Estimation of Evolutionary Distance between Nucleotide Sequences[1]

Fumio Tajima and Masatoshi Nei
University of Texas at Houston

A mathematical formula for estimating the average number of nucleotide substitutions per site (δ) between two homologous DNA sequences is developed by taking into account unequal rates of substitution among different nucleotide pairs. Although this formula is obtained for the equal-input model of nucleotide substitution, computer simulations have shown that it gives a reasonably good estimate for a wide range of nucleotide substitution patterns as long as δ is equal to or smaller than 1. Furthermore, the frequency of cases to which the formula is inapplicable is much lower than that for other similar methods recently proposed. This point is illustrated using insulin genes. A statistical method for estimating the number of nucleotide changes due to deletion and insertion is also developed. Application of this method to globin gene data indicates that the number of nucleotide changes per site increases with evolutionary time but the pattern of the increase is quite irregular.

Introduction

The evolutionary change of DNA sequences occurs by nucleotide substitution, deletion, and insertion. The change due to nucleotide substitution is measured in terms of the number of nucleotide substitutions per site between two homologous DNA sequences. Several statistical methods for estimating this number have been developed. Unfortunately, however, all of them have some deficiencies. Jukes and Cantor's (1969) method is the simplest one but gives underestimates when the rate of nucleotide substitution is not the same for all nucleotide pairs. Recently, Kimura (1980, 1981), Takahata and Kimura (1981), and Gojobori et al. (1982*a*) developed new methods for estimating the number of nucleotide substitutions, taking into account unequal rates of substitutions among different nucleotide pairs. However, these methods are all dependent on specific schemes of nucleotide substitutions, and if actual nucleotide substitution does not follow these schemes, the methods are expected to give biased estimates. Furthermore, they

1. Key words: nucleotide substitution, evolutionary distance, unequal substitution rates, deletion, insertion, globin genes, insulin genes.

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. 1(3):269–285. 1984.

0737-4038/84/0103-0003$02.00

are often inapplicable to actual data because of a negative argument in the logarithm of the formula used. In this paper we propose a new method that alleviates some of these deficiencies. We shall also consider the evolutionary changes of DNA arising from deletions and insertions and present a method for measuring the amount of these changes.

Number of Nucleotide Substitutions

Theory

Consider two homologous nucleotide sequences that diverged from a common ancestral sequence t years ago. We first consider the case where the rate of nucleotide substitution is the same for all pairs of nucleotides and equal to λ per site per year. The expected number of nucleotide substitutions per site between the two sequences for this case is given by

$$\delta = 2\lambda t. \tag{1}$$

If we know the proportion (π) of different nucleotides per site, δ can be estimated by

$$\delta = -\tfrac{3}{4} \log_e(1 - 4\pi/3), \tag{2}$$

where $0 \leqslant \pi \leqslant \tfrac{3}{4}$ (Jukes and Cantor 1969; Kimura and Ohta 1972).

At this point, we note that (2) can be written as

$$\delta = -b_1 \log_e(1 - \pi/b_1), \tag{3}$$

where $b_1 = 1 - \Sigma q_i^2$. Here q_i is the equilibrium frequency of the ith nucleotide ($i = 1, 2, 3, 4$ corresponding to the nucleotides A, T, G, C). When the rate of nucleotide substitution is the same for all nucleotide pairs, $q_i = \tfrac{1}{4}$, so that $b_1 = \tfrac{3}{4}$. We also note that $b_1 = \tfrac{3}{4}$ is the maximum value of π, which is attained at $t = \infty$.

Kimura (1980, 1981), Takahata and Kimura (1981), and Gojobori et al. (1982*a*) have shown that when the rate of nucleotide substitution varies with nucleotide pair, (2) gives an underestimate of δ. Part of the reason is that in this case the equilibrium value of π is generally smaller than $\tfrac{3}{4}$. Note that in any scheme of nucleotide substitution the value of π at $t = \infty$ is given by $b_1 = 1 - \Sigma q_i^2$. The value of q_i can be uniquely determined for any substitution scheme (Tajima and Nei 1982). This suggests that (3) may be used as an estimator of δ even for the case of unequal substitution rates. The estimate of δ obtained by (3) is always equal to or greater than that obtained by (2).

Equation (3) holds exactly for Tajima and Nei's (1982) equal-input model of nucleotide substitution with unequal rates. Let λ_{ij} be the rate of substitution of the jth nucleotide for the ith nucleotide per unit evolutionary time. This unit evolutionary time can be, for example, year, generation, or 1,000 years, depending on the purpose. In the equal-input model, $\lambda_{ij} = a_j$ for all i's except for λ_{jj}. In other words, the rate of substitution of the jth nucleotide for the ith nucleotide is the same, irrespective of the ith nucleotide. Therefore, the substitution rate matrix is given by (A1) in the Appendix, where $\lambda_{jj} = 1 - \Sigma_{i \neq j} a_i$. Using this substitution rate matrix, one can prove (3), as shown in the Appendix.

In practice, of course, the pattern of nucleotide substitution does not necessarily follow this scheme (see Gojobori et al. 1982*b*). When the substitution

scheme is different from the equal-input model, (3) is no longer valid, as is clear from the works of Kimura (1980, 1981), Takahata and Kimura (1981), and Gojobori et al. (1982*a*). In this case, however, a slight modification of (3) gives a quite reliable estimate, as will be shown later by computer simulation. This modification is based on the following observations. (i) In the equal-input model, $c_{ij} \equiv x_{ij}/(2q_iq_j)$ is constant for all i and j ($i < j$), where x_{ij} is the proportion of pairs of nucleotide i and j between the two homologous DNA sequences (see Appendix). (ii) Our computer simulations discussed in the next section have shown that when c_{ij} is not constant, (3) tends to give an underestimate. (iii) In the case of the equal-input model, δ can also be estimated by using information on the frequencies of nonidentical nucleotide pairs. Namely,

$$\begin{aligned}\delta &= -2\sum_{i=1}^{3}\sum_{j=i+1}^{4} q_iq_j\log_e(1-c_{ij}) \\ &= -b_2\log_e(1-\pi/b_2),\end{aligned} \tag{4}$$

where $b_2 = \pi^2/h$ and

$$h = \sum_{i=1}^{3}\sum_{j=i+1}^{4} x_{ij}^2/(2q_iq_j) \tag{5}$$

(see Appendix). When c_{ij} is not constant, however, (4) tends to give an overestimate of δ (results from our computer simulations). These observations suggest that an approximate estimate of δ is obtained by

$$\delta = -b\log_e(1-\pi/b), \tag{6}$$

where b is the average of b_1 and b_2 and given by

$$b = \left(1 - \sum_{i=1}^{4} q_i^2 + \pi^2/h\right)\Big/ 2. \tag{7}$$

It is desirable to know the accuracy of this formula for various patterns of nucleotide substitution. However, analytical evaluation of the accuracy is not easy, because the mathematical property of the most general substitution scheme requiring 12 parameters has not been studied. We have therefore conducted a computer simulation to examine this accuracy. As will be shown in the next section, this simulation indicates that (6) gives a quite reliable estimate as long as δ is smaller than 1. Needless to say, equation (6) holds exactly for the case of equal substitution rates or the equal-input model.

