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

Last compiled on Wednesday 18th January, 2017 at 15:30.

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. The new set of models, which we collectively call selac models, fit phylogenetic data substantially better than popular current models, suggesting more accurate inference of phylogenetic trees and branch lengths. Moreover, these models allow inference of population genetics parameters from data used for interspecific phylogenies.

Introduction

Phylogenetic analysis now plays a critical role in the fields of ecology, evolution, paleontology, medicine, conservation, and many others. While the scale and impact of phylogenetic studies has increased substantially over the past two decades, by comparison the realism of the mathematical models on which these analyses are based has changed relatively little. The simplest but most popular models are agnostic with regards to the different amino acid substitutions and may or may not include mutation bias (e.g. F81, F84, HYK85, TN93, and GTR for the former and JC69 and K80 for the latter, see Yang (2014) for an overview).

Another set of models attempt to include a 'selection' term ω, however 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 are 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 as 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* on 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 describes 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 likely only describe a subset of scenarios such as over/underdominance or frequency dependent selection (Hughes and Nei, 1988; **?**).

Fortunately, 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 while 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 different sites quickly becomes extremely challenging. Fortunately, under the simplifying assumptions of additivity between sites and alleles, calculating stationary and substitution probabilities are relatively straightforward, making fitting additive models of the evolutionary process to sequence data computationally feasible.

Another major advantage to our approach is that the parameters estimated are biologically meaningful. As with other phylogenetic methods, we generate estimates of branch lengths and nucleotide specific mutation rates. In addition, because the math behind our model, which we call SelAC (Selection on Amino acids and Codons) 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.

We model the substitution process as a classic Wright-Fisher process which includes the forces of mutation, selection, and drift (**?**Kimura, 1962; Wright, 1969; Iwasa, 1988; Berg and Lässig, 2003; Sella and Hirsh, 2005; **?**). For simplicity, we ignore linkage effects and, as a result of this and other assumptions, SelAC behaves in a site independent manner. SelAC 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 must assigned to a particular rate matrix category. Unlike these other researchers, 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 simplification, be extremely large. 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. We've utilized the same nested, population genetic based approach in more traditional genomic analyses (e.g. Gilchrist, 2007; Shah and Gilchrist, 2011; Gilchrist et al., 2015). That work and our current work illustrates how more information can be extracted from sequence data when more biologically based models are used.

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. The gene specific parameter ψ_g is multiplied by additional model terms to make a composite term ψ'_g which scales the strength and efficacy of selection for the optimal amino acid sequence.[1] SelAC also requires the designation 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

[1] should we introduce the a_* notation here?

substitution rates are allowed to vary independently (Halpern and Bruno, 1998; Lartillot and Philippe, 2004; Rodrigue and Lartillot, 2014). This, in turn, allows us to move beyond simply generating MLE estimates of parameters (c.f. Yang and Nielsen, 2008) and quantify our uncertainty in these values.

The work we present here contributes to the field of phylogenetics and molecular evolution in a number of ways. 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. 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. Finally, because our work fitness functions are well defined, we can provide estimates of key evolutionary statistics such as the distribution of effects on fitness and genetic load.

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 purifying selection. Specifically, we assume the strength of stabilizing selection for the optimal sequence, $\vec{a}_*$, is proportional to ψ, which we can estimate for each gene. In regards to model fit, our results clearly indicated that accounting for stabilizing selection for the optimal sequence substantially improves our model fit. Using both ΔAIC and AIC_w as our measure of model support, the SelAC models perform extraordinarily better than GTR+Γ, GY94, or FMutSel0. This is in spite of the need for estimating the optimal amino acid at each position in each protein, which adds more than 47,000 more parameters to the model than FMutSel0, the next most parameter rich codon model in our model set, and yet the SelAC model shows nearly 400,000 AIC unit improvement over FMutSel0. Comparing the two SelAC models in the model set, including the random effects term α_G not only provides greater biological realism by allowing site heterogeneity, it also provides substantially better model fit and improves the ΔAIC score by over 23,000 units than assuming $\alpha_G = \infty$.

