A New Method for Estimating Synonymous and Nonsynonymous Rates of Nucleotide Substitution Considering the Relative Likelihood of Nucleotide and Codon Changes

Wen-Hsiung Li, Chung-I Wu, and Chi-Cheng Luo
University of Texas Health Science Center at Houston

A new method is proposed for estimating the number of synonymous and nonsynonymous nucleotide substitutions between homologous genes. In this method, a nucleotide site is classified as nondegenerate, twofold degenerate, or fourfold degenerate, depending on how often nucleotide substitutions will result in amino acid replacement; nucleotide changes are classified as either transitional or transversional, and changes between codons are assumed to occur with different probabilities, which are determined by their relative frequencies among more than 3,000 changes in mammalian genes. The method is applied to a large number of mammalian genes. The rate of nonsynonymous substitution is extremely variable among genes; it ranges from $0.004 \times 10^{-9}$ (histone H4) to $2.80 \times 10^{-9}$ (interferon γ), with a mean of $0.88 \times 10^{-9}$ substitutions per nonsynonymous site per year. The rate of synonymous substitution is also variable among genes; the highest rate is three to four times higher than the lowest one, with a mean of $4.7 \times 10^{-9}$ substitutions per synonymous site per year. The rate of nucleotide substitution is lowest at nondegenerate sites (the average being $0.94 \times 10^{-9}$), intermediate at twofold degenerate sites ($2.26 \times 10^{-9}$), and highest at fourfold degenerate sites ($4.2 \times 10^{-9}$). The implication of our results for the mechanisms of DNA evolution and that of the relative likelihood of codon interchanges in parsimonious phylogenetic reconstruction are discussed.

Introduction

Until the late 1970s the most frequently used method for estimating the numbers of nucleotide substitutions between two homologous DNA sequences had been the one by Jukes and Cantor (1969; see also Holmquist 1972; Kimura and Ohta 1972). These authors made the simplifying assumption that nucleotide substitution occurs randomly. In recent years the rapid accumulation of DNA sequence data and the realization of nonrandom patterns of nucleotide substitution have stimulated the development of more elaborate methods (e.g., Holmquist and Pearl 1980; Kimura 1981; Gojobori et al. 1982a; Kaplan and Risko 1982; Lanave et al. 1984; Tajima and Nei 1984).

A different approach has also been taken in order to estimate the number of synonymous substitutions (which cause no amino acid replacement) and that of

1. Key words: synonymous substitution, nonsynonymous substitution, rate of substitution, nucleotide exchangeability, codon exchangeability.

Address for reprints and correspondence: Dr. Wen-Hsiung Li, P.O. Box 20334, Center for Demographic and Population Genetics, University of Texas, Houston, Texas 77225.

0733-4038/85/0202-0201$02.00

150
nonsynonymous substitutions separately (Kimura 1980; Miyata and Yasunaga 1980; Perler et al. 1980). This approach is desirable when the amino acid coding regions of genes are compared, for it is known that the rate of synonymous substitution is generally much greater than that of nonsynonymous substitution (e.g., Kaatos et al. 1977; Kimura 1977; Jukes and King 1979; Miyata et al. 1980). Currently, the most frequently used methods are those proposed by Miyata and Yasunaga (1980) and Perler et al. (1980). There are, however, several problems with these two methods. They are discussed fully later on, but for the present we will note that they involve the distribution of substitutions among sites, correction for multiple hits, and the weighting of alternative paths between codon pairs with more than one nucleotide difference. Because of these three problems, the two methods tend to give biased estimates when the degree of sequence divergence becomes relatively large (Brown et al. 1982; Miyata et al. 1982; Gojobori 1983). In an attempt to get better estimates, we have developed a new method. It is based on the pattern of codon degeneracy in the codon table and adopts Kimura’s (1980) method of distinguishing between transitional and transversional substitutions. It therefore allows for unequal rates of substitution among nucleotides caused by either differences between transitional and transversional rates or differences between synonymous and nonsynonymous rates. Correction for multiple substitutions can also be done in a more rigorous manner. Moreover, the relative likelihood of synonymous and nonsynonymous substitutions is determined from codon changes in mammalian genes and should therefore be more reliable than that determined from protein sequence data. As will be seen later, our method also provides more information about the substitution rates at nucleotide sites with different degrees of codon degeneracy.

After describing the method, we shall apply it to a large number of mammalian genes and compare the synonymous rate with the nonsynonymous rate in the same gene and synonymous (or nonsynonymous) rates among different genes. In addition, we shall discuss phylogenetic reconstruction and the implication of the relative likelihood of codon changes for the mechanisms of DNA evolution.

Method

In presenting the method, we shall use the nuclear genetic code as an example. Although the method is developed primarily for the study of nuclear genes, the regularity in codon degeneracy in the mammalian mitochondrial (mt) genetic code (Anderson et al. 1981) makes it even more suitable for studying mammalian mt genes.

We classify nucleotide sites into nondegenerate, twofold degenerate, and fourfold degenerate sites. A site is fourfold degenerate if all possible changes at the site are synonymous. The third positions of 32 of the 61 sense codons—e.g., GTN (Val), N being any base—are of this type. A site is twofold degenerate if one of the three possible changes is synonymous. The third positions of 24 of the 61 sense codons—e.g., CAY (His) (Y = T or C)—and the first positions of four leucine codons YAR (R = A or G) and four arginine codons CGR and AGR are of this type. (In mammalian mt genes, AGR are stop codons [Anderson et al. 1981] and hence the first positions of CGR [Arg] are nondegenerate.) We also include the third position of the three isoleucine codons in this class, though they are actually threefold degenerate sites. (In mammalian mt genes, there are only two codons for isoleucine and their third position is indeed a twofold degenerate site.) A site is
nondegenerate if all possible changes at this site are nonsynonymous or nonsense. The second positions of all sense codons and the first positions of most codons belong to this class, and so do the third positions of ATG (Met) and TGG (Trp). (In mammalian mt genes, there are two codons for Met and Trp, and their third positions are twofold degenerate sites.) We count the numbers of sites in these three classes in each of the two sequences compared and then compute the average numbers, denoting them by \( L_0 \) (nondegenerate), \( L_2 \) (twofold), and \( L_4 \) (fourfold), respectively.

The purpose of such classifications is to enable one to estimate synonymous and nonsynonymous rates of nucleotide substitution separately. Substitutions at nondegenerate sites and fourfold degenerate sites are nonsynonymous and synonymous, respectively. At the twofold degenerate sites, transversions (purine-pyrimidine changes) lead to nonsynonymous changes, whereas transitions (\( C \leftrightarrow T \) or \( A \leftrightarrow G \)) lead to synonymous changes. There is no exception to this rule in the mammalian mt genetic code. In the nuclear genetic code, there are two exceptions: the first positions of CGR and AGR (all for Arg) and the last positions of ATH (all for Ile, H = not-G). Here, we make an ad hoc adjustment to overcome this minor irregularity: we classify all synonymous changes in these two cases as “transitions” and all nonsynonymous changes as “transversions.” For example, \( C G G \) (Arg) \( \leftrightarrow \) \( A G G \) (Arg) is considered “transition,” whereas \( C G G \) (Arg) \( \leftrightarrow \) \( T G G \) (Trp) is considered “transversion,” though the contrary is actually true. With these modifications, the aforementioned rule about the twofold degenerate sites also applies to the nuclear genetic code.

