Nonsynonymous codon substitution — amino acid substitution

March 8, 2012

Consider a protein with a sequence of optimal amino acids $\hat{\mathbf{a}} = (\hat{a}_1, \hat{a}_2, \cdots \hat{a}_n)$ and the observed sequence of amino acids is $\mathbf{a} = (a_1, a_2, \cdots, a_n)$.

At position k, if the observed amino acid is different from the optimal amino acid, there is a selection coefficient s_k for the selection disadvantage. Let $\mathbf{s} = (s_1, s_2, \cdots, s_n)$.

Given the physiochemical distances (Grantham) between amino acids, the optimal protein and the selection strength $\mathbf{s}$, the of functionality of a protein $\mathbf{a}$ is

$$F(\mathbf{a}|\hat{\mathbf{a}}, \mathbf{s}) = \prod_{k=1}^{n} e^{-d(a_k, \hat{a}_k)s_k} = e^{-\mathbf{d}\cdot\mathbf{s}} \tag{1}$$

or other candidates:

$$F(\mathbf{a}) = \prod_{k=1}^{n} \frac{1}{1 + d_k s_k} \tag{2}$$

$$F(\mathbf{a}) = \frac{n}{\sum_{k=1}^{n} (1 + d_k s_k)} \tag{3}$$

$$F(\mathbf{a}) = \frac{n}{\sum_{k=1}^{n} \frac{1}{e^{-d(a_k, \hat{a}_k)s_k}}} \tag{4}$$

The condition $\hat{\mathbf{a}}, \mathbf{s}$ will be omitted from now on if there is no potential confusion.

The fixation probability of a single protein mutant $\mathbf{a}_j$ from a diploid population with wild type $\mathbf{a}_i$, if they differ at only one position k of the protein, is

$$\pi(\mathbf{a}_i \to \mathbf{a}_j) = \frac{1 - f(\mathbf{a}_i)/f(\mathbf{a}_j)}{1 - (f(\mathbf{a}_i)/f(\mathbf{a}_j))^{2N_e}} = \frac{1 - f_i/f_j}{1 - (f_i/f_j)^{2N_e}} \tag{5}$$

according to Sella-Hirsh (Add reference) where $f(\mathbf{a}_i)$ and $f(\mathbf{a}_j)$ are the fitnesses of $\mathbf{a}_i$ and $\mathbf{a}_j$. We assume that the mutation rate between 2 proteins who differ at more than one position is 0.

According to the canonical formula, we have

$$\pi(\mathbf{a}_i \to \mathbf{a}_j, p) = \frac{1 - e^{-2N_e ps}}{1 - e^{-2N_e s}} \tag{6}$$

where p is the initial frequency of the mutant, and $s = (f_j - f_i)/f_i$ is the selection advantage of $\mathbf{a}_j$ comparing to $\mathbf{a}_i$. When there is a single mutant in the population, i.e. $p = 1/(2N_e)$, the formula becomes $\frac{1-e^{-s}}{1-e^{-2N_e s}}$. Both formulae are valid under the same conditions: $s, \frac{1}{N}, Ns^2 \ll 1$.

As in Gilchrist 2007, fitness is a function of cost, functionality and some scaling factors:

$$f(\mathbf{a}) \propto \exp\{-\frac{C\Phi q}{F(\mathbf{a})}\}$$

where C is the expected cost of producing a single complete protein, q is the scaling constant seconds per ATP determining the relationship between the rate of ATP usage and fitness f, and Φ is a measure of gene expression, specifically protein production rate (protein per second). Further explanation of these constants and why the formula is like this?

If we combine $C\Phi q$ as one constant A, then

$$f(\mathbf{a}) \propto \exp\{-\frac{A}{F(\mathbf{a})}\}$$

In either S-H formula or the canonical formula of the fixation probability, the value that is of concern is f_i/f_j, now calculate f_i/f_j:

$$\begin{aligned}
\frac{f(\mathbf{a}_i)}{f(\mathbf{a}_j)} &= \frac{f_i}{f_j} (\text{consider } C \text{ as constant for now}) \\
&= \exp\Big[- \frac{A}{F(\mathbf{a}_i)} + \frac{A}{F(\mathbf{a}_j)} \Big] \\
&= \exp\Big[- A\Big(\frac{1}{F(\mathbf{a}_i)} - \frac{1}{F(\mathbf{a}_j)}\Big) \Big] \\
&= \exp\Big[- \frac{A}{\prod_{l\neq k}^{n} \exp(-d_l s_l)} \Big(\frac{1}{\exp(-d_k^i s_k)} - \frac{1}{\exp(-d_k^j s_k)}\Big) \Big] \\
&= \exp\Big[- \frac{A}{F_S} \Big(\exp(d_k^i s_k) - \exp(d_k^j s_k) \Big) \Big]
\end{aligned} \tag{7}$$