So far we have considered the deterministic change of DNA divergence. In practice, the numbers of nucleotide substitutions are studied by examining a finite number of nucleotides, and thus the estimate ($\hat{\delta}$) of δ is subject to sampling error. The sampling variance of $\hat{\delta}$ obtained from (6) is given by

$$\begin{aligned}V(\hat{\delta}) = &\left(\frac{\partial\delta}{\partial\pi}\right)^2 V(\pi) + \left(\frac{\partial\delta}{\partial b}\right)^2 V(b) \\ &+ 2\frac{\partial\delta}{\partial\pi}\frac{\partial\delta}{\partial b}\operatorname{cov}(\pi,b).\end{aligned} \tag{8}$$

It can be shown that the second and third terms of (8) are very small compared

with the first term unless n (number of nucleotide pairs examined) is unusually small, say, $n < 40$. Therefore, we have (approximately)

$$V(\hat{\delta}) = b^2\pi(1-\pi)/[(b-\pi)^2 n]. \quad (9)$$

Computer Simulation

In this section we shall examine two different aspects of the accuracies of the estimates of δ obtained by (3) and (6). One is the effect of deviation of nucleotide substitution from the equal-input model, and the other is the effect of sampling error when a relatively small number of nucleotides are examined. In the study of the former effect we assume that the DNA sequence under investigation is infinitely long.

Effect of Deviation from the Equal-Input Model of Substitution

Gojobori et al. (1982*b*) studied the relative rates of nucleotide substitution among the four nucleotides (A, T, G, C) for three functional genes (α and β globin genes and ACTH gene) and six pseudogenes (four globin pseudogenes, one Ig V_κ pseudogene, and one U1 snRNA pseudogene). These relative rates were quite different from the rates expected from any of the mathematical models studied so far. Therefore, it is interesting to know which statistical method gives the best estimate of δ when nucleotide substitution occurs according to these observed patterns. We therefore used the nine substitution schemes observed to simulate the evolutionary change of nucleotide sequences. In this simulation we followed Gojobori et al.'s (1982*a*) method and computed the δ values for the nine substitution schemes. That is, the matrix of relative substitution rates (P_{ij}; $i \neq j$) was first converted into the matrix of substitution rates (λ_{ij}) corresponding to $k \equiv \Sigma_i q_i \Sigma_{j \neq i} \lambda_{ij} = 0.0078125$, where k is the average number of nucleotide substitutions per unit evolutionary time. The values of x_{ij}'s for $\delta = 0.25, 0.5, 1.0$, and 2.0 were then obtained by squaring the matrix of substitution rates repeatedly (see Gojobori et al. [1982*a*] for details). Note that $\delta = 0.25, 0.5$, etc. are obtained by squaring the matrix five times, six times, etc. From the values of x_{ij}'s, δ was estimated by using seven different estimation methods, that is, (*a*) the Jukes-Cantor (JC) method, (*b*) Kimura's (1980) two-parameter (2P) method, (*c*) Kimura's (1981) three-substitution-type (3ST) method, (*d*) Takahata and Kimura's (1981) (TK) method, (*e*) Gojobori et al.'s (1982*a*) (GIN) method, (*f*) equation (3), and (*g*) equation (6). The deviation of the estimate from the true value of δ was measured by the following bias index:

$$B = \left[\sum_{i=1}^{r} (\hat{\delta}_i - \delta)^2 / r\right]^{1/2}, \quad (10)$$

where $\hat{\delta}_i$ is the estimate of δ for the ith substitution scheme and r is the number of substitution schemes used. In the present case $r = 9$.

The B values obtained are presented in table 1. It is clear that when δ is small, that is, $\delta \leq 0.5$, equation (6) gives an estimate of δ with the smallest amount of bias, whereas when $\delta \geq 1.0$, the TK and GIN methods tend to give a better estimate than equation (6). Equation (3) gives a smaller value of B than the JC, 2P, and 3ST methods for all values of δ, but the bias of the estimate obtained by

Table 1
Bias Indices (B)[a] of the Estimates of δ Obtained by Seven Different Methods for Various Schemes of Nucleotide Substitution

True δ	JC	2P	3ST	TK	GIN	(3)	(6)
Nine substitution schemes observed for actual genes:							
.25	.020	.019	.018	.016	.017	.014	.003
.50	.072	.069	.068	.022	.045	.053	.018
1.00	.240	.230	.223	.054	.135	.177	.108
2.00	.728	.698	.675	.381	.361	.514	.449
Thirty-one substitution schemes artificially generated:							
.25	.012	.012	.011	.009	.014	.007	.003
.50	.047	.046	.045	.027	.029	.026	.011
1.00	.178	.168	.164	.093	.060	.096	.061
2.00	.635	.582	.566	.322	.136	.316	.273

NOTE.—JC = Jukes and Cantor's (1969) method, 2P = Kimura's (1980) two-parameter method, 3ST = Kimura's (1981) three-substitution type method, TK = Takahata and Kimura's (1981) method, GIN = Gojobori et al.'s (1982*a*) method, (3) = eq. (3), and (6) = eq. (6).

[a] Bias indices were computed by (10).

(3) is larger than that of (6). As expected, the JC method gives an estimate of δ with the largest bias for all values of δ.

Since nine substitution schemes would not be sufficient for drawing a general conclusion, we used 31 more different schemes which were generated artificially by using random numbers. Seven substitution schemes were obtained by assuming that each element of the matrix of relative substitution rates (P_{ij}) takes one value of 0.001, 0.002, . . . , 0.009, and 0.01 with equal probability (1/10). The remaining 24 substitution schemes were obtained by assuming that P_{ij} takes one value of 0.001, 0.002, . . . , 0.009, and 0.01 with probabilities 0.19, 0.17, 0.15, . . . , 0.03, and 0.01, respectively. The P_{ij} matrices thus obtained covered a wide range of substitution patterns. The P_{ij} matrices were then converted into the substitution rate matrix corresponding to $k = 0.0078125$. Using these matrices, we again estimated δ's by using the seven statistical methods.

