Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions[1]

Masatoshi Nei and Takashi Gojobori†*

*Center for Demographic and Population Genetics, University of Texas at Houston; and †National Institute of Genetics, Mishima, Japan

Two simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions are presented. Although they give no weights to different types of codon substitutions, these methods give essentially the same results as those obtained by Miyata and Yasunaga's and by Li et al.'s methods. Computer simulation indicates that estimates of synonymous substitutions obtained by the two methods are quite accurate unless the number of nucleotide substitutions per site is very large. It is shown that all available methods tend to give an underestimate of the number of nonsynonymous substitutions when the number is large.

Introduction

In the study of the evolutionary divergence of DNA sequences, it is often required to estimate the numbers of synonymous (silent) and nonsynonymous (amino acid–altering) nucleotide substitutions separately. Since the rate of synonymous substitution is much higher than that of nonsynonymous substitution and is similar for many different genes, synonymous substitutions may be used as a molecular clock for dating the evolutionary time of closely related species (Kafatos et al. 1977; Kimura 1977; Perler et al. 1980; Miyata et al. 1980).

When the number of nucleotide substitutions between two DNA sequences is so small that there is no more than one nucleotide difference between each pair of homologous codons, the number of synonymous and nonsynonymous substitutions can be obtained simply by counting silent and amino acid–altering nucleotide differences. However, when more than one nucleotide difference exists between a pair of codons, the distinction between synonymous and nonsynonymous substitutions is no longer simple. For this reason, Perler et al. (1980) developed a statistical method for estimating synonymous substitutions. More elaborate methods were also developed by Miyata and Yasunaga (1980) and Li et al. (1985). The main difference between Perler et al.'s method and the latter two methods is that in the former an equal weight is given to two or more evolutionary pathways that are possible between a pair of condons whereas in the latter a larger weight is given to an evolutionary pathway involving silent substitutions than is given to a pathway involving amino acid–altering substitutions.

All of these methods, however, suffer from one technical deficiency—they are so complicated that one is often discouraged from using them. We therefore looked for a simpler method and discovered that an unweighted version of Miyata and Yasunaga's

1. Key words: synonymous substitution, nonsynonymous substitution, nucleotide substitution.

Address for correspondence and reprints: Dr. Masatoshi Nei, Center for Demographic and Population Genetics, University of Texas at Houston, P.O. Box 20334, Houston, Texas 77225.

Mol. Biol. Evol. 3(5):418–426. 1986.

0737-4038/86/0305-3504$02.00

(1980) method gives essentially the same estimates as those obtained by Miyata and Yasunaga's and by Li et al.'s methods. Furthermore, an even more crude method seems to give sufficiently accurate estimates when the number of nucleotide substitutions per site is relatively small. The purpose of this paper is to report the results of this study.

New Methods

Let us first discuss a modified (unweighted) version of Miyata and Yasunaga's (1980) method. The genetic code table indicates that all substitutions at the second nucleotide positions of codons result in amino acid replacement whereas a fraction of the nucleotide changes at the first and third positions are synonymous. Under the assumption of equal nucleotide frequencies and random substitution, this fraction is ~5% for the first position and ~72% for the third position.

We now compute the number of synonymous sites (s) and the number of nonsynonymous sites (n) for each codon, considering the above property of codon changes. We denote by f_i the fraction of synonymous changes at the ith position of a given codon ($i = 1, 2, 3$). The s and n for this codon are then given by $s = \sum_{i=1}^{3} f_i$ and $n = (3 - s)$, respectively (also see Kafatos et al. 1977). For example, in the case of codon TTA (Leu), $f_1 = 1/3$ (T → C), $f_2 = 0$, and $f_3 = 1/3$ (A → G). Thus, $s = 2/3$ and $n = 7/3$. For a DNA sequence of r codons, the total number of synonymous and nonsynonymous sites is therefore given by $S = \sum_{j=1}^{r} S_i$ and $n = (3r - S)$, respectively, where s_i is the value of s for the ith codon. When two sequences are compared, the averages of S and N for the two sequences are used.

