J Mol Evol (1981) 17:167–181

Journal of Molecular Evolution
© Springer-Verlag 1981

# The Estimation of Genetic Divergence

Richard Holmquist[1] and Thomas Conroy[2]

[1]University of California At Berkeley, Space Sciences Laboratory, Berkeley, California 94720, USA
[2]University of California at Berkeley, Department of Physics Undergraduate Division, Berkeley, California 94720, USA

**Summary.** We have independently repeated the computer simulations on which Nei and Tateno (1978) base their criticism of REH theory and have extended the analysis to include mRNAs as well as proteins. The simulation data confirm the correctness of the REH method. The high average value of the fixation intensity $\mu_2$ found by Nei and Tateno is due to two factors: 1) they reported only the five replications in which $\mu_2$ was high, excluding the forty-five replications containing the more representative data; and 2) the lack of information, inherent to protein sequence data, about fixed mutations at the third nucleotide position within codons, as the values are lower when the estimate is made from the mRNAs that code for the proteins. REH values calculated from protein or nucleic acid data on the basis of the equiprobability of genetic events underestimate, not overestimate, the total fixed mutations. In REH theory the experimental data determine the estimate $T_2$ of the time average number of codons that have been free to fix mutations during a given period of divergence. In the method of Nei and Tateno it is assumed, despite evidence to the contrary, that every amino acid position may fix a mutation. Under the latter assumption, the measure $X_2$ of genetic divergence suggested by Nei and Tateno is not tenable: values of $X_2$ for the $\alpha$ hemoglobin divergences are less than the minimum number of fixed substitutions known to have occurred.

Within the context of REH theory, a paradox, first posed by Zuckerkandl, with respect to the high rate of covarion turnover and the nature of general function sites in proteins is resolved.

*Offprint requests to:* R. Holmquist, University of California at Berkeley, Space Sciences Laboratory, 1414 Harbour Way South, Richmond, California 94804, USA

**Key words:** Evolution – Genetics – REH theory – Mutations – Natural selection – Nucleic acids – Proteins – Paleogenetics

## Introduction

In two papers Nei and Tateno conclude that 1) the maximum parsimony estimates as made by the augmentation algorithm (Moore 1977) systematically err by overestimating (Tateno and Nei 1978, p 72) the total number of nucleotide replacements; 2) the selectively constrained stochastic estimates as made from protein sequence data by REH theory (Holmquist et al. 1972; Jukes and Holmquist 1972; Holmquist 1976 b) also overestimate (Nei and Tateno 1978) this total; 3) these overestimates arise from the simplifying assumptions of the REH model which neglect nonrandom codon substitution; and 4) a more accurate, and lower, estimate can be made by a modification (Nei and Tateno 1978) of Dayhoff et al.'s (1972) mutation probability matrix.

Conclusion 1), above, is more than the facts warrant (Czelusniak et al. 1978; Holmquist 1978 a, 1979 a). Here we consider the criticism of Nei and Tateno with respect to REH theory and examine their claims for a better estimate of total fixed mutations, designated by them $X_2$.

## Nonuniform Amino Acid or Nucleotide Substitution

Does the assumption of nonuniform amino acid or nucleotide substitution cause stochastic theory to systematically overestimate the total number of fixed point mutations?

0022–2844/81/0017/0167/$ 03.00

In REH theory as originally published (Holmquist et al. 1972; Jukes and Holmquist 1972; Holmquist 1978 b) the simplifying assumptions were made that the four nucleotides A, C, G, and T are equally frequent and that the three possible one-step replacements away from any one of the four nucleotides to one of the remaining three have equal probability, namely, 1/3. These assumptions are the same as in the Jukes and Cantor (1969) estimation procedure used by Kimura and Ohta (1972). To the extent the claim by Nei and Tateno that allowing for deviations from these equiprobable events would give a lower estimate of the total number of nucleotide replacements is true, that claim must apply to all the above models. However, the claim is not justified.

To see this most easily, consider two structurally different genes *1, 2* differing in exactly *n* base loci. Let the probability that the base B at locus *i* in gene *1* changes to the base B' at locus *i* in gene *2* be $P_i$, with $P_i = P_{Bi} p_{Bi \to B'i}$, where $P_{Bi}$ is the probability that base B (= A, C, G, or T) occupies locus *i* in gene *1* and $p_{Bi \to B'i}$ is the conditional transition probability to base B' in gene *2*. Then if the *n* events are independent, the probability for the passage from gene *1* to gene *2* is

$$P = \prod_{i=1}^{n} P_i = \prod_{i=1}^{n} P_{Bi} \prod_{i=1}^{n} p_{Bi \to B'i} \,, \qquad (1)$$

The four $P_{Bi}$ at each locus must sum to unity, as must the three $p_{Bi \to B'i}$ for each B. Thus Equation (1) is maximal if and only if all $P_{Bi}$ are equal and all $p_{Bi \to B'i}$ are equal. Any deviation from these equiprobable values can only lower the average probability so that more, not fewer, nucleotide replacements will have been required to achieve the observable gene change by a stochastic mechanism. Direct calculation (Holmquist and Pearl 1980) gives the same result.

Because, in proteins, the frequency of occurrence of each amino acid and each amino acid interchange is *not* (Jukes et al. 1975; Holmquist 1978 c; Holmquist 1979 b) proportional to that expected from the genetic code table (in which, ignoring a small correction for the chain terminating codons, A:C:G:T $\sim$ 1:1:1:1), the observed types of amino acid substitution imply nonuniformity in the corresponding gene structures and nucleotide interchanges. Thus, nonuniform amino acid substitution must also increase stochastic estimates of the total mutations fixed.

Although Nei and Tateno (1978) state that we did not consider nonrandom amino acid substitution, this is not in fact the case. The conclusions of the preceding paragraphs follow directly from the complete analytical treatment, resulting in closed algebraic expressions for the observable effect of nonuniformities of the above sort published three years ago (Holmquist 1976 a) and recently reviewed (Ratner 1978). We also discussed the experimentally measured magnitudes of these types of nonuniformity (Holmquist 1976 b).

Now consider a nonuniform distribution of fixed mutations over the nucleotide loci. This is known to increase REH estimates both in practice (Jukes and Holmquist 1972; Holmquist and Pearl 1980) and in theory. Any concentration of fixed mutations at some loci results in more superimposed fixations at those loci. A greater total number of fixed mutations will thus be required to cause any given extent of observable change. REH estimates made from a simplified model in which each variable locus has an equal likelihood of fixing mutations will underestimate the total fixed mutations.

The conclusion of this section is that *any* type of nonuniformity must, on the average, increase REH.

## On the Measure $X_2$

In their paper Nei and Tateno (1978) suggest a new measure of genetic divergence which they designate $X_2$ and which is based on Dayhoff's PAM measure for amino acids, but so modified as to express an approximate minimal estimate of the total number of nucleotide substitutions for a given evolutionary period. They go on to state with reference to their Table 1, that *"REHC is always considerably larger than $X_2$"*. This is not so: in Table 1 below we compare, per 100 codons free to

**Table 1.**

| Observed amino acid differences in 100 variable codons | Evolutionary distance in | | | |
|---|---|---|---|---|
| | PAMs | $X_1$ | REHs | $X_2$ |
| 1 | 1 | 1.3 | 1.5 | 1.7 |
| 5 | 5 | 6.9 | 7.5 | 8.7 |
| 10 | 11 | 14.2 | 15 | 17.2 |
| 15 | 17 | 22.0 | 22 | 26.9 |
| 20 | 23 | 30.4 | 31 | 38.6 |
| 25 | 31 | 39.4 | 40 | 52.6 |
| 30 | 39 | 49.3 | 50 | 67.3 |
| 35 | 48 | 60.0 | 62 | 81.9 |
| 40 | 58 | 71.8 | 75 | 97.2 |
| 45 | 70 | 84.8 | 88 | 115.0 |
| 50 | 83 | 99.5 | 101 | 137.3 |
| 55 | 98 | 116.1 | 120 | 165.3 |
| 60 | 117 | 135.3 | 140 | 199.4 |
| 65 | 140 | 158.0 | 163 | 239.6 |
| 70 | 170 | 185.4 | 191 | 287.9 |
| 75 | 208 | 220.0 | 230 | 351.3 |
| 80 | 260 | 266.0 | 280 | 448.4 |
| 85 | 370 | 335.4 | 356 | 617.6 |
| 90 | -- | 463.1 | 516 | 929.1 |
| 94 | -- | -- | $\infty$ | -- |

