillot 2014).

As with other phylogenetic methods, we generate estimates of branch lengths and nucleotide specific mutation rates. In addition, because the math behind our model is mechanistically derived, our method can also be used to make quantitative inferences on the optimal amino acid sequence of a given protein as well as the average synthesis rate of each protein used in the analysis. The mechanistic basis of SelAC also means it can be easily extended to include more biological realism and test more explicit hypotheses about sequence evolution.

Mutation Rate Matrix μ

We begin with a 4x4 nucleotide mutation matrix that defines a model for mutation rates between individual bases. For our purposes, we rely on the general unrestricted model(Yang 1994, UNREST) because it makes no constraint on the instantaneous rate of change between any pair of nucleotides. In our view, the flexibility and potential for strong asymmetries in the transition among the different nucleotide states, and ultimately among the different codon states, is more consistent with our model. We note, however, that more constrained models, such as the Jukes-Cantor (JC), Hasegawa-Kishino-Yano (HKY), or the general time-reversible model (GTR), can also be used. The 12 parameter UNREST model defines the relative rates of change between a pair of nucleotides. Thus, we arbitrarily set the G→T mutation rate to 1, resulting in 11 free mutation rate parameters in the 4x4 mutation nucleotide mutation matrix. The nucleotide mutation matrix is also scaled by a diagonal matrix π whose entries, $\pi_{i,i} = \pi_i$, correspond to the equilibrium frequencies of each base. These equilibrium nucleotide frequencies are determined by analytically solving $\pi \times \mathbf{Q} = 0$. We use this $\mathbf{Q}$ to populate a 61×61 codon mutation matrix μ, whose entries $\mu_{i,j}$ describe the mutation rate from codon i to j under a "weak mutation" assumption. That is, the rate of allele fixation is greater than $N_e\mu$ and $N_e\mu \ll 1$, such that evolution is mutation limited, codon substitutions only occur one nucleotide at a time and, as a result, the rate of change between any pair of codons that differ by more than one nucleotide is zero.

Didn't we compare these models to UNREST and find UNREST better? If so we should mention this fact. Shouldn't

While the overall model does not assume equilibrium, we still need to scale our mutation matrices μ by a scaling factor S. As traditionally done, we rescale our time units such that at equilibrium, one unit of branch length represents one expected neutral substitution per site. More explicitly, $S = 1/\left(\sum_{i \in \text{codons}} \mu_i \pi_i\right)$. Thus, the final mutation rate matrix is the original mutation rate matrix multiplied by S, and, on average, the expected rate of neutral evolution is 1 substitution per unit time.

I assume the $-$? in the previous version was a typo, correct? Are my corrections here correct?

Protein Synthesis Cost-Benefit Function η

SelAC links fitness to the product of the cost-benefit function of a gene η and the organism's average target synthesis rate of the functionality provided by gene ψ. This is because the average flux energy an organism spends to meet its target functionality provided by the gene is $\eta \times \psi$. In order to link genotype to our cost-benefit function $\eta = \mathbf{C}/\mathbf{B}$, we begin by defining our benefit function $\mathbf{B}$.

Benefit.— Our benefit function $\mathbf{B}$ measures the functionality of the amino acid sequence $\vec{a}_i$ encoded by a set of codons $\vec{c}_i$, i.e. $a(\vec{c}_i) = \vec{a}_i$ relative to that of an optimal sequence $\vec{a}^*$. By definition, $\mathbf{B}(\vec{a}^*) = 1$ and $\mathbf{B}(\vec{a}_i|\vec{a}^*) < 1$ for all other sequences. We assume all amino acids within the sequence contribute to protein function and that this contribution declines as an inverse function of physicochemical distance between each amino acid and the optimal. Formally, we assume that

$$\mathbf{B}(\vec{a}_i|\vec{a}^*) = \left(\frac{1}{n}\sum_{p=1}^{n}\big(1 + G_p d(a_{i,p}, a_p^*)\right)\big)^{-1} \tag{1}$$

where n is the length of the protein, $d(a_{i,p}, a_p^*)$ is a weighted physicochemical distance between the amino acid encoded in gene i for position p and a_p^* is the optimal amino acid for that position of the protein. For simplicity, we define the distance between a stop codon and a sense codon as effectively infinite and, as a result, nonsense mutations are effectively lethal. The term G_p describes the sensitivity of the protein's function to physicochemical deviation from the optimimum at site position p. There are many possible measures for physiochemical distance; we use Grantham (1974) distances by default, though others may be chosen. We assume that $G_p \sim \text{Gamma}\,(\alpha = \alpha_G, \beta = \alpha_G)$ in order to ensure $\mathbb{E}(G_p) = 1$. Given the definition of the Gamma distribution, the variance in G_p is equal to $\alpha/\beta^2 = 1/\alpha_G$. Further, at the limit of $\alpha_G \to \infty$, the model becomes equivalent to assuming model uniform sensitivity site sensitivity where $G_p = 1$ for all positions p. Finally, we note that $\mathbf{B}(\vec{a}_i|\vec{a}^*)$ is inversely proportional to the average physicochemical deviation of an amino acid sequence $\vec{a}_i$ from the optimal sequence $\vec{a}^*$ weighted by each site's sensitivity to this deviation. $\mathbf{B}(\vec{a}_i|\vec{a}^*)$ can be generalized to include second and higher order terms of the distance measure d.

Cost.— Protein synthesis involves both direct and indirect assembly costs. Direct costs consist of the high energy phosphate bonds $\sim P$ of ATP or GTP's used to assemble the ribosome on the mRNA, charge tRNA's for elongation, move the ribosome forward along the transcript, and terminate protein synthesis. As a result, direct protein assembly costs are the same for all proteins of the same length. Indirect costs of protein assembly are potentially numerous and could include the cost of amino acid synthesis as well the cost and efficiency with which the protein assembly infrastructure such as ribosomes, aminoacyl-tRNA synthetases, tRNAs, and mRNAs are used. When these indirect costs are combined with sequence specific benefits, the probability of a mutant allele fixing is no longer independent of the rest of the sequence (Gilchrist et al. 2015) and, as a result, model fitting becomes substantially more complex. Thus for simplicity, in this study we ignore indirect costs of protein assembly that vary between genotypes and define,

$$\mathbf{C}(\vec{c_i}) = \text{Energetic cost of protein synthesis.} \tag{2}$$

$$= A_1 + A_2 n \tag{3}$$

where, A_1 and A_2 represent the direct cost, in high energy phosphate bonds, of ribosome initiation and peptide elongation, respectively, where $A_1 = A_2 = 4 \sim P$.

Defining Physicochemical Distances

Assuming that functionality declines with an amino acid a_i's physicochemical distance from the optimum amino acid a^* at each site provides a biologically defensible way of mapping genotype to protein function that requires relatively few free parameters. In addition, SelAC naturally lends itself to model selection since we can compare the quality of SelAC fits using different mixtures of physicochemical properties. Following Grantham (1974), we focus on using composition c, polarity p, and molecular volume v of each amino acid's side chain residue to define our distance function, but the model and its implementation can flexibly handle a variety of properties. We use the Euclidian distance between residue properties where each property c, p, and v has its own weighting term, α_c, α_p, α_v, respectively, which we refer to as 'Grantham weights'. Because physicochemical distance is ultimately weighted by a gene's specific average protein synthesis rate ψ, another parameter we estimate, there is a problem with parameter identifiablity. Ultimately, the scale of gene expression is affected by how we measure physicochemical distances which, in turn, is determined by our choice of Grantham weights. As a result, by default we set $\alpha_v = 3.990 \times 10^{-4}$, the value originally estimated by Grantham, and recognize that our our estimates of α_c and α_p and ψ are scaled relative to this choice for α_v. More specifically,

$$d(a_i, a^*) = \left(\alpha_c \left[c\left(a_i\right) - c\left(a^*\right)\right]^2 + \alpha_p \left[p\left(a_i\right) - p\left(a^*\right)\right]^2 + \\ \alpha_v \left[v\left(a_i\right) - v\left(a^*\right)\right]^2\right)^{1/2}.$$

Linking Protein Synthesis to Allele Substitution

Next we link the protein synthesis cost-benefit function η of an allele with its fixation probability. First, we assume that each protein encoded within a genome provides some beneficial function and that the organism needs that functionality to be produced at a target average rate ψ. By definition, the optimal amino acid sequence for a given gene, $\vec{a}^*$, produces one unit of functionality. Second, we assume that protein expression is regulated by the organism to ensure that functionality is produced at rate ψ. As a result, the realized average protein synthesis rate of a gene, ϕ, by definition, satisfies the equality $\phi = \psi/\mathbf{B}(\vec{a})$. In other words, the average production rate of a protein $\vec{a}$ with relative functionality $\mathbf{B}(\vec{a}) < 1$ must be $1/\mathbf{B}(\vec{a})$ times higher than the production rate needed if the optimal amino acid sequence $\vec{a}^*$ was encoded since, by definition, $\mathbf{B}(\vec{a}^*) = 1$. For example, a cell with an allele $\vec{a}$ where $\mathbf{B}(\vec{a}) = 0.9$ will have to produce $10/9 = 1.11\%$ times the proteins that a competitor cell with the optimal allele $\vec{a}^*$ would have to produce. Similarly, a cell with an allele $\vec{a}$ where $\mathbf{B}(\vec{a}) = 1/2$ will have to produce $2/1 = 2$ times the proteins that a cell with $\vec{a}^*$ would have to produce. Simply put, the fitness cost for a genotype encoding a suboptimal protein sequence stems from the need to produce suboptimal proteins at a higher rate in order to compensate for their lower functionality.

Third, we assume that every additional high energy phosphate bond, $\sim P$, spent per unit time to meet the organism's target function synthesis rate ψ leads to a slight and proportional decrease in fitness W. This assumption, in turn, implies

$$W_i\left(\vec{c}\right) \propto \exp\left[-A_0\,\eta(\vec{c}_i)\psi\right]. \tag{4}$$

where A_0, again, describes the decline in fitness with every $\sim P$ wasted per unit time. Because A_0 shares the same time units as ψ and ϕ and only occurs in SelAC in conjunction with ψ, we do not need to explicitly identify our time units.