To compute the number of synonymous and nonsynonymous nucleotide differences between a pair of homologous sequences, we compare the two sequences codon by codon and count the number of synonymous and nonsynonymous nucleotide differences for each pair of codons compared. When there is only one nucleotide difference, we can immediately decide whether the substitution is synonymous or nonsynonymous. For example, if the codon pairs compared are GTT (Val) and GTA (Val), there is one synonymous difference. We denote by s_d and n_d the number of synonymous and nonsynonymous differences per codon, respectively. In the present case, $s_d = 1$ and $n_d = 0$. When two nucleotide differences exist between the two codons compared, there are two possible ways to obtain the differences. For example, in the comparison of TTT and GTA, the two pathways are as follows: pathway I, TTT (Phe) ↔ GTT (Val) ↔ GTA (Val); pathway II, TTT (Phe) ↔ TTA (Leu) ↔ GTA (Val). Pathway I involves one synonymous and one nonsynonymous substitution, whereas pathway II involves two nonsynonymous substitutions. We assume that pathways I and II occur with equal probability. The s_d and n_d then become 0.5 and 1.5, respectively.

When there are three nucleotide differences between the codons compared, there are six different possible pathways between the codons, and in each pathway there are three mutational steps. Considering all these pathways and mutational steps, one can again evaluate s_d and n_d in the same way as in the case of two nucleotide differences. It is now clear that the total number of synonymous and nonsynonymous differences can be obtained by summing up these values over all codons; that is, $S_d = \sum_{j=1}^{r} s_{dj}$ and $N_d = \sum_{j=1}^{r} n_{dj}$, where s_{dj} and n_{dj} are s_d and n_d for the jth codon, respectively, and r is the number of codons compared. Note that $S_d + N_d$ is equal to the total number of nucleotide differences between the two DNA sequences compared.

We can, therefore, estimate the proportion of synonymous (p_s) and nonsynonymous (p_n) differences by the following equations:

$$p_S = S_d/S, \tag{1}$$

and

$$p_N = N_d/N, \tag{2}$$

where S and N are the average number of synonymous and nonsynonymous sites for the two sequences compared. To estimate the number of synonymous substitutions (d_S) and nonsynonymous substitutions (d_N) per site, we use the following formula developed by Jukes and Cantor (1969):

$$d = -\frac{3}{4}\log_e\left(1 - \frac{4}{3}p\right), \tag{3}$$

where p is either p_S or p_N. Of course, this method gives only approximate estimates of d_S and d_N because, strictly speaking, the Jukes-Cantor formula does not apply at some silent sites (twofold and threefold degenerate sites; see Perler et al. 1980).

Although the above method of estimating d_S and n_S is much simpler than the previous methods mentioned above, the computation is still time consuming when there are many codon pairs in which two or three nucleotide differences exist. David Lipman (personal communication) suggested a simpler way of computing S_d, N_d, S, and N. The idea behind his method of computing S_d and N_d is to apply information obtained from the single-base-change codons to the multiple-base-change codons.

We first consider only those codon pairs in which a single nucleotide difference exists and compute the proportion of synonymous (π_s) and the proportion of nonsynonymous (π_n) differences ($\pi_s + \pi_n = 1$) for each of the three nucleotide positions. Obviously, π_n is always 1 for the second position. For those codon pairs in which two or three nucleotide differences exist, we count the number of nucleotide differences for each position and multiply it by π_s or π_n.

Suppose that two sequences of r codons are compared and that there are m_{Si} codon pairs in which (1) the nucleotides differ only at the ith position (single base change) and (2) s_i of them are synonymous. The π_s and π_n for the ith position are then given by $\pi_{si} = s_i/m_{Si}$ and $\pi_{ni} = 1 - \pi_{si}$, respectively. Let m_{Mi} be the number of nucleotide differences at the ith position for those codon pairs in which two or three nucleotide differences exist. S_d and N_d are then computed by

$$S_d = (m_{S1} + m_{M1})\pi_{s1} + (m_{S3} + m_{M3})\pi_{s3}, \tag{3a}$$

and

$$N_d = (m_{S1} + m_{M1})\pi_{n1} + (m_{S2} + m_{M2}) + (m_{S3} + m_{M3})\pi_{n3}. \tag{3b}$$

Lipman suggested that S and N be computed by using the property that 5% of the first position changes and 72% of the third position changes are synonymous and all other changes are nonsynonymous under the assumption of equal nucleotide frequencies and random substitution. Thus, $S = (0.05 + 0.72)r$, and $N = 3r - S$.