PAM and REH values are from Table 1 in Holmquist (1972a). $X_1$ was calculated from the Jukes/Cantor (1969) formula and the Kimura/Ohta (1972) approximation (Eqs. 8 and 9, this paper). $X_2$ was calculated from Nei and Tateno's (1978) Eq. 11. The maximum error in $X_2$ can be greater than the 5% stated by Nei and Tateno: e.g., around $p_d = 0.1$ there is a 9% discrepancy between their Eqs. 11a and 11b. The reader should be aware that Nei and his coworkers are inconsistent with respect to their usage of $X_1$ and $X_2$, the two symbols meaning one thing in Nei and Chakraborty (1976) and another in Nei and Tateno (1978). In the present paper it is the usage in their 1978 paper which is referred to

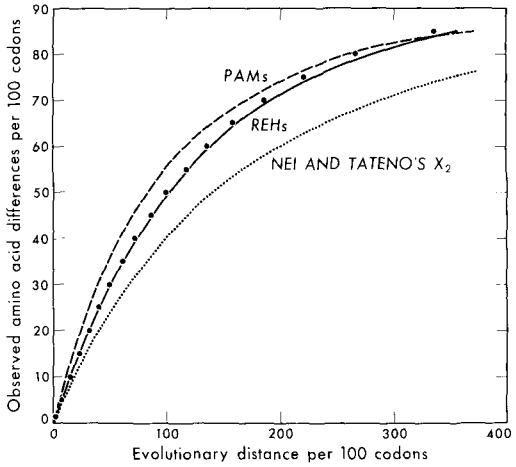


**Fig. 1.** The abscissa gives the total number of nucleotide substitutions necessary to explain a given number of observed amino acid differences between two homologous proteins 100 residues in length; – – – The PAM method of Dayhoff and her coworkers; —— The REH model of Holmquist, Cantor and Jukes; ···· The model of Nei and Tateno; • • • The Jukes/Cantor measure $X_1$ (note the solid circles were calculed independently of the solid REH curve and do not define the latter). The data for this Figure is in Table 1

accept mutations four measures of genetic divergence – the PAM measure of Dayhoff and her colleagues (1972); the Jukes and Cantor (1969) measure $X_1$; the REH measure of Holmquist et al. (1972); and the $X_2$ measure of Nei and Tateno (1978). The first three measures are similar in magnitude, whereas the Nei and Tateno measure $X_2$ is larger (Table 1) than any of them by a wide margin as graphically evident in Fig. 1. In this connection it should be emphasized that Table 1 in the Journal of Molecular Evolution 1, 211 (1972), which was calculated for that part of the gene free to fix mutations, and from which the REH values in Table 1 of the present paper were taken, is in agreement with the REH values as calculated by the method of Holmquist et al. (1972) and Jukes and Holmquist (1972). From the latter two papers, REHC = $100\,\mu_2 T_2/T$. When the entire gene is variable $T_2 = T$ so that REHC = $100\,\mu_2$. From Table A4 in the Journal of Molecular Biology *64*, 145 (1972), the relation between amino acid differences per 100 variable codons and REHC is quantitatively the same as in the Journal of Molecular Evolution. The reader can verify these statements by checking the relevant tables: as one illustration only, for 30 amino acid substitutions per 100 variable codons either Table 1 gives an REHC value of 50, from which $\mu_2$ = REHC/100 = 0.50 which from Table A4 would cause 30.10 amino acid substitutions. The very small difference between these values (30 *vs* 30.10) is due to the effect of chain terminating codons allowed for in the H/C/J theory. The Jukes/Cantor (1969) method used by Kimura and Ohta (1972), Nei and Chakraborty (1976), and Nei and Tateno (1978) is but a special limiting case of REH theory.

We do not understand Nei and Tateno's belief that REH and PAM values are not comparable: by direct calculation the numerical values of each are in close agreement in their estimates of genetic change for 100 variable codons both for proteins (Holmquist 1972 a and Table 1, this paper) and nucleic acids (Holmquist 1973).

The question then arises as how, in their Table 1, Nei and Tateno (1978) obtained values of $X_2$ smaller than REHC. The small values are due to two factors: 1) their mathematical model is based on the incorrect assumption that the entire structural gene of α-hemoglobin or of cytochrome *c* is free to fix mutations, and 2) they assume a different number of variable codons when calculating $X_2$ than when calculating REHC. We shall consider each of these in turn, but before doing so, it is helpful to briefly review the calculational methodology of REH theory for proteins.

### *A Resumé of REH Calculational Methodology for Proteins*

The total number of fixed nucleotide replacements REH separating two genes descendant from a common ancestral gene is related to the total number $T_2$, designated *varions*, of coding triplets which have been free to accept mutations in one or the other or both genes during some part of their period of divergence, and to the average number of fixations per varion $\mu_2$, designated the *fixation intensity*, by

$$\text{REH} = \mu_2 T_2 \quad . \tag{2}$$

REHC, the random evolutionary hits per codon, is obtained from Equation 2 by dividing by the total number T of coding triplets compared in the two genes.

In comparing pairs of homologous proteins the ratio $r_e$

$$r_e = \frac{n_e(2) + n_e(3)}{n_e(1)} \tag{3}$$

is determined experimentally from the observed numbers $n_e$ (1), $n_e$ (2), and $n_e$ (3) of amino acid replacements of the minimal 1-, 2-, and 3-base type, respectively. From this ratio $\mu_2$ is obtained either graphically (Jukes and Holmquist 1972) or analytically (Holmquist 1978 b). Approximately $\mu_2 \sim 6r_e$, which follows from the initial slope of Fig. A1 in Jukes and Holmquist (1972), and is useful for quick calculations (Nei and Tateno 1978; Holmquist 1978 b). Accurate values of $\mu_2$ are most readily calculated from the polynomial regression in Holmquist (1978 b).

However, it has been demonstrated (Nei and Tateno 1978) that the above procedure tends to overestimate $\mu_2$. This bias can be reduced to negligible proportions by replacing Eq. 3 by

$$\hat{r}_e = \frac{n_e(2) + n_e(3)}{n_e(1) + 1} \quad (4)$$

(Holmquist 1978 b).

From $\mu_2$ the proportion P (0) of unhit varions can be calculated (Jukes and Holmquist 1972; Holmquist 1978 b) so that

$$T_2 = \frac{AAD}{1 - P(0)} \quad (5)$$

In the original Jukes and Holmquist (1972) method $T_2$ was computed from the somewhat more complicated expression

$$T_2 = T\left[1 - P_e(0) + P_e(1)\frac{P(0)}{P(1)}\right] , \quad (6)$$

where $P_e(0) = n_e(0)/T$, $P_e(1) = n_e(1)/T$, $n_e(0)$ being the experimentally observed number of amino acid identities between the two proteins being compared. P (1) is the theoretically expected proportion of amino acid replacements of the minimal 1-base type among the $T_2$ varions. The two values of $T_2$ calculated from Eqs. 5 and 6 differ little numerically, but from a theoretical viewpoint Eq. 5 is preferable because it doesn't force agreement of the expected and observed numbers of amino acid replacements of the minimal 1-base type (Nei and Tateno 1978). A full discussion of the calculations is given elsewhere (Holmquist 1978 b).

We have calculated $\mu_2$, $T_2$, and REH for the simulation data in Nei and Tateno's (1978) Table 1 using both our original method (Eqs. 3 and 6, this paper) and with Eq. 4 replacing Eq. 3 to see how much $\mu_2$ is overestimated when using Eq. 3. These parameters were also calculated with Eq. 5 replacing Eq. 6 to see how much effect the two different ways of estimating $T_2$ had. The results are in Table 2 below. First, we note that the bias, $(r_e - f_e)/f_e$ overestimates r by an average of 6.5% (range 3.4–12.6%). Second, using Eq. 4, $f_e$ underestimates r by the negligible amount of about 0.76% ($\epsilon$) in Table 2. This has two corollaries: 1) The Taylor expansion approximations (keeping terms through the second order) from which Eqs. 6 and 14 in Holmquist (1978 b) were derived, and which give the expression for $\epsilon$ in footnote *a* to Table 2, are quite accurate and we do not need to worry about the effect of third and higher order terms; (2) $\hat{f}_e$ as defined in Eq. 4 is for all practical purposes an unbiased estimator of r and when used in conjunction with Fig. A1 in Jukes and Holmquist will give nearly unbiased values of $\mu_2$.

