SelAC Paper

Abstract

Methods used to infer phylogenies typically assume tractable but biologically naive time-reversible models. We develop a population genetics based model that uses biological parameters (energy cost of protein production, amino acid properties, and more) to create more realistic models for DNA substitution for protein-coding genes. These models incorporate selection on amino acids. 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.

Main Text

Introduction Phylogeny inference has become key to understanding processes in ecology, evolution, paleontology, medicine, conservation, and more. This is reflected in the expanding number of studies seeking to infer phylogenies (almost LARGE NUMBER studies per year (CITATION OF SOURCE – maybe Stolzfus et al.?)). However, though the scale and impact of these studies is increasing, the realism of the models used to infer the trees is fairly constant. In a recent sample of XXXXX studies inferring phylogenetic trees, VERY LARGE percent of them made use of simple Markov models that ignore biologically relevant factors such as differences in amino acid properties, selection on proteins, and species not being at equilibrium. There are models that deal more realistically with biological properties: LIST WITH SHORT DESCRIPTIONS. Here we propose a further advance to allow incorporation of amino acid physiochemical properties, selection on protein expression, selection-mutation balance, and other biological parameters into practical models for inferring phylogenies. An additional benefit is the inference of some of these properties from data gathered for inference.

Mutation The overall structure of our model involves a codon mutation model combined with a selection model on the codons based on the relative fitness (coming from explicit models of cost and benefits) at a site of the amino acid the codon encodes. We begin by defining a time reversible model for mutation rates between individual bases. We use existing substitution models (Jukes Cantor **?** or General Time Reversible **?**) to specify which rates are constrained to be equal for a mutation model. This results in a 4x4 mutation matrix, where each entry describes the instantaneous rate of change between a pair of nucleotides. This is converted to a 64x64 codon mutation matrix μ where entries $\mu_i j$ describe the mutation rate from codon i to j by making two assumptions: mutations are independent between nucleotides within a codon and that for any pair of codons that differ by more than one nucleotide, the instantaneous transition rate is zero, since changes involving two or more nucleotides during time deltaT have probabilities on the order of δt^2. Both assumptions are common in codon models **?**.

Benefit Benefits come from relative fitness of the proteins encoded by the sequence. At a given site, we set $\text{Benefit}(\vec{a}^*) = 1$; thus, $\text{Benefit}(\vec{a}_i|\vec{a}^*) \leqslant 1$ for all $\vec{a}_i$ other than the optimal sequence. We assume that the relative fitness of an overall amino acid sequence is the sum of the relative fitnesses at each site and that amino acids are independent. We also initially assume that each site is of equal importance in calculating fitness. This is obviously untrue for most proteins: for example, for a catalytic enzyme, we expect changes at an active site to matter much more for fitness than changes elsewhere in the molecule. However, we relax this assumption later.

For a protein differing from the optimal one at a particular amino acid, we expect the decline in fitness to correlate with magnitude of the difference in properties of the differing amino acid: changing one large hydrophobic amino acid to another probably has a smaller effect than changing it to a small, hydrophilic amino acid. Following **?**, 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 can be used (preliminary investigations using alternate properties found that the **?** properties provided the best fit overall). We use the Euclidian distance between residue properties where each property c, p, and v has its own estimated weighting term.

Under these assumptions,

$$\text{Benefit}(\vec{a}_i|\vec{a}^*) = \left(\frac{1}{n_g} \sum_{p}^{n_g} f\left(d\left(a_{i,p}, a_p^*\right)\right) \right)^{-1} \tag{1}$$

where n_g is the length of the protein, $d(a_{i,p}, a_p^*)$ is the 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, and $1/f(d)$ describes how the contribution of amino acid to protein function declines with d. The shape of the relationship between fitness, $f(d)$, and physiochemical distance, d, is generally unknown, so following **?** and **?** we use a Taylor series expansion to allow the data to define this curve.

Cost Generally speaking, protein synthesis involves both direct and indirect assembly costs. Direct costs consist of the high energy phosphate bonds in 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. Indirect costs consist of the cost of amino acid synthesis as well as synthesis of the protein assembly infrastructure such as ribosomes, aminoacyl-tRNA synthetases, tRNAs, mRNAs, etc. Direct synthesis costs are the same for all proteins of the same length. For simplicity, in this study we ignore any indirect costs of protein synthesis that vary between genotypes. As a result,

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

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

where, C_1 and C_2 represent the direct and indirect costs in ATPs of ribosome initiation and peptide elongation, respectively. For simplicity, in this study we only consider the direct costs of protein assembly and, thus, $C_1 = C_2 = 4\text{ATP}$.

