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

Last compiled on Friday 4th November, 2016 at 15:04.

Abstract

We present a phylogenetic approach rooted in the field of population genetics that more realistic models the evolution of protein-coding DNA under the assumption of stabilizing selection. The new set of models, which we collectively call selac models, fit phylogenetic data substantially better than current models, suggesting more accurate inference of phylogenies. 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 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 models assume neutrality between 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). The next set of models attempt to include a 'selection' term ω. Although it is rarely acknowledged, 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 example, ω is generally interpreted as indicating whether a sequence is under 'purifying' ($\omega < 1$) or 'positive' ($\omega > 1$) selection. However, the actually 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 over amino acids, including i, 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 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; Nowak, 2006).

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). This is desirable because population genetics is a mature, mathematically rigorous, and well established framework for describing biological evolution. One lesson from the field of population genetics is while there are only a few fundamental evolutionary forces at play, i.e. 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. As a result, fitting additive models of the evolutionary process to sequence data is computationally feasible. While critics may point out that such models are overly simplistic, we counter that these assumptions are commonly made and, in terms of fitting the data, our 'simplistic' models are still a fast improvement over standard approaches. Further, just as the GTR can serve as a null model for neutral evolution, our models are null hypotheses for more complex models that relax some of our more restrictive assumptions, such as amino acid substitution properties being independent of local structure or location.

Another major advantage to our approach is that the parameters estimated are biologically meaningful. As with outher phylogenetic methods, we generate estimats 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 quantiative 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 our model 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 using the classic the Wright-Fisher model 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). For simplicity, we ignore linkage effects and, as a result of this and other assumptions, our model behaves in a site independent manner. Our approach is developed in the same vein as previous phylogenetic applications of the Wright-Fisher model (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, our approach allows us to quantiatively evaluate the support for a particular amino acid being favored at a particular position within the protein encoded by a particular gene.

Because our approach requires twenty families of 61 × 61 matrices, the number of parameters needed to implement our model 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 physiochemical weighting parameters, and a shape parameter describing the distribution of site sensitivities. In addition to these genome wide parameters, our model requires a gene specific expression parameter ψ, which is also scaled by N_e and other genome wide terms, describes the average synthesis rate of a protein. The composite term $N_e\psi$ scales the strength and efficacy of selection for the optimal amino acid at a given site within a given gene. Our model 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, our model 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). 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. Our model 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), our approach 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 physio-chemical 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 physiochemical 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

Model Performance

1. Using ΔAIC as our measure, we see that even despite the need for estimating the optimal amino acid at each position in each protein, our model performs astronomically better than GTR, GY94, or YN08.

2. Including the random effects term G not only provides greater biological realism than assuming $G = 1$, it provides substantially better model fit and improves the ΔAIC score by over 16,000 units.