With respect to estimates of ψ within SelAC, they were strongly correlated with two separate empirical measurements of gene expression (See Figure). 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 proteins functionality, indicated a moderate level of variation in gene expression among sites. Our estimate of $\alpha_G = 1.40$, produced a distribution of sensitivity rates that ranged from 0.344-7.16, but with the 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 estimates were generally underestimated (see Supplemental). 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 in regards to composition weight, α_c, which is the distance between any two amino acids based on the ratio of noncarbon elements in the end groups to the number of side chains. We estimate the weighting factor of composition to be α_c=0.484, which suggests that the reduction in the distance between any given two amino acids and thus provides greater exchangeability between even the most distantly related residues.

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 ψ. In fact, ω showed similar, though slightly reduced correlations, with the same empirical estimates of gene expression described above (See Figure **??**). 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, when we evaluated each model according to the realistic nature of the data they produce, our SelAC model clearly performs better. When we simulated the sequence for *S. cervisieae*, starting from the ancestral sequence under both GTR+G 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 XX). In a way, this is somewhat unsurprising, given that both GTR+G 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 our SelAC provides. We do note that the adequacy of the SelAC model does vary among individual taxa, and does not always perfectly match the observed functionality. For instance, *S. castellii* is simulated with consistently higher functionality than observed (Figure XX). 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.