We then compare two sequences codon by codon and classify each difference as either transition or transversion. We denote \( P_i \) and \( Q_i \) as the number of observed transitional and transversional differences, respectively, at \( i \)-fold degenerate sites divided by \( L_i \) (the total number of \( i \)-fold degenerate sites; \( i = 0, 2, \) or 4). For example, suppose that the two sequences compared are TTT CTA and TCT CTG, respectively. In the first sequence, the first two positions of TTT (Phe) are nondegenerate sites and the third position is a twofold degenerate site; the first position of CTA (Leu) is a twofold degenerate site, the second position is a nondegenerate site, and the third position is a fourfold degenerate site. Therefore, in the first sequence, \( L_0 = 2 + 1 = 3 \), \( L_2 = 1 + 1 = 2 \), and \( L_4 = 0 + 1 = 1 \). Similarly, we have \( L_0 = 3 \), \( L_2 = 1 \), and \( L_4 = 2 \) for the second sequence. From these, we obtain the averages \( L_0 = (3 + 3)/2 = 3 \), \( L_2 = 1.5 \), and \( L_4 = 1.5 \). Furthermore, we note that the codon pair TTT and TCT have a transitional difference at a nondegenerate site and that the pair CTA and CTG have a transitional difference at a fourfold degenerate site. Therefore, \( P_0 = 1/3 \), \( P_4 = 1/1.5 \) and the other four \( P_i \) and \( Q_i \) values are 0.

Kimura (1980) developed a two-parameter method to estimate the number of transitional and transversional substitutions per site based on the observed proportions of transitional and transversional differences. To apply his method, we assume that the transitional and transversional rates of substitution are \( \alpha_i \) and \( 2\beta_i \) substitutions per \( i \)th type site per unit time, respectively. Using his formulas 8 and 9, we can readily show that the mean and approximate error variance of the number of transitional (\( A_i \)) and transversional (\( B_i \)) substitutions per \( i \)th type site are given by

\[
A_i = (1/2) \ln (a_i) - (1/4) \ln (b_i),
\]

\[
V(A_i) = [a_i^2P_i + c_i^2Q_i - (a_iP_i + c_iQ_i)^2]/L_i,
\]

for \( i = 0, 2, \) or 4.
\[ \begin{align*}
B_i &= (1/2) \ln (b_i), \\
V(B_i) &= b_i^2 Q_i (1 - Q_i) / L_i, 
\end{align*} \]  
(3)  
(4)

where \( a_i = 1/(1 - 2p_i - Q_i), b_i = 1/(1 - 2Q_i), \) and \( c_i = (a_i - b_i)/2. \) The total number \( (K_i) \) of substitutions per \( i \)th type site is given by

\[ K_i = A_i + B_i, \]  
(5)

with a variance approximately equal to

\[ V(K_i) = [a_i^2 p_i + d_i^2 Q_i - (a_i p_i + d_i Q_i)^2] / L_i, \]  
(6)

where \( d_i = b_i + c_i. \) We note that \( A_2 \) and \( B_2 \) denote the number of synonymous and nonsynonymous substitutions per twofold degenerate site, \( K_4 = A_4 + B_4 \) the number of synonymous substitutions per fourfold degenerate site, and \( K_0 = A_0 + B_0 \) the number of nonsynonymous substitutions per nondegenerate site.

Let \( K_S \) be the number of (synonymous) substitutions per synonymous site and \( K_A \) the number of (nonsynonymous) substitutions per nonsynonymous site. Following convention, we count each fourfold degenerate site as a synonymous site, each nondegenerate site as a nonsynonymous site, and each twofold degenerate site as one-third synonymous and two-thirds as nonsynonymous. (This is because, if a twofold degenerate site mutates randomly, only one-third of the time would it be scored as synonymous and two-thirds of the time as nonsynonymous.) We then obtain

\[ \begin{align*}
K_S &= (L_2 A_2 + L_4 K_4)/(L_2/3 + L_4) = 3(L_2 A_2 + L_4 K_4)/(L_2 + 3L_4), \\
V(K_S) &= 9[L_2^2 V(A_2) + L_4^2 V(K_4)]/(L_2 + 3L_4)^2, \\
K_A &= (L_2 B_2 + L_0 K_0)/(2L_2/3 + L_0) = 3(L_2 B_2 + L_0 K_0)/(2L_2 + 3L_0), \\
V(K_A) &= 9[L_2^2 V(B_2) + L_0^2 V(K_0)]/(2L_2 + 3L_0)^2. 
\end{align*} \]  
(7)  
(8)  
(9)  
(10)

There are also cases where nucleotide differences cannot be unambiguously classified. This happens when two codons differ by more than one nucleotide; for example, CCC (Pro) versus CAA (Gln). The change at the second position can be unambiguously included in \( Q_0 \) but that at the third position may be included in either \( Q_4 \) or \( Q_2, \) depending on whether the pathway of changes was CCC ↔ CCA ↔ CAA or CCC ↔ CAC ↔ CAA. (Note that there are one synonymous and one nonsynonymous change in the former pathway but that there are two nonsynonymous changes in the latter.) Since \( Q_4 \) is used to estimate the synonymous rate and \( Q_2 \) is used for the nonsynonymous rate, misclassification may result in erroneous estimates. The situation is more complex, though much rarer, when two compared codons differ at all three positions. For example, among the 6 (=3!) pathways between TCG (Ser) and CGT (Arg), three pathways require three nonsynonymous changes, one pathway requires two, and the remaining two pathways require only one nonsynonymous change. To determine the probabilities of different pathways in ambiguous cases, adequate information on the relative likelihood of codon changes is necessary. This is the subject of the following section.
A computer program for our method is available by sending us a floppy disk. Although the computation algorithm is somewhat complicated, the program is very easy to use.

Relative Likelihood of Codon and Amino Acid Changes

In the method of Perler et al. (1980), all codon changes are implicitly assumed to be equally probable. Miyata and Yasunaga (1980), on the other hand, determine the relative likelihood of codon changes by extrapolating from previous analyses of amino acid exchangeability (McLachlan 1971; Grantham 1974; Miyata et al. 1979). In this section, we evaluate such likelihoods from DNA sequence data by taking the following steps: (1) The codons at all ancestral nodes of a phylogenetic tree are inferred from the present-day DNA sequences, and the frequencies of codon changes between every node and its immediate descendant are recorded as observed frequencies. (2) The expected frequencies of codon changes are formulated. (3) The observed and expected codon changes are classified according to the differences in the physicochemical properties between the encoded amino acids. The relative likelihood of a particular codon or amino acid change is simply the ratio of the observed frequency to the expected one.

Inferring the Ancestral Codons

Fitch (1971) developed an algorithm that, given a phylogenetic tree, will find all the minimum distance solutions at any nucleotide position. (A minimum distance solution is an assignment to every ancestral node of a nucleotide that gives the smallest number of nucleotide substitutions for the given tree.)

If the DNA sequences available constitute an adequately branched tree, the nucleotide at each ancestral node can be uniquely determined; of course, uniqueness does not guarantee correctness (Holmquist 1979). The ancestral codon can certainly be uniquely determined if all three positions are. When the codon at any of the ancestral nodes has more than one solution, we take the average by assigning a probability to each solution as Fitch (1971) did and further assume that all positions are independent of one another. Therefore, if T has a probability of 1/4 at the first position, C is uniquely determined at the second position, and A has a probability of 1/4 at the third position, and the probability of having TCA is 1/16. Rules for connecting an ancestral codon to its immediate descendants can be similarly extended from the description of Fitch (1971). If the codons of adjacent nodes differ at more than one position, the “average” solution, i.e., each position assumed to have an equal probability of substitution prior to others, is used. This procedure favors nonsynonymous over synonymous substitutions. It will, however, introduce no serious bias, for the proportion of cases where the ancestral codon cannot be uniquely determined is small in the data we used.

Once the ancestral codons have been assigned, we can then record the frequencies of observed changes between any pair of codons differing by one nucleotide.

Expected Frequencies of Codon and Amino Acid Changes

If codon changes are determined solely by mutation, then the relative frequencies of changes from, say, TCA to CCA, ACA, GCA, TTA, TCT, TCC, and TCG should be $f_{TC}$, $f_{TA}$, $f_{TG}$, $f_{CT}$, $f_{AT}$, $f_{AC}$, and $f_{AG}$, respectively, where $f_{XY}$ is the relative
frequency of substituting nucleotide Y for X as defined by Gojobori et al. (1982b). Substitution of A or G for C at the second position of TCA results in one of the termination codons and is not counted.