3. Selac provides estimates of gene expression which are positively correlated with empirical estimates

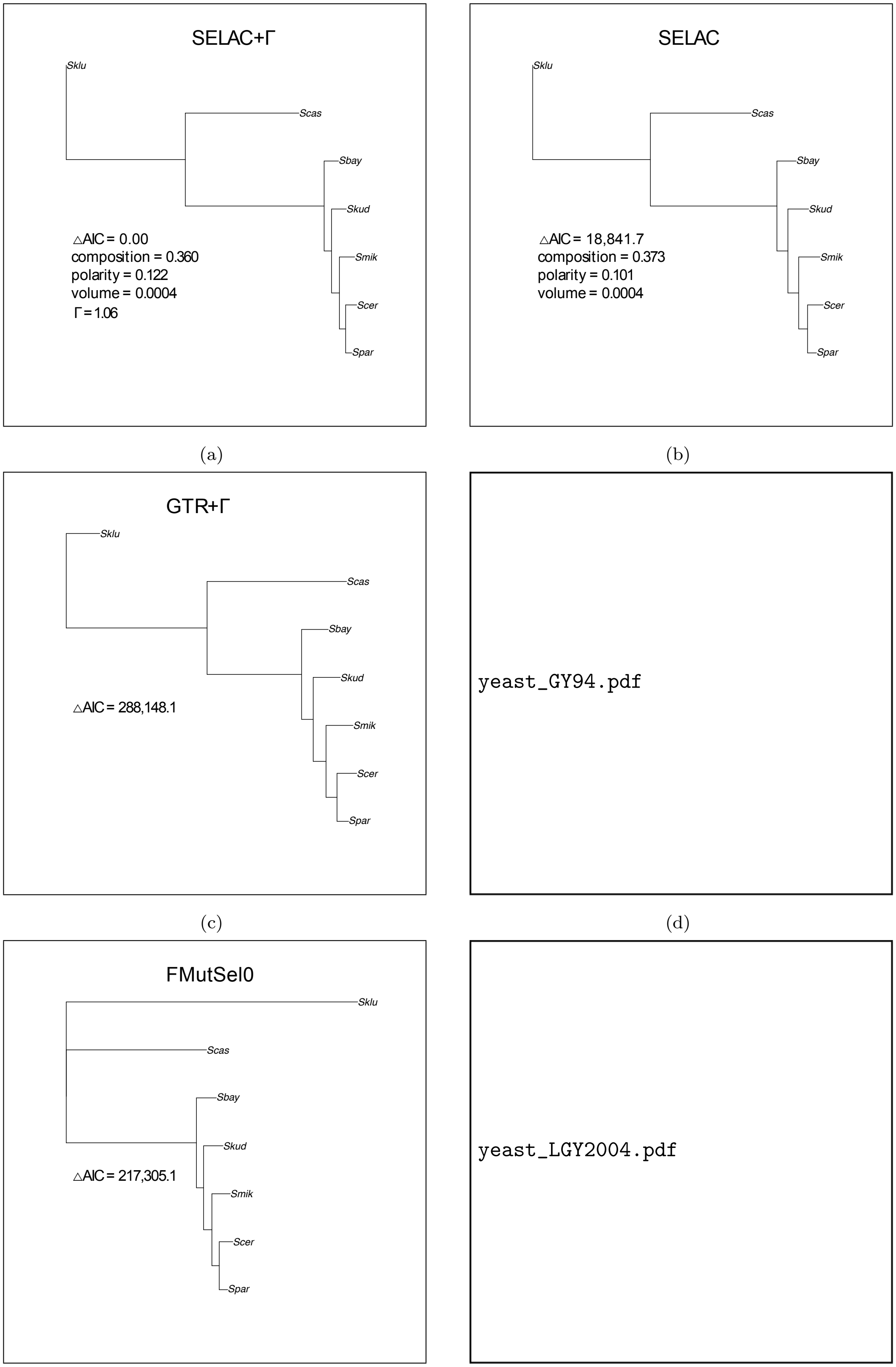
and explain 20-30% of the variation in the empirical measurements taken during log growth phase.[1]

4. By linking transition rates $q_{i,j}$ to gene expression ψ, our approach allows use the same model for genes under varying degrees of purifying selection. The traditional approach of concatenating gene sequences together is equivalent to assuming the same average protein synthesis rate ψ for all of the genes. By assuming the strength of stabilizing selection for the optimal sequence, $\vec{a}_*$, is proportional to ψ, our model allows us to estimate $\phi = \psi/\mathbf{B}$ for each gene. Our results clearly indicate that this information is available and accounting for it in our model substantially improves our model fit.

Figures

1. Branch Lengths

[1] mikeg: ~~We should replace the estimates of ψ with estimates of ϕ which is $\psi/\mathbf{B}(\vec{a}_{\text{obs}}|\vec{a}_*)$.~~ I've decided it's better to replace ϕ with ψ rather than the other way around. Thus, $\phi = \psi/\mathbf{B}(\vec{a})$ is the target synthesis rate of error free protein $\vec{a}$ and and $\phi = \psi$ when $\vec{a} = \vec{a}_*$ since, by definitioon, $\mathbf{B}(\vec{a}_*) = 1$.



