SelAC Paper

Log

- “Defining the Benefit Function” section began by mikeg on 7/23/15.
- Compiled on Thursday 23$^{\text{rd}}$ July, 2015 at 16:24.

Model

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, aminacyl-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,

$$\begin{aligned}\text{Cost}(\vec{c}_i) &= \text{Energetic cost of protein synthesis.} \qquad (1)\\ &= C_1 + C_2 n \qquad (2)\end{aligned}$$

where, C_1 and C_2 represent the direct and indirect costs in ATPs of ribosome initation and peptide elongation, respectively. We note that when sequence specific costs, such as ribosome pausing times, are included with variation in benefit between sequences, the liklihood of an allele with a particular mutation going to fixation is no longer independent of the rest of the sequence and, as a result, our site independent assumption is violated and the fitting of our model becomes much more complex.

Defining the Benefit Function

In order to link genotype to protein function, we define a benefit function Benefit, which describes 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

$$\text{Benefit}(\vec{a}_i|\vec{a}^*) = \left(\frac{1}{n_g}\sum_{p}^{n_g} f\left(d\left(a_{i,p}, a_p^*\right)\right)\right)^{-1} \qquad (3)$$

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, and so we use a Taylor Series expansion 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,

$$\begin{aligned} f(d) &= 1 + d + \sum_{j=2}^{j_{\max}} \frac{1}{j!} \frac{df^j}{d^j d} d^j + O(d^{j_{\max}+1}) && (4) \\ &= 1 + d + \sum_{j=2}^{j_{\max}} A_j d^j + O(d^{j_{\max}+1}) && (5) \\ & && (6) \end{aligned}$$

where we define $A_j = \frac{1}{j!} \frac{df^j}{d^j d}$ in order to emphasize the polynomial nature of our approximation to $f(d)$. Using the results from Liang (2007) and Elphinstone (1985), 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 $j_{\max}$ to multiples of 2. (Note that because $d > 0$, $f(d)$ is montonic and increasing when $j_{\max} = 1$.)

Linking Genotype Energetics to Fitness

If we assume that every ATP saved per unit time leads to some slight incremental increase in fitness W, then

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

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

$$\begin{aligned} W_i/W_j &= \exp\left[-q\eta(\vec{c}_i)\phi\right] / \exp\left[-q\eta(\vec{c}_j)\phi\right] && (8) \\ &= \exp\left[-q\left(\eta(\vec{c}_i) - \eta(\vec{c}_j)\right)\phi\right] && (9) \end{aligned}$$

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

$$\begin{aligned} W_i/W_j &= \exp\left[-q\left(C_1 + C_2 n\right)\frac{1}{n}\left(\sum_{p\in\mathbb{P}}\sum_{j=1}^{j_{\max}} A_j d\left(a_{i,p}, a_p^*\right) - d\left(ajp, a_p^*\right)\right)\phi\right] && (10) \\ & && (11) \end{aligned}$$

where $\mathbb{P}$ represents the codon positions in which $\vec{c}_i$ and $\vec{c}_j$ differ.

Notes for Jeremy

Since we are assuming A_0 and $A_1 = 1$, when using the recurrance relation in Lang 2007, we set $\lambda = 1$ and $\alpha_1 = -1/2$.

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).