Let us now number the codons from 1 to 61, denoted by n, by varying the third position first, followed by the second and then the first position. The three termination codons are skipped. We also label the amino acids, in alphabetical order, from 1 to 20. We may now define \( P_s(j) \) to be the probability that the \( n \)th codon, after one nucleotide substitution, codes for the \( j \)th amino acid (denoted aa\(_i\)). In the above example, TCA is the seventh codon, CCA codes for Pro (aa\(_{15}\)), ACA codes for Thr (aa\(_{17}\)), and so on. Therefore, \( P_s(15) = f_{TC} \), \( P_s(17) = f_{TA} \), and so on. \( P_s(j) = 0 \) if the \( n \)th codon cannot change to a codon for aa\(_i\) in one step. The expected relative frequency of having aa\(_i\) replaced by aa\(_j\) in the genes studied can then be expressed as

\[
e_{ij} = \sum_{n \in N_i} P_s(j)T_n,
\]

where \( T_n \) is the observed number of the \( n \)th codon in the inferred ancestral sequences and \( N_i \) is the set of codons coding for the \( i \)th amino acid. The quantity \( P_s(j)T_n \) is summed over all codons that code for aa\(_i\). Note that \( e_{ij} \) (the expected frequency of synonymous changes between codons for aa\(_i\)) is well defined and that \( e_{ij} \neq e_{ji} \). Similarly, we can calculate the observed frequency \( o_{ij} \) of changes from aa\(_i\) to aa\(_j\), from the changes between codons inferred from DNA sequences.

Classification of Codon Changes and Their Relative Likelihoods

Since there would be 400 elements in the \([e_{ij}]_{20 \times 20}\) matrix, each observed element \( o_{ij} \) may be subject to large sampling errors. We therefore classify all the amino acid changes into four classes on the basis of Grantham’s (1974) distance between amino acids. He assigns a distance, \( d_{ij} \), between aa\(_i\) and aa\(_j\) according to the physicochemical properties of the two amino acids. The value of \( d_{ij} \) ranges from 5 (Ile-Leu) to 215 (Cys-Trp). Li et al. (1984) define amino acid changes as conservative, moderately conservative, moderately radical, and radical if \( d_{ij} \) is less than 50, between 50 and 100, between 100 and 150, and greater than 150, respectively. Using this definition, we sum up all \( e_{ij} \)'s and \( o_{ij} \)'s with \( d_{ij} \)'s in the same range. In addition, we have the synonymous class in which \( e_{ij} \)'s and \( o_{ij} \)'s are summed from \( i = 1, 2, \ldots, 20 \).

Table 1 gives the results of the relative likelihood of codon changes for genes with moderate rates of substitution and for those with very high rates (see below). Conservative genes like histone and actin genes have too few amino acid replacements for such a survey. The observed changes are not whole numbers because they represent the average over all possible ancestral codons. The expected changes are normalized so that the total number of expected changes equals that of observed ones. Since only mammalian genes are used, multiple substitutions beyond the minimum distance solution should not be a serious problem for nonsynonymous substitutions. This assumption, however, tends to underestimate to some extent the number of synonymous substitutions because the chance of multiple synonymous substitutions at the same site may not be negligible between sequences from different mammalian orders (e.g., see Holmquist et al. 1982). In the case of genes evolving
Table 1
Relative Likelihoods of Codon Changes*

<table>
<thead>
<tr>
<th>Types of Changes</th>
<th>Genes with Moderate Substitution Rates $^b$</th>
<th>Genes with High Substitution Rates $^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous</td>
<td>960.5/429.8 = 2.235</td>
<td>477.6/292.1 = 1.635</td>
</tr>
<tr>
<td>Conservative</td>
<td>378.5/443.7 = 0.853</td>
<td>331.3/318.2 = 1.041</td>
</tr>
<tr>
<td>Moderately conservative</td>
<td>401.1/522.8 = 0.767</td>
<td>337.8/375.8 = 0.899</td>
</tr>
<tr>
<td>Moderately radical</td>
<td>89.2/313.1 = 0.285</td>
<td>108.5/222.7 = 0.487</td>
</tr>
<tr>
<td>Radical</td>
<td>11.7/131.5 = 0.089</td>
<td>36.8/83.2 = 0.442</td>
</tr>
<tr>
<td>Total</td>
<td>1841</td>
<td>1292</td>
</tr>
</tbody>
</table>

* The relative likelihood is the ratio of the observed frequency to the expected one. The expected frequencies are calculated according to the description in the text, with the assumption that substitution rates among all nucleotide pairs are equal.


at moderate rates, we choose only those for which DNA sequences from three or more mammalian species have been available. This helps to reduce ambiguity in inferring ancestral codons. For very rapidly evolving genes, we have to use also genes for which only two sequences are available.

From table 1, we find that the greater the physicochemical distance between two amino acids is, the less likely one will replace the other. For example, for every moderately radical replacement in a gene with a moderate nonsynonymous rate, 7.84 ($=2.235/0.285$) synonymous changes would be observed. (If there were no selective constraints on amino acid replacement, the ratio would be 1.) Not surprisingly, in rapidly evolving genes such as relaxin and immunoglobulin genes, amino acid exchanges are accepted to a much greater extent; even radical replacements occur at a substantial rate.

An interesting point from table 1 is that one should not extrapolate the relative likelihood (obs/exp) of amino acid replacement to estimate that of synonymous changes as suggested by Miyata and Yasunaga (1980). If we draw a regression line for the relative likelihood of amino acid exchanges on the amino acid distance ($d_{ij}$, using the midpoint of each class), we obtain good correlation coefficients for both types of genes ($r = -0.97$ and $-0.96$, respectively). However, the intercepts at $d_{ij} = 0$ are 1.05 and 1.16 for the two types of genes, respectively. These extrapolated values are considerably lower than the likelihoods (2.24 and 1.64, respectively) shown in table 1.

To apply the results of table 1, we should consider genes with slow, moderate, and high nonsynonymous rates separately. The distinctions are certainly not well
defined, but, generally speaking, this means genes with less than 0.6, between 0.6 and 1.2, and greater than 1.2 nonsynonymous substitutions per nonsynonymous site per $10^9$ years, respectively. Table 1 does not contain the relative likelihood of codon changes in slowly evolving genes. To estimate the substitution rates in conservative genes like histone, actin, and many hormone genes, we may use the relative likelihoods derived from “moderately conservative” genes. This is because conservative genes rarely require probability assignments to different substitution pathways, so that the results depend little on how we weight the pathways. The exceptions are codons for Leu and Arg. The two pathways between TTA and CTT (both for Leu) may pass through either CTA (Leu) or TTT (Phe). The former takes two synonymous substitutions, whereas the latter requires two exchanges between Leu and Phe. In such cases, we assume that two synonymous substitutions have taken place.

Finally, the relative likelihood of each pathway is simply the product of the relative likelihoods of the codon changes along the pathway (see Miyata and Yasunaga [1980] for details). For example, consider the pathway between CCC and CAA via CCA. Since CCC (Pro) ↔ CCA (Pro) is synonymous with a relative likelihood ($r$) equal to 2.235 and since CCA (Pro) ↔ CAA (Gln) is moderately conservative, with $r = 0.767$, this pathway is given a relative likelihood of 2.235 × 0.767 = 1.71. Similarly, the other pathway, CCC (Pro) ↔ CAC (His) ↔ CAA (Gln), is given the relative likelihood of 0.767 × 0.853 = 0.654. The probability of the former pathway is therefore 1.71/(1.71 + 0.654) = 0.723 and that of the latter is 0.277. In other words, between CCC and CAA, we assume that 0.723 synonymous changes and 0.723 + 2 × 0.277 = 1.276 nonsynonymous changes have occurred. Codons differing in all three positions can be similarly treated.

