How to calculate the non-synonymous to synonymous rate ratio of protein-coding genes under the Fisher–Wright mutation–selection framework

Mario dos Reis

Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK

First principles of population genetics are used to obtain formulae relating the non-synonymous to synonymous substitution rate ratio to the selection coefficients acting at codon sites in protein-coding genes. Two theoretical cases are discussed and two examples from real data (a chloroplast gene and a virus polymerase) are given. The formulae give much insight into the dynamics of non-synonymous substitutions and may inform the development of methods to detect adaptive evolution.

1. Introduction

Halpern & Bruno [1] devised a model to study the divergence of protein-coding genes based on the Fisher–Wright model of mutation, selection and random genetic drift [2,3]. In the model, each particular codon site in the gene is assigned its own set of amino acid fitnesses, and then the Fisher–Wright model is used to work out the evolutionary rate of the site. The model has seen a resurgence in recent years, and variations of it have been used, for example, to study performance of phylogenetic inference methods [4,5], to study codon usage [6] and to estimate the distribution of selection coefficients in protein-coding genes [7,8]. Perhaps surprisingly, the model has not been used to study the dynamics of the non-synonymous to synonymous rate ratio (also known as $\omega = d_N/d_S$) of protein-coding genes and its significance in the study of adaptive molecular evolution.

The purpose of this note is to propose a way to define and calculate an equivalent of the classical concept of the non-synonymous to synonymous rate ratio, in the context of the mutation–selection model of Halpern & Bruno [1]. It is hoped that by using first principles of population genetics, we can obtain an expression of $\omega$ as a function of the selection coefficients acting at codon sites in the protein-coding gene. This should provide much insight into the evolutionary dynamics of codon sites and it should be of advantage in the building of statistical models to detect adaptive evolution in protein-coding genes.

2. The site-wise mutation–selection model

Consider the evolution of a codon site $k$ in a protein-coding gene in a population with $N$ haploid genomes. Assume the site is currently fixed for codon $I$ (i.e. all $N$ alleles carry $I$ at site $k$). In the mutation–selection framework [1,8], the substitution rate (the rate at which novel mutant codons $J$ appear and eventually become fixed in the population) is

$$
q_{IJ,k} =
\begin{cases}
\mu_{ij} & \text{if the mutation is neutral (i.e. } S_{ij,k} = 0), \\
\mu_{ij} S_{ij,k} & \frac{1}{1 - e^{-\beta S_{ij,k}}} & \text{if otherwise}.
\end{cases}
$$

(2.1)
Here \( \mu_{ij} \) is the neutral mutation rate (per generation) from \( i \) to \( j \), and \( S_{ij,k} = F_{ij,k} - F_{ij} \) is the selection coefficient in favour of codon \( j \) and \( F_{ij,k} = 2N\mu_{ij} \) is the scaled Malthusian fitness of \( j \). Natural selection affects the relative substitution rate. When the mutation is advantageous (\( S_{ij,k} > 0 \)), the substitution rate is higher than the neutral rate (\( q_{ij,k} > \mu_{ij} \)), but if the mutation is deleterious (\( S_{ij,k} < 0 \)), then the substitution rate is reduced (\( q_{ij,k} < \mu_{ij} \)). Here, we assume that synonymous substitutions are neutral (\( S_{ij} = 0 \)), and thus, evolution at site \( k \) is determined by 20 amino acid fitnesses. The \( \mu_{ij} \) can be constructed from standard DNA substitution models (for example, if \( I = TTT \) and \( J = TTC \), then \( \mu_{ij} = \nu_k \pi_i^* \) under the HKY substitution model, see [8] for details).

Equation (2.1) describes codon substitution in populations as a continuous-time Markov process. This is sensible if the per generation mutation rate is small compared with the population size (\( N\mu_{ij} \leq 1 \)), so that there is little polymorphism in the population, and at most, two alleles segregate at a site at a time. The proportion of time, \( \pi_{ij,k} \), that site \( k \) spends fixed for \( I \) (i.e. the stationary frequency of \( I \)) is

\[
\pi_{ij,k} \propto \frac{\rho_{ij}}{\rho_k} \exp(F_{ij,k})
\]