Correspondingly, the ratio of fitness between two genotypes is,

$$\begin{aligned} W_i/W_j &= \exp\left[-A_0\,\eta(\vec{c}_i)\psi\right] / \exp\left[-A_0\,\eta(\vec{c}_j)\psi\right] && (5)\\ &= \exp\left[-A_0\left(\eta(\vec{c}_i) - \eta(\vec{c}_j)\right)\psi\right] && (6)\\ & && (7) \end{aligned}$$

Given our formulations of $\mathbf{C}$ and $\mathbf{B}$, the fitness effects between sites are multiplicative and, therefore, the substitution of an amino acid at one site can be modeled independently of the amino acids at the other sites within the coding sequence. As a result, the fitness ratio for two genotypes differing at a single site p simplifies to

$$\begin{aligned} \frac{W_i}{W_j} = \exp\Bigg[-&\frac{A_0\left(A_1 + A_2 n_g\right)}{n_g} \\ &\times \sum_{p\in\mathbb{P}} \left[d\left(a_{i,p}, a^*_p\right) - d\left(a_{j,p}, a^*_p\right)\right] G_p \psi\Bigg] \end{aligned}$$

where $\mathbb{P}$ represents the codon positions in which $\vec{c}_i$ and $\vec{c}_j$ differ. Fourth, we make a weak mutation assumption, such that alleles can differ at only one position at any given time, i.e. $|\mathbb{P}| = 1$, and that the population is evolving according to a Fisher-Wright process. As a result, the probability a new mutant, j, introduced via mutation into a resident population

i with effective size N_e will go to fixation is,

$$\begin{aligned} u_{i,j} &= \frac{1-(W_i/W_j)^b}{1-(W_i/W_j)^{2N_e}} \\ &= \frac{1-\exp\left\{-\frac{A_0}{n_g}\left(A_1+A_2n_g\right)\left[d\left(a_i,a^*\right)-d\left(a_j,a^*\right)\right]G_p\psi\, b\right\}}{1-\exp\left\{-\frac{A_0}{n_g}\left(A_1+A_2n_g\right)\left[d\left(a_i,a^*\right)-d\left(a_j,a^*\right)\right]G_p\psi\, 2N_e\right\}} \end{aligned}$$

where $b = 1$ for a diploid population and 2 for a haploid population (Kimura 1962; Wright 1969; Iwasa 1988; Berg and Lässig 2003; Sella and Hirsh 2005). Finally, assuming a constant mutation rate between alleles i and j, $\mu_{i,j}$, the substitution rate from allele i to j can be modeled as,

$$q_{i,j} = \frac{2}{b}\mu_{i,j}N_e u_{i,j}.$$

where, given our weak mutation assumption, $\mu_{i,j} = 0$ when two codons differ by more than one nucleotide. In the end, each optimal amino acid has a separate 64 x 64 substitution rate matrix $\mathbf{Q}_a$, which incorporates selection for the amino acid (and the fixation rate matrix this creates) as well as the common mutation parameters across optimal amino acids. This results in the creation of 20 $\mathbf{Q}$ matrices, one for each amino acid and each with $26,880$ unique rates, based on few parameters (one to 11 mutation rates, two free Grantham weights, the cost of protein assembly, A_1 and A_2, the gene specific target functionality synthesis rate ψ, and optimal amino acid at each position p, a^*_p), which can either be specified *a priori* or estimated from the data. SelAC can be generalized to allow transitions between optimal amino acids as well as between codons, which would result in a $(20 \times 64) \times (20 \times 64) = 1344 \times 1344$ matrix.

where does this 26,880 come from? Is this worth mentioning? If so why here

Given our assumption of independent evolution among sites, it follows that the probability of the whole data set is the product of the probabilities of observing the data at each individual site. Thus, the likelihood $\mathcal{L}$ of amino acid a being optimal at a given site position p is calculated as

$$\mathcal{L}\left(\mathbf{Q}_a|\mathbf{D}_p,\mathbf{T}\right) \propto \mathbf{P}\left(\mathbf{D}_p|\mathbf{Q}_a,\mathbf{T}\right) \tag{8}$$

In this case, the data, $\mathbf{D}_p$, are the observed codon states at position p for the tips of the phylogenetic tree with topology $\mathbf{T}$. For our purposes we take $\mathbf{T}$ as given but it could be estimated as well. The pruning algorithm of Felsenstein (1981) is used to calculate $\mathcal{L}\left(\mathbf{Q}_a|\mathbf{D}_p,\mathbf{T}\right)$. The log of the likelihood is maximized by estimating the genome scale parameters which consist of 11 mutation parameters which are implicitly scaled by $2N_e/b$, and two Grantham distance parameters, α_c and α_p, and the sensitivity distribution parameter α_G. Because A_0 and ψ_g always co-occur and are scaled by N_e, for each gene g we estimate a composite term $\psi'_g = \psi_g A_0 b N_e$ and the optimal amino acid for each position a^*_p of protein. When estimating α_G, the likelihood then becomes the average likelihood which we calculate using the generalized Laguerre quadrature with $k = 4$ points (Felsenstein 2001).

I thought we used 8 points.

Finally, we note that because we infer the ancestral state of the system, our approach does not rely on any assumptions of model stationary. Nevertheless, as our branch lengths grow the probability of observing a particular amino acid a at a given site approaches a stationary value proportional to $W(a)^{2N_e-b}$ (Sella and Hirsh 2005).

Implementation

All methods described above are implemented in the new R package, `selac` available through GitHub (`https://github.com/bomeara/selac`) [it will be uploaded to CRAN once peer review has completed]. Our package requires as input a set of fasta files that contain each coding sequence for a set of taxa, and the phylogeny depicting the hypothesized relationships among them. In addition to the SelAC models, we implemented the GY94 codon model of Goldman and Yang (1994), the FMutSel0 mutation-selection model of Yang and Nielsen (2008), and the standard general time-reversible nucleotide model that allows for Γ distributed rates across sites. These likelihood-based models represent a sample of the types of popular models often fit to codon data.

For the SelAC models, the starting guess for the optimal amino acid at a site comes from 'majority' rule, where the initial optimum is the most frequently observed amino acid at a given site (ties resolved randomly). Our optimization routine utilizes a four stage hill climbing approach. More specifically, within each stage a block of parameters are optimized while the remaining parameters are held constant. The first stage optimizes the block of branch length parameters. The second stage optimizes the block of gene specific composite parameters $\psi'_g = A_0 \psi_g N_e$. The third stage optimizes the model parameters shared across the genome α_c and α_p, and the sensitivity distribution parameter α_G. The fourth stage estimates the optimal amino acid at each site a^*. This entire four stage cycle is repeated six times, by which point the change in the log-$\mathcal{L}$ values for our analysis was less than XXX between cycles. For optimization of a given set of parameters, we rely on a bounded subplex routine (Rowan 1990) in the package `NLopt` (Johnson 2012) to maximize the log-likelihood function. To help the optimization navigate through local peaks, we perform a set of independent analyses with different sets of naive starting points with respect to the gene specific composite ψ' parameters, α_c, and α_p. Confidence in the parameter estimates can be generated by an 'adaptive search' procedure that we implemented to provide an estimate of the parameter space that is some pre-defined likelihood distance (e.g., 2 lnL units) from the maximum likelihood estimate (MLE), which follows Beaulieu and OMeara (2016); Edwards (1984).

Please replace XXX with appropriate numbers.

We note that our current implementation of SelAC is painfully slow, and is best suited for data sets with relatively few number of taxa (i.e. < 10). This is largely due to the size and quantity of matrices we create and manipulate to calculate the log-likelihood

Is this $<$ 10 accurate?

of an individual site. We have parallelized operations wherever possible, but the fact remains that, long term, this model may not be well-suited for R. Ongoing work will address the need for speed, with the eventual goal of implementing the model in popular phylogenetic inference toolkits, such as RevBayes (Hhna et al. 2016), PAML (Yang 2007) and RAxML (Stamatakis 2006).

Simulations

We evaluated the performance of our codon model by simulating datasets and estimating the bias of the inferred model parameters from these data. Our 'known' parameters under a given generating model were based on fitting SelAC to the 106 gene data set and phylogeny of Rokas et al. (2003). The tree used in these analyses is outdated with respect to the current hypothesis of relationships within *Saccharomyces*, but we rely on it simply as a training set that is separate from our empirical analyses (see). Bias in the model parameters were assessed under two generating models: one where we assumed a model of SelAC assuming uniform sensitivity across sites (i.e. $G_p = 1$ for all sites, i.e. $\alpha_G = \infty$), and one where we estimated the Gamma distribution parameter α_G from the data. Under each of these two scenarios, we used parameter estimates from the corresponding empirical analysis and simulated 50 five-gene data sets. For the gene specific composite parameter ψ'_g the 'known' values used for the simulation were five evenly spaced points along the rank order of the estimates across the 106 genes. The MLE estimate for a given replicate were taken as the fit with the highest log-likelihood after running five independent analyses with different sets of naive starting points with respect to the composite ψ'_g parameter, α_c, and α_p. All analyses were carried out in our `selac` R package.

Analysis of yeast genomes and tests of model adequacy

We focus our empirical analyses on the large yeast data set and phylogeny of Salichos and Rokas (2013). The yeast genome is an ideal system to examine our phylogenetic estimates of gene expression and its connection to real world measurements of these data within individual taxa. The complete data set of Salichos and Rokas (2013) contain 1070 orthologs, where we selected 100 at random for our analyses. We also focus our analyses only on *Saccharomyces sensu stricto*, including their sister taxon *Candida glabrata*, and we rely on the phylogeny depicted in Fig. 1 of Salichos and Rokas (2013) for our fixed tree. We fit the two SelAC models described above (i.e., SelAC and SelAC+Γ), as well as two codon models, GY94 and FMutSel0, and a standard GTR + Γ nucleotide model. The FMutSel0 model, which assumes that the amino acid frequencies are determined by functional requirements of the protein, but where XXXX, is the most similar to our model. In all cases, we assumed that the model was partitioned by gene, but with branch lengths linked across genes.

Can you fill in the XXX?

For SelAC, we compared our estimates of $\phi' = \psi'/\mathbf{B}$, which represents the average protein synthesis rate of a gene, to estimates of gene expression from empirical data. Specifically, we obtained gene expression data for five of the six species used - four species were measured during log-growth phase, whereas the other was measured at the beginning of the stationary phase (*S. kudriavzevii*) from the Gene Expression Omnibus (GEO). Gene expression in this context corresponds to mRNA abundances which were measured using either Microarray chips (*C. glabrata*, *S. castellii*, and *S. kudriavzevii*) or RNA-Seq (*S. paradoxus*, *S. mikatae*, and *S. cerevisiae*).