The B values for these new simulations are given in the lower half of table 1. When $\delta \leq 0.5$, equation (6) again gives the best result, the B value being considerably smaller than that for the other methods. When $\delta \geq 1$, however, the GIN method is superior to (6), though the latter is better than the TK method. Considering this case together with the case of empirical substitution schemes mentioned above, we can conclude that (6) is better than the other methods in estimating δ when δ is small, whereas the GIN method gives the best result when δ is large.

Although our bias index gives the average bias of the estimates of δ, it does not give information about the direction of the bias. This information is provided in figure 1, where the distribution of $\hat{\delta}$ is given in relation to δ for the four levels of δ. The distributions of $\hat{\delta}$ for the 2P and 3ST methods and equation (3) are not given here, because these are apparently inferior to equation (6). It is seen that when δ is 0.25 or 0.5, equation (6) gives a very narrow distribution around the true value of δ. The GIN method gives a mean value of $\hat{\delta}$ close to the true value,

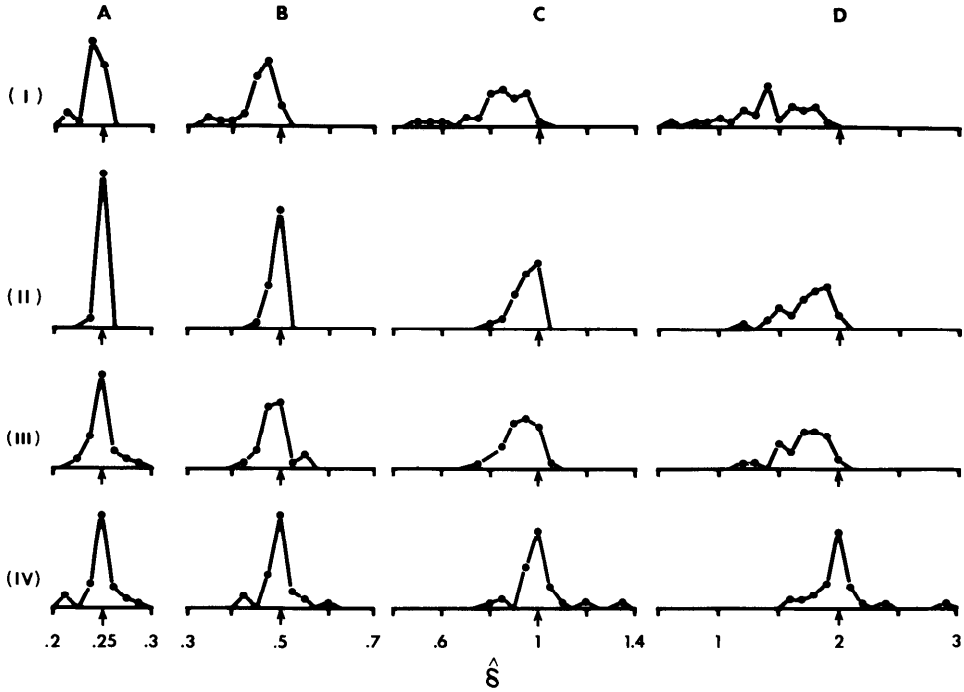

FIG. 1.—Distribution of the estimates ($\hat{\delta}$) of δ obtained by four different methods. *I*, JC method. *II*, eq. (6). *III*, TK method. *IV*, GIN method. A, B, C, and D represent the cases of $\delta = 0.25$, 0.5, 1, and 2, respectively. Arrows indicate the locations of the true values of δ. The scale of $\hat{\delta}$ varies with δ. The total number of observations is 40 in each case.

but the deviation from the true value is often large. However, the JC method almost always gives an underestimate of δ. The TK method also tends to give an underestimate, but the extent of underestimation is not as bad as that of the JC method. When $\delta \geq 1$, however, all methods except the GIN method give underestimates, but the extent of underestimation for equation (6) is small when $\delta = 1$. The GIN method generally gives an average estimate close to the true value of δ and a small value of B, though the B value for the case of $\delta = 1$ is slightly larger than that for equation (6). From figure 1, therefore, we may conclude that equation (6) is superior to the other methods when $\delta \leq 1$, but when $\delta > 1$ the GIN method is probably the best one.

Sampling Error

When the number of nucleotides compared is small, the estimates of q_i and x_{ij} may deviate from the expected values by chance, and this deviation is expected to affect the estimate of δ or produce cases to which equation (6) or other methods are inapplicable because of a negative argument in the logarithm involved. To examine the magnitude of this error, we conducted another computer simulation. In this simulation we considered three different numbers of nucleotides, that is, $n = 50$, 144, and 500. The latter two numbers were chosen to compare our results with those of Gojobori et al. (1982*a*). In Gojobori et al.'s computer simulation many inapplicable cases were produced when their six-parameter model of nucleotide substitution was used. Since we were primarily interested in the frequency

of inapplicable cases, we used the same substitution model. The substitution rates used were $\alpha = 0.00125$, $\alpha_1 = 0.008$, $\alpha_2 = 0.118$, $\beta = 0.005$, $\beta_1 = 0.004$, and $\beta_2 = 0.0059$ with $k = 0.01$, where the parameters α, α_1, etc. are identical to those given in Gojobori et al.'s (1982*a*) table 2. Ancestral sequences of 50, 144, and 500 nucleotides were generated by using pseudorandom numbers. From each of these ancestral sequences, 50 pairs of descendant nucleotide sequences were randomly produced for each of $\delta = 1.0$ and 2.0 by using the method described by Gojobori et al. (1982*a*). For each pair of descendant sequences, x_{ij}'s were computed, and $q_i = x_{ii} + \Sigma_{j \neq i} x_{ij}/2$ was obtained. Using these q_i's and x_{ij}'s, we estimated δ by the JC method and equations (3) and (6). In the case of $\delta = 2.0$, the $\hat{\delta}$ values for $n = 50$ were not computed, since in this case an estimate of δ is obviously unreliable because of a large sampling error.