Figures

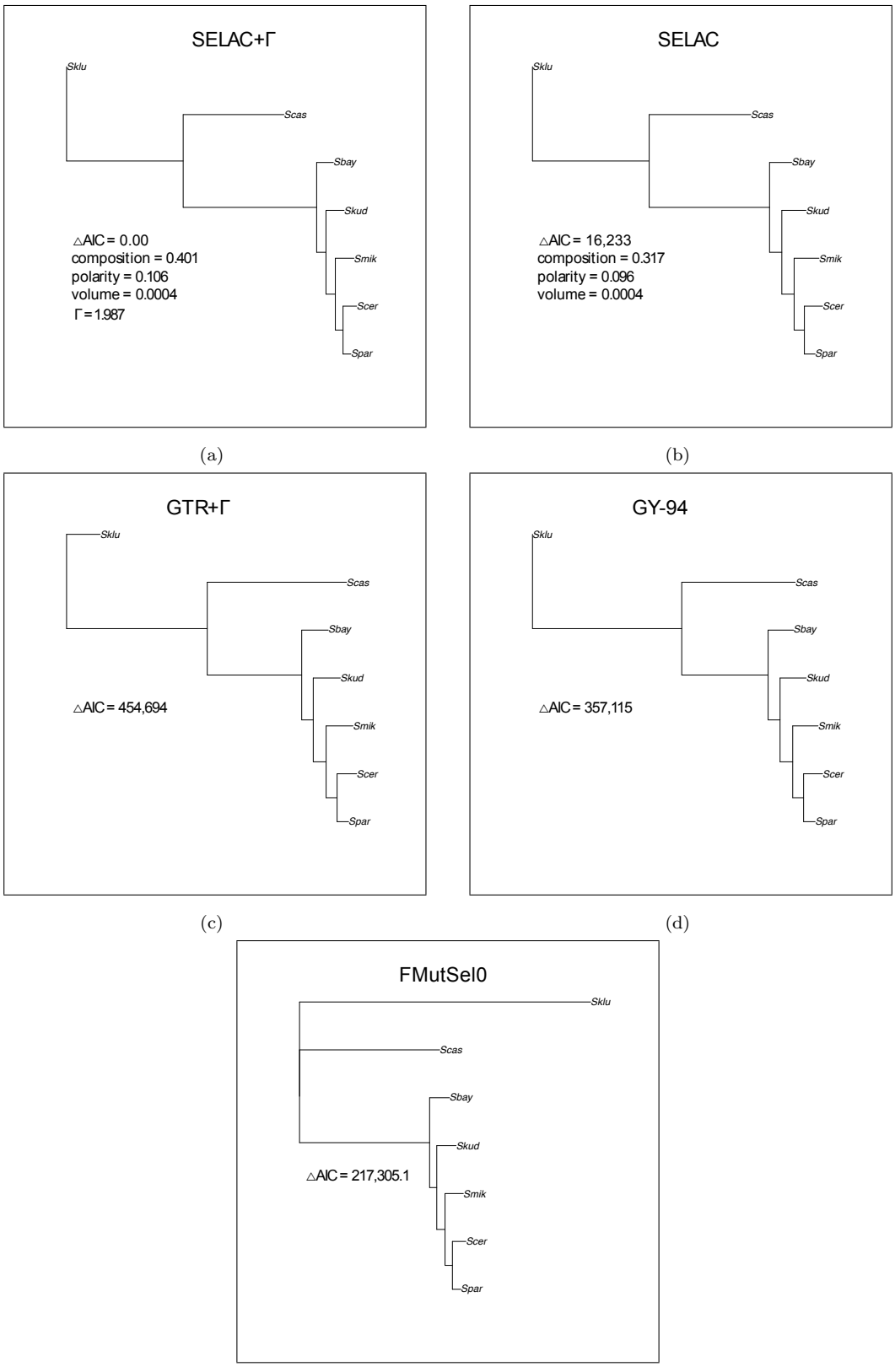


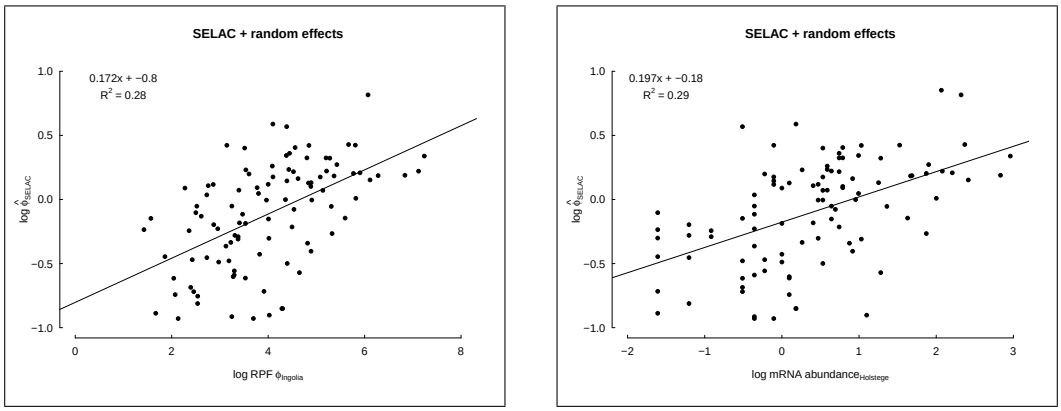
Figure 1: Maximum Likelihood Trees for (a) selac, (b) selac with uniform sensitivity $G = 1$, (c) GTR, (d) GY94, and (e) FMutSel0.

Figure 3: ω vs ϕ Figure goes here

2. Model Adequacy Illustrations (Brian or Jeremy?)

3. Gene Expression Comparisons (Jeremy now has this data)

   (a) SelAC vs. Empirical Measurements



(a) (b)

Figure 2: Comparison of log protein synthesis rate ψ for *S. cerevisiae* as predicted by selac to empirical estimates from (a) ribosome profile footprint data (Ingolia et al., 2009) and (b) mRNA abundance data (Holstege et al., 1998).

   (b) GTR ω vs. ϕ Measurements

4. Quantifying optimal AA (Mike once I get info from Jeremy)[2]

   (a) Represent using varying point size in 3D physicochemical space

   (b) Look for instances where there's a bimodal distribution across. This would suggest a shift in optimal AA.

5. Visualization of Evolutionary Landscapes (Brian)

   (a) Summary of Fitness Landscapes: frequency distribution of W_i with varying ψ.

   (b) Stationary Distribution of Fitness Values: Frequency distribution of $\exp[W_i]$ with varying ψ. Essentially above figure, but W_i values are evaluated based on their stationary probability distribution. These figures can be related to concepts about genetic load.