From the columns in Table 2 which list $\mu_2$, $T_2$, and REHC it can be seen that whenever $\mu_2$ is overestimated, $T_2$ is underestimated so that their product REH is relatively constant and in this sense robust to small errors in either $\mu_2$ or $T_2$. This is in agreement with a theoretical error analysis (Holmquist 1978 b). However, because between a fixation intensity $\mu_2$ of 2–6 there exists a nonlinearity in P (0) (see Fig. 3 in Holmquist et al. 1972), the underestimate in $T_2$ does not entirely compensate for the overestimate in $\mu_2$ so that the lower values of REHC in Table 2 are to be preferred. It should also be noted that it makes little difference whether $T_2$ is calculated from Eq. 5 or from Eq. 6. Finally we note that the values of $\mu_2$, $T_2$, and REHC in Table 2 are in essential agreement with those in Nei and Tateno's (1978) Table 1. The source of Nei and Tateno's error in concluding that the two measures they discuss, $X_1$ and $X_2$, are smaller than REH thus does not lie in any gross miscalculation of the REH values. The error must lie in the calculation of $X_1$ and $X_2$. In the following sections it is shown precisely how this error was introduced by the Nei/Tateno calculations.

*Random REH Theory As A Two-parameter Model – Stochastic Variance*

Two parameters, $\mu_2$ and $T_2$, suffice to describe REH theory under the assumptions of the equiprobability of genetic events. An alternative is to attempt to describe the evolution of proteins or nucleic acids in terms of a one-parameter theory as Nei and Tateno (1978) do in their Eq. 7, which defines $X_1$, and in their Eqs. 11, which define $X_2$. Such an approach cannot succeed in view of the many known selective constraints on protein structure. Nei and Tateno claim a smaller variance for the one-parameter measure $X_1$, as compared to REHC, but if one compares their Figs. 1 and 2, the variance in $X_1$ is hardly smaller than in REHC, and had $X_1$ been calculated on the basis of the same number of variable codons as REHC, or vice versa, there would be less difference still (See, this paper, Fig. 1 and Table 1, where both $X_1$ and REHC values are calculated for 100 variable codons). A correctly calculated $X_1$ is very nearly equal to REHC as evidenced in the 3rd and 6th column from the right in Table 2. Nei and Tateno (1978) state: "... an estimator of which the expected squared deviation from the population parameter is small is generally considered to be better than an unbiased estimator of which the squared deviation is large." This is true if the bias is negligibly small. But in the present case the bias in $X_1$ and $X_2$, as calculated by Nei and Tateno is so large that these estimates, as we shall see, are outside the realm of biological possibility, and one can state more generally that all one-parameter estimates are untenable.

The variances observed are to a nonnegligible extent inherent in the information content of the data (small observed number of substitutions for any pair of homologous proteins). We show this explicitly in the section on computer simulations later in this paper.

**Table 2.** Measures of genetic divergence

| PAM = 20 ($100X_2 = 34$) | $r_e$ | $\hat{r}_e$ | $\frac{r_e - \hat{r}_e}{\hat{r}_e}$ | $\epsilon$[a] | $\mu_2$ | $T_2$ T = 103 | REHC | $T_2P_d$[c] | $X_1$[d] | $T_2X_1$[e] | $X_2$[f] | $T_2X_2$[g] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Replication 1 | 0.466 | 0.437 | 0.066 | -0.0064 | | 27.6[b] | 0.74 | | | | | |
| | | | | | 2.77 | 27.6 | 0.74 | | | | | |
| | | | | | | 28.0 | 0.70 | 0.787 | 0.33 | 0.69 | 0.424 | 1.14 |
| | | | | | 2.59 | 28.1 | 0.71 | | | | | |
| Replication 2 | 0.419 | 0.387 | 0.083 | -0.0101 | | 22.1 | 0.53 | | | | | |
| | | | | | 2.48 | 22.1 | 0.53 | | | | | |
| | | | | | | 22.6 | 0.50 | 0.756 | 0.25 | 0.49 | 0.303 | 0.79 |
| | | | | | 2.29 | 22.7 | 0.51 | | | | | |
| Replication 3 | 0.729 | 0.668 | 0.091 | -0.0123 | | 21.4 | 0.97 | | | | | |
| | | | | | 4.65 | 21.4 | 0.97 | | | | | |
| | | | | | | 21.7 | 0.87 | 0.878 | 0.28 | 0.83 | 0.347 | 1.62 |
| | | | | | 4.14 | 21.8 | 0.88 | | | | | |
| Replication 4 | 0.540 | 0.501 | 0.078 | -0.0087 | | 24.0 | 0.75 | | | | | |
| | | | | | 3.24 | 24.0 | 0.75 | | | | | |
| | | | | | | 24.4 | 0.71 | 0.820 | 0.29 | 0.69 | 0.370 | 1.19 |
| | | | | | 2.99 | 24.5 | 0.71 | | | | | |
| Replication 5 | 0.753 | 0.669 | 0.126 | -0.0236 | | 15.6 | 0.73 | | | | | |
| | | | | | 4.86 | 15.6 | 0.73 | | | | | |
| | | | | | | 15.8 | 0.63 | 0.879 | 0.20 | 0.61 | 0.241 | 1.19 |
| | | | | | 4.15 | 15.9 | 0.64 | | | | | |
| PAM = 36 ($100X_2 = 62$) | $r_e$ | $\hat{r}_e$ | $\frac{r_e - \hat{r}_e}{\hat{r}_e}$ | $\epsilon^{\alpha}$ | $\mu_2$ | $T_2$ T = 141 | REHC | $T_2P_d$[c] | $X_1$[d] | $T_2X_1$[e] | $X_2$[f] | $T_2X_2$[g] |
| Replication 1 | 1.052[h] | 0.999 | 0.053 | -- | -- | -- | 0.42[i] | — | 0.45 | | 0.606 | -- |
| Replication 2 | 0.399 | 0.386 | 0.034 | -0.0016 | | 55.3 | 0.93 | | | | | |
| | | | | | 2.36 | 55.3 | 0.93 | | | | | |
| | | | | | | 55.8 | 0.91 | 0.753 | 0.49 | 0.88 | 0.667 | 1.41 |
| | | | | | 2.29 | 55.9 | 0.91 | | | | | |
| Replication 3 | 0.482 | 0.464 | 0.039 | -0.0020 | | 49.6 | 1.01 | | | | | |
| | | | | | 2.87 | 49.4 | 1.01 | | | | | |
| | | | | | | 50.0 | 0.98 | 0.801 | 0.46 | 0.95 | 0.626 | 1.60 |
| | | | | | 2.76 | 50.0 | 0.98 | | | | | |
| Replication 4 | 0.480 | 0.462 | 0.039 | -0.0023 | | 45.8 | 0.93 | | | | | |
| | | | | | 2.85 | 45.8 | 0.93 | | | | | |
| | | | | | | 46.2 | 0.90 | 0.799 | 0.42 | 0.87 | 0.562 | 1.46 |
| | | | | | 2.74 | 46.4 | 0.90 | | | | | |
| Replication 5 | 0.348 | 0.335 | 0.039 | -0.0022 | | 48.3 | 0.71 | | | | | |
| | | | | | 2.07 | 48.3 | 0.71 | | | | | |
| | | | | | | 48.9 | 0.69 | 0.715 | 0.39 | 0.67 | 0.521 | 1.06 |
| | | | | | 1.99 | 49.1 | 0.69 | | | | | |

[a] The magnitude of the relative error, as given by Eq. 14, in Holmquist (1976b), in the estimator $\hat{r}_e$. $\epsilon \cong -[2 - P(1)]/[T_2P(1)]^2$

[b] The first, second, third and fourth rows of numbers are for Eqs. 3 and 6, Eqs. 3 and 5, Eqs. 4 and 6, Eqs. 4 and 5, respectively

[c] Calculated from Eq. 10 from text and values of $P_e(0)$ reported by Nei and Tateno in their Table 1

[d] Calculated from Eq 7 in Nei and Tateno (1978), on assumption entire gene is free to fix mutations

[e] Calculated from Eq. 7 in Nei and Tateno (1978), on assumption $T_2$ coding triplets of the genes are free to fix mutations. Eq. 7 was multiplied by $T_2/T$ to make this distance directly comparable to the other values in the table

[f] Calculated from Eq 11 in Nei and Tateno (1978) on assumption entire gene is free to fix muations

