Population genetics models with selection for phylogenetic inference

Last compiled on Saturday 30th July, 2016 at 11:02.

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, the realism of the models used to infer the trees has changed relatively little by comparison. 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 **?** for an overview). The next set of models attempt to include a 'selection' term ω which is interpreted as describing stabilizing or diversifying selection depending on whether ω is less than or greater than 1, respectively. However, the link between ω and the key parameters found in standard population genetics models of Wright and Fisher or Moran, such as N_e, the distribution of s across genotype space, and mutation bias are far from clear andl likely only describe a subset of scenarios such as over/underdominance or frequency dependent selection (**?**). 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 reconstruction (e.g. **????**).

The field of population genetics is a mature, mathematically rigorous, and well established framework for describing biological evolution. Despite the fact that 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 overwhelmingly complex. In contrast, under the simplifying assumptions of additivity between and within sites, calculating stationary and substitution probabilities are relatively straightforward to calculate. As a result, fitting additive models of the evolutionary process to sequence data is computationally feasible. One major advantage to fitting models derived from population genetics over other approches is that the parameters estimated are biologically meaningful and of great interest to evolutionary biologists. For example in haploid population, if the product of effective population size N_e and the mutation rate of beneficial alleles μ_b is greater than one, then the waiting time for an adaptive allele to emerge after an environmental shift is much shorter than the substitution process itself and evolution itself is not mutation limited. Here we show how fitting a phylogenetic model derived from a population genetic model can be used to estimate $N_e\mu$ as well as the optimal amino acid sequence of a protein and the average protein synthesis rate of each gene using only coding sequence data from the tips of the tree.

The model we develop can be viewed as a special formulation of the Wright-Fisher (WF) model (**?????**). 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 WF model (e.g. **??????????**). Similar to **??** 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 othe 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 over all the 19 other at a particular position within the protein encoded by a particular gene.

Because our approach requires twenty families of 20×20 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. As a result, our model requires only a handful of parameters that are used genome wide such as nucleotide specific mutation parameters, the effective population size N_e, and side chain physiochemical weighting parameters. In addition to these genome wide parameters, our model requires two gene specific parameters: an expression parameter, ϕ which scale the average strength of stabilizing selection for the optimal amino acid at each position and a variance parameter which describes how the strength of stabilizing selection varies around the mean value between positions. 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 (**???**).

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 **?** studies of how models of molecular and evolutionary scales can be combined together in a nested manner. 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 they vary between sites. We've utilized the same nested, population genetic based approach in more traditional genomic analyses (e.g. **???**). That work and our current work illustrates how more information can be extracted from sequence data when more biologically based models are used. While the mapping between genotype and phenotype is more abstract than **?**, our approach has the advantage of not requiring a model of a protein's physical structure. By linking the strength of stabilizing selection to gene expression, specifically the average synthesis rate of a protein, we can weight the historical information encoded in different genes in a biologically defensible manner while simultaneously estimating their expression levels.

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.

3. By linking q to gene expression, our approach is an alternative to the more common approach of simply concatenation of protein coding sequences. More specifically, in our model simply concatenating gene sequences together is equivalent to assuming the same aver protein synthesis rate ϕ for all of the genes. By assuming the strength of stabilizing selection for the optimal sequence, $\vec{a}^*$, is proportional to ϕ, our model allows us to estimate ϕ for each gene. Our results clearly indicate that this information is

available and accounting for it in our model substantially improves our model fit.

Lots of other stuff

Discussion

- One important manner in which our approach moves beyond **?????** is in that we parameterize the model and fit branch lengths simultaneously rather than in two separate steps.[1]
- Assumptions of additivity and no epistasis are unrealistic but can be viewed as a first order approximation to these more complex scenarios and a starting point for later relaxing these assumptions.
- Implicitly, our model assumes that all genes are essential because an organism that is homozygous for alleles with zero activity (i.e. no benefit) would have to spend an infinite amount of energy to achive a target functionality production rate $\phi > 0$. This assumption can be relaxed by allowing ϕ to vary and fitness is function of that ϕ level and where there is an optimal ϕ level.
- 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 400 x 400 matrix instead of 20 separate 20x20

[1] 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?

matrices. The new elements would contain the rates at which the optimal amino acid at site switches from its current state to another amino acid. 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.

- We conjecture the stationary properties of the 400x400 matrix mentioned above should map to an expected 20 x 20 empirical matrix where we average across when each of the different amino acid are optimal.
- Statistical Physics model allows decoupling of N_e, μ, and strength selection.

Additional Points That Need to Be Mentioned

- In this study we develop a model where the substitution rate of an allele is based on the substitution probability of an allele under selection, mutation bias, and genetic drift, per standard models of population genetics.
- 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 assume the strength of selection for the optimal sequence increases with the target synthesis rate of the functionality the gene provides. That is genes with higher target expression levels are under stronger selection than genes with lower target expression levels.
- 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.
- We describe how functionality declines with physiochemical distance using a Taylor series expansion and a set of weighting terms, which we estimate.
- 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 other phylogenetic models.
- 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.