Our computer simulations (discussed below) have shown that S_d and N_d obtained by Lipman's method are generally very close to the values obtained by the method

mentioned earlier but that his method of estimating S and N is quite unreliable. The reason for this unreliability of the estimates of S and N is that the codon frequencies in real genes are often quite different from those expected given the assumption of equal nucleotide frequencies. Fortunately, however, S and N can be computed relatively easily by using the method described earlier. Therefore, it is possible to compute p_S and p_N by using the previous S and N and the S_d and N_d obtained by equations (3a) and (3b). In the following, we call this method the unweighted method II and the previous one the unweighted method I.

Previous Methods

The two methods presented here are quite different from Perler et al.'s (1980), since in the latter method twofold, threefold, and fourfold degenerate sites are treated separately. They are also different from Miyata and Yasunaga's (1980) method in two respects. (1) Miyata and Yasunaga give different weights for different pathways, assuming that the interchange of similar amino acids occurs with a higher probability than the interchange of dissimilar amino acids. (2) In our method, S and N are computed by using the two sequences to be compared. In Miyata and Yasunaga's method, these numbers are obtained by considering all intermediate codons involved in evolutionary pathways as well as those of the two sequences to be compared. Li et al.'s (1985) method is an extension of Miyata and Yasunaga's (1980) method, and nondegenerate, twofold degenerate, and fourfold degenerate nucleotide sites are considered separately. Transitional and transversional nucleotide substitutions are also considered separately. Actual computation is quite complicated, so that a computer is necessary.

Computer Simulation

To compare the accuracy of the methods proposed here with that of the previous ones, we did a computer simulation. In this simulation, we assumed that the mutational changes of the four nucleotides occur according to the relative frequencies obtained by Gojobori et al. (1982) for pseudogenes. The mutational changes were then translated into amino acid changes, and purifying selection was assumed to occur according to the amino acid changes. This selection eliminated a certain fraction of mutations, and the remainder were assumed to be "fixed" in the genome. We considered two different schemes of purifying selection, i.e., Gojobori's (1983) and Miyata and Yasunaga's (1980). In the former, all synonymous changes had a probability of fixation of 1, and all nonsynonymous changes had a probability of fixation of 0.2. We used these probabilities because actual data suggest that the ratio of d_S and d_N is approximately five (Li et al. 1985).

In the second scheme of purifying selection, the probability of fixation depended on the degree (δ) of polarity and volume differences of the amino acids interchanged. Miyata et al. (1979) computed this δ value (chemical distance) for all pairs of amino acids. We assumed that the relative probability of fixation (v) is given by

$$v = \begin{cases} 1 & \text{for } \delta = 0 \text{ (synonymous change)} \\ 1 - \delta/3.5 & \text{for } 0 < \delta < 3.465 \\ 0.01 & \text{for } \delta > 3.465. \end{cases}$$

This probability is exactly the same as the weight for each amino acid–altering substitution Miyata and Yasunaga (1980) used in their computation of d_S and d_N. Therefore, this simulation is favorable for their method.

The mathematical method used in our simulation is similar to that of Gojobori (1983). For mathematical simplicity, we used a discrete time model of nucleotide substitution, and for each unit of evolutionary time the mutational change of nucleotides was assumed to occur following Gojobori's 4 × 4 transition (mutation) matrix for the pseudogene scheme. The total mutation rate (not the substitution rate) used was 0.01 per nucleotide site per unit evolutionary time. Purifying selection was introduced immediately after mutation by using the 61 × 61 codon-substitution matrix. Using these two matrices, one can compute the probabilities of different amino acids occurring at a particular site for any number of evolutionary time units. To minimize the effect of stochastic errors, a sequence of 732 codons was used. This sequence length was close to the maximum length permitted by our computer capacity.

An ancestral sequence of 732 codons was generated by taking into account the equilibrium frequencies of 20 amino acids obtained from the mutation and selection matrices and by using pseudorandom numbers. Two descendant DNA sequences were then generated independently from the ancestral sequence by using the nucleotide-substitution scheme described above. After every 10 units of evolutionary time, the two DNA sequences were translated into amino acid sequences, and d_S and d_N were estimated by the new methods as well as by the three previous ones.