Under our model, each protein encoded within a genome carries out some beneficial function and that the organism needs that functionality to be produced at a target average rate ϕ (it has to catalyze a certain number of reactions per minute, for example) and that the organism regulates protein expression such that this target rate is achieved. As a result, the average protein production rate of a gene, ψ, is equal to $\phi/\text{Benefit}(\vec{a})$ and the total energy flux allocated towards meeting the target functionality of a particular gene is $\eta(\vec{c})\phi$. Every additional ATP spent per unit time to meet the organism's target function production rate ϕ leads to some slight proportional incremental decrease in fitness W, which implies

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

where q describes the decline in fitness with every ATP wasted per unit time ϕ and ψ are measured in. In other words, the cost of a suboptimal protein comes from the need to produce more proteins to get the

same overall function; this model of course breaks down when examining alleles that result in complete loss of function of a protein. Given that protein coding genes conserved across multiple species likely produce functional proteins, this assumption may not limit generality too severely.

Substitution We use a Fisher-Wright **?** model with weak mutation (only one difference between alleles at any instant in time within a population) and the fixation model of **?** to infer fixation probabilities given the cost and benefits above. The ratio of fitness between two genotypes is

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

Given our model of the Cost and Benefit functions above, this ratio simplifies to

$$W_i/W_j \approx \left[-q\left(C_1 + C_2 n\right)\frac{1}{n}\left(\sum_{p\in\mathbb{P}}\sum_{k=1}^{k_{\max}} A_k\left(d\left(a_{i,p}, a_p^*\right)^k - d\left(ajp, a_p^*\right)\right)\right)\phi\right] \tag{5}$$

where $\mathbb{P}$ represents the codon positions in which $\vec{c}_i$ and $\vec{c}_j$ differ. Given our 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 model, the probability a new mutant j introduced via mutation into a resident population i with effective size N_e will go to fixation is,

$$\begin{aligned} u_{i,j} &= \frac{1 - \left(W_i/W_j\right)^b}{1 - \left(W_i/W_j\right)^{N_e}} \\ &\approx \frac{1 - \exp\left[-bq\left(\eta(\vec{c}_i) - \eta(\vec{c}_j)\right)\phi\right]}{1 - \exp\left[-q\left(\eta(\vec{c}_i) - \eta(\vec{c}_j)\right)\phi 2N_e\right]} \\ &= \frac{1 - \exp\left[-bq\left(C_1/n + C_2\right)\left(\sum_{k=1}^{k_{\max}} A_k\left(d(\,a_{i,p}, a_p^*\right)^k - d\left(ajp, a_p^*\right)^k\right)\right)\phi\right]}{1 - \exp\left[-q\left(C_1/n + C_2\right)\left(\sum_{k=1}^{k_{\max}} A_k\left(d(\,a_{i,p}, a_p^*\right)^k - d\left(ajp, a_p^*\right)^k\right)\right)\phi 2N_e\right]}, \end{aligned}$$

where $b = 1$ for a diploid population and 2 for a haploid population **???**. Finally, assuming a constant mutation between alleles i and j, $\mu_{i,j}$, the transition rate from allele i to j can be modeled as,

$$q_{i,j} = \begin{cases} 2\mu_{i,j}N_e/b & \text{if alleles } i \text{ and } j \text{ differ by one nucleotide} \\ 0 & \text{else} \end{cases}$$

In the end, each optimal amino acid has a separate 64 x 64 substitution rate matrix, 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 a rich set of 21 such matrices (20 for the amino acids, and one for stop codons), based on few parameters (one to six mutation rates, three weights on physiochemical distances, the cost of protein production, target functionality, and optimal amino acid at each site), which can be inferred from the data. Future work will allow transitions between optimal amino acids as well as between codons, which would result in a 21 x 64 = 1344 by 1344 matrix. In the meantime, however, amino acids represented by six codons, even if they are not all within one mutational step of each other, like Leucine (L), are assumed to share a single substitution-rate matrix.