The mean ($\bar{\delta}$) and standard deviation ($\hat{\sigma}_\delta$) of $\hat{\delta}$ obtained and the frequency of inapplicable cases (f) are given in table 2. In this case the values for $n = \infty$, which can be obtained theoretically, are also presented. The JC method again gives underestimates of δ for both $\delta = 1$ and 2, but there are no inapplicable cases. Equation (3) gives a much better estimate of δ; however, there are a few inapplicable cases. Equation (6) gives an even better estimate of δ than equation (3), but the number of inapplicable cases is slightly larger than that for (3). Table 3 gives the results obtained by Gojobori et al. (1982*a*) for the TK and GIN methods. In both methods the frequency of inapplicable cases is very high compared with that of (3) and (6). If we remove inapplicable cases, however, the GIN method gives a relatively good estimate, though the variance is quite large. The TK method also gives a good estimate of δ when $\delta = 1$ but a serious underestimate when $\delta = 2$. From these results we can conclude that our equations (3) and (6) are less sensitive to sampling error than the TK and GIN methods.

Table 2 includes the observed and expected standard deviations of $\hat{\delta}$. The observed values were computed from replicate estimates of δ with the inapplicable cases excluded, whereas the expected values were obtained from (9). If we con-

Table 2
Results of Computer Simulation in Which Nucleotide Substitution Followed Gojobori et al.'s (1982*a*) Six-Parameter Model

True δ and n	JC Method				Equation (3)				Equation (6)			
	$\bar{\delta}$	$\hat{\sigma}_\delta$	σ_δ	f	$\bar{\delta}$	$\hat{\sigma}_\delta$	σ_δ	f	$\bar{\delta}$	$\hat{\sigma}_\delta$	σ_δ	f
1.0:												
50	.81	.21	.20	0/50	.96[a]	.30[a]	.31	2/50	1.16[a]	.47[a]	.34	5/50
144	.82	.16	.12	0/50	.99	.25	.18	0/50	1.08	.31	.20	0/50
500	.78	.05	.06	0/50	.92	.07	.10	0/50	.97	.08	.11	0/50
∞[b]	.79	.00	.00	0/50	.94	.00	.00	0/50	.97	.00	.00	0/50
2.0:												
144	1.22	.20	.20	0/50	1.87[a]	.63[a]	.70	8/50	2.04[a]	.48[a]	.73	11/50
500	1.22	.12	.11	0/50	1.95[a]	.58[a]	.38	2/50	2.02[a]	.60[a]	.40	3/50
∞[b]	1.20	.00	.00	0/50	1.80	.00	.00	0/50	1.83	.00	.00	0/50

NOTE.—$\bar{\delta}$ = average of the estimate ($\hat{\delta}$) of δ, $\hat{\sigma}_\delta$ = standard deviation of the estimate, σ_δ = expected standard deviation obtained from formula (9), f = proportion of inapplicable cases, and n = number of nucleotide pairs. The number of replications used is 50.

[a] These values were computed by excluding inapplicable cases.

[b] The values for $n = \infty$ were obtained theoretically.

Table 3
Results Obtained from Gojobori et al.'s (1982*a*) Computer Simulation

True δ and n	TK Method			GIN Method		
	$\bar{\delta}$	$\hat{\sigma}_{\delta}$	f	$\bar{\delta}$	$\hat{\sigma}_{\delta}$	f
1.0:						
144	1.00[a]	.24[a]	11/80	1.01[a]	.23[a]	22/80
500	1.07	...	0/16	1.06	...	0/16
2.0:						
144	1.36[a]	.30[a]	108/160	1.70[a]	.52[a]	129/160
500	1.53[a]	...	8/32	2.20[a]	...	23/32

NOTE.—$\bar{\delta}$ = average of the estimate ($\hat{\delta}$) of δ, $\hat{\sigma}_{\delta}$ = standard deviation of $\hat{\delta}$, f = proportion of inapplicable cases (denominator indicates the number of replications); n = number of nucleotide pairs.

[a] These values were computed by excluding inapplicable cases.

Table 4
Observed Numbers of the 10 Different Pairs of Nucleotides between the DNA Sequences for the Human and Rat Insulin A and B Chains

	AA	AT	AG	AC	TT	TG	TC	GG	GC	CC	Total
First position	9	1	0	0	14	0	1	13	0	13	51
Third position	2	3	5	1	3	1	5	8	2	21	51

NOTE.—The numbers at the first and third nucleotide positions of codons are listed separately. There are no nucleotide differences at the second position.

sider that the number of replications is only 50, the agreement between the observed and expected values seems to be reasonably good. Table 3 also gives the observed standard deviations for the TK and GIN methods. They are relatively small compared with those for (3) and (6) because there were many inapplicable cases excluded.

Numerical Example

Sures et al. (1980) determined the nucleotide sequence of the human preproinsulin mRNA and compared it with that of the rat preproinsulin-I mRNA. Preproinsulin consists of four polypeptide chains—the A and B chains, signal peptide, and C peptide. The A and B chains (51 amino acids) produce active insulin, whereas the signal and C peptides (54 amino acids) are removed before insulin is produced. Since the latter two polypeptides are considered to be subject to less stringent purifying selection than the former two polypeptides (Sures et al. 1980), we have analyzed them separately. Following Kimura (1981), we have also considered the first, second, and third nucleotide positions of codons separately. The numbers of 10 different pairs of nucleotides (n_{ij}) between the DNA sequences for the human and rat A and B chain genes are given in table 4. (The mRNA sequences were converted into the DNA sequences.) The relative frequency of nucleotide pair i and j (x_{ij}) can then be obtained by dividing these numbers (n_{ij}) by the total number, that is, 51. Once the x_{ij}'s are obtained, the average frequency of the ith nucleotide for the two sequences under comparison (q_i) is given by $q_i = x_{ii} + \Sigma_{j \neq i} x_{ij}/2$. Thus, we obtain $q_A = 0.186$, $q_T = 0.294$, $q_G = 0.255$, and $q_C = 0.265$ for the first nucleotide position. We also have $\pi \equiv \Sigma_{ij} x_{ij}$

$(i < j) = 0.0392$, $b_1 \equiv 1 - \Sigma q_i^2 = 0.7437$, $h = 0.005978$, $b_2 \equiv \pi^2/h = 0.2573$, and $b \equiv (b_1 + b_2)/2 = 0.5005$. Thus, the estimate of δ is $\hat{\delta} = 0.04$ from (6). However, the variance of $\hat{\delta}$ becomes 0.00087 from (9). Therefore, the standard error of $\hat{\delta}$ is 0.03. A similar computation for the third nucleotide position gives $\hat{\delta} = 0.55 \pm 0.20$. (There are no nucleotide differences at the second position.) It should be noted that in the present case application of the JC method gives $\hat{\delta} = 0.04 \pm 0.03$ for the first position and $\hat{\delta} = 0.44 \pm 0.12$ for the third position (table 5). Therefore, only when $\hat{\delta}$ is sufficiently large does the difference between the two methods become appreciably large. The estimates obtained by the TK and GIN methods are also presented in table 5. These methods again give essentially the same result for the first position, but the estimates for the third position are larger than the estimate from (6).