[g] Calculated from Eq. 11 in Nei and Tateno (1978) on assumption $T_2$ coding triplets of the genes are free to fix mutations. Eq. 11 was multiplied by $T_2/T$ to make this distance directly comparable to the other values in this table

[h] Exceeds theoretical maximum expectation value of 1.046. The number of minimal 1-base amino acid replacements in this simulation is 19 and the number of minimal 2-base replacements is 20. The high value of $\hat{r}_e$ indicates that either $\mu_2 > 8$ fixations per varion with $T_2 > 39$ or else this replication is a statistical outlier. As the expected number of hits per 100 varions is $100X_2 = 62$, the latter explanation is clearly correct

[i] A minimal value REHC = 1.5 MBDC (minimal base differences per codon) is given

*Restricted Nucleotide or Codon Mutabilities Require At Least A Two-Parameter Model*

It would seem this statement is obvious since the early demonstration of Fitch and Markowitz (1970) that two Poisson parameters plus two parameters which estimate the number of variable codons in each of the two groups of residues to which each Poisson parameter, respectively, applies better describe some of the data than a one Poisson parameter model. The Fitch and Markowitz model is thus a four-parameter model. However, because this model did not consider the detailed mechanisms of macromolecular divergence it necessarily stopped short at describing the data. Although in REH theory the incorporation of mechanistic constraints complicated the mathematical development, this was in part offset by reducing the number of parameters from four to two – $\mu_2$ and $T_2$. It turned out that the dynamic constraints were sufficiently important that the net result was an improved estimate of the total mutations fixed. The important point is that the parameter $T_2$ is an average measure of the restrictions on certain nucleotide or codon loci to accept mutations. In the next paragraphs the consequences of ignoring these restrictions are examined and found to explain the error in the Nei/Tateno calculations.

Nei and Tateno assumed that the entire structural gene is free to accept mutations. In their simulations of the divergence of α-hemoglobin, this gene corresponded to the 141 amino acid residues of that chain. It has been clear for a long time that the requirements of biological function limit the mutability of various sites in α-hemoglobin. Fitch (1972) has estimated that in any given α-hemoglobin chain an average of only 50 of these 141 residues are free to accept mutations at any point in time. For the five replications Nei and Tateno (1978) give in their Table 1 for an evolutionary period of PAM = 36, we found an average value of $T_2$ of 50 (T = 141) (see Table 2, this paper) in agreement with Fitch's estimate. For the shorter evolutionary period of PAM = 20 simulated by Nei and Tateno, our average value of $T_2$ is 22. One notes that the $T_2$ values estimated from protein sequence data are always somewhat larger than the number of amino acid differences (AAD) between the two chains, here averaging 38 and 18 respectively. This is in agreement with the common sense observation that some sites which could have varied did not because they were not hit, so that the total number of sites free to fix mutations should be larger than the AAD (see Eq. 5). Because of base replacements at the third position within codons, the true value of $T_2$ should be larger than the value calculated from protein data as such replacements are usually silent in amino acid interchanges. The high rate of acceptance of mutations in the fibrinopeptides B (Dickerson 1971) and snake venom toxins (Wilson, Carlson and White, 1977) argues that those sites which can change have done so. Experimentally it is unambiguously known that substantially less than all the gene is able to accept mutations.

When calculated on the basis of $T_2$ residues free to fix mutations, Nei and Tateno's $X_2$ values average 1.6 and 1.7 times *larger* than the REHC estimates for 36 and 20 PAMS respectively (See the columns in Table 1 labeled REHC and $T_2X_2$. The column labeled $X_2$ gives the $X_2$ value found by Nei and Tateno on the assumption the entire gene is free to accept mutations). That the $T_2X_2$ values are larger than the REHC values is in fact required, as we have seen, since Nei and Tateno do not assume equiprobability of genetic events.

It is instructive to consider the fact that only a portion of the gene is free to fix mutations from another viewpoint. The total number $X$ of fixed mutations separating two homologous genes under random substitution is given exactly by

$$X = \frac{\ln\left(1 - \frac{4}{3}\Pi\right)}{\ln\left(1 - \frac{4}{3L}\right)} \tag{7}$$

where $L$ is the number of nucleotides free to fix mutations in the gene and Π is the expected proportion of nucleotide differences between the gene pair (Eq. 16 in Holmquist 1972 b). Here the effect of $L$ on $X$ is explicit. If the covarion hypothesis of Fitch and Markowitz (1970) is correct $L$ may be small, and the smaller it is, the larger is $X$. If $L$ is sufficiently large then Eq. 7 becomes (since $L = 3T_2$)

$$X_1 \equiv \frac{X}{T_2} = -\frac{9}{4}\ln\left(1 - \frac{4}{3}\Pi\right) \tag{8}$$

which is simply the Jukes/Cantor formulation (1969), where Π is calculated from the proportion of amino acid differences $P_d$, $(1 - i_d)$, among those sites free to fix mutations:

$$i_d = (1 - \Pi)^2\left(1 - \frac{1}{4}\Pi\right) \tag{9}$$

(Kimura and Ohta 1972). To get this proportion one must know both the number of amino acid differences *and* the number of sites $T_2$ free to fix mutations, the correct proportion being the ratio between them. This proportion can be calculated from $T_2$ and the $P_e$ (0) data in Nei and Tateno's (1978) Table 1, and is listed in Table 2 of the present paper as

$$T_2 P_d = [1 - P_e(0)]\frac{T}{T_2} . \tag{10}$$

When the Jukes/Cantor measure $X_1$ is calculated on this basis (the column labeled $T_2X_1$ in Table 2, this paper) it is in agreement with the REHC value as would be anticipated from Fig. 1. The REHC values were not

adjusted in any way to force the observed agreement with the $_{T_2}X_1$ values, but follow from recognizing that all codon sites are not equally mutable.

The difference between $_{T}X_1$ and $_{T_2}X_1$ and between $_{T}X_2$ and $_{T_2}X_2$ Nei and Tateno believe to be irrelevant to the estimation of genetic distance. This distinction is in fact the central issue distinguishing our analyses. $_{T_2}X_2$ recognizes the known biology: sites are restricted in their variability. $_{T}X_2$ (Nei and Tateno's recommended measure) does not take cognisance of this indisputable fact. Although Nei and Tateno do not assume that all codons are equally mutable they do assume in calculating their measure $_{T}X_2$ that the same mutation probability matrix applies to each codon *site*. In particular, they assume that synonymous amino acid substitutions are for all codon sites given by the diagonal elements of this matrix. Nei and Tateno treat all amino acid sites as potentially variable all the time, and that is prohibited by the requirements of biological function. The essential distinction, pointed out by Fitch and others, is not made in their model between truly invariant sites and sites that are not varied either because they have not been "hit" or because of convergence and parallel evolution. There is thus an important distinction between site dependency and amino acid dependency. There is a high correlation between the two for *some* sites such as relatively immutable cysteine sites. However, the $_{T}X_2$ method assumes that *all* sites have the same probability of going from a given amino acid to any other (or of not changing at all). Since this assumption is certainly false, such a high correlation between the two dependencies is an exception rather than the rule and is *not* a property of the Nei and Tateno method.

The values of $X_1$ and $X_2$ in Table 1 of Nei and Tateno are too low because they assumed incorrectly, both in the simulation of the $\alpha$-hemoglobin divergences and in their theory, mutability of all gene sites. The agreement of an unrealistic simulation with a corresponding unrealistic theory cannot form a valid basis for understanding evolutionary divergence. Because of this they err in stating that *"the relative frequencies of $P_e$ (0), $P_e$ (1), $P_e$ (2), and $P_e$ (3) in* [Nei and Tateno's (1978)] *Table 1 are quite different from the theoretical values expected under random nucleotide substitution..."*. Nei and Tateno find this difference because they assumed all 141 codons of $\alpha$-hemoglobin were variable. When account is taken of the restricted variabilities, there is not disagreement between observed and expected values: the same experimentally observed number of amino acid differences can arise from a higher substitution rate over fewer sites or from a lower rate over many sites. Thus for PAM = 20, in Table 2, the same observable proportions of amino acid differences of minimal 0-, 1-, 2- and 3-base types can result from 34 fixed mutations over 103 amino acid sites or from 78 fixed mutations over about 23 sites.