The results obtained are presented in figure 1. Since the extent of stochastic errors was small in the present case, the results from only one replication are given for each selection scheme. The solid lines in the figure represent the expected values of d_S and

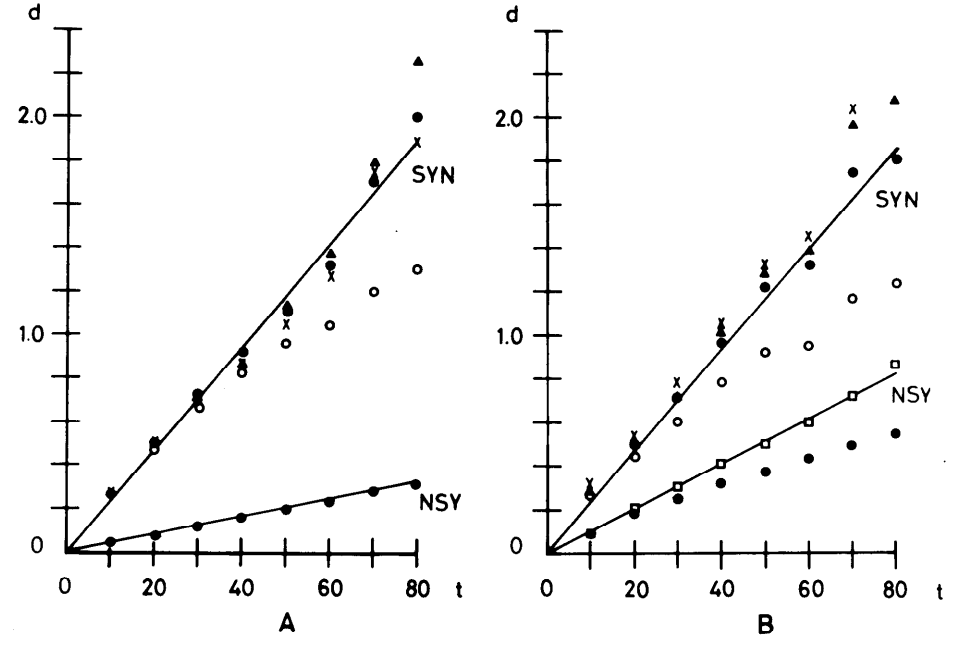

FIG. 1.—Relationships between the estimates of the numbers of synonymous (SYN = d_S) and nonsynonymous (NSY = d_N) substitutions per site with evolutionary time (t). These relationships were obtained by computer simulation (see text). Panel A, Gojobori's (1983) scheme of purifying selection; panel B, Miyata and Yasunaga's (1980) scheme of purifying selection. Circles = Perler et al.'s (1980) method; black dots = unweighted method I; triangles = Miyata and Yasunaga's (1980) method; ×'s = Li et al.'s (1985) method; squares = the value obtained by eq. (5). The estimates of d_N obtained by Perler et al.'s, Miyata and Yasunaga's, and Li et al.'s methods were virtually the same as those obtained by unweighted method I. The solid lines represent the expected numbers of substitutions.

d_N between two descendant nucleotide sequences. These expected values were obtained from the nucleotide-substitution matrices (see Gojobori 1983) and are, respectively, $d_S = 0.023t$ and $d_N = 0.004t$ for Gojobori's selection scheme and $d_S = 0.023t$ and $d_N = 0.010t$ for Miyata and Yasunaga's selection scheme, where t is the number of evolutionary time units. It is clear that the number of synonymous substitutions obtained from our unweighted method I increases almost linearly with evolutionary time and is close to the expected number for both selection schemes. (The results for unweighted method II are discussed below.) However, the estimate obtained by Perler et al.'s (1980) method is considerably lower than the expected number when $t \geqslant 40$. This confirms Takahata and Kimura's (1981) and Gojobori's (1983) earlier finding that Perler et al.'s method gives an underestimate of the number of synonymous substitutions.

Miyata and Yasunaga's method gives essentially the same estimate of d_S as ours when d_S is small but tends to give an overestimate when d_S is large. Li et al.'s method also gives essentially the same estimate as ours when d_S is small, but it has a tendency to give an overestimate of d_S when Miyata and Yasunaga's selection scheme is used. In figure 1B, the value for $t = 80$ is not given for Li et al.'s method because that method was inapplicable to this case.

When Gojobori's selection scheme is used, all methods give essentially the same estimate of d_N, and the estimate obtained is close to the expected number, though there is some tendency toward underestimation. When Miyata and Yasunaga's selection scheme is used, all methods again give virtually the same results. In this case, however, the estimates are substantially lower than the expectations when d_N is large. This underestimation is apparently caused by variation in the rate of amino acid substitution (λ) among different amino acid sites.

