

Nonsynonymous codon substitution — amino acid substitution

February 13, 2012

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

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 $\mathbf{s} = (s_1, s_2, \dots, s_n)$.

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

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

or other candidates:

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

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

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

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

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

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

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:

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

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

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

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 :

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

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}_j differ. F_S is the part of functionality shared by sites of the 2 proteins except site k .

Should we relax the condition so that the mutation between proteins with more than 1 different position is also possible?

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

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

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

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

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:

	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.

Next step: set the mutation rate to be a different number, add mutation bias, move from amino acid level to codon level.

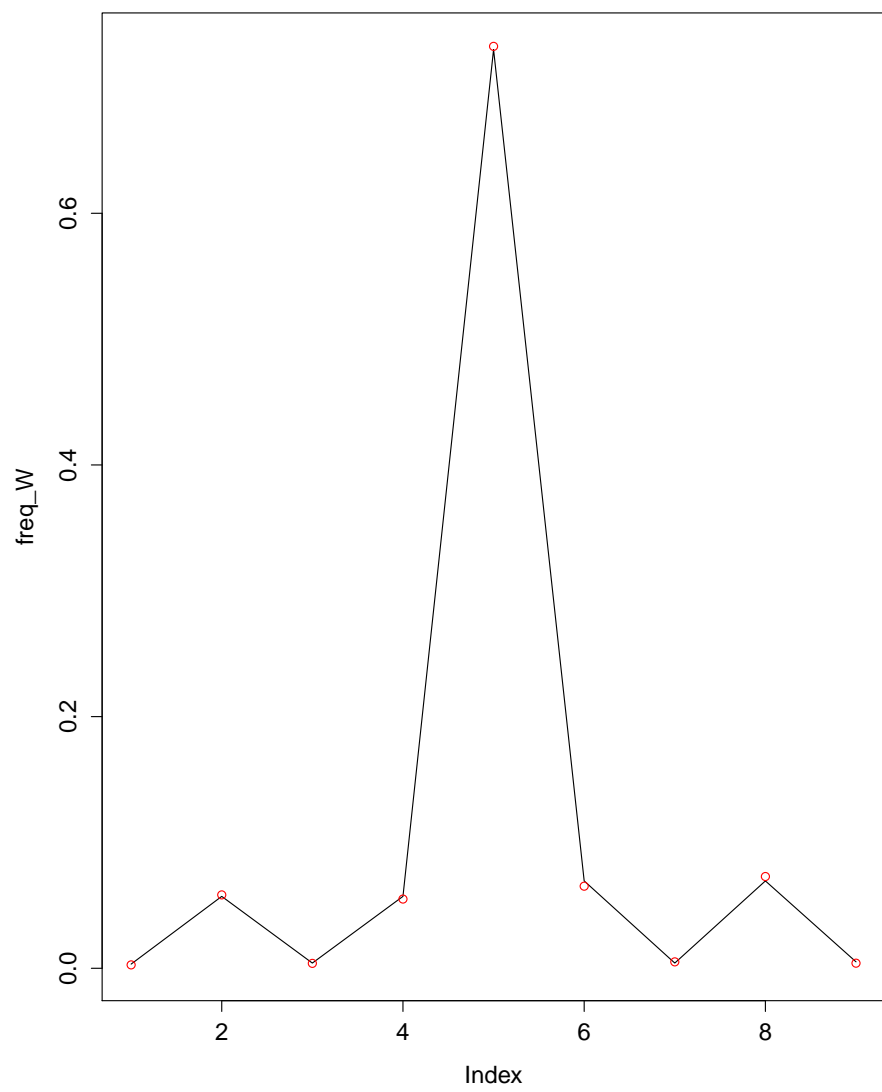


Figure 1: Stationary probabilities with 2 sites

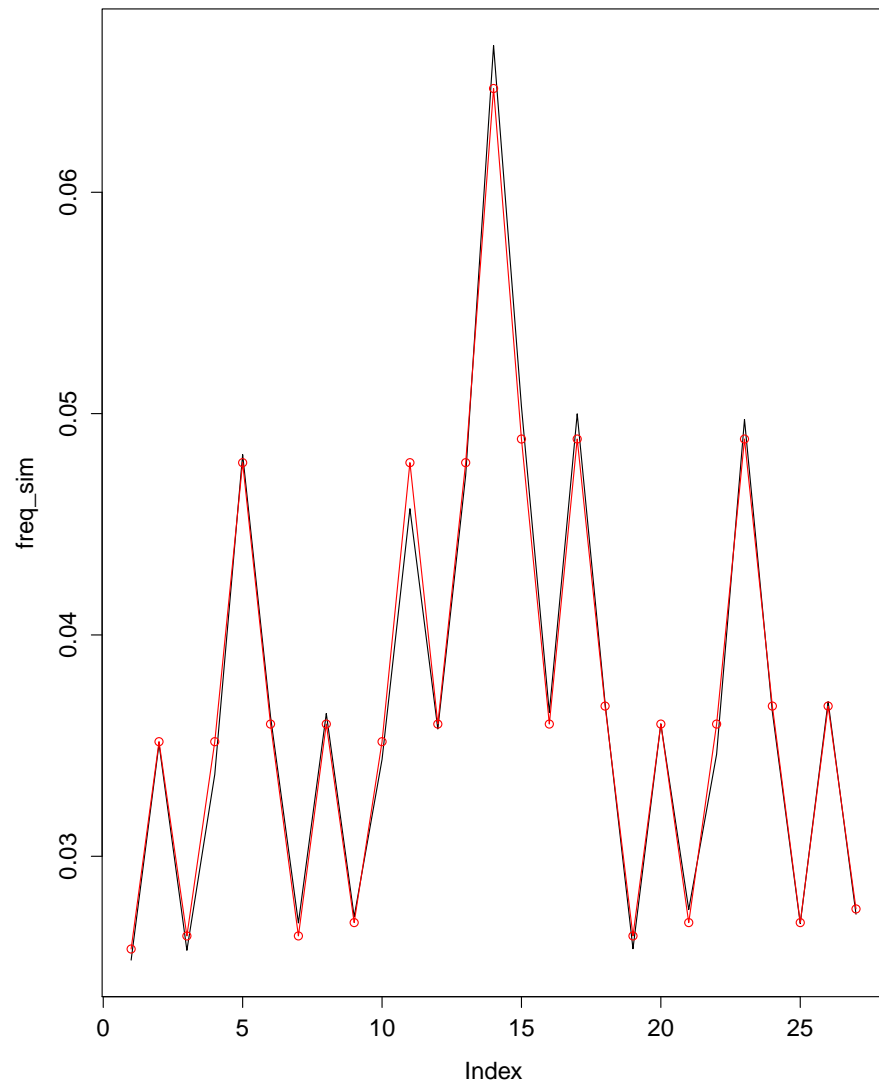


Figure 2: stationary probabilities with 3 sites