Nei and Tateno (1978) referring to their Table 2, state that Fitch's estimates (designated Y by Nei and Tateno) of the minimum numbers of nucleotide substitutions in hemoglobins and cytochromes *c* (given in Nei and Chakraborty, 1976) are in agreement with Nei and Tateno's (1978) X values. This is proof of the incorrectness of Nei and Tateno's X values as Fitch (1976) has shown that his minimum phyletic distances are nowhere near their final values: Although there were 23 nodes between pig and cotton cytochrome *c*'s, the minimum distances were still increasing linearly with no evidence of leveling off, a minimum average of 2.3 additional replacements being detected per added node. Further, their agreement with Fitch's values is not as good as the authors indicate. For 11 out of 13 of the $\beta$-hemoglobin estimates in their Table 2, Nei and Tateno's X is less than the biological minimum possible given by Fitch. This cannot be attributed to "sampling error". The same is true for 8/13 $\alpha$-hemoglobin comparisons and 4/10 cytochrome *c* comparisons.

The incorrectness of Nei and Tateno's X values is also brought out from the most recent parsimonious estimates of the genetic divergence between human $\alpha$-hemoglobin and the $\alpha$-hemoglobins of rabbit, horse, pig, llama, cow, kangaroo and chicken. The most recent unaugmented maximum parsimony estimates (Goodman, personal communication) and Nei and Tateno's comparable estimates X are 45/35.8, 46/24.9, 47/24.9, 53/32.2, 47/23.7, 59/39.6, and 84/56.6, the unaugmented maximum parsimony estimate being above the line. In each of these cases the estimates of Nei and Tateno are far below the minimum that must have occurred given any reasonable phylogeny. These seven species comprise half the data in Nei and Tateno's entries for $\alpha$-hemoglobin in Table 2 and indicate the Y values quoted there are obsolete. The much higher recent values just given reflect the information in new experimentally sequenced proteins. There is thus no agreement between Y values and X, but on the contrary serious disagreement.

*The Magnitude of the Number of Varions*

The adequacy of the above discussion (though not of the last two paragraphs) is thus seen to hinge on whether $T_2$ is large or small relative to T. Although in most cases less than the entire gene is free to fix mutations, that is $T_2 < T$, $T_2$ need not be as small as the protein sequence data would indicate. Silent changes at the third position within the codon would cause $T_2$ to be larger. For the fibrinopeptides, and for the third coding position in the histone IV gene, Kimura (1977) observed that the rate of fixed mutations was about $4 \times 10^{-9}$ per nucleotide site/year. Taking the chicken-human divergence at about 300 million years ago, and noting that there are $141 \times 3 = 423$ nucleotides in the $\alpha$-hemoglobin gene, and

taking the human-tuna divergence at about 400 million years ago, and noting that there are 103 × 3 = 309 nucleotides in the cytochrome *c* gene, it is readily calculated that had the evolutionary process been selectively neutral there would have been for either comparison about 500 fixed point mutations, enough to change most amino acids in either the alpha hemoglobin or cytochrome *c* chain, if each amino acid site had been free to fix mutations. The number of amino acid substitutions observed must therefore be not too far removed from the number of amino acid loci that were substitutable. Thus $T_2$, when estimated from amino acid sequence data, is an approximate minimal estimate of the true number of varions. This has been independently confirmed by the recent $S_v$ calculations of Noguchi (1977) who find that for the mammalian/bird and mammalian/amphibian divergences of cytochrome *c* the number of convarions $S_v$ is equal, to within experimental error, to the number of varions $T_2$ as calculated from REH theory for proteins and reported by Moore et al. (1976).

## Varions, Covarions, and Convarions

Nei and Tateno (1978) make the statement that "The concept of covarions is essentially a statistical one and has no rigid meaning like the variable codons in the J/H [Jukes and Holmquist, 1972] model". In fact, both concepts are statistical and neither is rigid. *Covarions* are defined as the number of codon positions in a gene which are free to accept mutations at some *point* in time. The number of *varions* $T_2$ is defined as the time average number of codon positions in one or the other or both of two homologous genes which have been free to accept mutations during some part of their *period* of divergence. The number of varions is thus simply the time average sum of the covarions, no covarion locus being counted more than once, in the two homologous genes during the period of their divergence.[1] Holmquist, Cantor and Jukes' (interactive REH) model and Fitch's covarion model have been independently studied by Karon (1979). Neither model requires the justification of the other; each can stand on its own merits. Both lead to similar conclusions. Very recently Tamaji Noguchi (1977, 1978) has extended these concepts and defined the term *convarion* to resolve a discrepancy between paleontological and molecular evolutionary estimates of the time of divergence between prokaryotes and eukaryotes.

[1] Varions are *not* defined as a group of codons which are subject to independent substitution with the same probability. The compensating charge-altering substitutions in the African papionine α-hemoglobins (Hewett-Emmett et al. (1976) are but one of many illustrations of the dependency of one substitution upon another. In estimating the *number* of varions by the REH model, the assumption of their independence is made for mathematical tractability. The number of variable codons $T_2$ is not fixed throughout evolutionary time. Table 7 in this paper or Table 5 in Moore et al. (1976) demonstrate conclusively that $T_2$ is not constant. Such constancy would be totally incompatible with a changing set of covarions

## Computer Simulations

We have independently repeated the computer simulations of evolutionary divergence described by Nei and Tateno (1978) and find that their criticism of REH theory based on those simulations cannot be supported, either logically, or experimentally from the simulation results.

### *A Logical Refutation of Nei and Tateno's Results*

The first claim by Nei and Tateno (1978) is that for a total of 2v = 10 mutations distributed over 50 varions the sample mean value of $\mu_2$ as given by REH theory is near 1.2, hence grossly overestimated as for this simulation the population mean $\mu_2$ is known to be 0.20.

In Table 3 we list some expected consequences, all experimentally observable, of taking 2v = 10, for which the expected $\mu_2$ = 10/50 = 0.20. In particular the expected number of amino acid differences, as calculated by two methods, is given in the last two columns of the table. So that the reader may verify the calculations in sufficient detail, the total number of expected amino acid differences is decomposed in the middle columns of Table 3 into the number of replacements of the minimal 0-, 1-, 2-, and 3-base types: n (i), i = 0, 1, 2, and 3. The first method of calculation is conventional REH theory under the assumption of the equiprobability of genetic events (Jukes and Holmquist 1972; Holmquist 1978 b). The second method of calculation was to use Eqs. 7 and 9 as a check because these equations have been verified independently in at least three laboratories (Holmquist 1972 b; Kimura and Ohta 1972; Nei and Chakraborty 1976). By either method, the expected number of amino acid differences is about 6.9 (Table 3). The expected error in this number is about $\sqrt{}$ 6.9 or 2.6. The approximate distribution of amino acid replacements Nei and Tateno would have had to observe in order to arrive at a value $\mu_2$ = 1.2 is readily calculated and shown in the second row of Table 3 (2v = 60). The approximate total number of amino acid substitutions required to obtain a value of $\mu_2$ = 1.2 is thus about 28 by either method of calculation (27.75 by REH theory; 28.11 by Equations 7 and 9). The error in this is about $\sqrt{}$ 28 or 5.2. There is no way 28 ± 5.2 can be reconciled with 6.9 ± 2.6. Moreover, a total of 60 fixed mutations would be necessary to explain this distribution arising from $\mu_2$ = 1.2. The distributions of n (0), n (1), n (2), and n (3) are so different for 2v = 10 and 2v = 60 that it is inconceivable that Nei and Tateno erred in counting (for 2v = 10, n (1) ∿ 7; for

Table 3. The computer simulations of Nei and Tateno (1978)

| Simulation | Total expected mutations 2v | $\mu_2$ | n(0)[c] | n(1) | n(2) | n(3) | AAD[d] (REH theory) | AAD (Eqs.7 and 9 with X = 2v) |
|---|---|---|---|---|---|---|---|---|
| 1 | 10[a] | 0.200 | 43.18 | 6.62 | 0.20 | 0 | 6.82 | 6.90 |
| | 60[b] | 1.200 | 22.25 | 23.17 | 4.50 | 0.08 | 27.75 | 28.11 |
| 2 | 20 | 0.400 | 37.45 | 11.81 | 0.74 | 0.00 | 12.56 | 12.69 |
| | 35 | 0.695 | 30.64 | 17.42 | 1.92 | 0.02 | 19.36 | 19.71 |
| 3 | 40 | 0.800 | 28.60 | 18.95 | 2.42 | 0.03 | 21.40 | 21.67 |
| | 38 | 0.760 | 29.36 | 18.39 | 2.23 | 0.02 | 20.64 | 20.91 |

[a] This is the total number of mutations randomly distributed over 50 codon sites as given by Nei and Tateno
[b] This is the total number of mutations that would be required to obtain the average $\mu_2$ values shown in the second row of each simulation
[c] n(i) is the number of amino acid replacements of the minimal 0-, 1-, 2-, and 3-base type expected for the given value of 2v
[d] AAD = amino acid differences

2v = 60, n (1) $\sim$ 23). There are three possibilities: 1) Nei and Tateno made an error in their simulation; 2) they made a calculational error in estimating $\mu_2$ and $T_2$; or 3) the data they averaged to obtain $\mu_2$ or $T_2$ were not representative of the total population of simulation results.