For further comparison, we also predicted protein synthesis rate (ϕ) by analyzing gene and genome-wide patterns of synonymous codon usage using ROC-SEMPPR (Gilchrist et al. 2015) for each individual genome. While, like SelAC, ROC-SEMPPR uses codon level information, it does not rely on any inter-specific comparisons and, unlike SelAC, uses only the intra- and inter-genic frequencies of synonymous codon usage as its data. Nevertheless, ROC-SEMPPR predictions of gene expression ϕ correlates strongly ($r = 0.53 - 0.74$) with a wide range of laboratory measurements of gene expression (Gilchrist et al. 2015).

While one of our main objectives was to determine the improvement of fit that SelAC has with respect to other standard phylogenetic models, we also evaluated the adequacy of SelAC. Model fit, measured with assessments such as the Akaike Information Criterion (AIC), can tell which model is least bad as an approximation for the data, but it does not reveal whether a model is actually doing a good job of representing the biological processes. An adequate model does the latter, one measure of which is that data generated under the model resemble real data (Goldman 1993). For example, Beaulieu et al. (2013) assessed whether parsimony scores and the size of monomorphic clades of empirical data were within the distributions of simulated under a new model and the best standard model; if the empirical summaries were outside the range for each, it would have suggested that neither model was adequately modeling this part of the biology.

For a given gene we first remove a particular taxon from the data set and the phylogeny. A marginal reconstruction of the likeliest sequence across all remaining nodes is conducted under the model, including where the attachment point of pruned taxon to the tree. The marginal probabilities of each site are used to sample and assemble the starting coding sequence. This sequence is then evolved along the branch, periodically being sampled and its current functionality assessed. We repeat this process 100 times and compare the distribution of trajectories against the observed functionality calculated for the gene. For comparison, we also conducted the same test, by simulating the sequence under the standard GTR + Γ nucleotide model, which is often used on these data but does not account for the fact that the sequence codes for a specific protein, and under FMutSel0, which includes selection on codons but in a fundamentally different way as our model.

The appropriate estimator of bias for AIC

As part of the model set described above, we also included a reduced form of each of the two SelAC models, SelAC and SelAC+Γ. Specifically, rather than optimizing the amino acid at any given site, we assume the the most frequently observed amino acid at each site is the optimal amino acid a^*. We refer to these 'majority rule' models as SelAC$_M$ and SelAC$_M + \Gamma$ and the majority rule parameterization greatly accelerates model fitting.

Since these majority rule models assume that the optimal amino acids are known prior to fitting of our model, it is tempting to reduce the number of parameters in the model by the number of total sites being analyzed. Despite having become standard behavior in the field of phylogenetics, this reduction is statistically inappropriate due to the fact that identification of the majority rule amino acid is made by examining the same data as we fit to our model. Because the difference in K when counting or not counting number of nucleotide sites drops out when comparing nucleotide models with AIC, this statistical issue does not apply to nucleotide models. It does, however, matter for AICc, where the number of parameters, K, and the sample size, n, combine in the penalty term. This also matters in our case, where the number of estimated parameters for the majority rule estimation differs based on whether one is looking at codons or single nucleotides.

In phylogenetics two variants of AICc are used. In comparative methods (e.g. Butler and King 2004; O'Meara et al. 2006; Beaulieu et al. 2013) the number of data points, n, is taken as the number of taxa. More taxa allow the fitting of more complex models, given more data. However, in DNA evolution, which is effectively the same as a discrete character model used in comparative methods, the n is taken as the number of sites. Obviously, both cannot be correct.

The original derivation of AICc by Hurvich and Tsai (1989) assumed a regression model, where the true model was in the set of examined models, as well as approximations in the derivation itself. The appropriatness of this approximation for phylogenetic data, where data points independence between taxa, is unclear. In any case, we argue that for phylogenetic data, a good estimate of data set size is number of taxa multiplied by number of sites. First of all, this is what is conventionally seen as the size of the dataset in the field. Second, when considering how likelihood is calculated, the likelihood for a given site is the sum of the probabilities of each observed state at each tip, and this is then multiplied across sites. It is arguable that the conventional approach in comparative methods is calculating AICc in this way: number of taxa multiplied by number of sites equals the number of taxa, if only one site is examined, as remains remarkably common in comparative methods. (One notable exception to this appoach to calculating AICc is the program SURFACE implemented by Ingram and Mahler (2013), which uses multiple characters and taxa. While its default is to use AIC to compare models, if one chooses to use AICc, the number of samples is taken as the product of number of sites and number of taxa.)

Recently, Jhwueng et al. (2014) performed an analysis that investigated what variant of AIC and AICc worked best as an estimator, but the results were inconclusive. Here, we have adopted and extended the simulation approach of Jhwueng et al. (2014) in order to examine a large set of different penalty functions and how well they approximate the remaining portion of Kullback-Liebler (KL) divergence between two models after accounting for the deviance (i.e., $-2\mathcal{L}$) (see Appendix 1 for more details).

Results

By linking transition rates $q_{i,j}$ to gene expression ψ, our approach allows use of the same model for genes under varying degrees of stabilizing selection. Specifically, we assume the strength of stabilizing selection for the optimal sequence, $\vec{a}^*$, is proportional to the average protein synthesis rate ϕ, which we can estimate for each gene. In regards to model fit, our results clearly indicated that linking the strength of stabilizing selection for the optimal sequence to gene expression substantially improves our model fit. Further, including the shape parameter α_G for the random effects term $G \sim \text{Gamma}(\alpha_G, \beta_g)$ to allow for heterogeneity in this selection between sites within a gene improves the ΔAICc of SelAC+Γ over the simpler SelAC models by over 22,000 AIC units. Using either ΔAICc or AIC_w as our measure of model support, the SelAC models fit extraordinarily better than GTR + Γ, GY94, or FMutSel0 (Table 1). This is in spite of the need for estimating the optimal amino acid at each position in each protein, which accounts for 49,881 additional model parameters. Even when compared to the next most parameter rich codon model in our model set, FMutSel0, SelAC+Γ model shows nearly 180,000 AIC unit improvement over FMutSel0.

With respect to estimates of ϕ within SelAC, they were strongly correlated with both our empirical (i.e. mRNA abundances) and model based (i.e. ROC-SEMPPR) measurements of gene expression (Figure 1 and Figures S1-S2, respectively). In other words, using only codon sequences, our model can predict which genes have high or low expression levels. The estimate of the α_G parameter, which describes the site-specific variation in sensitivity of the protein's functionality, indicated a moderate level of variation in gene expression among sites. Our estimate of $\alpha_G = 1.40$, produced a distribution of sensitivity terms G ranged from 0.344-7.16, but with nearly 90% of the weight for a given site-likelihood being contributed by the 0.344 and 1.48 rate categories. In simulation, however, of all the parameters in the model, only α_G showed a consistent bias, in that the MLE were generally lower than their actual values (see Supporting Materials). Other parameters in the model, such as the Grantham weights, provide an indication as to the physicochemical distance between amino acids. Our estimates of these weights only strongly deviate from Grantham's 1974 original estimates in regards to composition weight, α_c, which is the ratio of noncarbon elements in the end groups to the number of side chains. Our estimate of the composition weighting factor of α_c=0.484 is 1/4th the value estimate by Grantham which suggests that the substitution process is less sensitive to this physicochemical property when shared ancestry and variation in stabilizing selection are taken into account.

It is important to note that the nonsynonymous/synonymous mutation ratio, or ω, which we estimated for each gene under the FMutSel0 model strongly correlated with our estimates of $\phi' = \psi'/\mathbf{B}$ where $\mathbf{B}$ depends on the sequence of each taxa. In fact, ω showed similar, though slightly reduced correlations, with the same empirical estimates of gene expression described above (Figure 2). This would give the impression that the same conclusions could have been gleaned using a much simpler model, both in terms of the number of parameters and the assumptions made. However, as we discussed earlier, not only is this model greatly restricted in terms of its biological feasibility, SelAC clearly performs better in terms of its fit to the data and biological realism.

For example, when we simulated the sequence for *S. cervisieae*, starting from the ancestral sequence under both GTR + Γ and FMutSel0, the functionality of the simulated sequence moves away from the observed sequence, whereas SelAC remains near the functionality of the observed sequence (Figure 3b). In a way, this is somewhat unsurprising, given that both GTR + Γ and FMutSel0 are agnostic to the functionality of the gene, but it does highlight the improvement in biological realism in amino acid sequence evolution that SelAC provides. We do note that the adequacy of the SelAC model does vary among individual taxa, and does not always match the observed functionality. For instance, *S. castellii* is simulated with consistently higher functionality than observed (Figure 3c). We suspect this is an indication that assuming a single set of optimal amino acid across all taxa may be too simplistic, but we cannot also rule out other potential simplifying assumptions in our model, such as a single set of Grantham weights and α_G values or the simple, inverse relationship between physicochemical distance d and benefit $\mathbf{B}$.

Finally, we note that our simulation analysis suggested that the best measure of dataset size for AICc uses a scaled value of the product of number of sites and number of characters was the best at estimating KL distance. The model comparison approach described above included this assumption. For more details on the simulation approach, see Appendix 1.

Discussion

The work presented here contributes to the field of phylogenetics and molecular evolution in a number of ways. First, SelAC provides an complementary example to Thorne et al. (2012) studies of how models of molecular and evolutionary scales can be combined together in a nested manner. While the mapping between genotype and phenotype is more abstract than Thorne et al. (2012), SelAC has the advantage of not requiring knowledge of a protein's native folding. Second, our use of model nesting also allows us to formulate and test specific biological hypotheses. For example, we are able to compare a model formulation which assumes that physiochemical deviations from the optimal sequence are equally disruptive at all sites within a protein to one which assumes the effect of deviation from the optimal amino acid's physicochemical properties on protein function varies between sites. By linking the strength of stabilizing selection for an optimal amino acid sequence to gene expression, we can weight the historical information encoded in genes evolving at vastly different rates in a biologically plausible manner while simultaneously estimating their expression levels. Further, because our fitness functions are well defined, we can provide estimates of key evolutionary statistics such as the distribution of the effects of amino acid substitutions on fitness and genetic load. Finally, because our model is based on a mechanistic description of a sequence's cost-benefit function $\mathbf{C}/\mathbf{B}$, relaxing any given biological assumption is relatively straightforward.