Substitution Rates in Mammalian Genes

In this section, we apply our method to study the rates of synonymous and nonsynonymous substitution in various kinds of genes. For several reasons, we shall use mainly genes from mammals: man, rodents (mouse, rat, and Chinese hamster), and artiodactyls (cow, goat, and pig). First, the fossil record for mammals is better than those for other organisms; we assume that the above mammalian orders diverged 80 Myr ago. Second, many genes from mammals have been sequenced. Third, to get a reliable estimate of the rate of nucleotide substitution, the degree of sequence divergence should not be too small or too large. For this reason, genes from different mammalian orders are generally very suitable for estimating both the synonymous and nonsynonymous rates. For extremely conservative genes, such as those coding for histones and glucagon, estimates of nonsynonymous rates from comparisons of mammalian species may not be reliable, for the number of nonsynonymous substitutions is too small; and we have therefore also obtained estimates from comparisons between mammals and chicken or fish.

The results of our analysis are presented in table 2. In those cases where more than two sequences are used, the mean ($d$) and variance [$V(d)$] are obtained as follows. Suppose the $n$ sequences used are from $n$ mammalian orders. We assume that all mammalian orders diverged at the same time, i.e., 80 Myr ago. For each type of nucleotide site, we first compute the mean ($d_{ij}$) and variance [$V(d_{ij})$] of the number of substitutions between sequences $i$ and $j$. We then compute $d$ and $V(d)$ according to the following formulas:
<table>
<thead>
<tr>
<th>Genes</th>
<th>NONDEGENERATE SITES</th>
<th>TWOFOLD DEGENERATE SITES</th>
<th>FOURFOLD DEGENERATE SITES</th>
<th>NON-SYNONYMOUS RATE</th>
<th>SYNONYMOUS RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L&lt;sup&gt;ε&lt;/sup&gt;</td>
<td>Transition</td>
<td>Transversion</td>
<td>Total (±SE)</td>
<td>Transition</td>
</tr>
<tr>
<td>Histones:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td>101</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00 (0.00)</td>
<td>1.99</td>
</tr>
<tr>
<td>H3</td>
<td>101</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00 (0.00)</td>
<td>0.43</td>
</tr>
<tr>
<td>H2B</td>
<td>135</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04 (0.02)</td>
<td>1.24</td>
</tr>
<tr>
<td>H2A</td>
<td>120</td>
<td>0.03</td>
<td>0.04</td>
<td>0.07 (0.04)</td>
<td>1.44</td>
</tr>
<tr>
<td>Contractile system proteins:</td>
<td>126</td>
<td>0.02</td>
<td>0.06</td>
<td>0.08 (0.03)</td>
<td>0.75</td>
</tr>
<tr>
<td>Actin, α</td>
<td>376</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01 (0.01)</td>
<td>1.81</td>
</tr>
<tr>
<td>Hormones and active peptides:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Somatostatin-28</td>
<td>28</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00 (0.00)</td>
<td>0.52</td>
</tr>
<tr>
<td>Gastrin</td>
<td>82</td>
<td>0.00</td>
<td>0.09</td>
<td>0.09 (0.12)</td>
<td>0.78</td>
</tr>
<tr>
<td>Glucagon</td>
<td>29</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00 (0.00)</td>
<td>3.42</td>
</tr>
<tr>
<td>Insulin</td>
<td>51</td>
<td>0.02</td>
<td>0.09</td>
<td>0.11 (0.07)</td>
<td>2.34</td>
</tr>
<tr>
<td>ACTH</td>
<td>39</td>
<td>0.05</td>
<td>0.16</td>
<td>0.22 (0.11)</td>
<td>1.06</td>
</tr>
<tr>
<td>Parathyroid hormone</td>
<td>90</td>
<td>0.34</td>
<td>0.18</td>
<td>0.52 (0.15)</td>
<td>0.89</td>
</tr>
<tr>
<td>Corticotropin releasing factor</td>
<td>41</td>
<td>0.32</td>
<td>0.34</td>
<td>0.66 (0.13)</td>
<td>0.75</td>
</tr>
<tr>
<td>Glycoprotein hormone, α</td>
<td>92</td>
<td>0.36</td>
<td>0.42</td>
<td>0.78 (0.14)</td>
<td>2.99</td>
</tr>
<tr>
<td>Growth hormone</td>
<td>189</td>
<td>0.67</td>
<td>0.35</td>
<td>1.02 (0.12)</td>
<td>2.28</td>
</tr>
<tr>
<td>Insulin C-peptide</td>
<td>35</td>
<td>0.29</td>
<td>0.75</td>
<td>1.04 (0.29)</td>
<td>1.51</td>
</tr>
<tr>
<td>Prolactin</td>
<td>197</td>
<td>0.77</td>
<td>0.57</td>
<td>1.34 (0.14)</td>
<td>1.84</td>
</tr>
<tr>
<td>Relaxin C-peptide</td>
<td>99</td>
<td>1.19</td>
<td>0.85</td>
<td>2.03 (0.27)</td>
<td>1.79</td>
</tr>
<tr>
<td>Relaxin</td>
<td>54</td>
<td>1.51</td>
<td>1.30</td>
<td>2.81 (0.44)</td>
<td>2.54</td>
</tr>
<tr>
<td>Signal Peptides:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Somatostatin</td>
<td>23</td>
<td>0.00</td>
<td>0.31</td>
<td>0.31 (0.23)</td>
<td>0.70</td>
</tr>
<tr>
<td>Insulin</td>
<td>23</td>
<td>0.65</td>
<td>0.56</td>
<td>1.21 (0.39)</td>
<td>0.44</td>
</tr>
<tr>
<td>Relaxin</td>
<td>23</td>
<td>1.87</td>
<td>0.63</td>
<td>2.50 (0.71)</td>
<td>2.33</td>
</tr>
<tr>
<td>Hemoglobin:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>α chain</td>
<td>141</td>
<td>0.19</td>
<td>0.40</td>
<td>0.59 (0.10)</td>
<td>1.97</td>
</tr>
<tr>
<td>β chain</td>
<td>144</td>
<td>0.35</td>
<td>0.53</td>
<td>0.89 (0.13)</td>
<td>1.48</td>
</tr>
<tr>
<td>Protein</td>
<td>Value</td>
<td>Standard Error</td>
<td>t-Value</td>
<td>p-Value</td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------</td>
<td>----------------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>Immunoglobulin (Ig) and related proteins:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig Vα</td>
<td>100</td>
<td>0.51</td>
<td>1.97</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>Ig Vβ</td>
<td>105</td>
<td>0.41</td>
<td>0.52</td>
<td>3.13</td>
<td></td>
</tr>
<tr>
<td>Ig γ</td>
<td>99</td>
<td>0.41</td>
<td>0.72</td>
<td>4.50</td>
<td></td>
</tr>
<tr>
<td>Ig κ</td>
<td>321</td>
<td>0.60</td>
<td>1.86</td>
<td>2.44</td>
<td></td>
</tr>
<tr>
<td>Ig λ</td>
<td>106</td>
<td>1.18</td>
<td>1.65</td>
<td>4.95</td>
<td></td>
</tr>
<tr>
<td>Interferons:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>α1</td>
<td>166</td>
<td>0.70</td>
<td>0.60</td>
<td>2.14</td>
<td></td>
</tr>
<tr>
<td>α2</td>
<td>165</td>
<td>0.79</td>
<td>0.82</td>
<td>2.14</td>
<td></td>
</tr>
<tr>
<td>β1</td>
<td>159</td>
<td>1.33</td>
<td>1.03</td>
<td>3.15</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>136</td>
<td>1.85</td>
<td>1.45</td>
<td>4.63</td>
<td></td>
</tr>
<tr>
<td>Other proteins:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPRT</td>
<td>217</td>
<td>0.13</td>
<td>0.03</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>Neurophysin II</td>
<td>93</td>
<td>0.18</td>
<td>0.47</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>Metallothionein II</td>
<td>60</td>
<td>0.34</td>
<td>0.66</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>Fibrinogen γ</td>
<td>411</td>
<td>0.33</td>
<td>0.28</td>
<td>3.27</td>
<td></td>
</tr>
<tr>
<td>Metallothionein I</td>
<td>60</td>
<td>0.34</td>
<td>0.18</td>
<td>2.24</td>
<td></td>
</tr>
<tr>
<td>Albumin</td>
<td>590</td>
<td>0.35</td>
<td>0.64</td>
<td>4.02</td>
<td></td>
</tr>
<tr>
<td>α-Lactalbumin</td>
<td>123</td>
<td>0.55</td>
<td>0.50</td>
<td>2.55</td>
<td></td>
</tr>
<tr>
<td>α-Fetoprotein</td>
<td>586</td>
<td>0.76</td>
<td>0.55</td>
<td>2.69</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>1.83</td>
<td>0.42</td>
<td>2.26</td>
<td>1.68</td>
<td></td>
</tr>
</tbody>
</table>