It is the last possibility that is correct. In this respect the first simulation in Nei and Tateno's (1978) Table 1 for PAM = 36 also makes an atypical point (see Table 2, this paper, particularly footnote *h*), for of 1,596 comparisons between homologous pairs of real (as opposed to simulated) cytochrome *c* divergences, only 36 had such excessive divergences, that is less that 3% of the data (Moore et al. 1976). It is possible to confirm this assessment from Table 1 in the rebuttal of Nei and Tateno which follows this letter. In that Table, Nei and Tateno have averaged one set of numbers, the n (i), over all 50 replications: they obtain for n (0), n (1), n (2), and n (3), respectively, the values 43.0, 6.7, 0.3, and 0. REH theory predicted (Table 3, this letter) 43, 6.6, 0.2, and 0. It is worth emphasizing that the latter values were predicted *before* the simulation values were known to us. The agreement between the simulation and REH theory is good and shows 1) that Nei and Tateno's simulations are correct, and 2) that the simulation studies are in agreement with REH theory. In averaging the $\mu_2$ values, however, to obtain the average 1.36, Nei and Tateno did not use all the data, but only selected sets of the simulation. The two $\mu_2$ values Nei and Tateno report in their Table 1 (this issue) are for simulated sequence pairs with 8 amino acid replacements of the minimal 1-base type: n (1) = 8. It is obvious from the average $<$ n (1) $>$ = 6.7, of n (1) over all 50 replications in Table 1, that Nei and Tateno excluded most of the relevant data for which n (1) $<$ 6.7. Nei and Tateno state in footnote *b* of their Table 1 (this issue) that they averaged over all cases in which the J/H method could be used. This is not quite the case. They excluded those simulations (which comprise the majority of the 50 replications because of the small number (2v = 10) of fixed mutations) for which n (1) is finite and n (2) is zero. In such cases REH is estimated (Holmquist 1978 b) as 1.5 MBD (minimum base differences). Then $\hat{\mu}_2 = 1.5\,\text{MBD}/T_2$. When n (1) or n (2) is zero, there is insufficient information in the sequence data to calculate $T_2$ accurately, but clearly it must lie between the number of amino acid differences (AAD) and the total number T of codons in the aligned sequence pair. Thus a lower window for $\hat{\mu}_2$ is

$$\frac{1.5\,\text{MBD}}{T} \leqslant \hat{\mu}_2 \leqslant \frac{1.5\,\text{MBD}}{\text{AAD}} \quad . \tag{11}$$

Since in the simulations $T_2 = T = 50$, $\hat{\mu}_2 = 0.03$ MBD. It is these values of $\hat{\mu}_2$ that were excluded by Nei and Tateno and had they included these values in their average, the average $\hat{\mu}_2$ over the 50 replications would have been $\cong 0.2$ as it ought, rather than the very biased value 1.36 reported by them in their Table 1. Stated another way, had they calculated the average n (i) over the selected data set they used, they would have found REH theory predicted the value 1.36 which they found. Viewed in either manner the simulations confirm REH theory. It is not valid to compare the average of a limited set of selected simulations, not representative of the total population, with the average of the full set of population values.

*Experimental Refutation of Nei and Tateno's Criticism of REH Theory*

To confirm the above logical analysis, we independently repeated the Poisson simulation of gene divergence for

$2v = 10$, where v is the Poisson parameter used in generating daughter nucleic acid sequences from their common ancestral gene. We find that from the full set of 50 replications, Nei and Tateno (1978) reported only that minority of five replications that gave a high value of the fixation intensity $\mu_2$.

In Tables 4 and 5 we report the data for *all* fifty replications; both for the protein daughter sequence pairs (Table 4) generated as well as for the mRNA pairs (Table 5) which code for those proteins. A useful summary of these data is in Table 6. The individual REH values, the mean REH value, and the variance of the population of REH values are all in sensible agreement with both theoretical expectation (footnote *a* in Table 6) and the actual values found in the simulations. A discussion of each of these tables follows.

As an average of five total base replacements separate each of the two daughter sequences from their common parent ancestral sequence, the expected number of hits separating the two daughter sequences is $2v = 2 \times 5 = 10$, where v is the Poisson parameter for the distribution from which the actual number of hits were drawn. As for a Poisson distribution, the variance is equal to the mean, the expected standard deviation is $\sqrt{2v} = 3.2$. In Table 4, under the column Actual Hits, the observed mean and standard deviation for our 50 replications were 10.2 and 3.2. From the number of amino acid differences of the minimal 0-, 1-, 2-, and 3-base type separating daughter sequences (this information is given for each replication in columns 2 through 5 of Table 4), the REH estimate, calculated from Steps 1 through 8 in Holmquist (1978 b), of the total hits separating the daughter sequence pairs is given in the last column of the Table. Both the individual REH values and the mean REH value are in good agreement with the actual number of base replacements known to have occurred. The standard deviation of the population of REH values, 4.6, is but slightly larger than the true standard deviation of 3.2. *Nei and Tateno's (1978) claim for a grossly inflated variance for the REH values is not supported by the data.*

Turning to the estimated average total hits per varion, the fixation intensity $\mu_2$, from the sixth column of Table 4, is overestimated in 10 of the 50 replications. These are labeled by asterisks. Nei and Tateno (1978) reported a total of five out of fifty such overestimates with a mean of 1.2. The mean for the ten overestimates we find is 0.84. Both values are larger than the expected value $\mu_2 = 10/50 = 0.2$. For the remaining 40 replications in Table 4 the value of $\mu_2$ given by REH theory (column 7 of Table 4) is in reasonable agreement with the actual value (= Actual Hits/50) both for the individual replications and on the average. To illustrate, for replications 2, 17, 24, 39, and 43 the estimated/actual values of $\mu_2$ were, respectively: 0.12/0.18; 0.36/0.28; 0.27/0.18; 0.21/0.24; and 0.21/0.18. In some cases $\mu_2$ is higher than the true mean, in others less, the average value for these 40 replications being 0.19 with a standard deviation of 0.07 *vs.* an expected mean and standard deviation of 0.2 and 0.06. For all 50 replications, there were 21 $\mu_2$ values less than or equal to the true value, and 29 greater than the true values. These facts show conclusively that the few values reported by Nei and Tateno, though correct, were not representative of the performance of the REH method.[2]

In Table 5, a similar analysis was performed on the mRNA sequences which coded for the proteins from which Table 4 was constructed. The actual number of hits and their variance are in excellent agreement with the REH values. For those replications with asterisks, $\mu_2$ averages 0.38. Thus, for mRNA data, the magnitude of the overestimation is much less than for protein sequence data. The information content of the experimental data from which the estimation is made is an important limiting factor in the accuracy achievable. In the present case 72% (= [0.84–0.38]/[0.84–0.2]) of the overestimation is due to this limitation and not to deficiencies in the REH method. The remaining 28% of the overestimation is due to irreducible sampling variations that at low fixation rates can cause the point estimator $\hat{r}$ of r to exceed r.

Table 6 concisely summarizes the simulation data for both proteins and nucleic acids. The agreement between REH theory and experiment is obvious.