- 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.

- 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 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 production rate is 1/2 of the first protein.

- 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

Mutation Rate Matrix μ:

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 μ_ij 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 δt have probabilities on the order of δt^2. Both assumptions are common in codon models **?**.

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 $\eta_g \times \phi_g$.

Benefit: In order to link genotype to protein function, we first define a benefit function. This benefit function measures the functionality of the amino acid sequence $\vec{a}_i$ encoded by a set of codons $\vec{c}_i$, i.e. $a(\vec{c}_i) = \vec{a}_i$. For simplicity, we define the functionality of $\vec{a}_i$ relative to that of an optimal sequence $\vec{a}^*$. Thus by definition, $\mathbf{Benefit}(\vec{a}^*) = 1$ and, consequently, $\mathbf{Benefit}(\vec{a}_i|\vec{a}^*) < 1$ for all $\vec{a}_i$ other than the optimal sequence. While we assume that all amino acids contribute to protein function and that this contribution declines as an inverse function of physiochemical distance. More specifically,

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

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 we fit a Taylor series approximation of $f(d)$ based on the data. In order to allow for different sites to have different contributions to **Benefit**, we include a random effects term g in our definition of $f(d)$. Given our assumption that $\mathbf{Benefit}(\vec{a}^*) = 1$ and noting that d has its own free parameters $\vec{\alpha} = \{\alpha_c, \alpha_p, \alpha_v\}$, we define $f(d)$ as,

$$f(d) = 1 + g \sum_{k=1}^{k_{\max}} \frac{1}{k!} \frac{df^k}{d^k d} d^k + O(d^{k_{\max}+1}) \tag{2}$$

$$= 1 + g \sum_{k=1}^{k_{\max}} A_k d^k + O(d^{k_{\max}+1}) \tag{3}$$

$$\tag{4}$$

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 assume $g \sim \text{Gamma}\left(\alpha_g, \beta_g = 1/\alpha_g\right)$ in order to ensure $\mathbb{E}(g) = 1$. In addition, because ϕ and the A_k terms always co-occur in our model, 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 $d > 0 \forall a \neq a^*$, $f(d)$ is monotonic and increasing when $k_{\text{max}} = 1$.)

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 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,

$$\begin{aligned} \mathbf{Cost}(\vec{c}_i) &= \text{Energetic cost of protein synthesis.} \qquad (5) \\ &= C_1 + C_2 n \qquad (6) \end{aligned}$$

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, $C_1 = C_2 = 4\,\text{ATP}$.

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_c, α_p, α_v, respectively. Because of similar identifiability issues we have with A_1 and ϕ, we set $\alpha_v = 1$ and recognize that our our estimates of α_c and α_p are scaled relative to 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

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 $\phi/\mathbf{Benefit}(\vec{a})$ and the total energy flux allocated towards meeting the target functionality of a particular gene is $\eta(\vec{c})\phi$. In other words, the cost of a suboptimal protein comes from the need to produce more proteins to get the same overall functionality.

Third, we assume that every additional ATP spent per unit time to meet the organism's target function production rate ϕ leads to a slight and proportional decrease in fitness W. This assumption, in turn, implies

$$W_i\left(\vec{c}\right) \propto \exp\left[-q\eta(\vec{c}_i)\phi\right]. \tag{7}$$

where q describes the decline in fitness with every ATP wasted per unit time used to measure ϕ and ψ.

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

$$W_i/W_j \approx \exp\left[-q\left(C_1 + C_2 n\right)\frac{1}{n}\left(\sum_{p\in\mathbb{P}}\sum_{k=1}^{k_{\max}} A_k\left(d\left(a_{i,p},a_p^*\right)^k - d\left(a_{j,p},a_p^*\right)^k\right)\right)\phi\right] \tag{8}$$

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,

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

where $b = 1$ for a diploid population and 2 for a haploid population **?????**. Finally, assuming a constant mutation rate between alleles i and j, $\mu_{i,j}$, where $\mu_{i,j} = 0$ when two codons differ by more than one nucleotide, the transition rate from allele i to j can be modeled as,

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

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, three weights on physiochemical distances, the cost of protein production, target functionality, and optimal amino acid at each site), which we infer from the data. [MAG: PREVIOUS TEXT INCLUDED THE STOP CODONS AND REFERRED TO 21 MATRICES (AS IT STILL DOES BELOW). I THOUGHT WE ONLY MODELING THE SENSE CODONS.] [MAG: HOW DO WE DEAL WITH SERINE WHICH HAS TWO DISJOINT SETS OF 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. [MAG: I DON'T FOLLOW THIS.]

While the overall model does not assume equilibrium, we need 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. 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 $-\sum_i \mu_i * \pi_i = 1$, 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 Calculations 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 α_v 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, [MAG PLEASE USE THE STANDARD
labelNAME TO LABEL EQUATIONS AND OTHER OBJECTS AND USE

refNAME TO REFER TO THESE OBJECTS.]

$$putgammalikelihoodequationhere$$

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.

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).