Previously, Uzzell and Corbin (1971) showed that λ follows the gamma distribution among different sites. The gamma distribution is given by $f(\lambda) = [\beta^\alpha/\Gamma(\alpha)]e^{-\beta\lambda}\lambda^{\alpha-1}$, where $\alpha = \bar{\lambda}^2/V_\lambda$ and $\beta = \bar{\lambda}/V_\lambda$, $\bar{\lambda}$ and V_λ being the mean and variance of λ, respectively. If we use Jukes and Cantor's formulation, the probability of identity of nucleotides between two homologous sites for a given value of λ is given by $i = 1 - (3/4)[1 - e^{-8\lambda t/3}]$, where t is the time since divergence between two DNA sequences. Therefore, the mean probability of identity of nucleotides between the two sequences is

$$\bar{i} = \int_0^\infty i f(\lambda) d\lambda = \frac{1}{4} + \frac{3}{4}\left(\frac{\alpha}{\alpha + 4d/3}\right)^\alpha, \tag{4}$$

where $d \equiv 2\bar{\lambda}t$ is the expected number of nucleotide substitutions. Thus, if $\alpha = 1$, we obtain

$$d = \frac{p}{1 - (4/3)p}, \tag{5}$$

where $p = 1 - \bar{i}$. Interestingly, the d values estimated by this equation from observed values of p_N agree very well with the expected value of d_N (fig. 1B). This suggests that the coefficient of variation ($1/\sqrt{\alpha}$) of λ is close to 1 in the present case.

Above, we presented unweighted method II as a simplified version of method I. To see the accuracy of this method, we compared the estimates of d_S and d_N obtained by this method with those obtained by method I for the two computer simulations mentioned above. The results obtained for Gojobori's scheme of nucleotide substitution

Table 1
Estimates of the Number of Synonymous (d_S) and Nonsynonymous (d_N) Substitutions per Site Obtained by Methods I and II for Gojobori's Scheme of Nucleotide Substitution

	Time				
Type of Substitution	10	20	40	60	80
d_S:					
Method I	0.268	0.506	0.903	1.316	2.012
Method II	0.279	0.528	0.983	1.393	2.991
d_N:					
Method I	0.039	0.075	0.160	0.237	0.327
Method II	0.037	0.073	0.156	0.247	0.329

(purifying selection) are presented in table 1. Methods I and II give essentially the same estimates except for the d_S for $t = 80$. The d_S value (2.99) for $t = 80$ is considerably larger than the expected value (1.84). Similar results were obtained for Miyata and Yasunaga's scheme of nucleotide substitution, though d_S was sometimes substantially overestimated or underestimated when $d_S > 1.5$. The inaccuracy of method II for a large d_S is apparently caused by the fact that there are a small number of single-base-change codons for this case, which thus makes the estimate of π_{si} unreliable. This problem is expected to be more serious when the total number of codons compared is small. Nevertheless, method II is much easier to use than method I, and the results obtained are very similar when t is small.

Discussion

We have seen that our method I gives essentially the same estimates of d_S and d_N as those obtained by Miyata and Yasunaga's and Li et al.'s methods, though equal weights are given to all evolutionary pathways for a given pair of codons. There are two reasons for this. First, when the evolutionary time is relatively short, most of the codons compared have no more than one nucleotide difference, so that there is no effect of weighting. Second, whereas different weights are given to different pathways when there are two or three nucleotide differences, the differences in weight between different pathways are usually very small. For example, as mentioned above, there are two evolutionary pathways for the comparison of TTT (Phe) and GTA (Val). In pathway I, the relative probability (V) of change between TTT (Phe) and GTT (Val) is 0.59 according to Miyata and Yasunaga's weighting scheme, whereas the V between GTT (Val) and GTA (Val) is 1. Therefore, the V of pathway I is $V = 0.59 \times 1 = 0.59$. Similarly, the V of pathway II becomes $V = 0.82 \times 0.74 = 0.61$. Therefore, the two probabilities are very close, although pathway I involves a silent substitution and pathway II does not. We examined these V's for many different pairs of codons that were observed in our computer simulation. In most cases, different pathways showed similar V's or weights. Occasionally, different pathways showed quite different weights, but in these cases the absolute values of the weights were so small that they did not contribute very much to the total number of nucleotide substitutions [e.g., the change between TGG (Trp) and TCT (Ser)].