[2]mikeg: looked around, but can't find this data. Ideally it would be in the format of: Gene, Psi, Position, LLik AA A, LLik AA B, LLik AA C, ...

Model	logLik	Parameters Estimated	AIC	ΔAIC	Model Weight
GTR+Γ	-655166.4	610	1,311,553	504,151	<0.001
GY94	-612121.5	210	1,224,663	417,261	<0.001
FMutSel0	-598848.2	2810	1,203,316	395,914	<0.001
SelAC: UNREST	-465616.7	50,004	831,226	23,824	<0.001
SelAC: UNREST+Γ	-453706.0	50,005	807,402	0	0.999

Table 1: Comparison of model fits using ΔAIC.

(c) Distribution of Mutation Fitness Effects: Frequency distribution at which new mutants with fitness value $\exp[W_m]$ are introduced at stationarity.

Tables

Discussion

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 (**???**Halpern and Bruno, 1998; Lartillot and Philippe, 2004) [3]. Despite their widespread use, phylogenetic models based on purifying and diversifying selection,i.e. **?** and its 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.

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; **?**; 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 example, even though SelAC requires a large number of matrices, because of our assumption about protein functionality and physicochemical distance from the optimum, we are able to parameterize our substitution 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 between genes to the variation in the

[3] mikeg: UPDATED: Jeremy and Brian, can you provide some references that support this claim?

target expression level ψ. 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 **??**. Perhaps more convicingly, we find that the target expression level ψ and realized protein synthesis rate ϕ are reasonably well correlated with laboratory measurements of gene expression ($\rho = 0.33 - 0.44$). and Figure **??**)). The idea that quantiative 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 inter-genic 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, 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. Statistically speaking, it should be possible to estimate the rates at which this occurs on a course or fine scale depending on the given hypothesis and the amount of sequence data and computational resources available.

Additional advantages

1. More realistic behavior over time: Model adequacy (Figure **??**)
2. Improved fit (Table **??**)

3. Improved estimates of branch lengths and mutation: A better model gives you a better answer **??**

4. Better biological interpretation and more biological information.

5. Likelihood based estimate ancestor state rather consensus assumption.

6. Approach can even be expanded to other types of sequence data in which selection can be reasonably modeled, e.g. UCEs.

7. Allows us to describe evolutionary process using our inferred fitness landscapes (Figures **??**-**??**. Nevertheless, the idea that the strength of stabilizing selection and gene expression are positively correlated is well supported by other researchers (e.g. **?**) and

    - First, it indicates there is substantially more information in the coding sequences used for phylogenetic analysis than other methods acknowledge.
    - Second, it demonstrates how selection can be modeled as the product of two separate components. Here we use gene expression ψ and protein function $\mathbf{B}$, but more complex models could clearly be used.
    - Extensible to other researchers approaches that use structural and folding information.

Shortcomings in model implementation

1. Computationally expensive to fit model.

2. Estimating uncertainty is also expensive (though should be parallelizable further than fitting).

Shortcomings in model assumptions and extensions

1. Weak mutation which means that populations can get stuck on local sequence optima, especially so for high expression genes.

2. While we use a reasonable line of reasoning in developing our benefit model $\mathbf{B}$, it is not well supported by any particular set of experiments or data.

3. 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 domiance. Thus, our work can be considered a first order approximation to these more complex scenarios and a starting point for later relaxation these assumptions.

4. For example, because our current implementation ignores any selection on synonymous codon usage bias (CUB). Including such selection is tricky because introducing the site specific cost effects of CUB leads to non-additive (i.e. epistatic) interactions between sites. Relative to stabilizing selection on amino acid sequence, selection on CUB is thought to be substantially weaker. As a result, epstatic effects due to synonymous codon specific differences in assembly costs can likely ignored and selection on CUB incorporated into our current framework.