As phylogenetic methods become ever more ubiquitous in biology, and data set size and complexity increase, there is a need and an opportunity for more complex and realistic models (Goldman et al. 1996; Thorne et al. 1996; Goldman et al. 1998; Halpern and Bruno 1998; Lartillot and Philippe 2004). Despite their widespread use, phylogenetic models based on purifying and diversifying selection, i.e. Goldman and Yang (1994) and extensions, are very narrow categories of selection that mostly apply to cases of positive and negative frequency dependent selection at the level of a particular amino acid, not for tree inference itself.

Instead of heuristically extending population genetic models of neutral evolution for use in phylogenetics, it makes sense to derive these extensions from population genetic models that *explicitly* include the fundamental forces of mutation, drift, and natural selection. Starting with Halpern and Bruno (1998), a number of researchers have developed methods for linking site-specific selection on protein sequence and phylogenetics(e.g. Koshi et al. 1999; Dimmic et al. 2000; Koshi and Goldstein 2000; Robinson et al. 2003; Lartillot and Philippe 2004; Thorne et al. 2012; Rodrigue and Lartillot 2014). Our work follows this tradition, but includes some key advances. For instance, even though SelAC requires a large number of substitution matrices, because of our assumption about protein functionality and physicochemical distance from the optimum, we are able to parameterize these matrices using a relatively small number of genome-wide parameters and one gene specific parameter. We show that all of these parameters can be estimated simultaneously with branch lengths from the data at the tips of the tree.

By assuming fitness declines with extraneous energy flux, SelAC explicitly links the variation in the strength of stabilizing selection for the optimal protein sequence among genes, to the variation among genes in their target expression levels ψ. Furthermore, by linking expression and selection, SelAC provides a natural framework for combining information from protein coding genes with very different rates of evolution with the low expression genes providing information on shallow branches and the high expression genes providing information on deep branches. This is in contrast to more traditional approach of concatenating gene sequences together, which is equivalent to assuming the same average protein synthesis rate ψ for all of the genes, or more recent approaches where different models are fitted to different genes. Our results indicate that including a gene specific ψ value vastly improves SelAC fits (Table 1). Perhaps more convincingly, we find that the target expression level ψ and realized protein synthesis rate ϕ are reasonably well correlated with laboratory measurements of gene expression ($r = 0.34 - 0.65$; Figures 1, S1, and S2). The idea that quantitative information on gene expression is embedded within intra-genomic patterns of synonymous codon usage is well accepted; our work shows that this information can also be extracted from comparative data at the amino acid level.

Of course, given the general nature of SelAC and the complexity of biological systems, other biological forces besides selection for reducing energy flux likely contribute intergenic variation in the magnitude of stabilizing selection. Similarly, other physicochemical properties besides composition, volume, and charge likely contribute to site specific patterns of amino acid substitution. Thus, a larger and more informative set of Grantham weights might improve our model fit and reduce the noise in our estimates of ϕ. Even if other physicochemical properties are considered, the idea of a consistent, genome wide Grantham weighting of these terms seems highly unlikely. Since the importance of an amino acid's physicochemical properties likely changes with where it lies in a folded protein, one way to incorporate such effects is to test whether the data supports multiple sets of Grantham weights for either subsets of genes or regions within genes, rather than a single set.

Both of these points highlight the advantage of the detailed, mechanistic modeling approach underlying SelAC. Because there is a clear link between protein expression, synthesis cost, and functionality, SelAC can be extended by increasing the realism of the mapping between these terms and the coding sequences being analyzed. For example, SelAC currently assumes the optimal amino acid for any site is fixed along all branches.

This assumption can be relaxed by allowing the optimal amino acid to change during the course of evolution along a branch.

From a computational standpoint, the additive nature of selection between sites is desirable because it allows us to analyze sites within a gene largely independently of each other. From a biological standpoint, this additivity between site ignores any non-linear interactions between sites, such as epistasis, or between alleles, such as dominance. Thus, our work can be considered a first step to modeling to these more complex scenarios. For example, our current implementation ignores any selection on synonymous codon usage bias (CUB) (Yang and Nielsen 2008; Pouyet et al. 2016, c.f.). Including such selection is tricky because introducing the site specific cost effects of CUB, which is consistent with the hypothesis that codon usage affects the efficiency of protein assembly or **C**, into a model where amino acids affect protein funcdtion or **B**, results in a cost-benefit ratio **C**/**B** with epistatic interactions between all sites. These epistatic effects can likely be ignored under certain conditions or reasonably approximated based on an expectation of codon specific costs (e.g. Kubatko et al. 2016). Nevertheless, it is difficult to see how one could identify such conditions without modeling the way in which codon and amino acid usage affects **C**/**B**.

This work also points out the potential importance of further investigation into model choice in phylogenetics. For likelihood models, use of AICc has become standard. However, how one determines the appropriate number of parameters estimated in a model is more complicated than generally recognized. Common sense suggests that dataset size is increased by adding taxa and/or sites. In other words, a dataset of 1000 taxa and 100 sites must have more information on substitution models than a dataset of 4 taxa and 100 sites. Our simple analyses agree that the number of observations in a dataset (number of sites × number of taxa) should be taken as the sample size for AICc, but this conclusion likely only applies when there is sufficient independence between taxa. For instance, one could imagine a phylogeny where one taxon is sister to a polytomy of 99 taxa that have zero length terminal branches. Absent measurement error or other intraspecific variation, one would have 100 species but only two unique trait values, and the only information about the process of evolution comes from one happens on the path connecting the lone taxon to the polytomy. Although this is a rather extreme example, it seems prudent for researchers to use a simulation based approach similar to the one we take here to determine the appropriate means for calculating the effective number of data points in their data.

There are still significant deficiencies in the approach outlined here. Most worrisome are biological flaws in the model. For example, at its heart, the model assumes that suboptimal proteins can be compensated for, at a cost, simply by producing more of them. However, this is likely only true for proteins reasonably close to the optimal sequence. Different enough proteins will fail to function entirely: the active site will not sufficiently match its substrates, a protein will not properly pass through a membrane, and so forth. Yet, in our model, even random sequences still permit survival, just requiring more protein production. Other oversimplifications include the assumption of no selection on codon usage, no change of optimal amino acids through time, and no change of the effect of physiochemical properties on fitness through time. However, because we take a mechanistic approach, all of these assumptions can be relaxed through further extention of our model.

There are also deficiencies in our implementation. Though reasonable to use for a given topology with a modest number of species, it is too slow for practical use for tree search. It thus serves as a proof of concept, or of utility for targeted questions where a more realistic model may be of use (placement of particular taxa, for example). Future work will encode SelAC models into a variety of mature, popular tree-search programs. SelAC also represents a hard optimization problem: the nested models reduce parameter complexity vastly, but there are still numerous parameters to optimize, including the discrete parameter of optimal amino acid at each site. A different implementation, more parameter-rich, would optimize values of three (or more) physiochemical properties per site. This would have the practical advantage of continuous parameter optimization rather than discrete, and biologically would be more realistic (as it is the properties that selection ”sees”, not the identity of the amino acid itself).

Overall, SelAC represents an important step in uniting phylogenetic and population genetic models. It allows biologically relevant population genetic parameters to be estimated from phylogenetic information, while also dramatically improving fit and accuracy of phylogenetic models. Moreover, it demonstrates that there remains substantially more information in the coding sequences used for phylogenetic analysis than other methods can access. Given the enormous amount of efforts expended to generate sequence datasets, it makes sense for researchers to continue developing more realistic models of sequence evolution in order to extract the biological information embedded in these datasets. The cost-benefit model we develop here is just one of many possible paths of mechanistic model development.

Acknowledgements

This work was supported in part by NSF Awards MCB-1120370 (MAG and RZ) and DEB-1355033 (BCO, MAG, and RZ) with additional support from The University of Tennessee Knoxville and University of Arkansas (JMB). JMB was supported, in part, as a Postdoctoral Fellow at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation through NSF Award DBI-1300426, with additional support from UTK. The authors would like to thank Premal Shah, Todd Oakley, and our two anonymous for their helpful criticisms and suggestions for this work.

*

References

Beaulieu, J. M., B. C. O'Meara, and M. J. Donoghue. 2013. Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms. Systematic Biology 62:725–737.

Beaulieu, J. M. and B. C. OMeara. 2016. Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction. Systematic Biology 65:583–601.

Berg, J. and M. Lässig. 2003. Stochastic Evolution and Transcription Factor Binding Sites. Biophysics 48:S36–S44.

Butler, M. A. and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. American Naturalist 164:683–695.

Dimmic, M. W., D. P. Mindell, and R. A. Goldstein. 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. Pacific Symposium on Biocomputing 5:18–29.

Edwards, A. 1984. Likelihood. Cambridge science classics Cambridge University Press.

Felsenstein, J. 1981. Evolutionary trees from DNA-sequences - a maximum-likelihood approach. Journal of Molecular Evolution 17:368–376.

Felsenstein, J. 2001. Taking Variation of Evolutionary Rates Between Sites into Account in Inferring Phylogenies. Journal of Molecular Evolution 53:447–455.

Fisher, S., Ronald A. 1930. The Genetical Theory of Natural Selection. Oxford University Press, Oxford.

Gilchrist, M., P. Shah, and R. Zaretzki. 2009. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. Genetics 183:1493–1505.

Gilchrist, M. A. 2007. Combining Models of Protein Translation and Population Genetics to Predict Protein Production Rates from Codon Usage Patterns. Molecular Biology and Evolution 24:2362–2373.

Gilchrist, M. A., W.-C. Chen, P. Shah, C. L. Landerer, and R. Zaretzki. 2015. Estimating Gene Expression and Codon-Specific Translational Efficiencies, Mutation Biases, and Selection Coefficients from Genomic Data Alone. Genome Biology and Evolution 7:1559–1579.

Goldman, N. 1993. Statistical tests of models of DNA substitution. Journal of molecular evolution 36:182–198.

Goldman, N., J. L. Thorne, and D. T. Jones. 1996. Using Evolutionary Trees in Protein Secondary Structure Prediction and Other Comparative Sequence Analyses. Journal of Molecular Biology 263:196 – 208.

Goldman, N., J. L. Thorne, and D. T. Jones. 1998. Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution. Genetics 149:445–458.