In the present study, computer simulation was conducted by using specific schemes of mutation and purifying selection. However, the two selection schemes we used

Table 2
Estimates of the Number of Synonymous (d_S) and Nonsynonymous (d_N) Substitutions per Site Obtained by Four Different Methods for Three Pairs of Globin Genes

Globin Genes Compared and Type of Substitution	Method				
				New Method	
	Perler et al.	Miyata and Yasunaga	Li et al.	I	II
Human β vs. rabbit β:					
d_S	0.426	0.354	0.345	0.354	0.348
d_N	0.056	0.056	0.056	0.056	0.056
Human β vs. chicken β:					
d_S	0.671	0.752	0.747	0.709	0.712
d_N	0.233	0.211	0.223	0.218	0.233
Human β vs. human α1:					
d_S	0.924	1.043	1.055	0.994	1.149
d_N	0.456	0.447	0.443	0.455	0.443

Source.—Human β, Lawn et al. (1980); rabbit β (allele 1), Hardison et al. (1979); chicken β, Dolan et al. (1983); human α1, Michelson and Orkin (1980).

represent quite different patterns. Therefore, our method I and Miyata and Yasunaga's and Li et al.'s methods seem to give similar results for different types of purifying selection. Reexamination of the results of Gojobori's (1983) computer simulation on d_S (results not shown) also indicates that the three methods give more or less the same result for different mutation schemes. This seems to be true also with actual data. Table 2 shows the estimates of d_S and d_N for three pairs of globin genes. The estimates obtained by the three methods are more or less the same, though the estimates obtained by Perler et al.'s (1980) method are somewhat different. It is interesting that our method II gives essentially the same estimates as those obtained by method I.

Acknowledgments

We thank Wen-hsiung Li for giving us his computer program for Li et al.'s method and Takeo Maruyama for a helpful discussion. We are also grateful to David Lipman, who suggested a simple way of computing S_d and N_d. This study was supported by research grants from the National Institutes of Health and the National Science Foundation.

LITERATURE CITED

Dolan, M., J. B. Dodgeson, and J. D. Engel. 1983. Analysis of the adult chicken β-globin gene. J. Biol. Chem. **258**:3983–3990.

Gojobori, T. 1983. Codon substitution in evolution and the "saturation" of synonymous changes. Genetics **105**:1011–1027.

Gojobori, T., W.-H. Li, and D. Graur. 1982. Pattern of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18**:369–369.

Hardison, R. C., E. T. Butler III, E. Lacy, T. Maniatis, N. Rosenthan, and A. Efstratiadis. 1979. The structure and transcription of four linked rabbit β-globin genes. Cell **18**: 1285–1297.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. Munro, ed. Mammalian protein metabolism III. Academic Press, New York.

KAFATOS, F. C., A. EFSTRATIADIS, B. G. FORGET, and S. M. WEISSMAN. 1977. Molecular evolution of human and rabbit β-globin mRNAs. Proc. Natl. Acad. Sci. USA **74**:5618–5622.

KIMURA, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **38**:515–524.

LAWN, R. M., A. EFSTRATIADIS, C. O'CONNELL, and T. MANIATIS. 1980. The nucleotide sequence of the human β-globin gene. Cell **21**:647–651.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

MICHELSON, A. M., and S. H. ORKIN. 1980. The 3′ untranslated regions of the duplicated human α-globin genes are unexpectedly divergent. Cell **22**:371–377.

MIYATA, T., S. MIYAZAWA, and T. YASUNAGA. 1979. Two types of amino acid substitutions in protein evolution. J. Mol. Evol. **12**:219–236.

MIYATA, T., and T. YASUNAGA. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. **16**:23–26.

MIYATA, T., T. YASUNAGA, and T. NISHIDA. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc. Natl. Acad. Sci. **77**:7328–7332.

PERLER, F., A. EFSTRATIADIS, P. LOMEDICO, W. GILBERT, R. KOLODNER, and J. DODGESON. 1980. The evolution of genes: the chicken preproinsulin gene. Cell **20**:555–566.

TAKAHATA, N., and M. KIMURA. 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics **98**:641–657.

UZZELL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science **172**:1089–1096.

WALTER M. FITCH, reviewing editor

Received June 10, 1985; revision received May 9, 1986.