Table 5 also includes the estimates of δ for the first, second, and third nucleotide positions for the signal and C peptides. At the first and second positions the four methods used all give essentially the same estimate of δ. As expected, the $\hat{\delta}$ values for the signal and C peptides are larger than those for the A and B chains. At the third position of the signal and C peptides the JC method gives $\hat{\delta} = 0.63 \pm 0.16$ and equation (6), $\hat{\delta} = 0.91 \pm 0.39$. The other two methods are not applicable to this case. The value of $\delta = 0.91$ obtained by equation (6) is quite high compared with the corresponding value of the A and B chains. If we assume that the time since divergence between man and rat is 8×10^7 years, this gives a rate of nucleotide substitution of 5.7×10^{-9} per site per year. This is as high as Li et al.'s (1981) estimate (4.6×10^{-9}) of the rate of nucleotide substitution for pseudogenes. It is possible that there is little purifying selection operating at the third positions for these peptides.

Evolutionary Distance due to Deletion and Insertion

Recent data on nucleotide sequences of related genes indicate that a substantial proportion of evolutionary change of DNA sequence arises from deletion and insertion of nucleotides, particularly in noncoding regions of DNA. We note that most deletions and insertions are short and occur with an appreciable frequency (e.g., Efstratiadis et al. 1980; Langley et al. 1982; Cann and Wilson 1983). It is therefore possible to study the effects of these events on DNA divergence.

Table 5
Estimates ($\hat{\delta}$) of the Number of Nucleotide Substitutions per Site between the Human Preproinsulin and Rat Preproinsulin I Genes at the First, Second, and Third Nucleotide Positions of Codons

Gene Region and Position in Codon	$\hat{\delta}$			
	JC Method	GIN Method	TK Method	Equation (6)
A + B chains ($n = 51$):				
First	.04 ± .03	.04 ± .03	.04 ± .03	.04 ± .03
Second	0	0	0	0
Third	.44 ± .12	.60 ± .25	.79 ± .53	.55 ± .20
Signal + C peptides ($n = 54$):				
First	.17 ± .06	.19 ± .08	.15 ± .11	.18 ± .07
Second	.21 ± .07	.22 ± .08	.22 ± .07	.22 ± .08
Third	.63 ± .16	∞[a]	∞[a]	.91 ± .39

Source.—Data from Sures et al. (1980).

[a] ∞ = inapplicable case.

Nei et al. (1984) proposed a simple method of measuring the evolutionary distance between two homologous DNA sequences due to deletion and insertion: they compute the number of gap nucleotides per nucleotide site between a pair of DNA sequences compared. This quantity seems to be appropriate when a short period of evolutionary time is considered. When the evolutionary time considered is long, however, the following method seems to be better than that of Nei et al. (1984).

We again consider two homologous nucleotide sequences (X and Y) that diverged from a common ancestral sequence t evolutionary time units (e.g., years) ago. We assume that the length of a deletion or insertion is short compared with the total length of the DNA sequence (n) and that deletion and insertion occur independently. Let α be the proportion of DNA that is deleted during unit evolutionary time, i.e., $\alpha = m_D/n$, where m_D is the number of nucleotides deleted and n is the total number of nucleotides before deletion. Note also that α is the number of nucleotide deletions per nucleotide site and usually a very small quantity. Similarly, we denote by β the proportion of DNA that is inserted during unit evolutionary time, that is, $\beta = m_I/n$, where m_I is the number of nucleotides inserted. We assume that n remains more or less the same because of the compensating effects of deletion and insertion. In practice, α and β may vary with evolutionary time, and we denote the values of α and β for the ith evolutionary time unit by α_i and β_i, respectively. If we assume that deletion and insertion occur independently in sequences X and Y, the total number of nucleotide deletions and insertions per nucleotide site over the entire t is given by

$$\begin{aligned}\gamma &= 2\sum_{i=0}^{t-1}(\alpha_i+\beta_i)\\ &= 2(\bar{\alpha}+\bar{\beta})t,\end{aligned} \tag{11}$$

where $\bar{\alpha}$ and $\bar{\beta}$ are the averages of α_i and β_i over evolutionary time, respectively. In this connection it should be noted that γ measures only the DNA divergence due to deletion and insertion, and no consideration is given to the DNA changes due to nucleotide substitution.

The value of γ can be estimated in the following way. We first consider the evolutionary change of the number of nucleotides (n) in the lineage of X. Let $n_X(t)$ be the total number of nucleotides at time t in this lineage. We then have

$$\begin{aligned}n_X(t) &= n_X(t-1)(1-\alpha_{t-1})(1+\beta_{t-1})\\ &= n_X(0)\prod_{i=0}^{t-1}(1-\alpha_i)(1+\beta_i)\\ &\approx n_X(0)e^{-\Sigma\alpha_i+\Sigma\beta_i},\end{aligned} \tag{12}$$

where $n_X(0)$ is the initial number of nucleotides. A similar expression can be obtained for n for Y, that is, $n_Y(t)$. However, the total number of homologous nucleotides shared by X and Y is given by

$$\begin{aligned}n_{XY}(t) &= n_{XY}(t-1)(1-\alpha_{t-1})^2\\ &\approx n_X(0)e^{-2\Sigma\alpha_i},\end{aligned} \tag{13}$$

because insertions do not create any homologous DNA segments. Therefore, we have

$$P = \frac{n_{XY}}{\sqrt{n_X n_Y}} = \exp\left[-\sum_{i=0}^{t-1} (\alpha_i + \beta_i) \right], \tag{14}$$

where n_X, n_Y, and n_{XY} are the observed values of $n_X(t)$, $n_Y(t)$, and $n_{XY}(t)$. Thus, γ in (11) can be estimated by

$$\gamma = -2 \log_e P. \tag{15}$$

It is noted that P can also be defined as

$$P = 2n_{XY}/(n_X + n_Y). \tag{16}$$

This definition is simpler than (14), but when the rates (α and β) of deletion and insertion are not the same for sequences X and Y, (14) is more reasonable. In practice, however, (14) and (16) usually give very similar values.