where d_k^i is the distance between the amino acids at position k in $\mathbf{a}_i$ and $\hat{\mathbf{a}}$, which is the only position where $\mathbf{a}_i$ and $\mathbf{a}_i$ differ. F_S is the part of functionality shared by sites of the 2 proteins except site k.

If the functionality is defined as in Equation 3, then we have the following:

$$\frac{f(\mathbf{a}_i)}{f(\mathbf{a}_j)} = \prod_{k=1}^{n} \Big(\frac{f(\mathbf{a}_i^k)}{f(\mathbf{a}_j^k)}\Big)^{\frac{1}{n}} \tag{8}$$

i.e. the fitness ratio of the whole protein is the geometric mean of the fitness ratios between the two proteins for all sites. Therefore, when only one site (say, site k) is allowed to change, we have

$$\frac{f(\mathbf{a}_i)}{f(\mathbf{a}_j)} = \Big(\frac{f(\mathbf{a}_i^k)}{f(\mathbf{a}_j^k)}\Big)^{\frac{1}{n}}$$

and

$$\begin{aligned}
\frac{f(\mathbf{a}_i)}{f(\mathbf{a}_j)} &= \exp\Big[-\frac{A}{n}\Big(\frac{1}{F(\mathbf{a}_i)} - \frac{1}{F(\mathbf{a}_j)}\Big)\Big] \\
&= \exp\Big[-\frac{A}{n}(d_k^i s_k - d_k^j s_k)\Big] \qquad (9) \\
&= \exp\Big[-\frac{C\Phi q}{n}(d_k^i s_k - d_k^j s_k)\Big] \qquad (10) \\
&= \exp\Big[-\frac{C\Phi q s_k}{n}(d_k^i - d_k^j)\Big] \qquad (11)
\end{aligned}$$

this quantity is only related to site k and is independent of all other sites. Therefore, the sites are independent. In this case, the parameters C, Φ, q and s_k are all multiplied together, only C depends on n, so they should be treated as a composite parameter, as inseparable.

Instantaneous substitution rate from $\mathbf{a}_i$ to $\mathbf{a}_j$ is

$$u_{ij} = 2N_e\mu_{ij}\pi(\mathbf{a}_i \to \mathbf{a}_j|\hat{\mathbf{a}}, \mathbf{s}) \qquad (12)$$

where μ_{ij} is the mutation rate from $\mathbf{a}_i$ to $\mathbf{a}_j$. Now that the mutation and fixation are both at protein level, for simplicity, we assume that the mutation rates are all the same, as long as there is only one different position between the two proteins considered.

Questions:

1. Should the distance between a given protein and the optimal protein be a vector, with an entry for each site, or could it be simplified to a number as a (weighted) sum of the entries of the vector?
2. In the functionality expression, do we need to put a scaling constant somewhere?
3. For physiochemical distance, should the mean be 100 as in original Grantham's matrix or 1 as we did in the simulation?

To check if the simulation is done correctly, here are some results. For simplicity, only 3 states (amino acids) are considered now.

1. Check that the way to do simple simulation and to find the stationary probability is correct.

Given the instantaneous substitution rate matrix W for a Markov process with stationary probability vector $\mathbf{p}$, $\mathbf{p}$ can be calculated by taking any row of the matrix $\exp(Wt)$ where t is big enough to guarantee that the process has already reached stationarity.

On the other hand, if W comes from the mutation and fixation processes, then $\mathbf{p}$ can also be found by S-H's formula of stationary probabilities

$$\frac{p_i}{p_j} = (\frac{f_i}{f_j})^{\nu} = \frac{W_{ji}}{W_{ij}} \tag{13}$$

Comparing $\mathbf{p}$ calculated in both ways verifies that S-H's formula is accurate. Next we compare $\mathbf{p}$ calculated from the simulation with that from either of the formulae. The way I found $\mathbf{p}$ is as follows:

1. Do simulations with long enough time so that there are at least $10,000$ substitutions in the simulated chain.
2. Cut off the first 100 observations, before the time when the process reaches stationarity.
3. Find the average time $\bar{t}$ taken for a certain number of substitutions (for example, 10), and record the state of the chain when the amount $\bar{t}$ of time passed.
4. Find the frequency of the observed states, hence the approximation of $\mathbf{p}$.