Simulations

Empirical data

Discussion and Conclusion

Original Introduction

1. Phylogenetic methods play an important role in many fields of biology and medicine.
2. Essentially all phylogenetic approaches use a substitution matrix $Q = \{q_{i,j}\}$ to model evolution, where

$$q_{i,j} = \text{Substitution rate from state } i \text{ to } j.$$

3. Most models, e.g. F84, GTR, and GY94, use a 'time reversible matrix' where $\pi_i q_{i,j} = \pi_j q_{i,j}$ for all $i \neq j$.

4. TRM were initially derived under the assumption of neutrality but have been extended heuristically to describe non-neutral evolution.

5. However, non-neutral evolution is not a time reversible process, thus TRM models are unlikely to accurately describe evolutionary behavior when natural selection occurs. To illustrate the disconnect between time reversible models and non-neutral evolution we use a simplified version of the extremely popular Goldman and Yang (1994)**?** (GY94) codon level model. In their model,

$$q_{i,j} = \begin{cases} 0 & i \text{ and } j \text{ differ by more than one substitution} \\ \hat{\pi}_j & \text{Synonymous (S) substitution} \\ \omega\hat{\pi}_j & \text{Non-Synonymous (NS) substitution} \end{cases}$$

Where,

$$\omega = \text{'Selection' term applied to all NS substitutions}$$
$$\hat{\pi}_j = \text{Equilibrium frequency of codon } i$$

When $\omega < 1$ the GY94 model is purported to describe evolution under 'purifying' selection where S substitutions are favored over NS substitutions. However, the model has the following behavior

(a) If i is the current state, GY94 implies selection favoring i.

(b) However, if NS substitution occurs, **??** still applies and selection now favoring new state j!

Thus, the behavior of GY94 is actually not consistent with a constant selective environment, but instead is consistent with a system where the directionality of natural selection and a NS substitution occurs simultaneously. Similar inconsistencies occur when $\omega < 1$.

6. The counter argument to the fact that non-neutral evolution is not time reversible is the observation that TRM do a good job reconstructing phylogenetic trees. However, since the results of TRM models are rarely compared to non-TRM models (see [CITATIONS] for notable exceptions), how well they perform relative to more realistic models is an open question.

7. In this study we develop a non-TRM (NTRM) model where the substitution rate of an allele is based on the substitution probability of an allele in the presence of selection for reducing protein synthesis

costs and genetic drift, per standard models of population genetics.

8. In developing our model, we assume that for each protein coding gene there is a single amino acid sequence which executes its intended function better than any other sequence, i.e. is optimal. We also assume that the functionality of other amino acid sequences declines as the physiochemical properties of the sequence deviates from that of the optimal sequence.

9. We describe how functionality declines with physiochemical distance using a Taylor series expansion and a set of weighting terms, which we estimate.

10. Because we assume that a protein's functionality is a declining function of the product of the physiochemical distances of each of the protein's amino acid from the optimal, we can treat the evolution at each amino acid position in a site independent manner. An approach which is almost universally used in TRM models.

11. As a result, unlike most phylogenetic approaches, our model requires 20 different 20x20 rate matrices, one for when each amino acid is the optimal one.

12. Even though our model requires a large number of matrices, because of our assumption that a protein's functionality is a declining function of physiochemical distance from the optimum, we are able to parameterize our 20 matrices using only a handful of parameters which we estimate from the data.

13. Two additional key assumption of our model is that (a) the organism has an average target production rate ϕ for the functionality provided by each protein and (b) that protein synthesis is under some form of regulatory control such that the this average functionality production target is met. As a result, the relative rate of protein synthesis increases as the sequence's functionality declines due to deviation from the optimal sequence. This behavior, in turn, means that the energetic cost of protein synthesis for an allele deviating from the optimal sequence increases with the target production rate ϕ. For example, a protein encoding allele which has a 10% reduction in functionality will have the same energetic burden relative to its optimal sequence as a protein encoding allele of similar length which has a 20% reduction in functionality but whose target production rate is 1/2 of the first protein.

14. In its current formulation, our model is only applicable to protein coding sequences. However, it should be applicable to non-coding sequences so long as one has a mapping function between gene sequence and gene function.