Another claim by Nei and Tateno (1978) is that mean value of $\mu_2$ given by REH theory for $2v = 20$ is smaller than that for $2v = 10$. The claim is incorrect because expected values of $\mu_2$ are a strictly monotonic increasing function of the ratio $r = [n(2) + n(3)]/n(1)$, as Fig. A1 in Jukes and Holmquist (1972) unambiguously demonstrates. From Table 3 one notes that for $2v = 10$, and $2v = 20$, respectively, n (2) has the expected values 0.20

[2] In those cases where the $\mu_2$ values are greater than the mean, the reason is simple. For $2v/50 = 0.2$, the value of $[\langle n(2)\rangle + \langle n(3)\rangle]/\langle n(1)\rangle$, the angle brackets denoting expectation values, is 0.03453, and the expected number of amino acid differences for 50 codons is from Table A5 in Holmquist, Cantor and Jukes' 1972 paper 6.98. The mean number of amino acid differences for the 50 replications of Table 4 of this paper is 6.92 ± 2.79 (S.D.). The predictions of REH theory are thus in near perfect agreement with the simulation data. The point of this calculation is however, that when the expected number of amino acid differences is about 7 and there is a minimal 2-base type amino acid replacement, the point estimator $[n(2) + n(3)]/n(1)$ will have a value in the vicinity of $1/6 \cong 0.16$, as observed. Since the value 0.16 is greater than 0.03453, the higher $\mu_2$ values are explained. For a real gene divergence, we may have no easy way to determine that the higher values are incorrect. To see this, we need only note that for the ten replications with asterisks in Table 4, the average fixation intensity is $\mu_2 = 0.84$. This value of the fixation intensity would give rise to the same observable pattern of replacements as in Table 4 (columns 2 through 5) provided that natural selection had restricted the number of codons able to fix mutations to 16 rather than the 50 assumed in the simulation. In such a case, the higher value would be correct, even though it is an outlier for the simulation

**Table 4.** Simulated Protein Sequence Data. Independent replications of the computer simulation reported by Nei and Tateno (1978). The replications are for 2v = 10. Comparision of actual number of base replacements with the number calculated from REH theory for equiprobable genetic events. Calculated values of the fixation intensity $\mu_2$ are also given

| Replication | n(0)[a] | n(1) | n(2) | n(3) | $\mu_2$[b] | $\mu_2$ min[c] | $\mu_2$ max[d] | Actual hits | REH |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 43 | 7 | 0 | 0 | – | 0.21 | 1.50 | 9 | 10.5 |
| 2 | 46 | 4 | 0 | 0 | – | 0.12 | 1.50 | 9 | 6.0 |
| 3* | 43 | 6 | 1 | 0 | 0.88 | 0.24 | 1.50 | 9 | 13.5 |
| 4* | 38 | 11 | 1 | 0 | 0.53 | 0.39 | 1.50 | 15 | 20.1 |
| 5 | 44 | 6 | 0 | 0 | – | 0.18 | 1.50 | 10 | 9.0 |
| 6 | 45 | 5 | 0 | 0 | – | 0.15 | 1.50 | 7 | 7.5 |
| 7 | 45 | 5 | 0 | 0 | – | 0.15 | 1.50 | 13 | 7.5 |
| 8* | 41 | 8 | 1 | 0 | 0.69 | 0.30 | 1.50 | 13 | 16.1 |
| 9* | 41 | 9 | 0 | 0 | – | 0.27 | 1.50 | 10 | 13.5 |
| 10 | 42 | 8 | 0 | 0 | – | 0.24 | 1.50 | 11 | 12.0 |
| 11 | 45 | 5 | 0 | 0 | – | 0.15 | 1.50 | 7 | 7.5 |
| 12 | 45 | 5 | 0 | 0 | – | 0.15 | 1.50 | 6 | 7.5 |
| 13 | 47 | 3 | 0 | 0 | – | 0.09 | 1.50 | 7 | 4.5 |
| 14* | 47 | 2 | 1 | 0 | 1.98 | 0.12 | 1.50 | 6 | 8.4 |
| 15* | 42 | 7 | 1 | 0 | 0.78 | 0.27 | 1.50 | 12 | 14.8 |
| 16 | 41 | 9 | 0 | 0 | – | 0.27 | 1.50 | 14 | 13.5 |
| 17 | 38 | 12 | 0 | 0 | – | 0.36 | 1.50 | 14 | 18.0 |
| 18 | 46 | 4 | 0 | 0 | – | 0.12 | 1.50 | 7 | 6.0 |
| 19 | 42 | 8 | 0 | 0 | – | 0.24 | 1.50 | 10 | 12.0 |
| 20 | 46 | 4 | 0 | 0 | – | 0.12 | 1.50 | 9 | 6.0 |
| 21 | 43 | 7 | 0 | 0 | – | 0.21 | 1.50 | 8 | 10.5 |
| 22 | 44 | 6 | 0 | 0 | – | 0.18 | 1.50 | 10 | 9.0 |
| 23 | 42 | 8 | 0 | 0 | – | 0.24 | 1.50 | 13 | 12.0 |
| 24 | 41 | 9 | 0 | 0 | – | 0.27 | 1.50 | 9 | 13.5 |
| 25 | 44 | 6 | 0 | 0 | – | 0.18 | 1.50 | 8 | 9.0 |
| 26* | 44 | 5 | 1 | 0 | 1.02 | 0.21 | 1.50 | 11 | 12.2 |
| 27 | 37 | 13 | 0 | 0 | – | 0.39 | 1.50 | 16 | 19.5 |
| 28 | 46 | 4 | 0 | 0 | – | 0.12 | 1.50 | 7 | 6.0 |
| 29 | 46 | 4 | 0 | 0 | – | 0.12 | 1.50 | 8 | 6.0 |
| 30 | 43 | 7 | 0 | 0 | – | 0.21 | 1.50 | 11 | 10.5 |
| 31 | 45 | 5 | 0 | 0 | – | 0.15 | 1.50 | 5 | 7.5 |
| 32* | 40 | 9 | 1 | 0 | 0.63 | 0.33 | 1.50 | 12 | 17.5 |
| 33 | 44 | 6 | 0 | 0 | – | 0.18 | 1.50 | 8 | 9.0 |
| 34* | 39 | 10 | 1 | 0 | 0.57 | 0.36 | 1.50 | 16 | 18.8 |
| 35 | 39 | 11 | 0 | 0 | – | 0.33 | 1.50 | 14 | 16.5 |
| 36 | 42 | 8 | 0 | 0 | – | 0.24 | 1.50 | 13 | 12.0 |
| 37 | 42 | 8 | 0 | 0 | – | 0.24 | 1.50 | 13 | 12.0 |
| 38 | 46 | 4 | 0 | 0 | – | 0.12 | 1.50 | 8 | 6.0 |
| 39 | 43 | 7 | 0 | 0 | – | 0.21 | 1.50 | 12 | 10.5 |
| 40* | 43 | 6 | 1 | 0 | 0.88 | 0.24 | 1.50 | 9 | 13.5 |
| 41 | 45 | 5 | 0 | 0 | – | 0.15 | 1.50 | 11 | 7.5 |
| 42 | 42 | 8 | 0 | 0 | – | 0.24 | 1.50 | 15 | 12.0 |
| 43 | 43 | 7 | 0 | 0 | – | 0.21 | 1.50 | 9 | 10.5 |
| 44 | 44 | 6 | 0 | 0 | – | 0.18 | 1.50 | 9 | 9.0 |
| 45 | 43 | 7 | 0 | 0 | – | 0.21 | 1.50 | 12 | 10.5 |
| 46 | 48 | 2 | 0 | 0 | – | 0.06 | 1.50 | 3 | 3.0 |
| 47 | 48 | 2 | 0 | 0 | – | 0.06 | 1.50 | 5 | 3.0 |
| 48 | 44 | 6 | 0 | 0 | – | 0.18 | 1.50 | 7 | 9.0 |
| 49 | 42 | 8 | 0 | 0 | – | 0.24 | 1.50 | 10 | 12.0 |
| 50* | 35 | 14 | 1 | 0 | 0.42 | 0.48 | 1.50 | 18 | 24.2 |
| Total | | 336 | 10 | 0 | | | | 508 | 546.1 |
| Average | 43.08 | 6.72 | 0.25 | 0 | 0.84 | 0.21 | 1.50 | 10.2 | 10.9 |
| S.D. | 2.79 | 2.68 | 0.44 | – | 0.14 | 0.09 | 0.00 | 3.2 | 4.6 |

[a] n(i) is the observed number of amino acid differences of the minimal i-base type between the two daughter sequences

[b] Equation 15 in Holmquist (1978b)

[c] Equation 11, this paper

[d] For the replications with asterisk (*), REH = $\mu_2 T_2$. For the other replications the minimal value of REH is given: $REH_{min}$ = N(1) + 2(N)2 + 3(N)3

