Estimating Synonymous and Nonsynonymous Substitution Rates

Spencer V. Muse
Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University

Partitioning the total substitution rate into synonymous and nonsynonymous components is a key aspect of many analyses in molecular evolution. Numerous methods exist for estimating these rates. However, until recently none of the estimation procedures were based on a sound statistical footing. In this paper, the evolutionary model of Muse and Gaut (1994) is used as the basis for two sets of parameters quantifying silent and replacement substitution rates. The parameters are shown to be equal when the four nucleotides are equally frequent and unequal otherwise. Maximum-likelihood estimation of these parameters is described, and the performance of these estimates is compared to that of existing estimation procedures. It is shown that the estimates of Nei and Gojobori (1986) are not unbiased for either set of parameters, although they provide very good estimates for one set as long as sequence divergence is not too high. However, some disturbing properties are found for the Nei and Gojobori estimates. In particular, it is shown that the expected value of the Nei and Gojobori estimate of silent substitution rate is a function of both the silent and replacement substitution rates. The maximum-likelihood estimates have no such problems.

Introduction

An important component of many molecular evolutionary analyses is the estimation of synonymous (silent) and nonsynonymous (replacement) nucleotide substitution rates. Even a quick search of the literature reveals a plethora of estimation methods (Miyata and Yasunaga 1980; Perler et al. 1980; Li, Wu, and Luo 1985; Nei and Gojobori 1986; Lewontin 1989; Li 1993; Pamilo and Bianchi 1993). Although the details of these methods vary, they all use the same basic approach:

1. Count the number of silent and replacement sites in the observed sequences.
2. Compare each pair of homologous codons site by site and infer the number of silent and replacement differences, using the most parsimonious evolutionary pathways between the two codons.
3. Adjust for multiple hits using a standard evolutionary model, such as the Jukes and Cantor (1969) or Kimura (1980) models.

Thus, each method begins by inferring \( p_s \) and \( p_r \), the proportions of silent and replacement differences per silent and replacement site. The proportions are then corrected to give estimates, \( \hat{D}_s \) and \( \hat{D}_r \), of the expected numbers of silent and replacement substitutions per silent and replacement site. These approaches have weaknesses at each stage, as has been pointed out by several authors (Li, Wu, and Luo 1985; Nei 1987; Lewontin 1989; Muse and Weir 1992; Muse and Gaut 1994; Goldman and Yang 1994). Specifically, the number and location of silent/replacement sites change over time, so it is difficult to count or classify these sites—today’s silent site may have been a replacement site for most of its evolutionary history. It follows that the methods used for multiple hit correction are logically flawed, since the corrections are made using models and parameters other than those being estimated. For example, the use of a Jukes–Cantor correction cannot be correct, since properties of the genetic code must be incorporated. By using the Jukes–Cantor correction one implicitly assumes that the sites being analyzed have evolved with a single, constant substitution rate throughout time. However, the only sites in coding regions for which this might be a valid assumption are second codon positions. First and third positions alternate between being silent and replacement sites. As we will see, this becomes especially problematic when sequences are highly diverged.

The shortcomings of these methods rise from a common source: the lack of an explicitly stated evolutionary model. It is the precise definition of an evolutionary model that provides the form of multiple-hit corrections, and the differences between models lead to different estimates of evolutionary distance. Since the traditional methods for estimating silent and replacement substitution rates do not make corrections according to a well-specified model, it is impossible to make rigorous claims about their statistical properties, including bias (Li 1993) and variance (Ota and Nei 1994). This does not, however, exclude the possibility that the methods have good properties.

The purpose of this paper is twofold. First, using an evolutionary model that is appropriate for coding sequences, I suggest two different sets of evolutionary parameters describing synonymous and nonsynonymous substitution rates and discuss the differences between
the two. Second, I describe a procedure for computing maximum-likelihood estimates of these parameters. Using simulation, the behavior of these estimates is compared with the behavior of the Nei and Gojobori (1986) estimates. These estimates were chosen for comparison because of their very frequent use in molecular evolutionary studies. Some problems inherent to such applications are discussed at length.