Goldman, N. and Z. H. Yang. 1994. Codon-based model of nucleotide substitution for protein-coding DNA-sequences. Molecular Biology and Evolution 11:725–736.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862–864.

Halpern, A. L. and W. J. Bruno. 1998. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. Molecular Biology And Evolution 15:910–917.

Hughes, A. L. and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class-i loci reveals overdominant selection. Nature 335:167–170.

Hurvich, C. M. and C.-L. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76:297–307.

Hhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. Systematic Biology 65:726.

Ingram, T. and D. L. Mahler. 2013. SURFACE: detecting convergent evolution from data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. Methods in ecology and evolution 4:416–425.

Iwasa, Y. 1988. Free fitness that always increases in evolution. Journal of Theoretical Biology 135:265–281.

Jhwueng, D.-C., H. Snehalata, B. C. O'Meara, and L. Liu. 2014. Investigating the performance of AIC in selecting phylogenetic models. Statistical applications in genetics and moleculr biology 13:459–475.

Johnson, S. G. 2012. The NLopt nonlinear-optimization package. Version 2.4.2 – Released 20 May 2014.

Kimura, M. 1962. on the probability of fixation of mutant genes in a population. Genetics 47:713–719.

Koshi, J. M. and R. A. Goldstein. 1997. Mutation matrices and physical-chemical properties: Correlations and implications. Proteins-Structure Function And Genetics 27:336–344.

Koshi, J. M. and R. A. Goldstein. 2000. Analyzing site heterogeneity during protein evolution. Pages 191–202 *in* Biocomputing 2001. World Scientific.

Koshi, J. M., D. P. Mindell, and R. A. Goldstein. 1999. Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes. Molecular biology and evolution 16:173–179.

Kubatko, L., P. Shah, R. Herbei, and M. A. Gilchrist. 2016. A codon model of nucleotide substitution with selection on synonymous codon usage. Molecular Phylogenetics and Evolution 94:290 – 297.

Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology And Evolution 21:1095–1109.

Mayrose, I., N. Friedman, and T. Pupko. 2005. A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21:ii151–ii158.

McCandlish, D. M. and A. Stoltzfus. 2014. Modeling evolution using the probability of fixation: History and implications. The Quarterly Review of Biology 89:225–252.

Muse, S. V. and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Molecular Biology and Evolution 11:715–724.

Nowak, M. A. 2006. Evolutionary Dynamics: Exploring the Equations of Life. Belknap of Harvard University Press, Cambridge, MA.

O'Meara, B. C., C. Ane, M. J. Sanderson, and W. P.C. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Pouyet, F., M. Bailly-Bechet, D. Mouchiroud, and L. Guguen. 2016. SENCA: A Multilayered Codon Model to Study the Origins and Dynamics of Codon Usage. Genome Biology and Evolution 8:2427–2441.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Molecular Biology And Evolution 20:1692–1704.

Rodrigue, N. and N. Lartillot. 2014. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. Bioinformatics 30:1020–1021.

Rodrigue, N., N. Lartillot, D. Bryant, and H. Philippe. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347:207–217.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Rowan, T. 1990. Functional Stability Analysis of Numerical Algorithms. Ph.D. thesis University of Texas, Austin.

Salichos, L. and A. Rokas. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Sella, G. and A. E. Hirsh. 2005. The application of statistical physics to evolutionary biology. Proceedings of the National Academy of Sciences of the United States of America 102:9541–9546.

Shah, P. and M. A. Gilchrist. 2011. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proceedings of the National Academy of Sciences of the United States of America 108:10231–10236.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Thorne, J. L., N. Goldman, and D. T. Jones. 1996. Combining protein evolution and secondary structure. Molecular Biology and Evolution 13:666–673.

Thorne, J. L., N. Lartillot, N. Rodrigue, and S. C. Choi. 2012. Codon models as a vehicle for reconciling population genetics with inter-specific sequence data. Codon Evolution: Mechanisms And Models Pages 97–110 D2 10.1093/acprof:osobl/9780199601165.001.0001 ER.

Wright, S. 1969. Evolution and the genetics of populations. Vol. 2. The theory of gene frequencies. vol. 2. University of Chicago Press.

Yang, Z. 2014. Molecular Evolution: A Statistical Approach. Oxford University Press, New York.

Yang, Z. H. 1994. Maximum-likelihood phylogenetic estimation from DNA-sequences with variable rates over sites - approximate methods. Journal Of Molecular Evolution 39:306–314.

Yang, Z. H. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology And Evolution 24:1586–1591.

Yang, Z. H. and R. Nielsen. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Molecular Biology and Evolution 25:568–579.

Table

		Parameters				Model
Model	logLik	Estimated	AIC	AICc	ΔAICc	Weight
GTR+Γ	-655166.4	610	1,311,553	1,311,554	287,415	<0.001
GY94	-612121.5	210	1,224,663	1,224,663	200,524	<0.001
FMutSel0	-598848.2	2810	1,203,316	1,203,362	179,223	<0.001
SelAC$_M$	-478282.7	50,004	1,056,573	1,073,290	49,151	<0.001
SelAC	-465616.7	50,004	1,031,241	1,047,958	23,819	<0.001
SelAC$_M + \Gamma$	-465089.7	50,005	1,030,189	1,046,906	22,767	<0.001
SelAC+Γ	-453706.0	50,005	1,007,422	1,024,139	0	>0.999

Table 1: Comparison of model fits using AIC, AICc, and AIC$_\mathrm{w}$. Note the subscripts M indicate model fits where the most common or 'majority rule' amino acid was fixed as the optimal amino acid a^* for each site. As discussed in text, despite the fact that a^* for each site was not fitted by our algorithm, its value was determined by examining the data and, as a result, represent an additional parameter estimated from the data and are accounted for in our table.

Figures

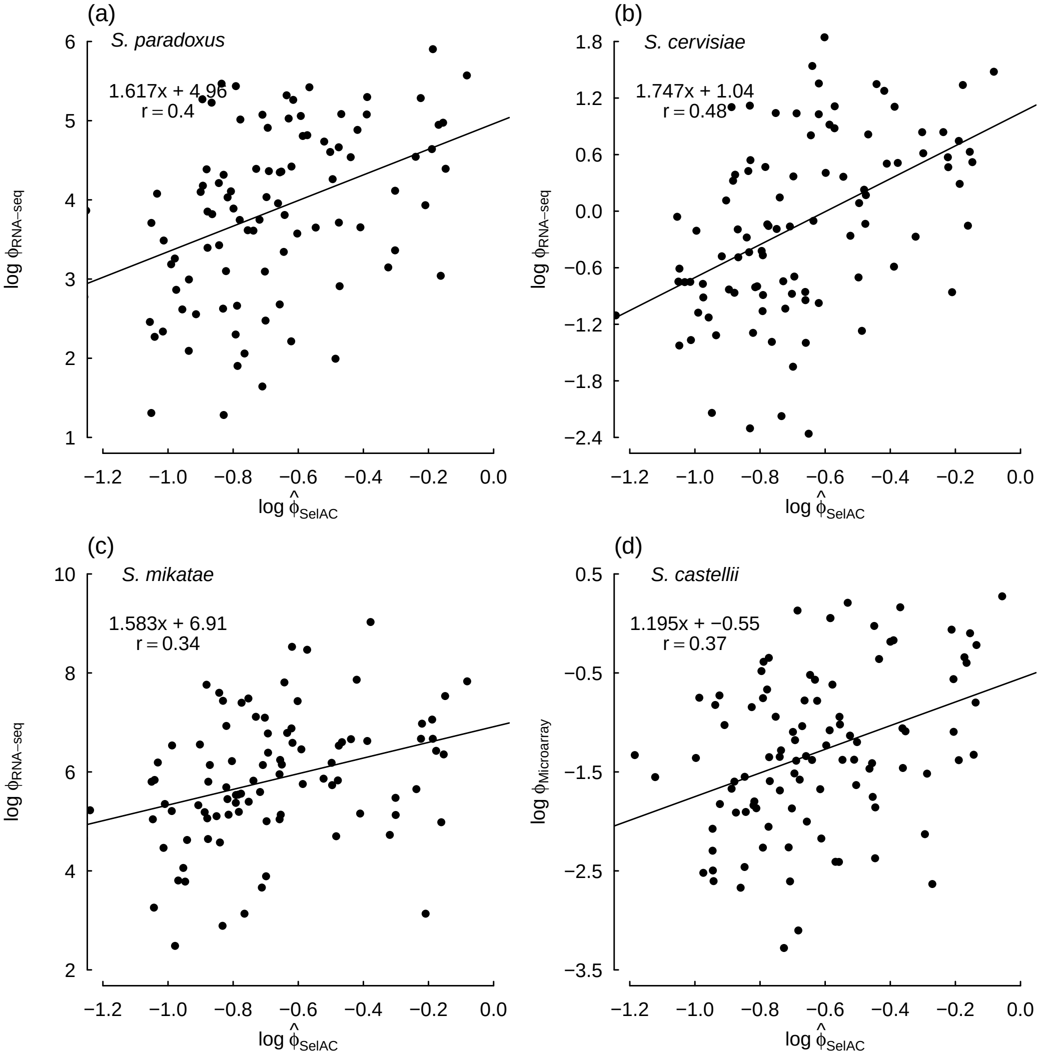

Figure 1: Comparisons between estimates of ϕ obtained from SelAC+Γ and direct measurements of expression for individual yeast taxa across the 100 selected genes from Salichos and Rokas (2013). Estimates of ϕ were obtained by solving for ψ based on estimates of ψ', and then dividing by $\mathbf{B}(\vec{a}_i|\vec{a}^*)$. Gene expression was measured using either RNA-Seq (a-c) or Microarray chips (d), and the equations in the upper left hand corner of each panel represent the regression fit and correlation coefficient r.