5. SelAC implicitly assumes that all genes are essential because an organism that is homozygous for null alleles with zero activity (i.e. no benefit) would have to spend an infinite amount of energy to achive a target functionality synthesis rate $\psi > 0$. The only way to generate such null alleles is through the evolution of a premature stop codon which, given this model behavior, can never go to fixation. Two ways this assumption of essentiality could be relaxed are by making fitness W a function of ψ such that $W(\psi = 0) > 0$ or by incorporating functional overlap between proteins into our calculations. While SelAC assumes that functionality declines with physicochemical distance from the optimum amino acid, as it stands this decline is less dramatic than one might intuitively expect. For example, for the yeast dataset, the average taxon has a functionality of _____ under our framework, while a random sequence has an average functionality of _____, rather than values close to zero as one would expect. Thus, while our assumption of about the link between functionality and physicochemical distance may be reasonable for sequences close to the optimum sequence $\vec{a}_*$, it clearly fails for protein sequences far from the optimum. [4] This implies that the difference in expression between a random sequence and the observed sequence would only be on the order of _____-fold. It is important to note that the this does not mean that selection is ineffective in our model. The log of the selection coefficient between two alleles depends on both their functionality and the gene's target expression level ψ. Thus for high expression genes, slight changes in functionality can still lead to large differences in fitnesses between alleles.

6. PROBLEM: Yeast having problems with estimating kluyveri branch. Are any of these gene in CLeft? According to Cedric, "C-Left goes from IDSAKL0C00110 to (including) SAKL0C10846.Every mapping hit you have in your dataset with an ID in between (numerically) is on C-Left."

[4] mikeg: Two thoughts. One, because we are integrating over G rather than using the MLE estimate in this calculation, I'm wondering if we are substantially under estimating the effect of a few sites with high senstivities would have on protein function. If it is not hard to do, we could evaluate the model at the MLE parameters and then find the optimal G for each position. Two, if the true values of G follow a heavy tailed distribution, it may make more sense to try and model G using an inverse-gamma distribution.

7. PROBLEM: GTR is scaled at nt level so likely 3 times selac rate of codon substitution. [verify with sims of phi of zero][5] Math should clarify things.

8. Not currently integrated with other approaches

9. Identifiability issues

10. Issues with discreteness of amino acids

Lots of sequences available and in pipeline, let's get to it!

Methods

We link genotype, phenotype, fitness, drift, and fixation, by extending the approach we have successfully used to quantify the evolutionary forces of fitness, drift, and fixation on to the evolution codon usage bias based on an organism's coding sequences (Gilchrist and Wagner, 2006; Gilchrist, 2007; Shah and Gilchrist, 2011; Gilchrist et al., 2015). More specifically, in order to link genotype, phenotype, and fitness, we assume that organisms have set of fixed, but *a priori* unspecified, metabolic requirements and the organism meets these requirements through the appropriate translation of its proteome. We assume that each protein has, on average, a target synthesis rate of ψ and, for now, that ψ is fixed over the tree. We also assume that natural selection favors genotypes that are able to synthesize their proteome 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 q. 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, a complete, error free peptide consisting of a_* provides one unit of the gene's functionality. Thus ψ for a given protein is determined by both the organism's metabolic requirements and the functionality of the protein encoded by $\vec{a}_*$. SelAC allows us to link amino acid sequence and gene expression directly to genotype fitness and, in turn, substitution rate in a general, yet simple and biologically plausible, manner.

The overall structure of SelAC involves a codon mutation model combined with a selection model based on the cost and benefits of translating a given genotype and the target gene expression rate of a gene.

[5] mikeg: any progress on this? Regarding the math, if we are scaling time by nt substitutions (synoymous or non-synonymous), it would seem like both SelAC and GTR's mutation models are comparable.

Allele Substitution Model

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 (UNREST) (**?**) because it makes no constraint on the instantaneous rate of change between any pair of nucleotides. 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 correspond to the equilibrium frequencies of each base. These equilibrium nucleotide frequencies are determined through analytically solving $\pi \times \mathbf{Q} = 0$. We use this $\mathbf{Q}$ to populate a 64×64 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 much 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.