An Evolutionary Model for Coding Regions

Muse and Gaut (1994) presented a model for coding regions in the context of constructing relative rate tests using three sequences. The model (dubbed the codon model) also allows estimates of pairwise evolutionary distances, so its form will be repeated here. The model is similar in form to those of Hasegawa, Kishino, and Yano (1985) and Adachi and Hasegawa (1992) for the evolution of nucleotide and amino acid sequences, respectively. An alternative evolutionary model for codons was developed by Goldman and Yang (1994). Goldman and Yang also gave several empirical examples of computing evolutionary distances with their model.

Define $R_{ij}$ to be the rate of change from codon $i$ to codon $j$. Numbering of the codons may be done in any convenient fashion. Let $\alpha$ and $\beta$ denote the synonymous and nonsynonymous substitution rates, respectively. If we assume that in time $t$ only one nucleotide substitution event can occur in any particular codon, we can define the substitution process among the 61 nontermination codons as follows:

$$R_{ij} = \begin{cases} \alpha \pi_{nj} & \text{synonymous} \\ \beta \pi_{nj} & \text{nonsynonymous} \\ 0 & \text{multiple substitutions needed} \end{cases}$$

where $\pi_{nj}$, $n_j \in \{A, C, G, T\}$, accounts for the frequency of the “target nucleotide” when considering a change from codon $i$ to codon $j$. In practice, these values are estimated using the base frequencies observed in the data. For example, if $i = \text{AGG}$ and $j = \text{AGA}$, then $R_{ij} = \alpha \pi_A$. The replacement of the $G$ by an $A$ results in no amino acid substitution, since both codons encode arginine. On the other hand, the instantaneous rate for an AGG codon being replaced by a CGA is 0. Even though the change is synonymous, it requires two nucleotide substitutions. Note that the model still allows the possibility of multiple changes within a codon, it simply requires that they occur through a series of steps rather than in a single step. Termination codons are excluded from the model since their occurrence would result in considerable terminal length variation in the functional proteins. Thus, we have a $61 \times 61$ instantaneous rate matrix, $R$. The diagonal elements of $R$ are defined to make the elements of each row sum to zero. That is, $R_{ii} = -\sum_{j \neq i} R_{ij}$. It can be shown using the approach of Muse (1995) that this model would be exactly equivalent to the Jukes and Cantor (1969) model if there were no differences between synonymous and nonsynonymous rates ($\alpha = \beta$), base frequencies were all 0.25, and termination codons were not excluded.

It seems doubtful that closed-form expressions for transition probabilities of the form $P_{ij}(t)$ (the probability of changing from codon $i$ to codon $j$ in time $t$) can be found for this model. Instead, we must use the fact that the instantaneous rates in $R$ provide the transition probabilities in $P(t)$, the $61 \times 61$ matrix of $P_{ij}(t)$, via the expression $P(t) = e^{Rt} = I + Rt + (Rt)^2/2! + (Rt)^3/3! + \ldots$. Even though expressions for transition probabilities are intractable, the asymptotic frequencies of the 61 codons can be computed. Standard methods show that the equilibrium frequency of the codon consisting of nucleotides $i, j, k$, and $k$ is $\pi_i \pi_j \pi_k (1 - \Pi_{\text{nuc}})$, where $\Pi_{\text{nuc}} = \pi_T \pi_A \pi_G + \pi_T \pi_G \pi_A + \pi_T \pi_A \pi_A$. (I will use the notation $\Pi_j$ to refer to the equilibrium frequency of codon $i$, $i = 1, 2, \ldots, 61$.) The intuitive modifications are successful in providing asymptotic frequencies for genetic codes other than the universal code. None of the results produced below depend on any particular genetic code.

Biological Parameters

It now remains to demonstrate the relationship between the parameters in this model, $\alpha$ and $\beta$, and the parameters biologists want to estimate. Typically, rates are expressed in units of the number of silent (replacement) substitutions per silent (replacement) site. However, we have already seen that the precise definitions of these “parameters” are not clear. I will now suggest two definitions based on the model of the previous section.