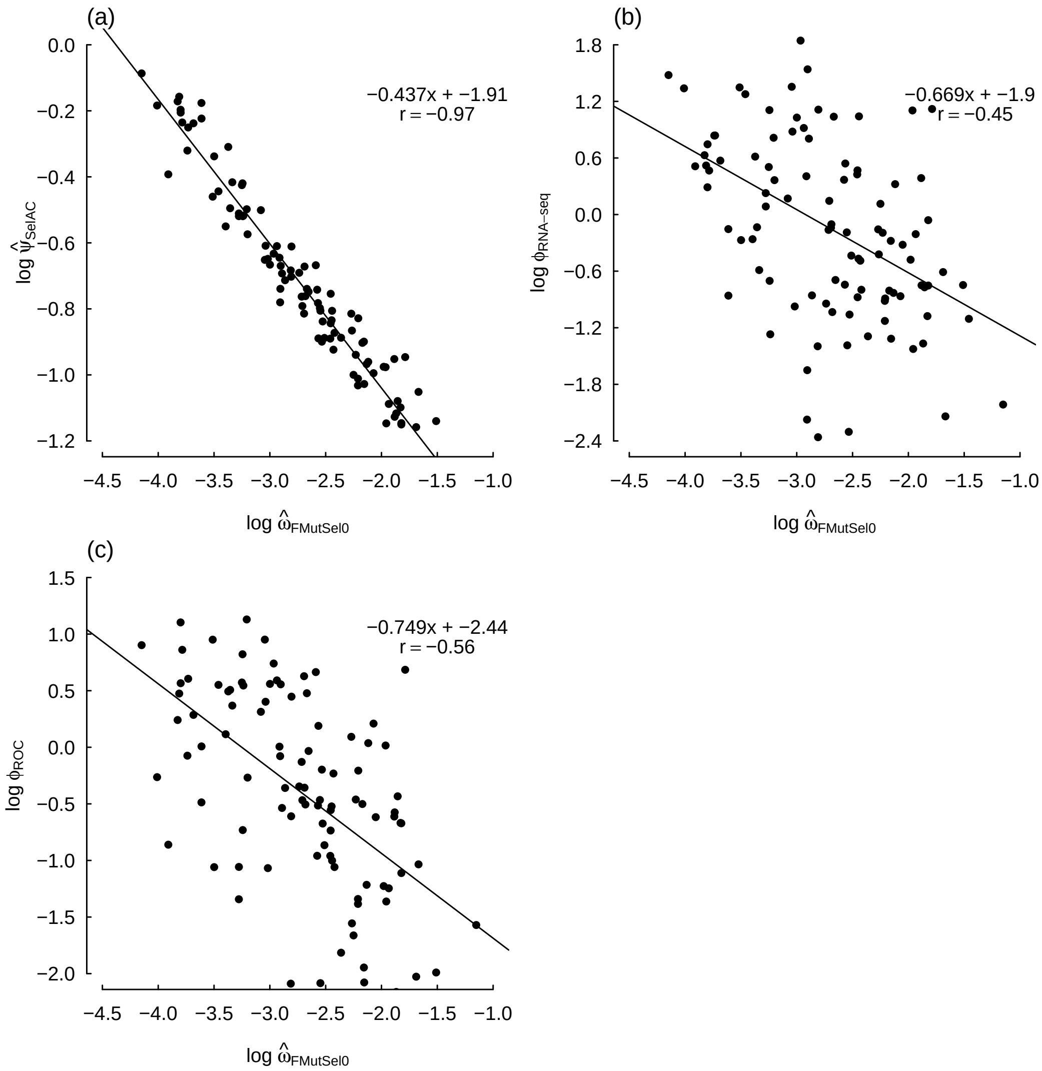

Figure 2: Comparisons between ω, which is the nonsynonymous/synonymous mutation ratio in FMutSel0, ψ obtained from SelAC+Γ (a), a direct measurement of expression (b), and a model-based prediction of gene expression that does not account for ancestry (c), for *S. cerevisiae* across the 100 selected genes from Salichos and Rokas (2013). As in Figure 1, the equations in the upper left hand corner of each panel provide the regression fit and correlation coefficient. Estimates of ψ were solved from estimates of ψ'.

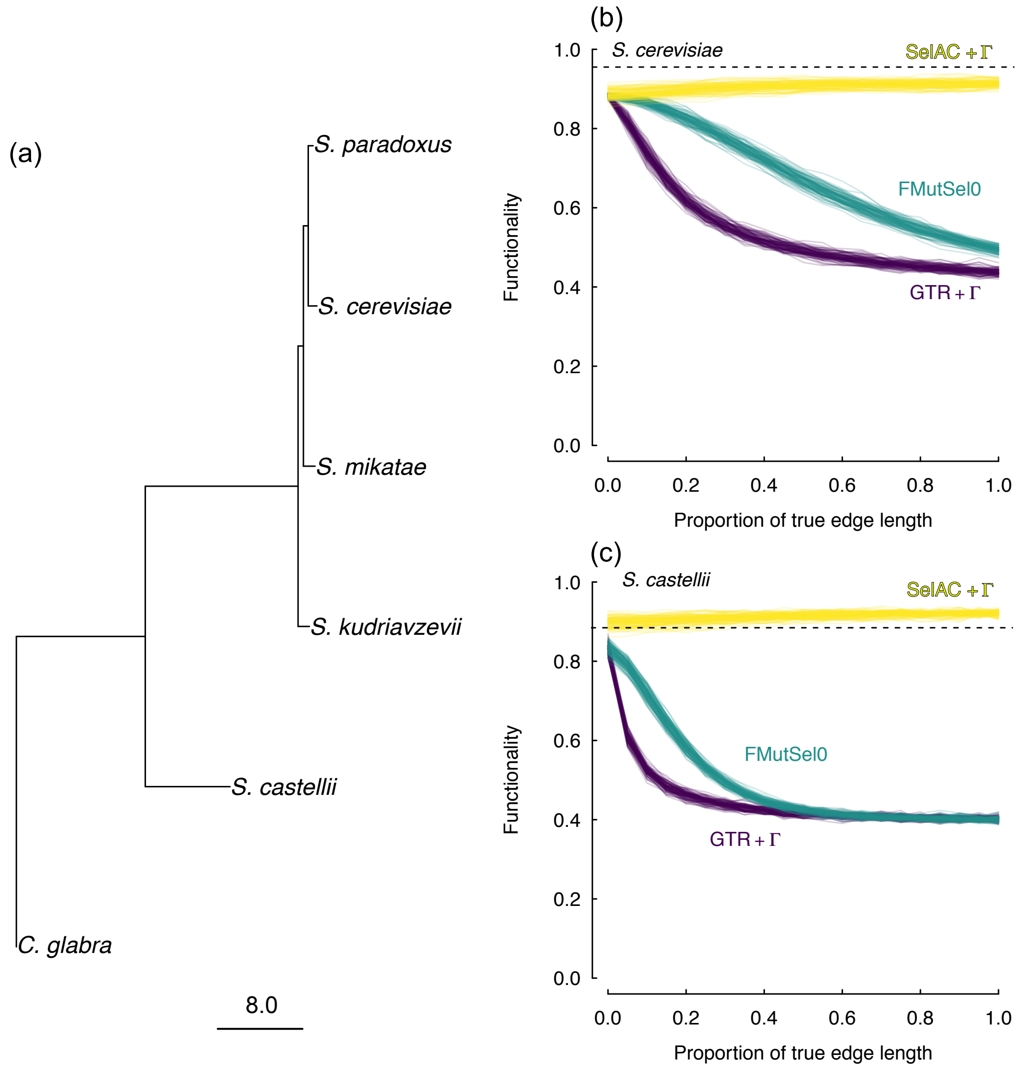

Figure 3: (a) Maximum likelihood estimates of branch lengths under SelAC+Γ for 100 selected genes from Salichos and Rokas (2013). Tests of model adequacy for *S. cerevisiae* (b) and *S. castellii* (c) indicated that, when these taxa are removed from the tree, and their sequences are simulated, the parameters of SelAC+Γ exhibit functionality that is far closer to the observed (dashed black line) than data sets produced from parameters of either FMutSel0 or GTR + Γ.

Part I

Supporting Materials

Supporting Materials

Comparisons of SelAC gene expression estimates with empirical measurements

In our model, the parameter ϕ measures the realized average protein synthesis rate of a gene. We compared our estimates of ϕ to two separate measures of gene expression, one empirical (Figure S1), and one model-based prediction that does not account for shared ancestry, for individual yeast taxa across the same set of genes. Our estimates of ϕ are positively correlated both measures, which are also strongly correlated with each other (Figure 1 - S2) On the whole, these comparisons indicate not only a high degree of consistency among all three measures, but also, importantly, that estimates of ϕ obtained from SelAC provide real biological insight into the expression level of a gene.