While the overall model does not assume equilibrium, we still need to scale our mutation matrices μ. Traditionally, it is rescaled such that at equilibrium, one unit of branch length represents one expected substitution per site. Here, a scaling factor is calculated as the average rate $-\sum_i \mu_i \pi_i = 1$, where i indexes a particular codon in a given gene. The final mutation rate matrix is the original mutation rate matrix multiplied by 1/scaling factor.

Protein Synthesis Cost-Benefit Function η:

SelAC links fitness to the product of the cost-benefit function of a gene g, η_g, and the organism's average target synthesis rate of the functionality provided by gene g, ψ_g. This is because the average flux energy an organism spends to met its target functionality provided by gene g is $\eta_g \times \psi_g$. In order to link genotype to our cost-benefit function η, we begin by defining our benefit function.

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

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

where n_g 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 infinite and, as a result, nonsense mutations are always lethal. The term G_p describes the sensitivity of the protein's function to deviation in Grantham's physicochemical space. We assume that $G_p \sim \text{Gamma}\,(\alpha = \alpha_G, \beta = \alpha_G)$ in order to ensure $\mathbb{E}(G_p) = 1$.

At the limit of $\alpha_G \to \infty$, the model collapses to a model with uniform sensitivity of $G_p = 1$ for all positions p. $\mathbf{B}(\vec{a}_i|\vec{a}_*)$ 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 sites senstivity 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 any 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 emphasize that other properties could be used. 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, 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,

$$d(a_i, a_*) = \sqrt{\alpha_c \left(c\left(a_i\right) - c\left(a_*\right)\right)^2 + \alpha_p \left(p\left(a_i\right) - p\left(a_*\right)\right)^2 + \alpha_v \left(v\left(a_i\right) - v\left(a_*\right)\right)^2}.$$

Linking Cost of 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, is equal to $\psi/\mathbf{B}(\vec{a})$ and the total energy flux allocated towards meeting the target functionality of a particular gene is $\eta(\vec{c})\psi$. As we shall show below, the fitness cost for a genotype encoding a suboptimal protein sequence stems from the need to produce $1/\mathbf{B}(\vec{a})$ proteins in order to produce the equvalent functionality of one protein consisting of the optimal amino acid sequence a_*. For example, a protein encoding allele which has a 10% reduction in functionality relative to the optimal sequence, i.e. $\mathbf{B}(\vec{a}) = 0.9$, will have the same energetic burden and selective cost relative to its optimal sequence as a protein encoding allele of similar length which has a 20% reduction in functionality but whose target synthesis rate is 1/2 of the first protein.

Third, we assume that every additional high energy 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 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 explicilty identify our time units.

Correspondingly, the ratio of fitness between two genotypes is,

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

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

$$W_i/W_j = \exp\left\{-A_0\left(A_1 + A_2 n\right)\frac{1}{n}\sum_{p\in\mathbb{P}}\left[d\left(a_{i,p}, a_{*,p}\right) - d\left(a_{j,p}, a_{*,p}\right)\right]\psi\right\} \tag{5}$$

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,

$$\begin{aligned} u_{i,j} &= \frac{1 - \left(W_i/W_j\right)^b}{1 - \left(W_i/W_j\right)^{2N_e}} \\ &= \frac{1 - \exp\left\{-\frac{A_0}{n}\left(A_1 + A_2 n\right)\left[d\left(a_i, a_*\right) - d\left(a_j, a_*\right)\right]\psi\, b\right\}}{1 - \exp\left\{-\frac{q}{n}\left(A_1 + A_2 n\right)\left[d\left(a_i, a_*\right) - d\left(a_j, a_*\right)\right]\psi\, 2N_e\right\}} \end{aligned}$$

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 $\mathbf{Q}_a$ matrices, one for each amino acid, with up to 26,880 unique rates, based on few parameters (one to 11 mutation rates, two free Grantham weights, the cost of protein production $\mathbf{C}$, the target functionality synthesis rate, and optimal amino acid at each site position p, $a_{*,p}$), which we infer 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.