Comparison with Nei et al.'s Formula

Nei et al. (1984) proposed to measure the DNA divergence due to deletion and insertion by

$$\gamma_m = g/m_T, \tag{17}$$

where g is the number of nucleotides in the gaps between two DNA sequences and m_T is the total number of nucleotides compared. This gives a minimum estimate of DNA divergence due to deletion and insertion. This can be seen from figure 2, in which an artificial example of evolutionary change of DNA due to deletion and insertion is presented. In this example sequence X at time I has a deletion of 60 nucleotides (nt) starting from nucleotide position 301, whereas sequence Y has a deletion of 40 nt starting from position 601. Therefore, the divergence between X and Y is properly measured by γ_m, which becomes 100/1,000 = 0.1. In practice, however, we do not know the ancestral sequence of X and Y, so it is difficult to determine whether the two gaps between X and Y are due to deletion or insertion. If they are caused by insertion, the ancestral sequence should have had 900 nt instead of 1,000. In this case the DNA divergence should be 100/900 = 0.111. This indicates that γ_m gives an underestimate of DNA changes if both deletion and insertion occur. Our formula (15) takes care of both deletion and insertion, though it depends on the model used. In the present case the estimate ($\hat{\gamma}$) obtained by equation (15) is 0.108, which is intermediate between the two estimates obtained above.

Another advantage of γ over $\hat{\gamma}_m$ is that it takes care of multiple events of deletion and insertion at least to some extent. In figure 2 sequence X experienced an insertion during the evolutionary period between time I and time II, whereas sequence Y experienced another deletion involving positions from 351 to 380. The latter deletion is overlapped with the deletion in X, so that γ_m gives an underestimate of DNA changes. It becomes 180/1,060 = 0.170. In (15) deletions and insertions are assumed to occur independently, and multiple deletions and insertions are taken into account. Indeed, $\hat{\gamma}$ becomes $-2 \log_e(880/\sqrt{1{,}010 \times 930}) = 0.193$, which is larger than γ_m.

Numerical Example

Efstratiadis et al. (1980) compared the nucleotide sequences of various parts of the noncoding regions of globin genes from diverse organisms. This comparison indicates that a majority of deletions/insertions involve a small number of nucleotides, but there are a few deletions/insertions in which a large number of nucleotides (more than 50) are involved. However, amino acid sequence data suggest that deletions and insertions are much less frequent in the coding regions of globin genes than in the noncoding regions (Hunt et al. 1978). To see the pattern of accumulation of DNA changes due to deletion/insertion, we computed the evolutionary distance given by (15) for the 5′ flanking region (including about 120 nt upstream starting from the cap site), 5′ leader region (about 50 nt between the cap site and the initiation codon), intron I (about 130 nt), and 3′ tail (noncoding) region (about 130 nt) of globin genes as well as for the coding region (about 438 nt or 146 codons). We used Efstratiadis et al.'s (1980) data for the noncoding region and Hunt et al.'s (1978) data for the coding region. In the latter case we used a codon rather than a nucleotide as a unit of change, because this does not change the numerical value of our measure. In both cases we assumed that the authors' alignment of sequences was correct.

The values of n_X, n_Y, and n_{XY} for the coding region (amino acid sequence) are presented in table 6. From these values we can estimate γ by using (15). For example, in the case of human (X) and newt (Y) α chain genes $n_X = 141$, $n_Y = 142$, and $n_{XY} = 141$. Therefore, $\hat{\gamma}$ becomes 0.007. Table 6 indicates that $\hat{\gamma}$ is small when the two sequences compared are closely related but tends to increase as the time since divergence (t) increases. Thus, the comparison of human and shark α chains gives a value of $\hat{\gamma} = 0.084$. However, $\hat{\gamma}$ does not seem to be linearly related to evolutionary time (fig. 3). Namely, $\hat{\gamma}$ is 0 up to $t = 300$ million years (Myr) and then increases slightly. This reflects the fact that the length of the coding region of DNA is strongly conserved in the evolutionary process.

The noncoding region of DNA undergoes a much more rapid change due to deletion/insertion. However, the four different parts of the noncoding region seem

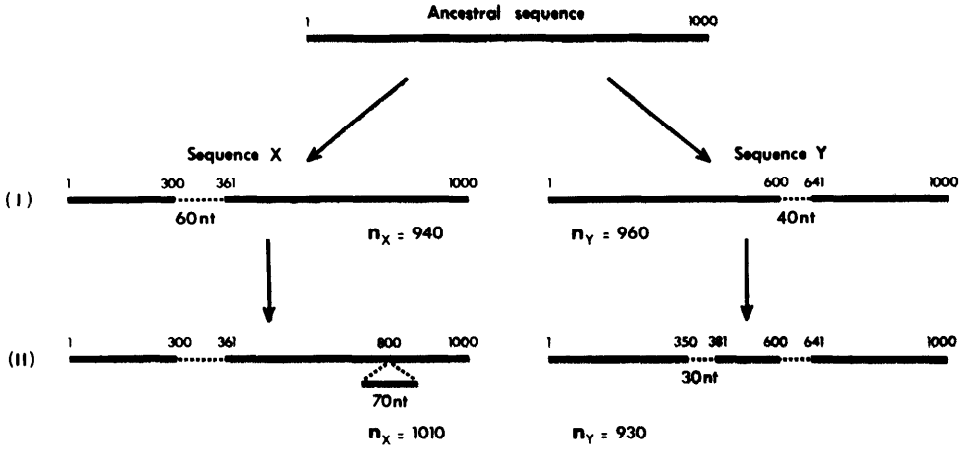

FIG. 2.—A hypothetical example of evolutionary changes of DNA sequences due to deletion and insertion. Solid lines stand for DNA sequences, and broken lines, gaps. The numbers on DNA sequences represent nucleotide positions. See text for further explanation.

Table 6
Estimates of Evolutionary Distances ($\hat{\gamma}$) due to Deletion and Insertion among the Coding Region Sequences of Various Globin Genes (below the diagonal)

Gene	1	2	3	4	5	6	7	8
1. Human α.....	(141)	141	141	140	139	139	139	134
2. Chicken α.....	0	(141)	141	140	139	139	139	134
3. Newt α	.007	.007	(142)	140	140	140	140	134
4. Carp α.....	.021	.021	.028	(142)	139	140	140	134
5. Shark α.....	.084	.084	.076	.091	(149)	140	140	134
6. Human β.....	.063	.063	.056	.056	.104	(146)	146	140
7. Chicken β.....	.063	.063	.056	.056	.104	0	(146)	140
8. Frog β	.095	.095	.102	.102	.150	.042	.042	(140)

NOTE.—In this table $\hat{\gamma}$ represents the distance per codon rather than per nucleotide. The values above the diagonal are the numbers of codons shared (n_{XY}) by the two sequences compared (the total number of codons compared minus the number of codons in the gaps). The values on the diagonal are the number of codons in the sequence concerned (n_X or n_Y).