Following are some results.

1. Let W be the substitution rate matrix between all the protein with 2 sites. Since there are only 3 states for each site, total number of protein is 9. Hence the dimension of the matrix W is 9×9.

   In Figure 1 on page 5 the solid line connects the values of stationary probabilities for each state from theoretical calculation, and the red circles from simulation.

2. Another simulation is done with 2 sites (3 states for each site) for each protein, starting with $(1, 2)$, with optimal protein $(2, 2)$, selection coefficient 0.1, running time 10^9, sites dependent. The stationary probabilities are:

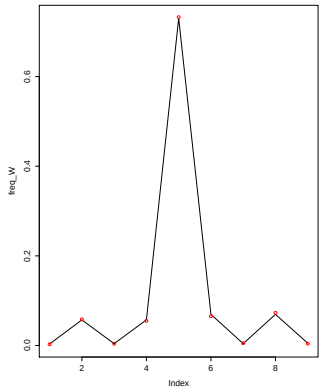

Figure 1: Stationary probabilities with 2 sites

1	2	3	4	5	6	7	
0.003364677	0.057514946	0.004235888	0.058025656	0.727881756	0.069636795	0.004130742	0.070
0.003324449	0.057134781	0.004132447	0.057134781	0.730153540	0.069429813	0.004132447	0.069

First row is from simulation, second row is from Sella-Hirsh's formula. They are very close to each other. The number of steps in this simulation is 669061, which is relatively big.

3. A simulation with 3 sites and 3 states for each site confirmed the correctness of simulation. The chain starts with $(1, 1, 1)$, with optimal protein $(2, 2, 2)$, selection coefficient 0.01, running time 10^8, sites dependent. The number of observations is 216126. Figure 2 on page 6 is a plot of stationary probabilities from both simulations and Sella-Hirsh's formula.

Questions to answer next:

1. How to construct phylogeny from the simulated data?

2. How to approximate the functionality (and other variables of concern) in the site-dependent case using site-independent case?

   In site-independent case, it's equivalent to assuming that every other site is at the optimal amino acid hence the functionality at those sites is 1. When the process has reached stationarity, if Φ is high and selection is strong, then the frequency of optimal amino acids will be high hence the

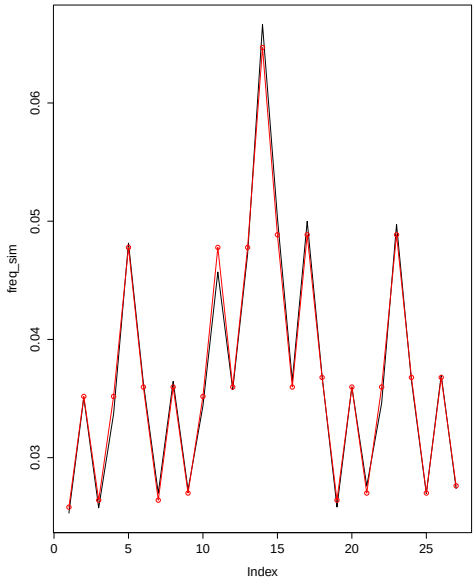

Figure 2: stationary probabilities with 3 sites

assumption is reasonable. However, when Φ is small or selection is very weak, there tend to be more non-optimal amino acids at many sites.

Consider the case when there are only 2 sites (amino acids) in the protein, fix the second site at a certain amino acid and let only site 1 change. How does site 2 affect the stationary distribution at site 1?