Finally, given our assumption of independent evolution among sites, the probability of the whole data set is the product of the probabilities of observing the data at each individual site. Thus, the log likelihood of amino acid a being optimal at a given site position p is calculated as

$$\mathcal{L}\left(\mathbf{Q}_a|\mathbf{D}_p, \mathbf{T}\right) \propto \mathbf{P}\left(\mathbf{D}_p|\mathbf{Q}_a, \mathbf{T}\right) \tag{6}$$

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 **?** is used to calculate $\mathcal{L}(\mathbf{Q}_a)$. The log 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 (**?**).

Implementation

All methods described above are implemented in the new R package, `selac` available through CRAN (`http://cran.us.r-project.org`). 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 **?**, 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, we initiate the optimal amino acid at each site by using a 'majority' rule, where the initial optimum is the most frequently observed amino acid at a given site. Our optimization routine then proceeds by cycling though multiple phases. The first phase optimizes the branch lengths while holding the model parameters constant. The second phase optimizes the gene specific composite parameter $\psi' = A_0 \psi N_e$ across genes, while holding constant both the branch lengths and the model parameters shared across the genome (i.e., α_c and α_p, and the sensitivity distribution parameter α_G). This is followed by a third phase that optimizes the parameters across the genome, while keeping the branch lengths and the composite parameters constant. Finally, the fourth phase estimates the optimal amino acid at each site while keeping the branch lengths and all model parameters at their current values. This entire procedure is repeated six times. The difference in loglikelihood between the 5th and 6th iteration was less than XXXXX. For optimization of a given set of parameters, we rely on a bounded subplex routine `NLopt ?` 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 *alphap*. Confidence in the parameter estimates are 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 from **?**.

We note that our current implementation is painfully slow, and is particularly suited for smaller data sets in terms of numbers of taxa. This is largely due to the size and quantity of matrices we create and manipulate just to calculate the log-likelihood of an individual given 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 MrBayes (**??**), PAML (**?**) and RAxML (**?**).

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 **?**.[6] 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 section on Analyzing Yeast Genome). Bias in the model parameters were assessed under two generating models: one where we assumed a model of SelAC assuming $\alpha_G = \infty$, and one where we estimated 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.[7] 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 *alphap*. All analyses were carried out in our `selac` R package.

Analysis of yeast genome and tests of model adequacy

We focus our empirical analyses on the large yeast data set and phylogeny of **?**. The yeast genome is an ideal system to examine our phylogenetically estimates of gene expression and its connection to real world measurements of these data within individual taxa. The complete data set of **?** contain 1070 orthologues, where we selected 100 at random for our analyses. We also focus our analyses only on the Saccharomyces *sensu stricto*, including their sister taxon *Candida glabrata*, and we rely on the phylogeny depicted in Fig. 1 of **?** for our fixed tree. We fit both the new models described in this paper, 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, 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.

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. In other words, we examined whether data simulated under the MLE produces properties similar to the input sequence data (e.g., **?**. 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

[6] mikeg: Should this be updated to reference the **?** dataset? JMB: No. The sims were already done before we decided to change data sets. It's stupid to redo. The data set used to come up with parameters to simulate under is really irrelevant.

[7] mikeg: Ibid

to sample and assemble the starting coding sequence, from which we calculate functionality relative to the optimal sequence. The current state of the simulated coding sequence and its current functionality is examined at equidistant points along the length of the pruned branch. 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.

References

Berg, J. and M. Lässig. 2003. Stochastic evolution and transcription factor binding sites. Biophysics 48:S36–S44.

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.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences of the United States of America 102:14338–14343.

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. 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.

Gilchrist, M. A. and A. Wagner. 2006. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. Journal of Theoretical Biology 239:417–434.

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.

Holstege, F. C. P., E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, et al. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95:717 – 728.

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.

Ingolia, N. T., S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman. 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science 324:218–223.

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

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. *In* Biocomputing 2001. World Scientific, 191–202.

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.

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.

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.

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.

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.

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 :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. 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.