Methods

Allele Substitution Model

Defining the Mutation Rate Matrix μ:

We begin by defining a time reversible model for mutation rates among codons, μ, a 64x64 matrix, where entries $\mu_i j$ describe the mutation rate from codon i to j. We seed these rates according to a pre-defined substitution-rate model (e.g., JC, GTR) for a four-state nucleotide model, which describe the instantaneous rate of change from nucleotide i to j. For simplicity we assume that the mutations occur independent between nucleotides within a codon. For codons that differ only by one nucleotide, the rate between codons is equal to the rate between the pair of nucleotides. For any pair of codons that differ by more than one nucleotide, the rates are set to zero, since changes involving two or more nucleotides during time deltaT have probabilities on the order of δt^2.

Defining Protein Synthesis Cost-Benefit Function η:

Because our model assumes that natural selection favors genotypes that are able to meet their metabolic requirements more efficiently than their competitors, our framework centers on the cost-benefit function of a gene g, η_g, and the organisms average target production rate of the functionality provided by gene g, ϕ_g. This is because the average amount of energy an organism spends to met its target functionality for a gene g is $\eta_g \times \phi_g$.

Defining the Cost Function Generally speaking, protein synthesis involves both direct and indirect assembly costs. Direct costs consist of the high energy phosphate bonds in 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. Indirect costs are many and consist of the cost of amino acid synthesis as well as synthesis of the protein assembly infrastructure such as ribosomes, aminoacyl-tRNA synthetases, tRNAs, mRNAs, etc. Direct synthesis costs are the same for all proteins of the same length. For simplicity, in this study we ignore any indirect costs of protein synthesis that vary between genotypes. As a result,

$$\text{Cost}(\vec{c_i}) = \text{Energetic cost of protein synthesis.} \tag{6}$$

$$= C_1 + C_2 n \tag{7}$$

where, C_1 and C_2 represent the direct and indirect costs in ATPs of ribosome initiation and peptide elongation, respectively. When sequence specific costs, such as ribosome pausing times, are included with sequence specific benefits, the probability of a mutant allele fixing is no longer independent of the rest of the sequence. As a result, our site independent assumption is violated and the fitting of our model becomes much more complex. For simplicity, in this study we only consider the direct costs of protein assembly and, thus, $C_1 = C_2 = 4\text{ATP}$.

Defining the Benefit Function: In order to link genotype to protein function, we define a benefit function which measures the functionality of the peptide encoded by $\vec{c}_i$, i.e. $a(\vec{c}_i) = \vec{a}_i$ relative to the the optimal sequence $\vec{a}^*$. By definition, we set $\text{Benefit}(\vec{a}^*) = 1$ and assume $\text{Benefit}(\vec{a}_i|\vec{a}^*) < 1$ for all $\vec{a}_i$ other than the optimal sequence. How protein functionality declines with deviation from $\vec{a}^*$, is an overwhelmingly complex problem and likely varies between different categories of proteins. Instead of claiming to accurately model this relationship between genotype and protein function, we will fit a Taylor Series expansion to our data in order to approximate its general behavior. We will also assume a form that results in independent evolution between sites within a gene. Alternative forms of $\text{Benefit}(\vec{a}_i|\vec{a}^*)$ can, of course, be explored by other researchers.

To begin, we assume that each amino acid makes a similar contribution to protein function (an assumption that can be relaxed) and that this contribution declines as an inverse function of physiochemical distance. More specifically, we assume

$$\text{Benefit}(\vec{a}_i|\vec{a}^*) = \left(\frac{1}{n_g} \sum_{p}^{n_g} f\left(d\left(a_{i,p}, a_p^* \right) \right) \right)^{-1} \tag{8}$$

where n_g is the length of the protein, $d(a_{i,p}, a_p^*)$ is the 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, and $1/f(d)$ describes how the contribution of amino acid to protein function declines with d.

How $f(d)$ changes with d is unknown, as a result we use a combination of a Taylor Series expansion and random effect to describe the relationship between f and d. Given our assumption that $\text{Benefit}(\vec{a}^*) = 1$ and

noting that d has its own free parameters α and β, we define $f(d)$ as,

$$f(d) = 1 + g \sum_{k=1}^{k_{\max}} \frac{1}{k!} \frac{df^k}{d^k d} d^k + O(d^{k_{\max}+1}) \tag{9}$$

$$= 1 + g \sum_{k=1}^{k_{\max}} A_k d^k + O(d^{k_{\max}+1}) \tag{10}$$

$$\tag{11}$$

where we define $A_k = \frac{1}{k!} \frac{df^k}{d^k d}$ in order to emphasize the polynomial nature of our approximation to $f(d)$ and use g to represent a random effect. Here we assume $g \sim \text{Gamma}\,(\alpha_g, \beta_g = 1/\alpha_g)$ in order to ensure $\mathbb{E}(g) = 1$, but other functions could be used.