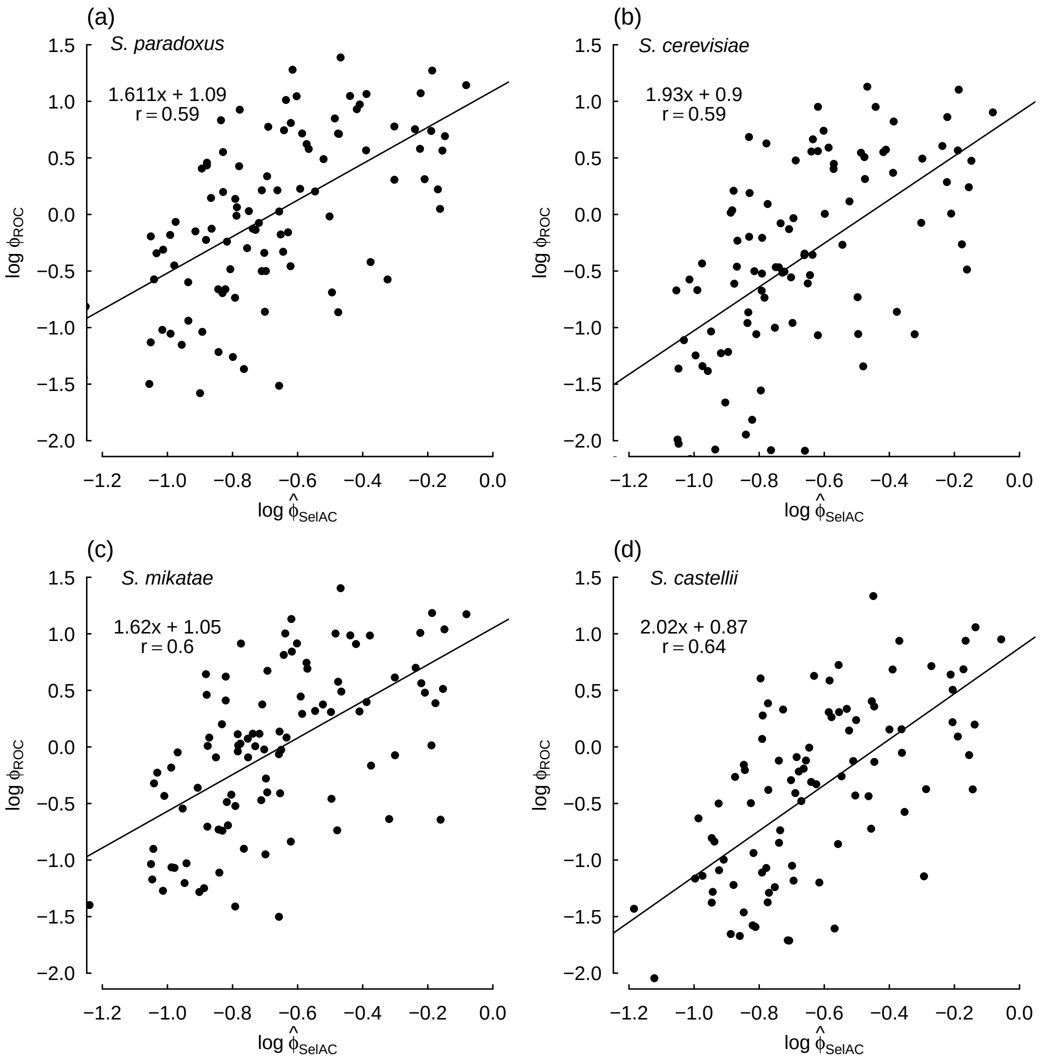

Figure S1: Comparisons between estimates of ϕ obtained from SelAC+Γ and the predicted gene expression from the ROC SEMPER model (Gilchrist et al. (2015)) for individual yeast taxa across the 100 selected genes from Salichos and Rokas (2013). As with figures in the main text, estimates of ϕ were obtained by solving for ψ based on estimates of ψ', and then dividing by $\mathbf{B}(\vec{a}_i|\vec{a}^*)$. The equations in the upper left hand corner of each panel represent the regression fit and correlation coefficient.

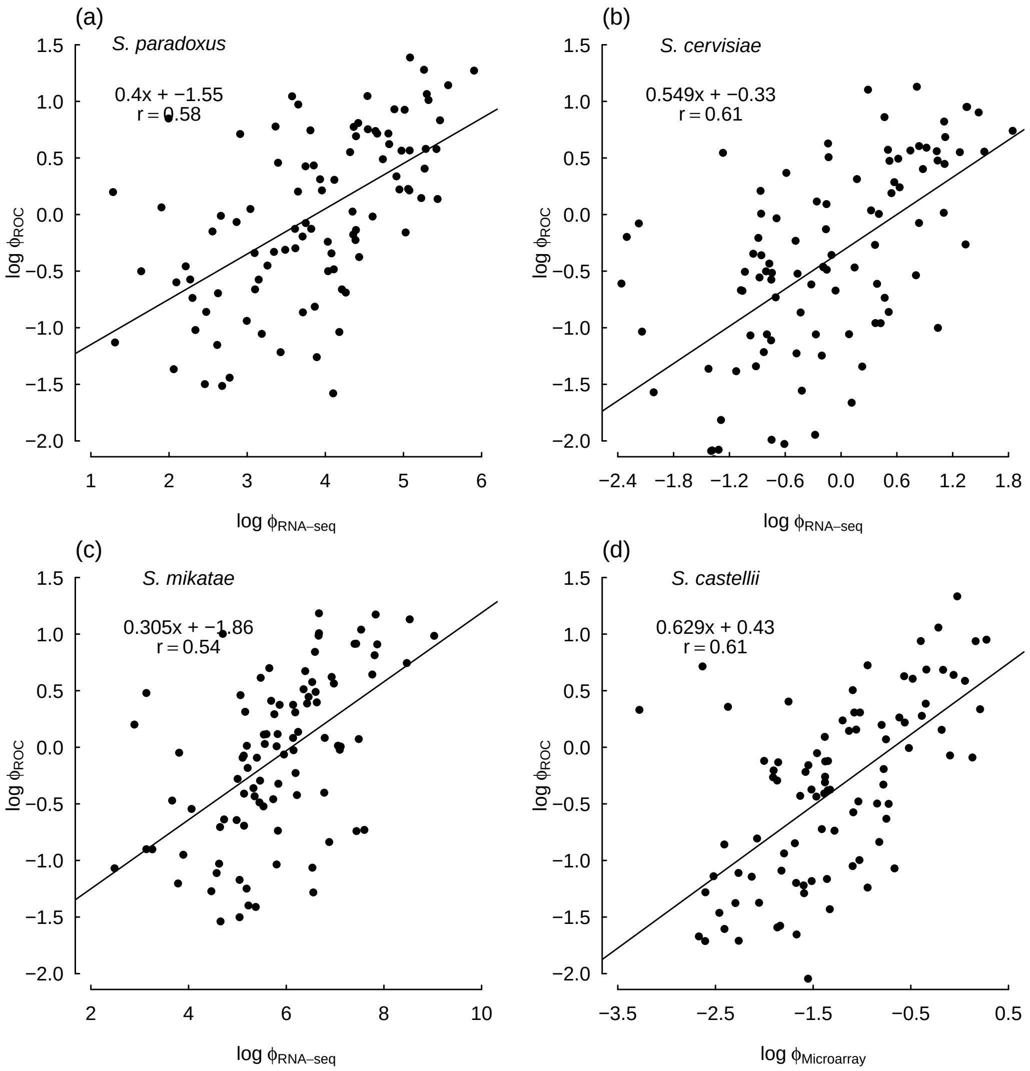

Figure S2: Comparisons of predicted gene expression from the ROC SEMPER model (Gilchrist et al. (2015)) and direct measurements of expression from RNA-Seq or Microarray data for individual yeast taxa across the 100 selected genes from Salichos and Rokas (2013). The equations in the upper left hand corner of each panel represent the regression fit and correlation coefficient.

Simulations

Overall, the simulation results indicate that SelAC model can reasonably recover the known values of the generating model (Figure S3 - S6). This includes not only the parameters in the model, but also the optimal amino acids for a given sequence as well as the estimates of the branch lengths. There are a few observations to note. First, the ability to accurately recover the true optimal amino acid sequence will largely depend on the magnitude of ϕ. This is, of course, intuitive, given that ϕ sets the strength of stabilizing selection towards an optimal amino acid at a site. However, the inclusion of α_G into the model, appears to generally increase values of ϕ and generally improves the ability to recover the optimal amino acids even for the gene with the lowest baseline ϕ. Second, we found a strong downward bias in estimates of α_G, which actually translates to greater variation among the rate categories. The choice of a gamma distribution to represent site-specific variation in sensitivity was based on mathematical convenience and convention, rather than on biological reality. Nevertheless, we suspect that this bias is in large part due to the difficulty in determining the baseline ψ for a given gene and the value of α_G that globally satisfies the site-specific variation in sensitivity across all genes, as indicated by the slight upward bias in estimates of ψ. A reviewer pointed out that it may also be difficulty for the model to account for changing amino-acid, which we agree may also play a role. It has been suggested, in studies of the behavior of the gamma distribution in applications of nucleotide substitution model, that increasing the number of rate categories can often improve accuracy of the shape parameter (Mayrose et al. (2005)). Future work will address this issue.

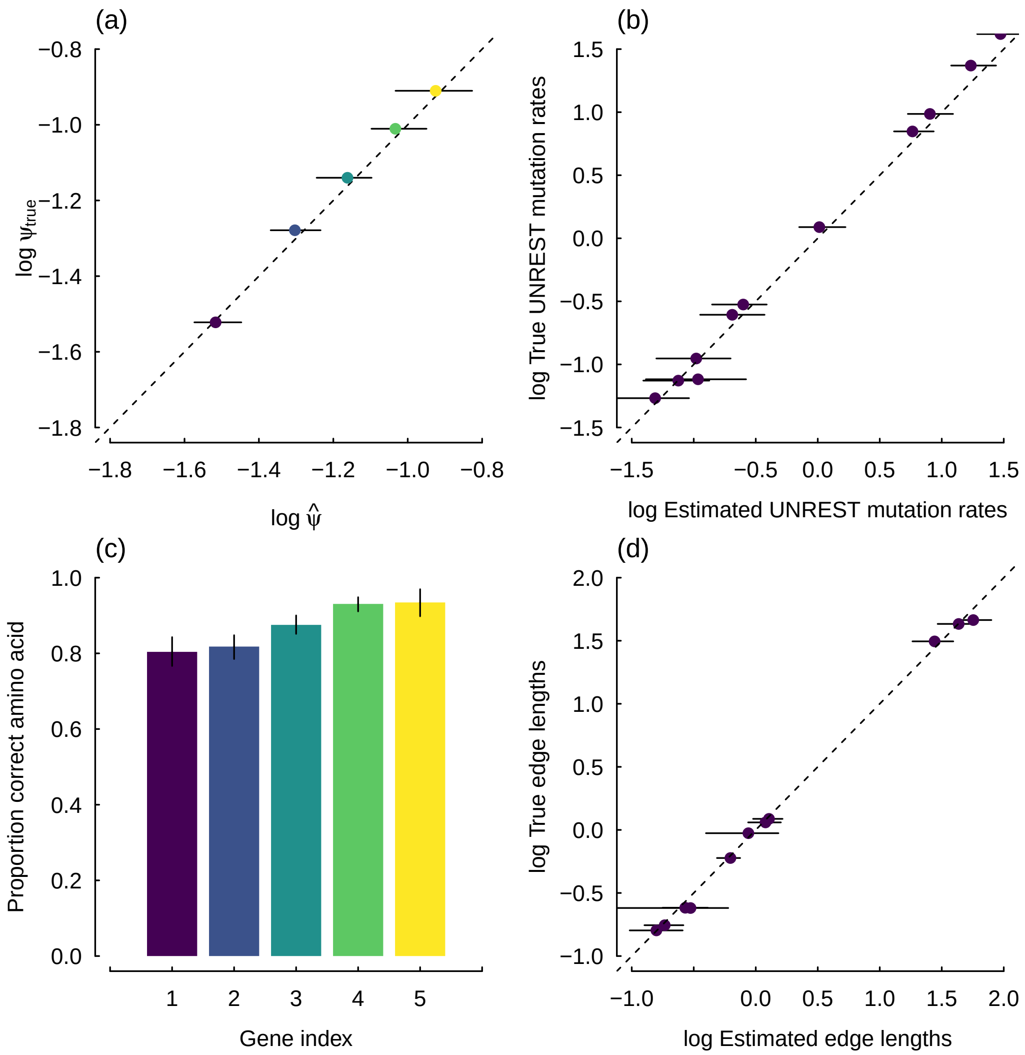

Figure S3: Summary a 5-gene simulation for a SelAC model where we assume $\alpha_G = \infty$, and thus, no site-specific sensitivity in the generating model. The 'known' parameters were based on fitting the same SelAC to the 106 gene data set and phylogeny of Rokas et al. (2003), with gene choice being based on five evenly spaced points along the rank order of the gene specific composite parameter ψ'_g. The points and associated uncertainty in the estimates of the gene-specific average protein synthesis rate, or ψ (calculated from ψ')(a), nucleotide mutation rates under the UNREST model (b), proportion of correct optimal amino acids for a given gene (c), and estimates of the individual edge lengths are based the mean and 2.5% and 97.5% quantiles across on 50 simulated datasets (d). Gene index on the x-axis refers to the arbitrary number assigned to the simulated gene.