[8], where \( \pi_{ij} \) is the frequency for a neutrally evolving sequence (i.e. a pseudo-gene). Thus, the substitution rate at \( k \), averaged over time, is

\[
\rho_k = \sum_{i \neq j} \pi_{ij,k} \rho_{ij,k}
\]

where the sum is over all codon pairs \( i \neq j \). This rate can be partitioned into its non-synonymous and synonymous component rates, \( \rho_k = \rho_{N,k} + \rho_{S,k} \), where

\[
\rho_{N,k} = \sum_{i \neq j} \pi_{ij,k} q_{ij,k} I_{ij,k} \quad \text{and} \quad \rho_{S,k} = \sum_{i \neq j} \pi_{ij,k} q_{ij,k} (1 - I_{ij,k}),
\]

and where the indicator function \( I_{ij,k} = 1 \) if the substitution is non-synonymous and \( = 0 \) if otherwise. Note that the synonymous rate \( \rho_{S,k} \) varies among sites (for example, if a site is conserved for methionine, then the synonymous rate is zero). For a neutrally evolving sequence, the rates are given by

\[
\rho_{N} = \frac{\sum_1^i \pi_{ij} \mu_{ij}(1 - I_{ij})}{\sum_1^i \pi_{ij} \mu_{ij}} \quad \text{and} \quad \rho_{S} = \frac{\sum_1^i \pi_{ij} \mu_{ij}}{\sum_1^i \pi_{ij} \mu_{ij}}
\]

Note that equation (2.1) gives the instantaneous substitution rate, that is, the rate conditioned on site \( k \) being fixed for \( I \) at the present time. On the other hand, \( \rho_k \) is the rate at equilibrium, averaged over all codons and weighted by their stationary frequencies.

3. The relative non-synonymous substitution rate

The absolute non-synonymous to synonymous substitution rate ratio at site \( k \) is \( \rho_{N,k}/\rho_{S,k} \). However, because synonymous rates vary over sites, we need to normalize the ratio by the site’s synonymous rate ratio, \( \rho_{S,k}/\rho_S \), and then normalize by \( \rho_{S,k}/\rho_S \) (to correct for the different proportions of synonymous and non-synonymous substitutions at neutrality). This leads to the following definition:

\[
\omega_k = \frac{\rho_{N,k}}{\rho_{S,k}} \times \frac{\rho_{S,k}}{\rho_S} \times \frac{\rho_S}{\rho_{S,k}} = \frac{\rho_{N,k}}{\rho_S}.
\]

Alternatively, we can define \( \omega_k \) as the relative non-synonymous rate \( \omega_k = \rho_{N,k}/\rho_S \) where constant \( c \) is set so that the ratio is one for neutrally evolving sequences, that is, under the constraint \( \omega_k = \rho_{N,k}/\rho_S = 1 \). The obvious solution is \( c = 1/\rho_S \), leading to the same definition as above. Note that \( c \) has the desirable property of being constant over sites. The reader should not be surprised that the synonymous rate drops out from equation (3.1). When doing statistical inference, synonymous substitutions have information about the neutral mutation rates, and thus inform the value of \( \rho_{S,k} \). Similarly, the relative synonymous rate at site \( k \) is

\[
\gamma_k = \frac{\rho_{S,k}}{\rho_S}.
\]
Figure 1a shows an example for the rbcL gene of flowering plants. Fitness values were estimated under the Halpern–Bruno model by Tamuri et al. [9], and we use their values to calculate $\omega_0$ and $\gamma$ here. The average rates across sites are $\omega = 0.109$ and $\gamma = 1.031$. Note that for many sites, synonymous rates are faster than for a neutrally evolving sequence (i.e. $\gamma > 1$). This is owing to the quirky nature of the genetic code coupled with the mutational biases ($\nu$, $\kappa$, $\pi_t$).

4. The non-synonymous rate during adaptive evolution

When the fitnesses of amino acids are constant through time, sites will spend most of the time fixed for the optimal amino acid. Occasionally, suboptimal amino acids may become fixed, and then substituted after a short period of evolutionary time. This means that the non-synonymous rate at sites is reduced compared with the rate for neutrally evolving sequences (i.e. $\omega_0 < 1$). However, when fitnesses at sites vary over time (for example, after an environment shift or under intense frequency-dependent selection [10]), the non-synonymous rate may be accelerated compared with the rate for neutrally evolving sequences ($\omega_0 > 1$). We now study the case where fitnesses change as an adaptation to a novel environment.

Consider a site $k$ where the fitness of $I$ is $F_{IA}$ in environment $A$. The stationary frequencies and instantaneous substitution rates are $\pi_{IA}^s$ and $q_{IA}^s$. Now, imagine that the environment shifts (for example, a population of mammals living in a suddenly colder climate, or a virus colonizing a new host, where the intracellular environment in the new host is different from the reservoir host). The fitness of $I$ in the new environment $B$ is now $F_{IB}$. The probability that the site is currently fixed for $I$ at the moment of the environment shift is $\pi_{IA}^s$, but the substitution rate is now that of the new environment $q_{IB}^s$. Thus, the expected absolute and relative non-synonymous rates at the environment shift are

$$\rho_{NIA} = \sum_{I \neq J} \pi_{IA}^s q_{IJ}^s I_N$$

and

$$\omega_k^s = \frac{\rho_{NIA}^s}{\rho_{N}^s}$$

(4.1)

If the shift in fitness values is large, then the rate will be much accelerated ($\omega_k^s > 1$). This occurs because the site is likely to find itself fixed for a suboptimal amino acid in the new environment, and novel mutations to optimal amino acids will become fixed quickly. However, if the fitness shift is moderate, the rate may still be lower than the neutral rate ($\omega_k^s < 1$).