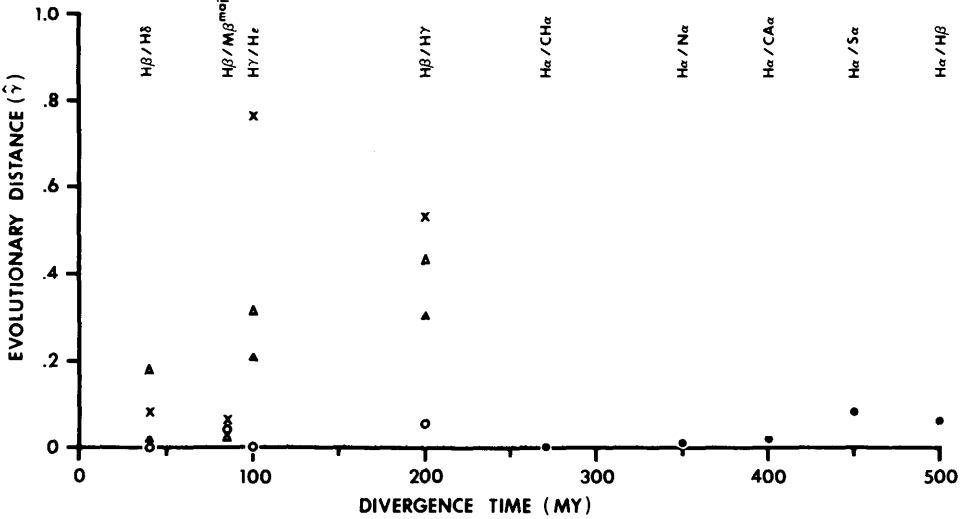

FIG. 3.—Relationships between the evolutionary distances (γ) for various parts of globin genes and evolutionary time. ● = coding regions, ○ = 5′ leader region, ▲ = intron I, △ = 5′ flanking region, X = 3′ tail (noncoding) region. To avoid overcrowding of data points, we present only the results for the comparisons involving human globins. Hα = human α globin, Hβ = human β globin, Hδ = human δ globin, $M\beta^{maj}$ = mouse β^{maj} globin, Hγ = human γ globin ($^{A}\gamma$ and $^{G}\gamma$). Hε = human ε globin, CHα = chicken α globin, Nα = newt α globin, CAα = carp α globin, and Sα = shark α globin. The evolutionary times used are identical with those used by Efstratiadis et al. (1980) and Dayhoff (1972). MY = million years.

to have different rates of accumulation of DNA changes (fig. 3). The 3′ tail region apparently has the highest rate, whereas the 5′ leader region has the lowest rate. This is probably because the 5′ leader region plays an important role for mRNA processing and translation and thus the DNA sequence is not very flexible. The relationship between $\hat{\gamma}$ and evolutionary time is again nonlinear, though $\hat{\gamma}$ generally increases as t increases. This nonlinear relationship is mainly due to the fact that a deletion or insertion occasionally involves a large number of nucleotides. Thus,

the large value of $\hat{\gamma}$ for the comparison of the 3′ tail regions of the human ϵ and γ chains is caused by the fact that the γ chain has a long stretch of deletion (44 nt) compared with the ϵ chain.

Discussion

We have seen that our new formulas, particularly equation (6), give a good estimate of nucleotide substitutions as long as the true value of δ is less than 1. For $\delta > 1$, the GIN method seems to be better than equation (6), if we exclude the cases where the formulas are inapplicable. However, when $\delta > 1$, the GIN method is very often inapplicable because of a negative argument in the logarithm involved. Therefore, if we take into account this property as well as the simplicity of equation (6) compared with the GIN formula, (6) seems to be generally preferable to the GIN method. It should also be noted that in most studies of molecular evolution δ is smaller than 1, so that equation (6) can be applied to a wide variety of cases.

It should be noted, however, that equation (6) depends on the assumption that all nucleotide sites examined are subject to the same pattern of nucleotide substitution irrespective of the location of the nucleotide. In practice, this assumption does not seem to hold in many cases. It is well known that functionally important parts of genes are subject to nucleotide substitution less often than unimportant parts. Amino acid–altering nucleotide substitutions are also known to occur less frequently than synonymous substitutions. When the number of nucleotide substitutions per site (δ) is small, this causes no problem, since there will be few backward and parallel substitutions in this case. As δ increases, however, backward and parallel substitutions may accumulate at functionally less important sites, whereas functionally more important sites may remain substitution free. In this case the method proposed here is expected to give underestimates of δ. At the present time, it is not easy to take into account this factor properly, though some approximate treatment of the problem has been proposed (Nei and Li 1979). To make a general formulation of this problem, a more detailed knowledge of nucleotide substitution in various genes is required.

Our formulation of γ in (15) was presented to quantify the effect of deletion and insertion on the evolutionary change of DNA sequences. As we have seen from data on globin genes, the evolutionary change of DNA arising from these factors occurs in a less regular fashion than that arising from nucleotide substitution. This is because there is a small proportion of large deletions and insertions that involves a large number of nucleotides. These deletions and insertions apparently occur haphazardly but affect the DNA sequences substantially once they occur. Because of this, γ generally does not increase linearly with evolutionary time and thus cannot be used as a molecular clock. Nevertheless, γ gives a quantitative measure of DNA change due to deletion and insertion and would be useful for evolutionary studies of DNA sequences.

Acknowledgments

We thank Dr. Clay Stephens for his comments on the manuscript. This work was supported by research grants from the National Institutes of Health and the National Science Foundation.

APPENDIX

Nucleotide Substitution under the Equal-Input Model

Let us denote nucleotides A, T, G, and C by 1, 2, 3, and 4, respectively. Let λ_{ij} be the rate of substitution of the jth nucleotide for the ith nucleotide per unit evolutionary time (e.g., year) and q_i be the equilibrium frequency of the ith nucleotide. In the equal-input model (Tajima and Nei 1982), $\lambda_{1j} = \lambda_{2j} = \lambda_{3j} = \lambda_{4j} = a_j$ is assumed for all λ_{ij} except λ_{jj}, which is equal to $1 - \Sigma_{i=1}^{4} a_i$ for $i \neq j$. Therefore, the transition matrix for the four nucleotides may be written as

$$\mathbf{P} = \begin{bmatrix} 1-(a_2+a_3+a_4) & a_2 & a_3 & a_4 \\ a_1 & 1-(a_1+a_3+a_4) & a_3 & a_4 \\ a_1 & a_2 & 1-(a_1+a_2+a_4) & a_4 \\ a_1 & a_2 & a_3 & 1-(a_1+a_2+a_3) \end{bmatrix}, \quad \text{(A1)}$$

and the equilibrium frequency of the ith nucleotide is given by

$$q_i = a_i \Big/ \sum_{j=1}^{4} a_j \quad \text{(A2)}$$

(Tajima and Nei 1982).