Supporting Materials

Figures

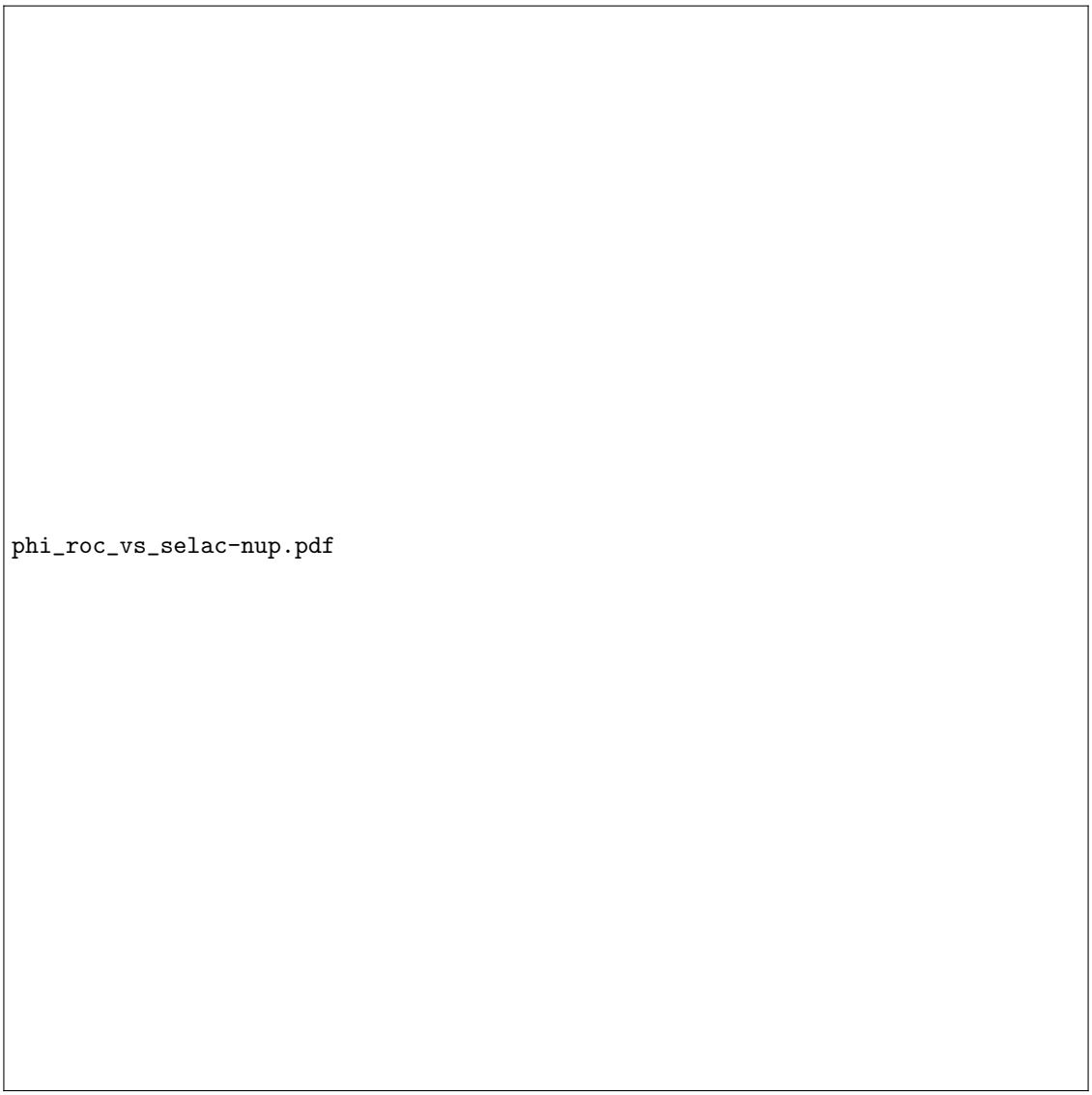


Figure S1: Comparison of ROC and SelAC estimates of ϕ. ROC estimates are inferred from genome scale patterns of synonymous codon usage driven by a combination of mutation bias and natural selection for synonymous codons with shorter pausing times which, as here, is scaled by gene expression ϕ. See Gilchrist et al. (2015) for more details.