* Unless stated otherwise, all rates are obtained from comparisons between different mammalian orders, which are assumed to have diverged 80 Myr ago. All rates are in substitutions per site per 10^9 year.

| References: | histone H4 (Zhong et al. 1983); Seiler-Tuyns and Birnstiel 1981; Sugarman et al. 1983; histone H3 (Zhong et al. 1983); Sittman et al. 1983; Engel et al. 1982; histone H2A (Zhong et al. 1983); Harvey et al. 1982; histone H2B (Zhong et al. 1983); Sittman et al. 1983; actin-α (Hanauer et al. 1983; Zakut et al. 1982); somatostatin-28 (Shen et al. 1982; Goodman et al. 1983); gastrin (Kato et al. 1983; Yoo et al. 1982); glucagon (Bell et al. 1983b; Lopez et al. 1983; Bell et al. 1983a); insulin (Bell et al. 1980; Wetkam et al. 1982; Kwok et al. 1983; Lomendico et al. 1979); ACTH (Takahashi et al. 1983; Nakamishita et al. 1981; Notake et al. 1983); parathyroid hormone (Hendy et al. 1981); corticotropin releasing factor (Shibahara et al. 1983); glycoprotein hormone, α subunit (Fiddes et al. 1981; Nilson et al. 1983; Chin et al. 1981; Godine et al. 1982); growth hormone (Miller and Eberhardt 1983; Seeberg et al. 1983); proactin (Miller and Eberhardt 1983); relaxin (Hudson et al. 1983; Hudson et al. 1981; Haley et al. 1982); hemoglobin α (Liebhaber et al. 1980; Liebhaber and Begley 1983; Schon et al. 1982; Heindell et al. 1978; Nishioka and Leder, 1979); hemoglobin β (Lawn et al. 1980; Schon et al. 1981; Hardison et al. 1979; Konkel et al. 1979); immunoglobulin (Ig) Vα (Rochavi et al. 1982); Ig γ (Ellison et al. 1982; Honjo et al. 1979); Ig κ (Hieter et al. 1980; Sheppard and Gutman 1981; Emorine et al. 1983); interferon-α1 (Shaw et al. 1983); interferon-α2 (Lawn et al. 1981; Shaw et al. 1983); interferon-β1 (Higashi et al. 1983); interferon-γ (Gray and Goeddel 1983); HPRT (Jolly et al. 1983; Konecki et al. 1982); neurophysin II (Land et al. 1982; Schmale et al. 1983); metallothionein II (Karim and Richards 1982; Schmidt and Hamer 1983; Griffith et al. 1983); metallothionein I (Andersen et al. 1983; Schmidt and Hamer 1983; Griffith et al. 1983); Fibrinogen γ (Chung et al. 1983; Crabtree and Kant 1982); Albumin (Dugacizky et al. 1982; Sargent et al. 1981); α-lactalbumin (Hall et al. 1982; Dandekar et al. 1981); and α-fetoprotein (Morinaga et al. 1983; Law and Dugacizky 1981; Jagodziński et al. 1981). |

- Number of codons compared.
- All rates in these cases are obtained from comparisons between mammals and birds or fish. The divergence time between mammals and birds is assumed to be 270 Myr and between mammals and fish to be 400 Myr.
- This rate is the mean and variance of the rate of nonsynonymous substitution are obtained from comparisons between mammals and birds or fish.
- The average for all proteins is obtained only from comparisons between mammalian orders.
\[ \bar{d} = \sum_{i < j}^{n} \hat{d}_{ij}/[n(n-1)/2], \]

\[ V(d) = 4 \sum_{i < j}^{n} V(d_{ij})/[n^2(n-1)]. \]

In the case where more than one sequence from the same mammalian order are used, formulas for computing \( \bar{d} \) and \( V(d) \) can also be obtained (see Wu and Li 1985). For simplicity, however, we have computed \( \bar{d} \) and \( V(d) \) by taking a simple average. For example, if sequence 1 is from man and sequences 2 and 3 are from mouse and rat, then the relations \( \bar{d} = (\bar{d}_{12} + \bar{d}_{13})/2 \) and \( V(d) = [V(d_{12}) + V(d_{13})]/2 \) are used. In each gene group of table 2, the results are arranged in the increasing order of the rate of nonsynonymous substitution. In the following we discuss a number of interesting features.

First, the rate of nonsynonymous substitution is extremely variable among different kinds of genes; it ranges from \( 0.004 \times 10^{-9} \) (histone H4) to 2.80 \( \times 10^{-8} \) (interferon-\( \gamma \)) substitutions per nonsynonymous site per year, a more than hundredfold difference (see the next to last column of table 2). As is well known, histone and actin genes are extremely conservative. Some hormone genes, e.g., somatostatin-28, glucagon, gastrin, and insulin genes, are also extremely conservative, but some others, e.g., relaxin and prolactin, evolve rapidly. The insulin C peptide has often been used as an example of rapid evolution, but it actually evolves considerably more slowly than relaxin and the relaxin C peptide. The nonsynonymous rate observed in the signal peptide region of the insulin gene (1.2 \( \times 10^{-9} \)) is about the average for signal peptide regions (results not shown); the extremely high and low rates in the signal peptide regions of the relaxin and somatostatin genes, respectively, are probably extreme random deviates, for signal peptide regions are very short. Hemoglobin genes evolve at an intermediate rate, while immunoglobulin and interferon genes evolve rapidly. The average rate of nonsynonymous substitution for all the gene comparisons from mammals is 0.88 \( \times 10^{-9} \) substitutions per nonsynonymous site per year (see the last row). Interestingly, the median of the nonsynonymous rates in table 2, 0.87 \( \times 10^{-9} \) for \( \beta \)-globin, is almost identical with the mean.

Second, the rate of synonymous substitution varies greatly from gene to gene (see the last column of table 2), though not so much as the rate of nonsynonymous substitution. The variation is especially great among genes with fewer than 50 fourfold degenerate sites; for example, in \( \beta \)2-microglobulin the rate is as high as 11.8 \( \times 10^{-9} \), while in parathyroid hormone it is as low as 1.7 \( \times 10^{-9} \). It appears that our results do not support Miyata et al.'s (1980) suggestion, based on a considerably smaller number of genes, that the rate of synonymous substitution is similar among genes. From table 2, the average rate of synonymous substitution for all the comparisons of genes from mammals is 4.7 \( \times 10^{-9} \) substitutions per synonymous site per year. This is five times higher than the average rate of nonsynonymous substitution. Our estimate is not much different from Hayashida and Miyata's (1983) estimate of 5.5 \( \times 10^{-9} \). (Their estimate reduces to 5.1 \( \times 10^{-9} \) if 80 Myr instead of 75 Myr is used as the divergence time between mammalian orders. Note also that we used 36 genes, while they used only 11 genes.) Interestingly, in histones H4 and H3 the synonymous rates are very high, though the nonsynon-
ymous rates are extremely low. There is, however, a tendency for a gene with a high nonsynonymous rate to have a high synonymous rate; in the majority of genes with a nonsynonymous rate higher than $1.0 \times 10^{-9}$, the synonymous rate is higher than the average.