Second codon positions are often referred to as perfectly replacement sites, since all changes at these sites result in an amino acid substitution. (Be forewarned that I will introduce a slightly more stringent definition of a “perfectly replacement site” momentarily.) This suggests that a reasonable interpretation of the “number of replacement substitutions per replacement site” might be the expected number of substitutions at a perfectly replacement site in time $t$. Perfectly silent sites do not exist. Contrary to the popular notion that fourfold degenerate sites fall into this category, it is possible for changes at a first or second codon position to change the degeneracy class of a (currently) fourfold degenerate third codon site. Nonetheless, it is still reasonable to define as a parameter the number of substitutions expected at a (hypothetical) perfectly silent site in time $t$. 
Refer to these parameters as $\Delta_s$ and $\Delta_n$. After accounting for unequal nucleotide frequencies, it is seen that

$$\Delta_s = \left(1 - \sum \pi_n^2\right) \alpha t \quad \text{and} \quad \Delta_n = \left(1 - \sum \pi_n^2\right) \beta t. \quad (1)$$

A second possibility for the “number of silent substitutions per silent site” is to define the parameter in terms of the actual number of events expected for a given sequence, incorporating the effects of such factors as base and codon composition. To facilitate the description, introduce the following quantities:

- $s_i$: the number of potential one-step silent changes from codon $i$.
- $E_i$: the “effective number” of silent sites per codon.
- $C_i$: the expected number of silent substitutions per codon.

The parameter $E_i$ (and $E_n$, its counterpart for nonsynonymous changes) deserves explanation. A perfectly silent site should offer three silent substitutions. Likewise, a perfectly replacement site should offer three replacement substitutions. This definition differs from the traditional one by requiring that three potential changes exist, not simply that all changes are of the same type. Note that with this definition neither perfectly silent nor perfectly replacement sites exist. Perfectly silent sites fail to exist for the reasons mentioned previously; perfectly replacement sites are hypothetical entities because of the effects of stop codons, changes to which are not considered replacements (and which are not allowed under the present evolutionary model). Because of the difficulties surrounding the designation of silent and replacement sites, it seems reasonable to define $E_i$ as one-third of the total number of potential silent substitutions:

$$E_s = \sum_{i=1}^{61} f_{i} s_i / 3. \quad (2)$$

Using these definitions, along with their analogs for replacement substitutions, the expected number of silent and replacement substitutions per silent and replacement site can now be defined as

$$D_s = C_s/E_s \quad \text{and} \quad D_n = C_n/E_n. \quad (3)$$

I now have defined two potential sets of parameters. The first, $\Delta_s$ and $\Delta_n$, might be called “perfect-site” measures, because they denote the expected numbers of changes at (hypothetically) perfectly silent and replacement sites. The second pair, $D_s$ and $D_n$, might be called “actual site” measures, because they describe the expected numbers of changes making explicit account for asymmetries in the genetic code. Indeed, this is the primary source of differences between the two measures. I would argue that the actual-site measures are the parameters that existing distance methods are attempting to estimate. Simulation results presented later reinforce this conjecture. Notice the similarity between the form of these parameter definitions in equation (3) and the form of the standard statistics.

By expanding both the numerator and denominator, we see below that the actual site measures, $D_s$ and $D_n$, describe precisely the same parameters as the perfect-site measures, $\Delta_s$ and $\Delta_n$, when base frequencies are equal but different parameters otherwise. The expansions rely on the fact that the expected number of substitutions per codon is

$$C_i = t \sum_{j \neq i} f_i s_{ij} = C_n + C_s. \quad (4)$$

In this notation, $s_{ij}$ is defined to be equal to $R_{ij}$ if the indicated change is silent and 0 otherwise, while $N_{ij} = R_{ij}$ if the indicated change is a replacement and 0 otherwise. It now follows that

$$D_s = \frac{C_s}{E_s} = \frac{t \sum_{j \neq i} f_i s_{ij}}{\sum_i f_i s_i / 3} = \frac{\alpha t \sum_i f_i s_i \pi_n}{\sum_i f_i s_i / 3}, \quad (5)$$

and if all base frequencies are equal,

$$D_s = \frac{\alpha t \sum_i f_i s_i / 4}{\sum_i f_i s_i / 3} = \frac{3}{4} \alpha t. \quad (6)$$

The analogous result for $D_n$ can be derived in the same way.