(a)

(b)

(c)

(d)

(e)

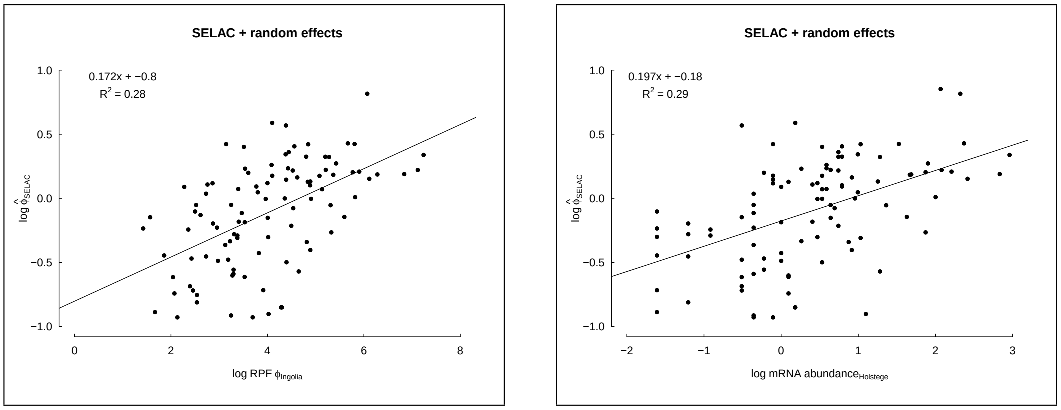
(f)

Figure 1: Maximum Likelihood Trees for (a) selac, (b) selac with uniform sensitivity $G = 1$, (c) GTR, (d) GY94, and (e) YN08, (f) Lartillot and Philippe (2004).

2. Model Adequacy Illustrations (Brian or Jeremy?)

3. Gene Expression Comparisons (Mike, trying to get data for additional yeast species.)

   (a) SelAC vs. Empirical Measurements

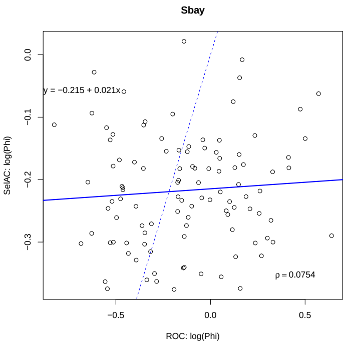


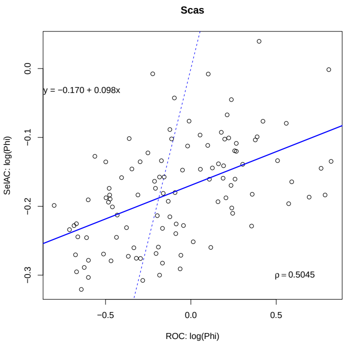
(a)

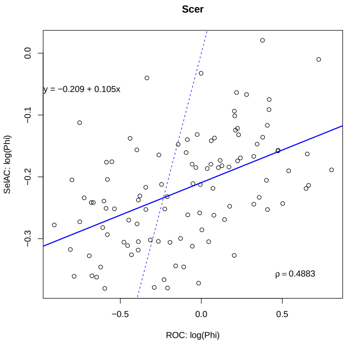
(b)

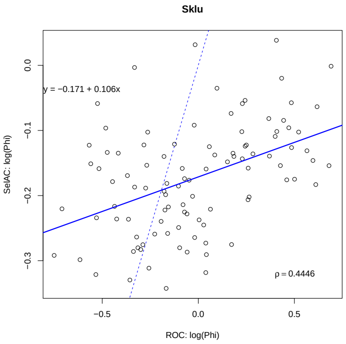
Figure 2: Comparison of log protein synthesis rate ψ for *S. cerevisiae* as predicted by selac to empirical estimates from (a) ribosome profile footprint data (Ingolia et al., 2009) and (b) mRNA abundance data (Holstege et al., 1998).