Because ϕ and A_1 always co-occur, we cannot identify them separately from one another; as a result, we set $A_1 = 1$ and recognize that our estimates of ϕ are scaled relative to this term. Using the results from **?** and **?**, we can ensure that $f(d)$ is a monotonic, increasing function of d by fitting our model using a transformation of variables α and β and by restricting $k_{\max}$ to multiples of 2. (Note that because $d > 0$, $f(d)$ is monotonic and increasing when $k_{\max} = 1$.)

Defining Physiochemical Distances between Amino Acids : 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 linking comparing genotypes 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 **?**, 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, β, γ, respectively, [NOTE: WE MAY WANT TO USE $\alpha_c, \alpha_p, \ldots$ INSTEAD]. Because of similar identifiability issues we have with A_1 and ϕ, we set $\gamma = 1$ and recognize that our our estimates of α and β are scaled relative to γ. More specifically,

$$d(a_i, a^*) = \sqrt{\alpha \left(c\left(a_i\right) - c\left(a^*\right)\right)^2 + \beta \left(p\left(a_i\right) - p\left(a^*\right)\right)^2 + \gamma \left(v\left(a_i\right) - v\left(a^*\right)\right)^2}.$$

Linking Cost of Protein Synthesis to Allele Fixation

In order to link the protein synthesis cost-benefit function η of an allele with its fixation probability, we must make a number of assumptions. First, we assume that each protein encoded within a genome carries out some beneficial function and that the organism needs that functionality to be produced at a target average rate ϕ. 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 ϕ. As a result, the average protein production rate of a gene, ψ, is equal to $\phi/\text{Benefit}(\vec{a})$ and the total energy flux allocated towards meeting the target functionality of a particular gene is $\eta(\vec{c})\phi$. Third, we assume that every additional ATP spent per unit time to meet the organism's target function production rate ϕ leads to some slight proportional incremental decrease in fitness W. This assumption, in turn, implies

$$W_i\left(\vec{c}\right) \propto \exp\left[-q\eta(\vec{c}_i)\phi\right]. \tag{12}$$

where q describes the decline in fitness with every ATP wasted per unit time ϕ and ψ are measured in. Correspondingly, the ratio of fitness between two genotypes is,

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

Given our assumptions about the Cost and Benefit functions above, this ratio simplifies to

$$W_i/W_j \approx \left[-q\left(C_1 + C_2 n\right)\frac{1}{n}\left(\sum_{p\in\mathbb{P}}\sum_{k=1}^{k_{\max}} A_k\left(d\left(a_{i,p}, a_p^*\right)^k - d\left(ajp, a_p^*\right)\right)\right)\phi\right] \tag{13}$$

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 model. As a result, the probability a new mutant j introduced via

mutation into a resident population i with effective size N_e will go to fixation is,

$$\begin{aligned} u_{i,j} &= \frac{1-(W_i/W_j)^b}{1-(W_i/W_j)^{N_e}} \\ &\approx \frac{1-\exp\left[-bq\left(\eta(\vec{c}_i)-\eta(\vec{c}_j)\right)\phi\right]}{1-\exp\left[-q\left(\eta(\vec{c}_i)-\eta(\vec{c}_j)\right)\phi 2N_e\right]} \\ &= \frac{1-\exp\left[-bq\left(C_1/n+C_2\right)\left(\sum_{k=1}^{k_{\max}} A_k\left(d(\,a_{i,p},a_p^*\right)^k - d\left(ajp,a_p^*\right)^k\right)\phi\right]}{1-\exp\left[-q\left(C_1/n+C_2\right)\left(\sum_{k=1}^{k_{\max}} A_k\left(d(\,a_{i,p},a_p^*\right)^k - d\left(ajp,a_p^*\right)^k\right)\phi 2N_e\right]}, \end{aligned}$$

where $b = 1$ for a diploid population and 2 for a haploid population **???**. Finally, assuming a constant mutation between alleles i and j, $\mu_{i,j}$, the transition rate from allele i to j can be modeled as,

$$q_{i,j} = \begin{cases} 2\mu_{i,j}N_e/b & \text{if alleles } i \text{ and } j \text{ differ by one nucleotide} \\ 0 & \text{else} \end{cases}$$