Given that $D_s$ and $\Delta_s$ (and $D_n$ and $\Delta_n$) are typically unequal, it is important to decide which (if either) set of parameters is most suited for a particular task. For instance, both sets should be appropriate for phylogeny construction, since both are linear functions of time. (However, it is likely that the two sets of distance estimates will differ in their efficiency of recovering the
correct tree.) When using distance estimates to make inferences about the evolutionary process, such as comparisons of silent and replacement rates or comparisons of rates between genes, the effect of parameter choice is more critical. The above derivation points out a disturbing feature connected to the use of the actual-site measures. Like $\Delta_{a}$ and $\Delta_{n}$, $D_{s}$ and $D_{n}$ are linear functions of time. Unlike the perfect-site measures, however, the multipliers of the two actual-site measures are not the same because of asymmetries in the genetic code: it is seen that $D_{s} = \kappa_{1}a_{t}$, while $D_{n} = \kappa_{2}b_{t}$. The symbols $\kappa_{1}$ and $\kappa_{2}$ refer to the multipliers of $a_{t}$ and $b_{t}$ from equation (5) and its nonsynonymous analog. As a result of this, $D_{s} \neq D_{n}$ when $\alpha = \beta$. This suggests that the difference between unbiased estimators of $D_{s}$ and $D_{n}$ may not be a wise choice of test statistic when addressing the null hypothesis that silent and replacement substitution rates are equal. This type of test is performed frequently to provide evidence of positive Darwinian selection (This topic is treated more fully in the Discussion.)

**Maximum-likelihood Estimation**

The codon model allows for the computation of maximum-likelihood estimates (MLEs) of $a_{t}$ and $b_{t}$, and, by the invariance property of MLEs (Edwards 1992), it also allows for maximum-likelihood estimation of both the actual- and perfect-site measures. Given two homologous nucleotide sequences, the framework of Felsenstein (1981) can be used to express the likelihood function. If two sequences have codons $i$ and $j$ at codon position $k$, then the contribution of that position to the likelihood is $L_{k} = f_{ij}P_{ij}(t)$, and the complete likelihood is $L = \prod L_{k}$. This result uses the fact that the codon model is reversible (Musc and Gaut 1994). Standard numerical methods (Press et al. 1992) can be used to find the values of $\hat{a}_{t}$ and $\hat{b}_{t}$ that maximize the likelihood function.

The statistical properties of MLEs are well studied (Edwards 1992). Two important properties are those of consistency and asymptotic efficiency. A consistent estimator converges to the true (unknown) parameter value as the sample size increases. An estimator is asymptotically efficient if its variance converges to the smallest value possible for an estimator with the same expected value as the sample size increases. Maximum-likelihood estimators are both consistent and asymptotically efficient. Not only do we have an optimality result for the variance, but we also have a method to calculate it because the asymptotic variance of MLEs is related to the curvature of the likelihood surface near the maximum. Finally, MLEs are also known to have distributions that are asymptotically normal, so it is a simple matter to compute (large-sample) confidence intervals or to conduct hypothesis tests.

**Simulation Study**

Although the asymptotic properties of MLEs should hold for very long sequences, we have no theory available to predict the behavior of the estimates for sequences of finite length. In order to explore this issue and to examine the performance of existing methods, a simulation study was performed. For a series of increasing substitution rates, 100 random data sets (pairs of sequences) were generated using the codon model, and evolutionary distances were estimated first using maximum likelihood and then with the method of Nei and Gojobori (1986) (the NAG estimates, denoted $d_{a}$ and $d_{n}$). Some results are shown in table 1 and figure 1.

The results in table 1 show that the MLEs of $D_{s}$ and $D_{n}$ have negligible bias throughout the studied portion of the parameter space. Because $\Delta_{a}$, $\Delta_{n}$, $D_{s}$, and $D_{n}$ are all linear functions of $a_{t}$ and $b_{t}$, the MLEs of these parameters are biased to the same degree. It remains difficult to make assertions regarding the (un)biasedness of the NAG estimates, since it is not clear what parameters they are designed to estimate. However, the results in table 1 suggest that the expectations of the NAG estimates are close to $Q$ and $D_{n}$, especially for low to moderate levels of sequence divergence. The NAG estimates are seen to perform especially well when base frequencies are equal. Although the results are not shown in table 1, the NAG estimates are considerably worse estimators of the perfect-site measures. This reinforces the earlier suggestion that the actual-site measures are the targets of the NAG estimates.