To examine the issue of uniform synonymous rate in greater detail, we have used only those genes in table 2 for which there are more than 50 fourfold degenerate sites (to reduce stochastic noise) and compared only sequences from different mammalian orders. Another source of variation may be introduced if DNA sequences from rodents are used in the studies of some genes but not in others. There is evidence that synonymous rates as well as substitution rates in weakly constrained DNA regions are significantly higher in both mouse and rat than in other mammals (Wu and Li 1985). One plausible explanation for this is the very short generation time of these rodents. For each gene, we therefore separate the comparisons of DNA sequences that involve a rodent sequence from those that do not. We then compute the averages for these two types of comparisons separately. (A gene may thus have two different synonymous rates, one involving rodents and the other involving only other mammalian orders.)

There are 21 genes from table 2 that meet the above criteria and have at least one rodent sequence. The synonymous rates of these genes involving rodent sequences are distributed as follows: one less than $2.5 \times 10^{-9}$, eight between $2.5 \times 10^{-9}$ and $5.0 \times 10^{-9}$, and 12 between $5.0 \times 10^{-9}$ and $7.5 \times 10^{-9}$; the mean of these rates is $4.97 \times 10^{-9}$. There are only 10 genes that have sequences from two nonrodent mammalian orders; the distribution of their synonymous rates are as follows: three less than $2.5 \times 10^{-9}$, six between $2.5 \times 10^{-9}$ and $5.0 \times 10^{-9}$, and one between $5.0 \times 10^{-9}$ and $7.5 \times 10^{-9}$; the mean is $3.41 \times 10^{-9}$. Therefore, the synonymous rates in comparisons that involve rodents tend to be considerably higher than those in comparisons that do not. In both sets of comparisons, however, synonymous rates among different genes do not appear to be as uniform as suggested by Miyata et al. (1980) and Hayashida and Miyata (1983): the highest rates are about three times the lowest ones in both sets. Caution therefore should be used when synonymous rates are used as a molecular clock to date the divergence time between genes or species.

Third, at fourfold degenerate sites the rate of transitional substitution tends to be higher than the rate of transversional substitution, though at each site two types of transversional change and only one type of transitional change can occur. The averages for these two rates are $2.5 \times 10^{-9}$ and $1.7 \times 10^{-9}$, respectively. This observation can be explained by the fact that transitional mutation occurs more frequently than transversional mutation (Li et al. 1984). At twofold degenerate sites, the rate of transitional substitution is on the average about the same as that at fourfold degenerate sites, but the rate of transversional substitution is usually considerably lower than the corresponding rate at fourfold degenerate sites; the averages for the two rates at twofold degenerate sites are $1.8 \times 10^{-9}$ and $0.44 \times 10^{-9}$, respectively. This lower rate of transversional substitution occurs because all transversional changes at twofold degenerate sites are nonsynonymous. At nondegenerate sites, the rates of transitional and transversional substitution are on the average about the same ($0.48 \times 10^{-9}$ and $0.45 \times 10^{-9}$, respectively) and are usually considerably lower than the corresponding values at fourfold degenerate sites. These low rates occur because all changes at nondegenerate sites are nonsy-
nononymous. Thus, the (total) rate of nucleotide substitution is lowest at nondegenerate sites, intermediate at twofold degenerate sites, and highest at fourfold degenerate sites; the averages are $0.94 \times 10^{-9}$, $2.3 \times 10^{-9}$, and $4.2 \times 10^{-9}$, respectively.

Fourth, the rate of nucleotide substitution at fourfold degenerate sites tends to be lower than the rate of synonymous substitution (last column in table 2), though all changes at fourfold degenerate sites are synonymous. Indeed, the average value is only $4.2 \times 10^{-9}$ for the former rate but $4.7 \times 10^{-9}$ for the latter rate. This difference occurs because in computing the rate of synonymous substitution, the transitional changes at twofold degenerate sites are also included. As mentioned above, only one-third of a twofold degenerate site is counted as synonymous but transitional (synonymous) substitution at a twofold degenerate site occurs at a rate higher than one-third of the total rate at a fourfold degenerate site. For this reason, the synonymous rate obtained by the conventional definition tends to be higher than the rate of nucleotide substitution at fourfold degenerate sites. In inferring the stringency of functional constraint on synonymous changes, it has been customary to compare the rate of synonymous substitution with the substitution rate in pseudogenes, which is taken as the neutrality standard. It now appears that for this purpose one should compare instead the substitution rate at fourfold degenerate sites with the substitution rate in pseudogenes. On the other hand, the rate of nucleotide substitution at nondegenerate sites is usually quite similar to the nonsynonymous rate. This is because the rate of transversional (nonsynonymous) substitution at twofold degenerate sites is on the average quite similar to the rates of transitional and transversional substitution at nondegenerate sites and also because in an average gene the number of twofold degenerate sites is about only one-fourth of the number of nondegenerate sites, so that a different rate at these sites would not affect the (weighted) average much. Therefore, there is no need to make a distinction between the substitution rate at nondegenerate sites and the nonsynonymous rate.

In computing the rates presented in table 2, we have assumed that the divergence time between two genes is the same as the divergence time between the two species from which the genes were obtained. In those genes that belong to multigene families there is a possibility that the genes compared were actually derived from a duplication before the species divergence or, in other words, that we compared paralogous rather than orthologous genes. This possibility exists for the histone genes, immunoglobulin genes, and the interferon-α and -β genes. Another possibility that can inflate the estimated rates is gene conversion between nonallelic genes of a multigene family.

Discussion
Estimation Methods

In our method, correction for multiple substitutions at the same site is made by using Kimura’s (1980) two-parameter method, which takes into account only the possibility of unequal rates of transitional and transversional substitutions. Although this is usually the largest deviation from random substitution, other possibilities of nonrandom substitution may also be important. Our method will therefore tend to give underestimates when the degree of sequence divergence becomes large. In fact, this might be the reason why the rates of synonymous substitution in histones H4 and H3 obtained from comparisons between mammals
and chickens are considerably lower than those obtained from comparisons among mammals (table 2). Although it is not difficult to incorporate into our model more possibilities of nonrandom substitution, the computational procedure becomes complicated and the results are difficult to interpret. Moreover, incorporation of many parameters in a model tends to reduce its applicability when the degree of sequence divergence becomes large (Gojobori et al. 1982a; Tajima and Nei 1984). Of course, these are common difficulties in the estimation of the number of nucleotide substitutions (Gojobori et al. 1982a; Li et al. 1985).

We should also mention that our formula for the variances of $A_i$ and $B_i$ are approximations. We have neglected the covariance between $P_i$ and $Q_i$ because we assumed that $P_i$ and $Q_i$ are independent binomial variables. The covariance is, however, usually small compared with the variances of $P_i$ and $Q_i$. We also note that the classification of a site as synonymous or not depends on the sites adjacent to it. Therefore, changes in adjacent sites can affect the status of a site. This effect would tend to increase the variances of $A_i$ and $B_i$. The effect is probably negligible when $A_i$ and $B_i$ are small but may not be so when $A_i$ and $B_i$ are large.