In the end, each optimal amino acid has a separte 64 x 64 substitution-rate matrix, which incorporates selection for the amino acid as well as the common mutation parameters across optimal amino acids. Thus, have overall 21 of such matrices, 20 for the amino acids, and one for stop codons. Future work will allow transitions between optimal amino acids as well as between codons, which would result in a 21 x 64 = 1344 by 1344 matrix. In the meantime, however, amino acids represented by six codons, even if they are not all within one mutational step of each other, like Leucine (L), are assumed to share a single substitution-rate matrix.

While the overall model does not assume equilibrium, we need to scale the overall matrix in some way. 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. Here the scaling is done jointly across all the 21 matrices to allow branch lengths under the fixed optimal amino acid model to be comparable to the branch lengths under the global model. 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 $-\sum_i \mu_i i * \pi_i = 1$, 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.

Likelihood calculation on a tree:

Given our assumption of independent evolution among sites, the probability of the whole data set is the product of the probabilities of data at each individual sites. Thus, the log likelihood is taken as

$$putlikelihoodequationhere$$

The log likelihood is maximized by estimating the combined parameter for $C * q * \Phi$, Ne, two of the three Grantham distance parameters, α_c, α_p (again, we hold $\alpha_g amma$ constant – see above), the free mutation rate parameters (i.e., five free parameters if assuming GTR) and their three free nucleotide frequency parameters, π_i, given an alignment and a fixed tree topology. We also assume each optimal acid across all sites as free parameters to be estimated in the model. There are two ways in which we estimate the optimal amino acid at a given site: 1) use the majority rule of the amino acids observed across different species at a homologous site, and 2) numerically optimizing the choice of optimal amino acid at each site.

In the case of including a random effect as described in Eq.(5) by specifying a discrete gamma, the log likelihood function becomes,

$$putgammalikelihoodequationhere$$

where k specifies the number of discrete categories. Note that this would add an additional free parameter, α_g, which describes the shape of the distribution.

Results

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 the standard GTR model.

2. In addition, we are able to generate estimates of gene expression which are well correlated with empirical estimates.

Lots of other stuff

Discussion

- Note that our definition of ϕ and our scaling of functionality differ slightly from our previous work (**????**). In our previous work, we were concerned with how changes in synonymous codons affected error rates and synthesis costs and, as a result, defined functionality relative to an error free protein, rather than an optimal one, and conflated ϕ and ψ.

- Our approach requires relatively few parameter.

    1. Distance function $d(a_i, a^*)$: If n_d is the number of physiochemical properties examined, the number of parameters estimated is $n_d - 1$
    2. Benefit function Benefit: If n_A is the order of our Taylor Series approximation, the number of parameters is $n_A - 1$.
    3. Gene expression ϕ: One ϕ for each gene analyzed.
    4. Mutation bias: Depends on the model used it is either equal to the number of parameters in the model n_μ or $n_\mu - 1$.

- Our approach can be expanded by allowing the optimal amino acid to change during the course of evolution. This should allow us to use a large, single matrix 400 x 400 matrix instead of 20 separate 20x20 matrices. Further, if we may be able to compare the statistical properties of this extended transition matrix to the single transition matrix used in other approaches.

- Statistical Physics model allows decoupling of N_e, μ, and strength selection.

References

Elphinstone, C. 1985. A method of distribution and density estimation. Ph.D. thesis, University of South Africa.

Liang, L. 2007. A semi-parametric approach to estimating item response functions. Ph.D. thesis, Ohio State University. Provides approach for creating positive, monotonically increasing polynomials after Elphinstone (1985).

Notes for Jeremy

Log

- "Defining the Benefit Function" section began by mikeg on 7/23/15.
- Methods expanded to include "Defining Benefit Function" on ?
- Linking Genotype Energetics to Fitness and Fixation subsection added to Model.
- Compiled on Saturday 19$^{\text{th}}$ September, 2015 at 23:48.