A proper estimate of synonymous distance, $\hat{D}_{s}$, should have the property that $E(\hat{D}_{s}) = ket$ for some value of $k$. That is, the estimate's expected value should be a linear function of $a_{t}$ passing through the origin. To see if the MLE and NAG estimates have this property, the models $E(d_{a}) = b_{0} + b_{1}a_{t}$ and $E(\hat{d}_{a}) = b_{0} + b_{1}a_{t}$ were applied to the simulated data summarized in table 1. The analogous models were fit for $\hat{D}_{n}$ and $\hat{D}_{n}$. (Again, because both sets of parameters are linear functions of $a_{t}$ and $b_{t}$, results will hold for both.) The analyses were performed separately for the equal and unequal nucleotide frequency simulations because the parameter values differ as shown in equation (5). The null hypothesis $H_{0}$: $b_{0} = 0$ is rejected only for the case of $\hat{D}_{a}$ when base frequencies are unequal ($P = 0.0336$). This provides only weak evidence that the intercepts for the NAG estimates are non-zero. Regression analyses checking for nonlinearity also failed to uncover problems with either the NAG estimates or with the MLEs.
Fig. 1.—Simulation results. In each plot, the appropriate NAG estimate, \( \hat{d}_c \) or \( \hat{c}_n \), is plotted against the corresponding MLE, \( \hat{D}_c \) or \( \hat{D}_n \). The solid line is the line \( y = x \). Note the tendency for the NAG estimates to underestimate, especially when base frequencies are unequal. The unusual dichotomous behavior near the origin, particularly noticeable in the silent rate plots, seems to be caused by the NAG estimates' dependence on both \( \alpha \) and \( \beta \). This behavior would not have been as obvious had values of \( \beta t \) between 0.1 and 0.5 been used. The more uniform choice of \( \alpha t \) values masks the dichotomy in the replacement rate plots.
A more troubling observation is that the expectation of \( \hat{d}_s \) is a function of \( \beta \), and that \( \text{E}(\hat{d}_s) \) is a function of \( \alpha \). One would want an estimate of the silent substitution rate to have an expectation that depends only on the primary parameter, \( \alpha \); the secondary parameter, \( \beta \), should not contribute to the expectation of an estimate of silent rate. Evidence that \( \hat{d}_s \) has this problem is clearly visible by examining the pattern of expectations within columns or rows in table 1. Regression analyses adding a regressor for the second parameter soundly reject, in all cases, the null hypothesis that only the value of the primary parameter affects the expectation of the NAG estimates (\( P \)-values are 0.0372 and 0.0322 for \( \hat{d}_s \), 0.0003 and 0.0001 for \( \hat{d}_u \)). This effect seems to be quite strong. Notice that within each row in the bottom half of table 1 the mean value of \( \hat{d}_u \) increases with the value of \( \alpha \). A similar result is seen with \( \hat{d}_r \). When \( \alpha > \beta \), the mean of \( \hat{d}_r \) decreases as \( \beta \) increases. Note that when \( \alpha < \beta \), as tends to be the case in the first two rows, the relationship is reversed. This suggests a nonlinear relationship between \( \text{E}(\hat{d}_s) \) and \( \beta \). One would suspect that the same type of behavior would also be found for \( \hat{d}_u \) if a larger portion of the parameter space was studied. Because it is typically the case that silent changes are much more prevalent than replacements, the nonlinearity is probably of little concern. However, the fact that \( \text{E}(\hat{d}_s) \) and \( \text{E}(\hat{d}_u) \) are each functions of both \( \alpha \) and \( \beta \) is very troubling. No such problems were observed for the MLEs.

The effects of base frequency on the NAG estimates are also seen to be quite strong. The simulation data were split into two subsets, one for the equal frequency case and one for the unequal case. Within each of these sets the same linear regression model cited above was used, for both \( \hat{d}_s \) and \( \hat{d}_u \). The slopes of the regression lines are very close to the multipliers from the actual-site measures when base frequencies are equal. Recall that when base frequencies are equal, the value of \( \hat{d}_s \) is 0.75\( \alpha \) and the value of \( \hat{d}_u \) is 0.75\( \beta \). The estimated regression coefficients are 0.758 and 0.719 for the equal frequency simulation data. For the base frequencies used in the unequal case (\( \pi_A = \pi_C = 0.15, \pi_G = \pi_T = 0.35 \)), use of equation (5) shows the true regression coefficients to be 0.682 for silent rates and 0.722 for replacement rates. The estimated coefficients for \( \hat{d}_s \) and \( \hat{d}_u \) are 0.613 and 0.674, respectively. The