We now compare our method with those of Miyata and Yasunaga (1980) and Perler et al. (1980). In our view, the latter two methods have three drawbacks. First, both methods assume that nucleotide substitution occurs randomly. This assumption tends to underestimate the number of substitutions, particularly if the degree of sequence divergence is large (Takahata and Kimura 1981; Gojobori et al. 1982a; Gojobori 1983). By contrast, our method allows for the difference between transitional and transversional rates, which is usually the largest deviation from random substitution (Li et al. 1984). Second, correction for multiple substitutions at the same site has not been done satisfactorily in previous methods. Perler et al.'s formulas are equivalent to assuming that, at twofold degenerate sites, substitutions can occur only between two states when considering synonymous substitutions and only among three states when considering nonsynonymous substitutions. For example, TTT is allowed to change to TTC, but not to TTA or TTG, in the estimation of synonymous rates. This assumption is incorrect. In Miyata and Yasunaga's method, synonymous substitutions at twofold and fourfold degenerate sites are considered together and so are nonsynonymous substitutions at twofold degenerate and nondegenerate sites. They then use Jukes and Cantor's (1969) method to make corrections for multiple substitutions. This is rather ad hoc. In our method, twofold degenerate sites, fourfold degenerate sites, and nondegenerate sites are considered separately, so that corrections for multiple substitutions can be done more rigorously. Moreover, in our method the error variances can also be derived in a more rigorous manner. (Actually, using $P_i$, $Q_i$, and $L_i$, we have also included in our computer program a computational procedure similar to that of Miyata and Yasunaga. Of course, our weights for the possible paths between codons are different from theirs.) Third, Perler et al.'s assumption of equal probability for different pathways tends to underestimate the rate of synonymous substitution, particularly when the degree of sequence divergence becomes large. Miyata and Yasunaga do assign different weights for different pathways, but, as mentioned above, their weights are more in favor of nonsynonymous than synonymous substitutions. Therefore, their method also tends to underestimate the rate of synonymous substitution. As noted above, our weights are based on direct comparisons of DNA sequences and should therefore be more reasonable than those of Miyata and
Yasunaga. We should, however, emphasize that these three drawbacks would become serious only when the degree of sequence divergence becomes relatively large. In fairly closely related sequences—say, $K_S < 0.5$—their methods and ours are expected to give similar results.

In addition to the above, our method has two more advantages. First, it is easier to understand. The computational procedures in the other two methods are not easy to understand, and the classification of sites in Perler et al.’s method is quite complicated. Second, it provides more information on the substitution rates at sites with different degrees of coding degeneracy. This is useful for understanding the mechanisms of DNA evolution (see below).

The three methods have been applied to β-globin genes, and the results are shown in table 3. All three methods give very similar estimates for the number ($K_A$) of nonsynonymous substitutions per nonsynonymous site. This is to be expected, because the $K_A$ values are only 0.5 or smaller, so that the chance for multiple substitutions to occur at a site is small—or, in other words, the corrected values are close to the uncorrected values regardless of the method used. In the comparisons between mammals and in those between mammals and birds, the three methods also give similar estimates for the number ($K_S$) of synonymous substitutions per synonymous site. This is also not surprising, for the $K_S$ values are only about 0.8 or smaller. In the comparisons between mammals and frog, however, the three methods give very different estimates. As expected, the estimate by Perler et al.’s method is the smallest, while that by our method is the largest. Of course, one cannot conclude from these examples that our method performs better than the other two methods. To compare the relative performances, one will need to conduct a simulation study.

<table>
<thead>
<tr>
<th>Comparisons</th>
<th>Methods</th>
<th>$K_A$</th>
<th>$K_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between mammal species</td>
<td>1</td>
<td>0.121</td>
<td>0.489</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.123</td>
<td>0.435</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.121</td>
<td>0.443</td>
</tr>
<tr>
<td>Mammals vs. birds</td>
<td>1</td>
<td>0.247</td>
<td>0.777</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.238</td>
<td>0.831</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.245</td>
<td>0.843</td>
</tr>
<tr>
<td>Mammals vs. frog</td>
<td>1</td>
<td>0.502</td>
<td>1.359</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.483</td>
<td>1.822</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.478</td>
<td>2.233</td>
</tr>
</tbody>
</table>

**Table 3**

Numbers ($K_A$) of Nonsynonymous Substitutions per Nonsynonymous Site and Numbers ($K_S$) of Synonymous Substitutions per Synonymous Site between β-Globin Sequences

**Note.**—In this case the numbers of synonymous substitutions are obtained by taking the average of $P_i$'s and $Q_i$'s of the four comparisons before using eqs. (1) and (3) for correction.

* Perler et al. (1980); 1; Miyata and Yasunaga (1980), 2; and present authors, 3.
* Pairwise comparisons among human (Law et al. 1980), mouse (Konkel et al. 1979), rabbit (Hardison et al. 1979), and goat (Schon et al. 1981).
* Mammals (man, mouse, rat, and goat) and Xenopus (Williams et al. 1980).
Estimating Synonymous/Nonsynonymous Rates of Evolution 165

Estimation of ancient divergence is inherently unreliable, and most known factors tend to result in underestimation. For this reason, Perler et al.'s conclusion that, in the evolution of globin and insulin genes, synonymous substitution was much slower in the period before mammalian divergence than in the more recent period has been challenged by Brown et al. (1982), Gojobori (1983), and Kimura (1983a). The results in table 3 suggest that Perler et al.'s method can indeed give a serious underestimate of $K_S$ when comparing distantly related genes. Therefore, their conclusion is not warranted.

All the methods discussed here have to classify nucleotide sites in some way. These methods would be unreliable if the number of sites in each class ($L_0$, $L_2$, and $L_4$ in the present model) changed considerably from time to time and from species to species. Fortunately, this does not appear to be the case. In our study, the number of sites for the same gene usually remains fairly stable across the species examined, even for distantly related ones.

Relative Likelihoods of Codon Changes

Many previous studies have been devoted to the subject of amino acid exchangeability, primarily in an attempt to understand the biochemical roles of different amino acids in various proteins (e.g., Zuckerandl and Pauling 1965; Epstein 1967; McLachlan 1971). As a consequence, many amino acid distance matrices have been proposed to correlate the evolutionary exchangeability with known physicochemical differences between amino acids (Sneath 1966; McLachlan 1971; Grantham 1974; Miyata et al. 1979). All these studies have been based on amino acid sequence data and hence have two limitations: (1) It is difficult to formulate accurately the expected number of amino acid changes without knowing the underlying codons. Previous studies assumed that the expected number of changes between aa$_1$ and aa$_2$ is proportional to $f_1 f_2$, where $f_i$ is the frequency of aa$_i$. This is obviously only an approximation. (2) Amino acid sequence data per se provide no information on synonymous changes. Our analysis of DNA sequence data was motivated to overcome these two limitations and, in our opinion, provides a more reliable estimate of codon or amino acid exchangeability.

From table 1, we see a strong correlation between relative likelihood of amino acid interchanges and amino acid distances on the part of nonsynonymous substitution. This suggests that Grantham's (1974) indices are quite adequate in predicting amino acid exchangeability. It is also clear that purifying (negative) selection arising from functional constraints is a major force at this level of evolution. On the other hand, the relative likelihood of synonymous changes is much greater than the value extrapolated from amino acid changes in protein sequences. For this reason, Miyata and Yasunaga's (1980) method would give a pathway with synonymous substitutions a lower weight than expected.

The results of table 1 are obtained by taking the average of all minimum distance solutions, when there is more than one solution. We note that all minimum distance solutions require the same smallest number of substitutions, regardless of the nature of codon changes involved. In reality, evolution is more likely to have taken a pathway that requires more synonymous codon changes and fewer radical amino acid replacements. Taking the average means that we may still underestimate the likelihood of synonymous (and perhaps even conservative) changes and overestimate that of the radical changes. The advantage of using many sequences and
inferring their ancestral codons is that such ambiguous cases can be reduced. It also enables us to formulate the expected number of codon changes more precisely.