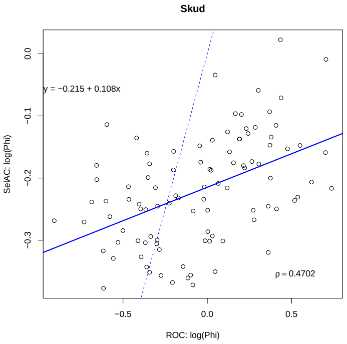
   (b) SelAC vs. ROC Measurements (should probably go in Supporting Materials)

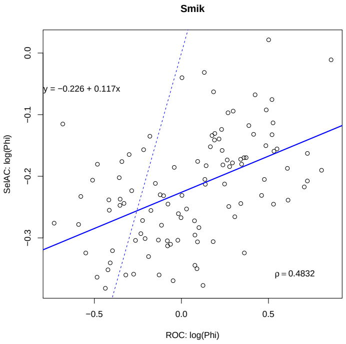


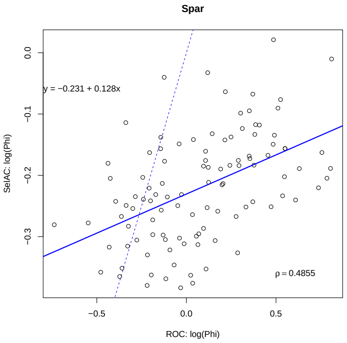












4. Quantifying optimal AA (Mike once I get info from Jeremy)

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

		Parameters			Model
Model	logLik	Estimated	AIC	ΔAIC	Weight
GTR + Gamma	-557990.3	648	1,117,277	454,694	<0.001
GY94	-509625.1	223	1,019,698	357,115	<0.001
GY08	-518090.6		1,120,945	458,362	<0.001
SelAC: GTR	-382725.8		850,391	187,808	<0.001
SelAC: UNREST	-381881.4		678,816	16,233	<0.001
SelAC: UNREST + Gamma	-373765.6		662,583	0	0.999

Table 1: Model comparisons

(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.

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

Tables

1. Model Fit Summary: Table of number of parameters, estimates for key parameters (or their summaries), and ΔAIC values. (Jeremy, please complete)

Discussion

As phyogenetic methods become ever more ubiquitous in biology, there is a real need to move beyond the cartoon, single matrix models most researchers use (**???**Halpern and Bruno, 1998) [2]. Despite their widespread use, phylogenetic models based on purifying and diversifying selection,i.e. Goldman and Yang (1994) 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.

[2] mikeg: Jeremy and Brian, can you provide some references that make this statement but only make minor advances?

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 Wright-Fisher based methods (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, we parameterize the model and fit branch lengths simultaneously rather than in two separate steps.[3] Even though our model requires a large number of matrices, because of our assumption about protein functionality and physiochemical 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, our model explicitly links the variation in the strength of stabilizing selection for the optimal protein sequence between genes to the variation in the target expression level ψ. In turn, our model allows us to generate quantiative estimates of gene expression and amino acid optimality (Figures 3 and 4). We believe this is an important advance that helps us link molecular evolution and function a very general manner. Our results indicate that including a gene specific ψ value vastly improves our model fits (Table 1 and Figure 3)). Of course, given the general nature of our model and the complexity of biological systems, other biological forces also contribute to gene level variation in natural selection. Nevertheless, the idea that the strength of stabilizing selection and gene expression are positively correlated is well supported by other researchers (e.g. Drummond et al., 2005) and we find that the target expression level ψ and realized protein synthesis rate ϕ are reasonably well correlated with laboratory measurements of gene expression.

Additional advantages

1. More realistic behavior over time: Model adequacy (Figure 2)

2. Improved fit (Table 1)

3. Improved estimates of branch lengths and mutation: A better model gives you a better answer 1

4. Better biological interpretation and more biological information.

5. Likelihood based estimate ancestor state rather consensus assumption.

[3] mikeg: Jeremy or Brian: I know this is true for Halpern and Bruno and am 95% sure it applies to Jeff Thorne's work. Can you confirm its true for the Lartillot references?

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 5a-5c.

The fact that our approach allows us to estimate gene expression from phylogenetic data has a number of important implications by itself.