### Table 1

**Simulation Results**

<table>
<thead>
<tr>
<th>Parameter Values</th>
<th>Silent Rates, Equal Frequencies</th>
<th>Silent Rates, Unequal Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha = 0.100 )</td>
<td>( \beta = 0.02 )</td>
<td>0.0741</td>
</tr>
<tr>
<td>( \delta_r = 0.075 )</td>
<td>( \beta = 0.05 )</td>
<td>0.0749</td>
</tr>
<tr>
<td>( \alpha = 0.300 )</td>
<td>( \beta = 0.225 )</td>
<td>0.2267</td>
</tr>
<tr>
<td>( \delta_r = 0.250 )</td>
<td>( \beta = 0.500 )</td>
<td>0.5100</td>
</tr>
<tr>
<td>( \alpha = 1.000 )</td>
<td>( \beta = 0.750 )</td>
<td>0.7562</td>
</tr>
<tr>
<td>( \delta_r = 0.750 )</td>
<td></td>
<td>0.7570</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter Values</th>
<th>Replacement Rates, Equal Frequencies</th>
<th>Replacement Rates, Unequal Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha = 0.10 )</td>
<td>( \beta = 0.02 )</td>
<td>0.0151</td>
</tr>
<tr>
<td>( \delta_r = 0.015 )</td>
<td>( \beta = 0.05 )</td>
<td>0.0144</td>
</tr>
<tr>
<td>( \alpha = 0.50 )</td>
<td>( \beta = 0.05 )</td>
<td>0.0372</td>
</tr>
<tr>
<td>( \delta_r = 0.038 )</td>
<td>( \beta = 0.08 )</td>
<td>0.0356</td>
</tr>
<tr>
<td>( \alpha = 0.100 )</td>
<td>( \beta = 0.100 )</td>
<td>0.1070</td>
</tr>
<tr>
<td>( \delta_r = 0.100 )</td>
<td>( \beta = 0.075 )</td>
<td>0.0706</td>
</tr>
<tr>
<td>( \alpha = 0.500 )</td>
<td>( \beta = 0.500 )</td>
<td>0.5000</td>
</tr>
<tr>
<td>( \delta_r = 0.500 )</td>
<td>( \beta = 0.375 )</td>
<td>0.3502</td>
</tr>
</tbody>
</table>

Notes.—Headings divide table 1 into four quadrants. Within each quadrant, the first column contains the true parameter values used for generating the simulated data. The table entries are the mean values of the MLE (top number) and NAG estimates (bottom number) for 100 replicate data sets generated with the indicated parameter values. The labels at the top show the values for the secondary parameter. For example, when 100 data sets were generated with equal base frequencies \( \pi_A = \pi_C = 0.15, \pi_G = \pi_T = 0.35 \), the true value of \( \hat{d}_s \) was 0.038. The true value of \( \hat{d}_u \) for these parameter values, 0.036. Moving to the lower half of table 1, we see that the average value of \( \hat{d}_u \) is 0.0356. The true value of \( \hat{d}_u \) was 0.0356. The right half of the table presents results for unequal base frequencies. The frequencies used for the simulations were \( \pi_A = \pi_C = 0.15, \pi_G = \pi_T = 0.35 \).
NAG estimates seem to be systematically underestimating the actual-site measures. However, it must be noted that the effect of the secondary parameter blurs this claim: the apparent underestimation could be an artifact of the bias introduced by the secondary parameter. In either case, the estimates are demonstrating an undesirable property. Again, the MLEs showed none of these problems.