Figure 1b shows an example for the phb2 gene of the influenza virus. Fitness values were estimated under the Halpern–Bruno model by Tamuri et al. [9]. A subset of 25 adaptive sites (where fitnesses are different for viruses evolving in human versus avian hosts [8,11]) were identified by Tamuri et al. [11], and their fitnesses estimated by Tamuri et al. [8]. We use the estimates to calculate $\omega_k$, $\gamma$, and $\omega_k^s$ here. The classical lineage of human influenza probably originated from a host shift from an avian to a mammal reservoir in the early-twentieth century [12]. We calculate $\omega_k^s$ at the putative host shift. The average rate at adaptive sites is $\omega^s = 1.670$ (across all sites $\omega = 0.195$ and $\gamma = 0.960$). Note that for 16 sites for which fitnesses are different between hosts, we find that $\omega_k^s < 1$. This indicates that the criterion $\omega_k^s > 1$ to detect adaptive evolution is conservative in this case.

The probability that the site is fixed for $I$, time $t$ after the environment shift is

$$\Pr_{t}^{(B)}(l) = \sum_{I} \pi_{IA}^s p_{JI}^{(B)}(t),$$

(4.2)

where $p_{JI}^{(B)}(t)$ are the transition probabilities obtained using standard Markov theory, i.e. by calculating $P^{(B)}(t) = \exp(tQ^{(B)})$. Thus, the absolute and relative non-synonymous rates, time $t$ after the shift, are

$$\rho_{NIA}^s(t) = \sum_{I \neq J} \Pr_{t}^{(B)}(l) q_{IJ}^s I_N$$

and

$$\omega_k^s(t) = \frac{\rho_{NIA}^s(t)}{\rho_{N}^s}.$$
The transition probabilities in equation (4.2) are exponential decay functions of time, and so \( \omega_k(t) \) is also an exponential decay. Initially, the value of \( \omega_k(t = 0) \) will be high, and as the time goes to infinity, \( \omega_k(t \to \infty) \) will approach the stationary value given by equation (3.1). In other words, soon after an environment shift, a burst of adaptive substitutions will occur at sites where fitnesses have changed, and substitutions will accumulate until the protein-coding gene reaches a state of adaptive equilibrium. For example, figure 2a shows the decay of \( \omega_k(t) \) for the 25 adaptive sites in the pb2 gene after a host shift.

5. Conclusion

Previous authors have shown that the relationship between the non-synonymous rate and the selection coefficient is approximately \( \omega = S/(1 - \exp(-S)) \) [13,14], but the approximation relies either on the infinite-sites model or assumes that all mutant amino acids have the same fitness. Equations (3.1) and (4.1) provide more realistic approximations but are hard to visualize. Consider a site fixed for \( I \). The probability that the next mutation will be 1 is \( P_I = \mu I / \Sigma \mu I \) for \( I \neq J \). Over time, the proportion of \( I \) to \( J \) mutations at the site will be \( \pi_{IJK}P_{IJK} \). Thus, the average selection coefficient on mutations at site \( k \) is

\[
S_k = \sum_{I \neq J} \pi_{IJK}P_{IJK}\frac{S_{IJK}}{C_{IJK}}.
\]

Figure 2b shows \( \omega_k \) as a function of \( S_k \) for simulated sites when fitnesses are constant or when they shift with the environment. Note that the approximation \( \omega = S/(1 - \exp(-S)) \) provides a reasonable lower bound on \( \omega_k \). In general, \( \omega_k \) increases with \( S_k \), but the relationship is not as simple as in the previous approximations [13,14].

In the site-wise mutation–selection model, one calculates the selection coefficients first, and hence one may know whether a site has been under positive selection without calculating \( \omega_k \) [8]. However, the model is over-parametrized, computationally expensive, and fitnesses can be well estimated only in large datasets [9]. Instead, the model should be of advantage in evolutionary reasoning and in model building. For example, the behaviour of \( \omega_k \) under more complex models (such as frequency-dependent selection [10], adaptation to gradual environment changes [10] or selection on codon usage [6]) can also be studied under the site-wise mutation–selection framework. This will be a worthwhile effort as it will shed light on our ability to detect adaptive evolution in molecular sequences.

Data accessibility. The data accompanying with this study are available at Dryad doi:10.5061/dryad.3r3q4.

Acknowledgements. I thank Ziheng Yang, Richard Goldstein and Asif Tamuri for valuable comments.

Funding statement. M.d.R. is supported by BBSRC (UK) grant no. BB/J009709/1 awarded to Ziheng Yang.

Conflicts of interest. I have no competing interests.

References