Finally, the results of table 1 also have a bearing on parsimonious phylogenetic reconstruction using amino acid or DNA sequence data (e.g., Moore et al. 1973; Czeluzniak et al. 1982). In using the maximum parsimony method, authors usually make no distinction between the types of codon changes inferred. We point out that considering the relative likelihoods of different codon changes can increase the discriminative power of the maximum parsimony principle. To see this point, let us consider a simple hypothetical example in which the codons of four DNA sequences at a particular homologous codon position are, respectively, CCT (Pro), CCG (Pro), CGT (Arg), and CGG (Arg). Designate the four sequences as a, b, c, and d. If all nucleotide substitutions are given the same weight, the two phylogenetic trees \([(a, b), (c, d)]\) (which means that a joins b, c joins d, and then their ancestors join each other) and \([(a, c), (b, d)]\) are both most parsimonious, each requiring three nucleotide substitutions. However, the first tree requires two synonymous substitutions and only one nonsynonymous substitution (Pro for Arg or Arg for Pro), whereas the second tree requires one synonymous substitution and two nonsynonymous substitutions (both are replacements of Arg by Pro or replacements of Pro by Arg). Since the likelihood of a synonymous substitution is nearly three times that of the replacement of Pro by Arg or Arg by Pro (a moderately conservative replacement), the first tree may be considered more parsimonious if greater weights (distance values) are given to nonsynonymous substitutions.

To illustrate how to apply the “weighted” parsimony method to real data, let us use the α-globin genes from man (h), mouse (m), rabbit (r), and goat (g, and g₂ representing alleles 1 and 2) as an example. Consider the trees \([(h, m), (r, g₁, g₂)], [(h, r), m], (g₁, g₂)] and \([(m, h), (g₁, g₂)]\). At the nucleotide level, the first two trees require 159 substitutions and the third 160 substitutions. At the codon level, the first tree requires 95.6 synonymous changes, 21.3 conservative, 32.1 moderately conservative, 10 moderately radical, and 0 radical amino acid replacements. The numbers for the second tree are 91.3, 19.9, 35.7, 11.6, and 0.5, respectively, and those for the third tree are 93.8, 19.7, 34.9, 11.0 and 0.5, respectively. Since a conservative amino acid change is equivalent to 2.235/0.853 = 2.62 synonymous changes, a moderately conservative one is equivalent to 2.235/0.767 = 2.91 synonymous changes, and so on, we may convert all the nucleotide substitutions into the equivalent of synonymous ones. The first tree, therefore, requires the equivalent of 323.4 synonymous substitutions, the second tree requires 351.0, and the third tree requires 345.8. One may therefore choose the first tree over the other two on the ground of codon changes. This choice is, of course, quite tentative, for only one gene is used.

Mechanisms of DNA Evolution

The results in table 2 are, in our opinion, quite compatible with the neutral mutation–random drift hypothesis (Kimura 1968; King and Jukes 1969). This hypothesis predicts that the rate of evolution will be higher for molecules or parts of a molecule with weak or no selective constraints and low for those that are tightly constrained. Among the coding regions of a gene, nondegenerate sites are expected to be most tightly constrained because all mutations at these sites are nonsynonymous, so that they are subject to constraints at both the RNA and
protein levels. The constraints at the protein level would be particularly strong, for natural selection acts on phenotypes of the organism for which the protein structure and function play a decisive role. Thus, the substitution rate is expected to be lowest at nondegenerate sites. This is indeed the case (table 2). The rate should be higher at twofold degenerate sites, for a fraction of the mutations at these sites are synonymous, so that these sites should be less stringently constrained than nondegenerate sites. The rate should be even higher at fourfold degenerate sites, for all mutations at these sites are synonymous. These two predictions are also seen to be true.

Under the neutralist view, we can also give a plausible explanation for why the rate of nonsynonymous substitution varies greatly among different kinds of genes. For those genes, such as histone and actin genes, that have extremely low nonsynonymous rates (table 2), we may assume that they are subject to tight selective constraints because their functions are important and require very specific protein structures. According to this view, nearly every nonsynonymous mutation in histone and actin genes is deleterious and would not become fixed in the population. On the other hand, for the genes, such as those coding for relaxin, interferons, and immunoglobulins, that have high nonsynonymous rates (table 2), we may assume that they are subject to fewer or weaker selective constraints because their functions do not require very specific protein structures or because their functions are not very important or can be carried out by some other genes, i.e., these genes are “dispensable” (Wilson et al. 1977). We also noted above that the variation in synonymous rate among genes is large, though much smaller than that in nonsynonymous rate. Under the neutralist view, this variation can be explained by assuming that it is largely due to chance effects and partly due to variation in selective constraints among genes. It now appears that synonymous mutations are not completely free of selective constraints but may be constrained by such factors as tRNA availability (Ikemura and Ozeki 1983) and the binding strength between codon and anticodon (Grosjean and Fiers 1982).

Although some of the above arguments were based on assumptions that remain to be substantiated by experimental data, the neutralist view can apparently give a consistent explanation for the various features of DNA evolution. In contrast, the selectionist view cannot give a consistent explanation. To explain the negative correlation between rate and selective constraints by this view, one would have to assume that the probability of a mutation being selectively advantageous is higher in functionally less important DNA regions. It is, however, difficult to imagine how the function of a gene can be constantly improved by synonymous mutations. One may concede that neutral or nearly neutral mutations are predominant at synonymous sites while concurrently maintaining that the majority of amino acid substitutions in protein evolution have been due to advantageous mutations (see, e.g., Goodman 1982). But it is again difficult to imagine why advantageous mutations should occur more often in less important coding regions, such as those coding for the signal peptides and the insulin C and relaxin C peptides—or, in other words, why these latter regions have constantly been improved by new mutations. Furthermore, it is difficult to imagine why the rate of improvement in function should be faster in a less important gene than in a more important one. In short, the selectionist view cannot give a satisfactory explanation for the negative correlation between rate and functional importance.
Table 4
Average Substitution Rates in Pseudogenes and at Fourfold Degenerate, Twofold Degenerate, and Nondegenerate Sites of Functional Genes

<table>
<thead>
<tr>
<th>Parameter of Comparison</th>
<th>Pseudogenes</th>
<th>Fourfold Degenerate Sites</th>
<th>Twofold Degenerate sites</th>
<th>Nondegenerate Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate ($\times 10^9$)</td>
<td>4.85</td>
<td>4.18</td>
<td>2.26</td>
<td>0.94</td>
</tr>
<tr>
<td>Ratio to the pseudogene rate</td>
<td>1.00</td>
<td>0.86</td>
<td>0.47</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Under the neutralist view, one can use the substitution rate in pseudogenes as the neutrality standard to infer the stringency of selective constraints in various gene regions. From the studies of Li et al. (1981), Li (1983), and Gojobori and Nei (1984), we obtain an average rate of $4.85 \times 10^{-9}$ substitutions per nucleotide site per year. The ratios of the average substitution rates at fourfold degenerate sites, twofold degenerate sites, and nondegenerate sites to the pseudogene rate are shown in table 4. Fourfold degenerate sites are on the average subject to only very weak constraints, because their substitution rate is only 14% lower than the pseudogene rate. Twofold degenerate sites are subject to certain constraints, for they evolve only half as fast as pseudogenes. Nondegenerate sites are generally tightly constrained, for their average substitution rate is only one-fifth of the pseudogene rate. If we assume that the nucleotide substitutions at nondegenerate sites are totally due to neutral mutations, then the proportion of neutral mutations at nondegenerate sites is approximately 20%. This is quite similar to Kimura's (1983b) estimate of 0.14 ± 0.06. Of course, the proportion would vary from gene to gene, ranging from 62% in interferon-γ to almost zero in histone H4 (table 2). Similarly, we can infer the stringency of selective constraints in various genes. For example, albumin has previously been used as an example of rapid evolution at the amino acid level, and the argument for this is that albumin is a dispensable protein (Wilson et al. 1977). Although the nonsynonymous rate in the gene for albumin is not low, being $0.92 \times 10^{-9}$ (table 2), it is only 19% of the pseudogene rate. Thus, it appears that albumin is subject to fairly strong selective constraints. Among the genes listed in table 2, interferon-γ has the highest nonsynonymous rate, $2.80 \times 10^{-9}$. This is 58% of the pseudogene rate. Thus, it appears that, even in a rapidly evolving gene, a substantial fraction of nonsynonymous mutations are deleterious.

Acknowledgments

This work was supported by research grants from the National Institutes of Health and the National Science Foundation. We thank W. M. Fitch for suggestions.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received July 26, 1984; revised October 29, 1984.