From (7),

$$f_{ij} = \frac{f_i}{f_j} = \exp\Big(- \frac{A}{F_2}(e^{d_i s_1} - e^{d_j s_1})\Big) \tag{14}$$

In site-independent case, we assume site 2 is at the optimal amino acid, i.e. $F_2 = 1$, therefore

$$f_{ij}^{\text{ind}} = \exp\Big(- A(e^{d_i s_1} - e^{d_j s_1})\Big) \tag{15}$$

The relationship between the two fitness ratios is

$$f_{ij}^{\text{ind}} = (f_{ij})^{F_2} \tag{16}$$

If site 2 is not at the optimal amino acid, then $F_2 < 1$. Now consider the fitness ratios between an amino acid (aa_i) and the optimal amino acid

(aa_o).

$$f_{io} = \exp\Big(-\frac{A}{F_2}(e^{d_i s_1} - 1)\Big) \quad (17)$$

$$f_{io}^{\text{ind}} = (f_{io})^{F_2} \quad (18)$$

Since $f_{io} < 1$, $f_{io}^{\text{ind}} > f_{io}$, the fixation probability from an arbitraty amino acid to the optimal amino acid is higher in the site-dependent case. In other words, the selection strength at one site is **higher** when other sites are not at the optimal amino acids.

Let's see what happens with stationary distributions under site-independent and -dependent cases. From (13),

$$\frac{p_i}{p_o} = \big(\frac{f_i}{f_o}\big)^{\nu} = (f_{io})^{\nu} \quad (19)$$

Therefore

$$(\frac{p_i}{p_o})^{\text{ind}} = (f_{io}^{\text{ind}})^{\nu} = (\frac{p_i}{p_o})^{F_2} > \frac{p_i}{p_o} \quad (20)$$

This result could easily be generalized when the functionality F_S is known for othere sites and only one site is changing, with F_2 replaced by F_S. Simulations also verified the relation.

Rewriting the equation (7), we can see

$$\frac{f_i}{f_j} = \exp\Big[-A\Big(e^{d_k^i s_k - \ln F_S} - e^{d_k^j s_k - \ln F_S}\Big)\Big] \quad (21)$$

What does this reflect the effect of F_S on selection strength at site k?

The most essential and what we are most interested in are the fixation rates from one protein to another, under both site-dependent and site-independent cases. If this is clear, the substitution rates and stationary distribution will follow, also the mean fitnesses. On the other hand, fixation probability is a function of fitness ratio, which also determines the stationary probabilities.

Now we already know how to express the fitness ratio in site-dependent case in terms of that in site-independent case, with functionality at other sites as exponent, when there is only one site that is different between two proteins.

To get the fitness ratio between any two proteins, we could use proteins that are one site away from them as bridges and represent the ratio as a product of several ratios we already know. For example:

$$\begin{aligned}
\frac{f_{AA}}{f_{BB}} &= \frac{f_{AA}}{f_{AB}} \cdot \frac{f_{AB}}{f_{BB}} \\
\frac{f_{AA}}{f_{AB}} &= (\frac{f_A^2}{f_B^2})^{-F_A} \\
\frac{f_{AB}}{f_{BB}} &= (\frac{f_A^1}{f_B^1})^{-F_B}
\end{aligned}$$

If furthermore, the selection coefficients at the first and second positions are the same, then $\frac{f_A^2}{f_B^2} = \frac{f_A^1}{f_B^1} = \frac{f_A}{f_B}$ in site-independent case. Therefore

$$\frac{f_{AA}}{f_{BB}} = (\frac{f_A}{f_B})^{-F_A - F_B}$$

If there are more than 2 sites that are different, then there need to be more than one intermediate proteins to relate them together. For example,

$$\frac{f_{AAA}}{f_{BBB}} = \Big(\frac{f_A}{f_B}\Big)^{-F_{AA} - F_{AB} - F_{BB}}$$

However, if the selection coefficients are not the same across the sites, this relationship does not hold any more. The reason is that even fitness ratio in site-independent case depends on the selection coefficient at that particular site.

Parameters to investigate during the optimization (optimx):

1. s, Φ, C, q, N_e, some of them are correlated to each other and cannot be split separately
2. tree topologies (at different extremes), branch lengths, number of tips
3. ancestral sequences — stationary, uniform, same sequence as used in the simulation

Hessian matrix, variance, information; confidence interval, bootstrapping; ...

Meeting on Thursday, Mar 8, 1012.

1. Do the plots of MLE bias with respect to number of sites (200, 400, 800 or more in between) : first, fix other parameters, estimate s; second, estimate Φ while fix other parameters. Do this for trees with 4 tips and 8 tips.

2. Estimate the composite parameter.

3. Instead of giving the equilibrium frequencies, make it depend on the parameters as from Sella-Hirsh's formula, and do the optimization that way.

4. Let s vary according to Gamma distribution instead of being fixed, and estimate the Gamma parameter(s).