- 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. Finally, it provides a framework for estimating both shallow and deep branch lengths through the use of coding sequences with potentially very different rates of evolution.

- 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 optima.

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, CUB based epstasis can likely ignored and selection on CUB incorporated into our current framework.

5. Our model implicilty 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 $\psi > 0$. It is worth noting that in its current formulation, the only way to generate such null alleles is through the evolution of a premature stop codon. 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.

6. Currently, our model assumes the optimal amino acid for any site is fixed along all branches. This assumption could be relaxed by by allowing the optimal amino acid to change during the course of evolution along a branch. This would result in the need of additional parameters describing the rates at which the optimal amino acid switches at a site. To allow for changes to the optimal amino acid across all branches makes the model non-time reversible. While such behavior might be desirable to modle the effect of a particular widespread environmental change, incorporating such behavior in a general manner would introduce a whole new set of challenges.

7. We use a universal set of Grantham weights for all sites. Since the importance of an amino acid's physiochemical 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. This would allow the rank order of amino acid functionality given a particular optimal AA to change between sites. Because ψ is determined, in part, by our choice of a reference distance weighting $\alpha_v = 4 \times 10^{-4}$. A larger and more informative set of Grantham weights might reduce the noise in our estimates of ϕ.

8. 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."

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

10. Not currently integrated with other approaches

11. Identifiability issues

12. 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 $\vec{a}_*$. Our approach 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. [4]

The overall structure of our model 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. We fit the model using maximum likelihood and . . . [5]

Allele Substitution Model

Mutation Rate Matrix μ:

We begin by defining a time reversible model for mutation rates between individual bases. We use the existing drift driven substitution model JC (Jukes and Cantor, 1969), also referred to as a General Time

[4] mikeg: Moved from methods.

[5] mikeg: Jeremy, care to elaborate here?

Reversible (GTR) (Tavare, 1986; Yang, 2014) which has 9 independent parameters, the most possible for a model that assumes mutations occur independently between nucleotides. This results in a 4x4 mutation matrix, where each entry describes the instantaneous rate of change between a pair of nucleotides for the most recent common ancestor at that site. This is converted to a 64 × 64 codon mutation matrix μ where 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. [6]

Protein Synthesis Cost-Benefit Function η:

Our model 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. This

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 physiochemical 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 physiochemical 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 physiochemical space. We assume that $G_p \sim \text{Gamma}\,(\alpha = \alpha_G, \beta = \alpha_G)$ in order to ensure $\mathbb{E}(G_p) = 1$. [7] At the limit of $\alpha_G \to \infty$, the model collapses to a model with uniform sensitivity

[6] mikeg: Jeremy, please update for UNREST.

[7] mikeg: I just realized we had a mistake here. I had written that $\beta = 1/\alpha_G$. I'm hoping that Jeremy did not implement this incorrectly. If he did, I'm hoping his implementation was the equivalent of the 'shape' and 'scale' formulation, not the 'shape' and 'rate' formulation we refer to above.

of $G_p = 1$ for all positions p. $\mathbf{B}(\vec{a}_i|\vec{a}_*)$ is inversely proportional to the average physiochemical 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,

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

$$= C_1 + C_2 n \tag{3}$$

where, C_1 and C_2 represent the direct cost, in high energy phosphate bonds, of ribosome initiation and peptide elongation, respectively, where $C_1 = C_2 = 4 \sim P$. [8]

Defining Physiochemical Distances : Assuming that functionality declines with an amino acid a_i's physiochemical 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, our approach naturally lends itself to model selection since we can compare the quality of our model fits using different mixtures of physiochemical 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 physiochemical 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.

[8] mikeg: Jeremy, can we let C_1 vary as a factor of C_2 and then refit the model?. Answer: leave for later.

Ultimately, the scale of gene expression is affected by how we measure physiochemical distances which, in turn, is determined by our choice of Grantham weights. As a result, we set $\alpha_v = 4 \times 10^{-4}$, 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. [9] 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 equivalent 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

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

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

[9] mikeg: Jeremy, is this correct? If not, please correct.