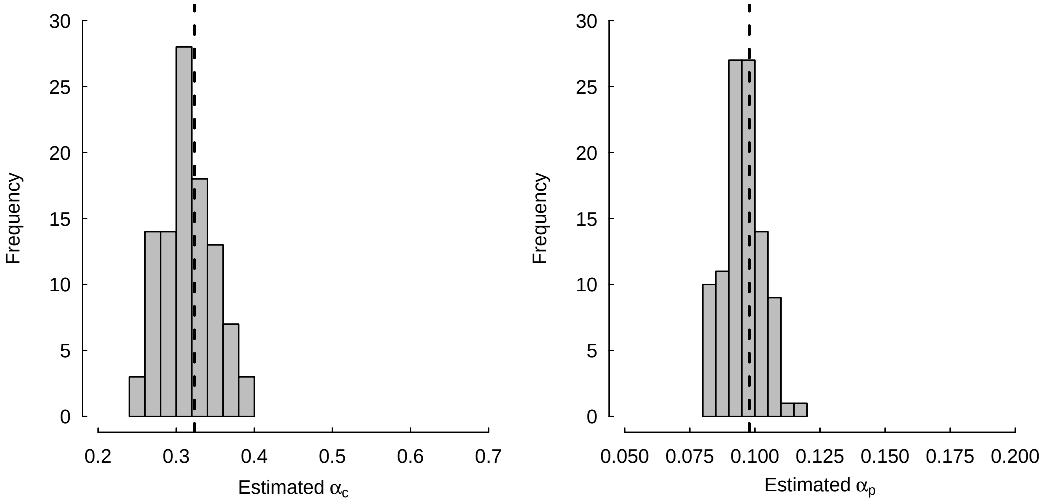

Figure S4: The distribution of estimates of the Grantham weights, α_c and α_p, in a SelAC model, where we assume $\alpha_G = \infty$, and thus no site-specific sensitivity in the generating model. The dashed line represents the value used in the generating model.

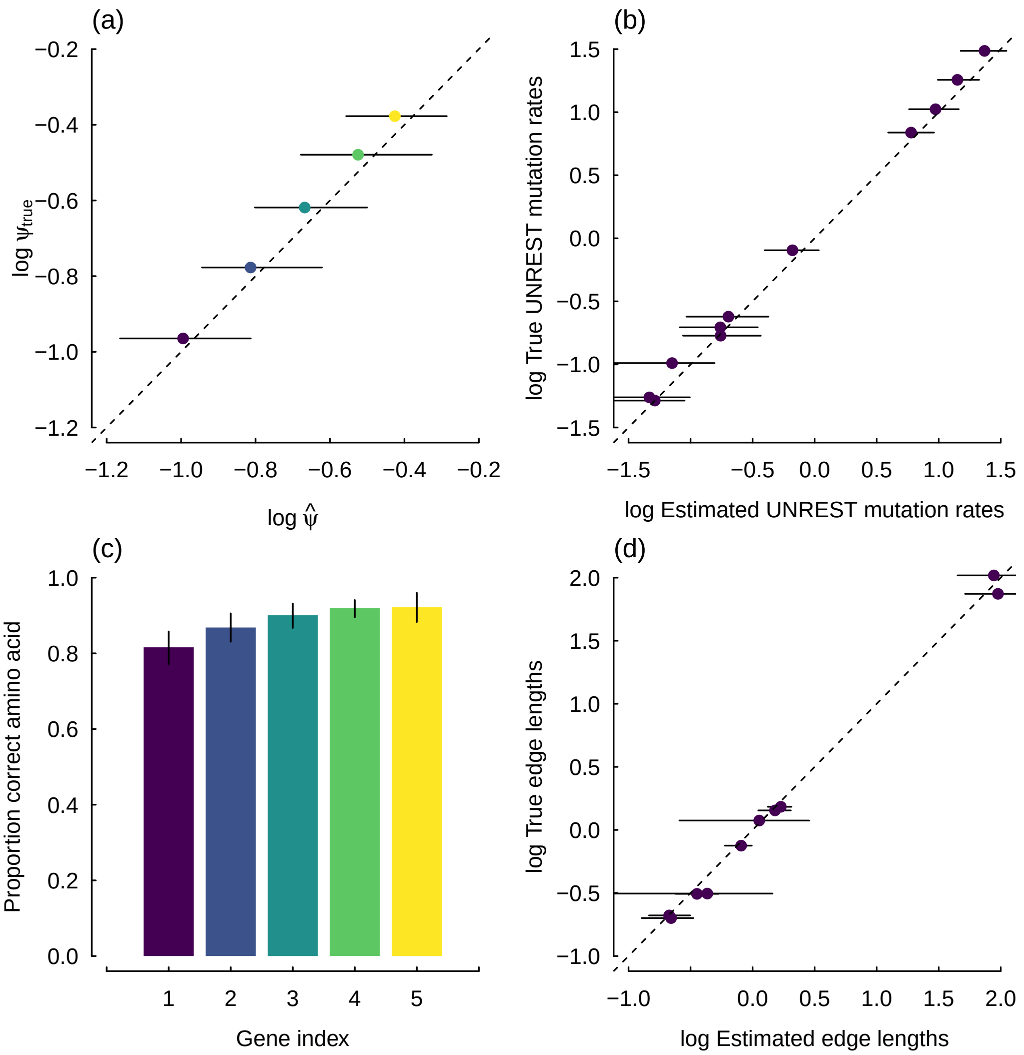

Figure S5: Same figure as in Figure S3, except the generating model includes site-specific sensitivity in the generating model (i.e., α_G).

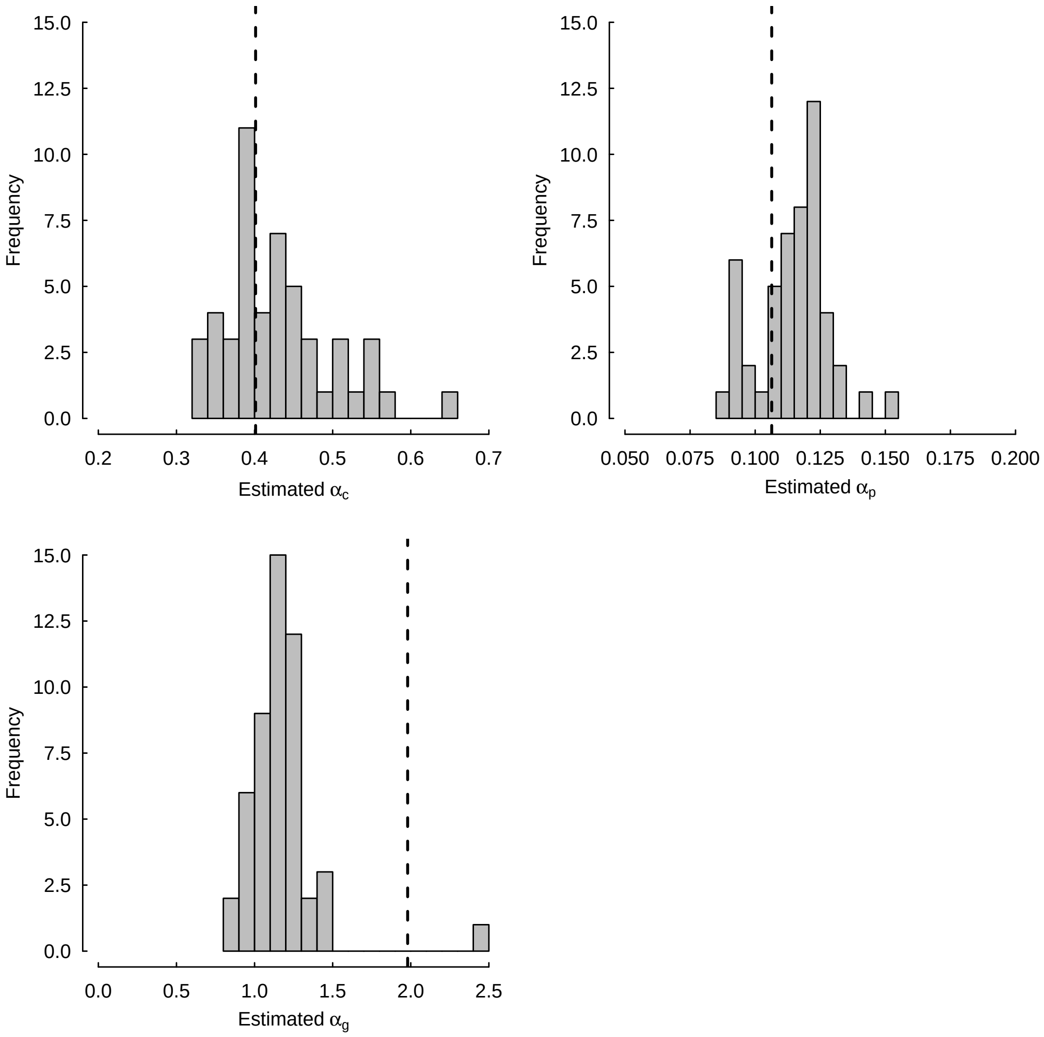

Figure S6: Same figure as in Figure S4, except the generating model includes site-specific sensitivity in the generating model (i.e., α_G). Unlike, Grantham weights, which showed no systematic bias, there is a downward bias in estimates of α_G.