**Empirical Study**

Controlled simulation studies are useful to gain an understanding of statistical properties of estimators, but the relevance of such results to real data is often unclear. Empirical studies at least allow us to see if different methods provide similar results. Of course, consistency among methods does not insure correct results, but it does provide some amount of comfort. The absence of consistent results forces us to examine the assumptions of each method and to test the validity of the assumptions for the data at hand. Presumably, the method with assumptions that more closely match the system under study should provide more reliable results. A moderately large set of actual sequence data was used to search for consistency between the maximum likelihood and Nei and Gojobori (1986) methods. Gaut et al. (1992) examined the issue of substitution-rate heterogeneity at the
rbcL locus among a diverse group of monocots. The data from that study have been supplemented with five outgroup sequences. In figure 2a, for each pairwise comparison the MLE of \( D_s \) is plotted against the corresponding estimate, \( \hat{D}_s \), from the Nei and Gojobori (1986) method. Figure 2b presents the same type of plots for \( D_r \) and \( \hat{D}_r \). The plots are consistent with the simulation results. The estimates are highly correlated for closely related sequences, with the correlation diminishing as sequence divergence increases. Note the tendency for the NAG estimates to provide lower estimates than the corresponding MLEs, especially for silent sites. This effect is presumably a result of the use of most parsimonious evolutionary pathways between codons by the NAG estimates, leading to an undercount of substitution events.

**Discussion**

The primary goal of the work presented above was to compare MLEs of synonymous and nonsynonymous substitution rates with the much simpler estimates of Nei and Gojobori (1986). For the most part, the results are encouraging: the NAG estimates are virtually identical to MLEs of appropriate parameters when low to moderate levels of sequence divergence are present, although the NAG estimates become increasingly biased as sequence divergence increases. Intuitively, one would predict such behavior. The expected values of the NAG estimates were defined as the expected number of silent (replacement) sites per codon. An advantage of the perfect-site measures over the actual-site measures is that the functional forms of both \( \Delta_1 \) and \( \Delta_r \) are the same. That is, \( \Delta_1 = \kappa \sigma_1 \) and \( \Delta_r = \kappa \beta_1 \) (both use the same multiplier, \( \kappa \)). This allows silent and replacement rates to be compared directly using raw distance estimates. On the other hand, the multipliers for \( D_s \) and \( D_r \) are different. The problem this presents is seen most easily by considering the difference, \( D_s - D_r \). Functions of this sort are often compared to zero in efforts to detect natural selection (Hughes and Nei 1988; Hughes, Ota, and Nei 1990). Because the multipliers do not cancel unless base frequencies are equal, the value of this difference is not equal to 0 under strict neutrality. This suggests that one should not compare the difference of unbiased estimates of \( D_s \) and \( D_r \) to zero as a test for natural selection.

Because the problem of testing for natural selection is an important one, additional simulations were performed to see if the difference of the NAG estimates has a mean of zero when silent and replacement rates are equal (\( \alpha = \beta \)). Some results are shown in table 2. For a series of increasing substitution rates, 1,000 replicate data sets (pairs of sequences) were generated using the codon model described earlier. For each pair of sequences \( \hat{D}_s \) and \( \hat{D}_r \) were computed. It is seen in table 2 that the difference of the two estimates does not appear to be zero when base frequencies are unequal. Each of the 10 parameter settings lead to rejection (\( P \leq 0.001 \)) of the null hypothesis that the mean difference equals zero, and in all 10 cases the mean of \( \hat{D}_s \) is less than that of \( \hat{D}_r \). The bias of \( \hat{D}_s - \hat{D}_r \) is substantial when
The observed base frequencies are used in the proper calculation of the MLEs, \( \alpha \) and \( \beta \). This is similar in spirit to the ad hoc procedure used by Kondo et al. (1993).

The recent advances in computing speed have eliminated the need for evolutionary analyses to be limited to models that have simple closed form solutions. Complex evolutionary models allow us to study the more subtle details of the evolutionary process. While it may be the case that these details can be uncovered using simple models and statistics based primarily on intuition, we should have little confidence in such analyses until it can be shown that the statistical methods provide reasonable results under more realistic models. Evolutionary models such as the one used here allow this type of testing. It is important that we recognize both the utility and the accessibility of many computationally intensive procedures, rather than simply dismissing all of them as being prohibitive.

**Acknowledgments**

This work has been supported by NIH fellowship GM16250 to the author and NIH grant GM45876 to Andrew G. Clark. Computing resources were provided by the Pennsylvania State University Center for Computational Biology.

**Literature Cited**


Stanley Sawyer, reviewing editor

Accepted August 18, 1995