Correspondingly, the ratio of fitness between two genotypes is,

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

Given our formulations of **C** and **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

$$W_i/W_j = \exp\left\{-q\left(C_1 + C_2 n\right)\frac{1}{n}\sum_{p\in\mathbb{P}}\left[d\left(a_{i,p}, a_{*,p}\right) - d\left(a_{j,p}, a_{*,p}\right)\right]\psi\right\} \qquad (5)$$

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,

$$\begin{aligned} u_{i,j} &= \frac{1 - \left(W_i/W_j\right)^b}{1 - \left(W_i/W_j\right)^{N_e}} \\ &= \frac{1 - \exp\left\{-\frac{q}{n}\left(C_1 + C_2 n\right)\left[d\left(a_i, a_*\right) - d\left(a_j, a_*\right)\right]\psi\, b\right\}}{1 - \exp\left\{-\frac{q}{n}\left(C_1 + C_2 n\right)\left[d\left(a_i, a_*\right) - d\left(a_j, a_*\right)\right]\psi\, 2N_e\right\}} \end{aligned}$$

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 XXXXX unique rates, based on few parameters (one to six mutation rates, two free Grantham weights, the cost of protein production **C**, the target functionality synthesis rate, and optimal amino acid at each site), which we infer from the data. Our model can be generalized to allow transitions between optimal amino acids as well as between codons, which would result in a $(21 \times 64) \times (21 \times 64) = 1344 \times 1344$ matrix.

While the overall model does not assume equilibrium, we still need to scale our substitution matrices $\mathbf{Q}$. Traditionally, it is rescaled such that at equilibrium, one unit of branch length represents one expected substitution per site. In our case, we want to do this scaling across all the matrices, since the branch lengths are used in common across the gene. One wrinkle is that this must be done taking optimal amino acid frequency into account. Because we are using 21 64×64 matrices rather than a single 1344×1344 matrix, our scaling done jointly across all the 21 matrices to allow branch lengths under the fixed optimal amino acid model to ensure that branch lengths are comparable. We calculate from the data a vector of 1344 empirical frequencies, π for each of the 64 codons observed when assuming each of 21 possible as the optimal amino acid (including stop codons). A scaling factor is then calculated as the average rate $-\sum_i \mu_i \pi_i = 1$, where i indexes a particular codon for a particular optimal amino acid. The final substitution-rate matrix is the original substitution-rate matrix multiplied by this scaling factor. This matrix can then be applied to all the sites to calculate the likelihood. Finally, in the usual manner the diagonal elements of $\mathbf{Q}$ defining substitution for a given optimal amino acid are fixed so that the rows sum to zero, which allows for $P(t) = \exp\left[\mathbf{Q}_a t\right]$ to be calculated for the matrix π_j whose elements give the probabilities that codon j replaces codon i after time t.

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 an individual site is calculated as

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

In this case, the data, $\mathbf{D}$, are the observed codon states at the tips of a phylogeny $\mathbf{T}$, whose topology are known. The pruning algorithm of Felsenstein (1981) is used to calculate $\mathbf{L}(\mathbf{Q}_a)$. The log likelihood is maximized by estimating the global parameters: $C\,q\,N_e$, 8 mutation parameters which are scaled by $2N_e/b$, and two Grantham distance parameters, α_c and α_p, and the sensitivity distribution parameter α_G. For each gene, we also estimate its target functionality synthesis rate ψ and the optimal amino acid for each position in the protein. When estimating α_G, the likelihood then becomes the average likelihood which we calculate using Laguerre quadrature with $k = 8$ points (Yang, 1994; Felsenstein, 2001).

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. 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.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. *In* H. N. Munro, ed., Mammalian Protein Metabolism, vol. III. Academic Press, New York, 21–132.

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., 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.

Tavare, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *In* R. M. Miura, ed., Some Mathematical Questions In Biology: DNA Sequence Analysis, vol. 17 of *Lectures On Mathematics In The Life Sciences*. The American Association For The Advancement Of Science, American Mathematical Society, Providence, RI.

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.