Let us now consider two long homologous nucleotide sequences (X and Y) that diverged from a common ancestral sequence t years (or evolutionary time units) ago. We denote by $y_{ij}(t)$ the proportion of homologous nucleotide pairs where X and Y have nucleotides i and j, respectively, at time t. Then we have

$$y_{ij}(t) = \sum_{m=1}^{4} \sum_{n=1}^{4} \lambda_{mi} \lambda_{nj} y_{mn}(t-1). \quad \text{(A3)}$$

Under the equal-input model (A3) is approximately given by

$$\begin{aligned} y_{ij}(t) = &\left(1 - \sum_{k \neq i} a_k - \sum_{k \neq j} a_k\right) y_{ij}(t-1) \\ &+ a_j \sum_{n \neq j} y_{in}(t-1) + a_i \sum_{m \neq i} y_{mj}(t-1). \end{aligned} \quad \text{(A4)}$$

Using (A2), we obtain

$$\begin{aligned} y_{ij}(t) &= [y_{ij}(0) - q_i q_j]\left(1 - 2\sum_{k=1}^{4} a_k\right)^t + q_i q_j \\ &\approx [y_{ij}(0) - q_i q_j] \exp\left(-2\sum_{k=1}^{4} a_k t\right) + q_i q_j. \end{aligned} \quad \text{(A5)}$$

First consider the case of $i \neq j$. In this case $y_{ij}(0) = 0$, because at time 0 the two sequences must have been the same. Therefore, we have

$$y_{ij}(t) = q_i q_j \left[1 - \exp\left(-2\sum_{k=1}^{4} a_k t\right)\right]. \quad \text{(A6)}$$

When $i = j$, we have $y_{ii}(0) = q_i$ and

$$y_{ii}(t) = q_i(1-q_i)\exp\left(-2\sum_{k=1}^{4} a_k t\right) + q_i^2. \tag{A7}$$

Let us denote by x_{ij} the proportion of pairs of nucleotides i and j ($i < j$) between sequences X and Y. When $i \neq j$ (A6) gives

$$\begin{aligned} x_{ij} &= y_{ij}(t) + y_{ji}(t) \\ &= 2q_iq_j\left[1-\exp\left(-2\sum_{k=1}^{4} a_k t\right)\right]. \end{aligned} \tag{A8}$$

This equation indicates that $x_{ij}/(2q_iq_j)$ is constant for all combinations of i and j ($i < j$).

The average number of nucleotide substitutions per site between sequences X and Y is

$$\delta = 2\sum_{i=1}^{4} q_i(1-\lambda_{ii})t. \tag{A9}$$

Under the equal-input model it becomes

$$\delta = 2\left(1-\sum_{i=1}^{4} q_i^2\right)\sum_{i=1}^{4} a_i t. \tag{A10}$$

Substitution of (A10) into (A8) gives

$$x_{ij} = 2q_iq_j\left\{1-\exp\left[-\delta\Big/\left(1-\sum_{k=1}^{4} q_k^2\right)\right]\right\}. \tag{A11}$$

Since $\pi = \Sigma_{ij}x_{ij}$ and $b_1 = 2\Sigma_{ij}q_iq_j = 1 - \Sigma_i q_i^2$ for $i < j$, we obtain (3) in the text. From (A11) we also have

$$\delta = -\left(1-\sum_{k=1}^{4} q_k^2\right)\log_e[1-x_{ij}/(2q_iq_j)].$$

Since $x_{ij}/(2q_iq_j)$ is constant for all values of $i < j$, we obtain (4) in the text.

LITERATURE CITED

CANN, R. L., and A. C. WILSON. 1983. Length mutations in human mitochondrial DNA. Genetics **104:**699–711.

DAYHOFF, M. O., ed. 1972. Atlas of protein sequence and structure. Vol. **5.** National Biomedical Research Foundation, Silver Spring, Md.

EFSTRATIADIS, A., J. W. POSAKONY, T. MANIATIS, R. M. LAWN, C. O'CONNELL, R. A. SPRITZ, J. K. DERIEL, B. G. FORGET, S. M. WEISSMAN, J. L. SLIGHTOM, A. E. BLECHL, O. SMITHIES, F. E. BARALLE, C. C. SHOULDERS, and N. J. PROUDFOOT. 1980. The structure and evolution of the human β-globin gene family. Cell **21:**653–668.

GOJOBORI, T., K. ISHII, and M. NEI. 1982*a*. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J. Mol. Evol. **18:**414–423.

GOJOBORI, T., W.-H. LI, and D. GRAUR. 1982*b*. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18:**360–369.

HUNT, L. T., S. HURST-CALDERONE, and M. O. DAYHOFF. 1978. Globins. Pp. 229–249 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. Vol. **5,** suppl. 3. National Biomedical Research Foundation, Silver Spring, Md.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–123 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:**111–120.

———. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. **78:**454–458.

KIMURA, M., and T. OHTA. 1972. On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. **2:**87–90.

LANGLEY, C. H., E. A. MONTGOMERY, and W. F. QUATTLEBAUM. 1982. Restriction map variation in the Adh region of Drosophila. Proc. Natl. Acad. Sci. **79:**5631–5635.

LI, W.-H., T. GOJOBORI, and M. NEI. 1981. Pseudogenes as a paradigm of neutral evolution. Nature **292:**237–239.

NEI, M., and W.-H. LI. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. **76:**5269–5273.

NEI, M., F. TAJIMA, and T. GOJOBORI. 1984. Classification and measurement of DNA polymorphism. *In* A. CHAKRAVARTI, ed. Methods in human population genetics. Hutchinson Ross, Stroudsburg, Pa.

SURES, I., D. V. GOEDDEL, A. GRAY, and A. ULLRICH. 1980. Nucleotide sequence of human preproinsulin complementary DNA. Science **208:**57–59.

TAJIMA, F., and M. NEI. 1982. Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J. Mol. Evol. **18:**115–120.

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.

WALTER M. FITCH, reviewing editor

Received August 29, 1983; revision received October 17, 1983.