**Table 5.** Simulated mRNA Sequence Data. Independent replications of the computer simulation reported by Nei and Tateno (1978). The replications are for 2v = 10. Comparison of actual number of base replacements with the number calculated from REH theory. Calculated values of the fixation intensity $\mu_2$ are also given. The mRNA sequences from which this data came were those which coded for the protein sequences of Table 4

| Replication | N(0)[a] | N(1) | N(2) | N(3) | $\mu_2$[b] | $\mu_2$[c] | $T_2$[d] | Actual hits | REH[e] |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 41 | 9 | 0 | 0 | – | 0.18 | – | 9 | 9 |
| 2 | 43 | 7 | 0 | 0 | – | 0.14 | – | 9 | 7 |
| 3* | 43 | 5 | 2 | 0 | 0.82 | 0.18 | 13.2 | 9 | 10.9 |
| 4* | 36 | 13 | 1 | 0 | 0.20 | 0.30 | 79.0 | 15 | 16.0 |
| 5 | 40 | 10 | 0 | 0 | – | 0.20 | – | 10 | 10 |
| 6 | 43 | 7 | 0 | 0 | – | 0.14 | – | 7 | 7 |
| 7 | 39 | 11 | 0 | 0 | – | 0.22 | – | 13 | 11 |
| 8* | 39 | 9 | 2 | 0 | 0.53 | 0.26 | 27.6 | 13 | 14.7 |
| 9 | 40 | 10 | 0 | 0 | – | 0.20 | – | 10 | 10 |
| 10* | 40 | 9 | 1 | 0 | 0.28 | 0.22 | 41.9 | 11 | 11.8 |
| 11 | 43 | 7 | 0 | 0 | – | 0.14 | – | 7 | 7 |
| 12 | 44 | 6 | 0 | 0 | – | 0.12 | – | 6 | 6 |
| 13 | 43 | 7 | 0 | 0 | – | 0.14 | – | 7 | 7 |
| 14* | 45 | 4 | 1 | 0 | 0.53 | 0.12 | 12.6 | 6 | 6.7 |
| 15* | 39 | 10 | 1 | 0 | 0.25 | 0.24 | 49.3 | 12 | 12.7 |
| 16* | 38 | 11 | 1 | 0 | 0.24 | 0.26 | 58.1 | 14 | 13.7 |
| 17 | 37 | 13 | 0 | 0 | – | 0.26 | – | 14 | 13 |
| 18 | 43 | – | 0 | 0 | – | 0.14 | – | 7 | 7 |
| 19 | 40 | 10 | 0 | 0 | – | 0.20 | – | 10 | 10 |
| 20 | 45 | 5 | 0 | 0 | – | 0.10 | – | 9 | 5 |
| 21 | 42 | 8 | 0 | 0 | – | 0.16 | – | 8 | 8 |
| 22* | 41 | 8 | 1 | 0 | 0.31 | 0.20 | 34.9 | 10 | 10.8 |
| 23 | 39 | 11 | 0 | 0 | – | 0.22 | – | 13 | 11 |
| 24 | 41 | 9 | 0 | 0 | – | 0.18 | – | 9 | 9 |
| 25 | 44 | 6 | 0 | 0 | – | 0.12 | – | 8 | 6 |
| 26* | 40 | 9 | 1 | 0 | 0.28 | 0.22 | 41.9 | 11 | 11.8 |
| 27* | 36 | 12 | 2 | 0 | 0.42 | 0.32 | 42.4 | 16 | 17.7 |
| 28 | 43 | 7 | 0 | 0 | – | 0.14 | – | 7 | 7 |
| 29 | 42 | 8 | 0 | 0 | – | 0.16 | – | 8 | 8 |
| 30* | 40 | 9 | 1 | 0 | 0.28 | 0.22 | 41.9 | 11 | 11.8 |
| 31 | 45 | 5 | 0 | 0 | – | 0.10 | – | 5 | 5 |
| 32* | 40 | 8 | 2 | 0 | 0.58 | 0.24 | 23.8 | 12 | 13.8 |
| 33 | 42 | 8 | 0 | 0 | – | 0.16 | – | 8 | 8 |
| 34* | 37 | 11 | 2 | 0 | 0.45 | 0.30 | 36.9 | 16 | 16.7 |
| 35* | 37 | 12 | 1 | 0 | 0.22 | 0.28 | 68.1 | 14 | 14.9 |
| 36 | 38 | 12 | 0 | 0 | – | 0.24 | – | 13 | 12 |
| 37 | 39 | 11 | 0 | 0 | – | 0.22 | – | 13 | 11 |
| 38 | 46 | 4 | 0 | 0 | – | 0.08 | – | 8 | 4 |
| 39 | 38 | 12 | 0 | 0 | – | 0.24 | – | 12 | 12 |
| 40* | 43 | 6 | 1 | 0 | 0.39 | 0.16 | 22.5 | 9 | 8.8 |
| 41* | 42 | 7 | 1 | 0 | 0.34 | 0.18 | 28.5 | 11 | 9.8 |
| 42* | 38 | 11 | 1 | 0 | 0.24 | 0.26 | 58.1 | 15 | 13.7 |
| 43 | 41 | 9 | 0 | 0 | – | 0.18 | – | 9 | 9 |
| 44 | 43 | 7 | 0 | 0 | – | 0.14 | – | 9 | 7 |
| 45* | 40 | 8 | 2 | 0 | 0.58 | 0.24 | 23.8 | 12 | 13.8 |
| 46 | 47 | 3 | 0 | 0 | – | 0.06 | – | 3 | 3 |
| 47 | 45 | 5 | 0 | 0 | – | 0.10 | – | 5 | 5 |
| 48 | 43 | 7 | 0 | 0 | – | 0.14 | – | 7 | 7 |
| 49 | 41 | 9 | 0 | 0 | – | 0.18 | – | 10 | 9 |
| 50* | 34 | 15 | 1 | 0 | 0.18 | 0.34 | 103.2 | 18 | 18.3 |
| Total | | 427 | 25 | 0 | | | | 508 | 498.4 |
| Average | 40.96 | 8.54 | 0.50 | 0 | 0.38 | 0.19 | 42.5 | 10.2 | 10.0 |
| S.D. | 2.85 | 2.64 | 0.71 | – | 0.17 | 0.06 | 23.0 | 3.2 | 3.6 |

[a] N(i) is the observed number of codons with base differences between the two daughter mRNAs

[b] Calculated from Table 1 for nucleic acids in Holmquist, Cantor, and Jukes (1972) following the procedure used to construct Table A5 in that reference and the methodology in Holmquist (1978b). See Holmquist (1980)

[c] $\mu_{2min}$ = [N(1) + 2N(2) + 3N(3)]/50

[d] See *b*. For $T_2$ values not given, $T_{2min}$ = N(1) + N(2) + N(3)

[e] For the replications with asterisk (*), REH = $\mu_2 T_2$. For the other replications $REH_{min}$ = N(1) + 2N(2) + 3N(3)

Table 6. Comparison of the expected[a] mean number $\mu$ of total base replacements and the expected[a] population standard deviation $\sigma$ with the corresponding sample quantities $\hat{\mu}$, $\hat{\sigma}$ from the actual simulation and as calculated from REH theory using protein and mRNA sequence data[b] as the experimental basis

| Sample size N[c] | Parameter estimated | Actual | REH |
|---|---|---|---|
| | From protein data | | |
| 50[d] | $\hat{\mu}$ | 10.2 | 10.9 |
| | $\hat{\sigma}$ | 3.2 | 4.6 |
| 40[e] | $\hat{\mu}$ | 9.7 | 9.7 |
| | $\hat{\sigma}$ | 3.0 | 3.7 |
| 10[f] | $\hat{\mu}$ | 12.1 | 15.9 |
| | $\hat{\sigma}$ | 3.6 | 4.5 |
| | From mRNA data | | |
| 50[d] | $\hat{\mu}$ | 10.2 | 10.0 |
| | $\hat{\sigma}$ | 3.2 | 3.6 |
| 31[e] | $\hat{\mu}$ | 8.8 | 8.1 |
| | $\hat{\sigma}$ | 2.6 | 2.5 |
| 19[f] | $\hat{\mu}$ | 12.4 | 13.1 |
| | $\hat{\sigma}$ | 2.9 | 3.0 |

[a]For the Poisson simulation with 2v = 10, the expected mean is $\mu = 10$, and the expected standard deviation is $\sigma = \sqrt{10} = 3.16$. For multiple large samples of N replications each only 1% of the samples would have means $\hat{\mu}$ in excess of $\mu + 2.326\sigma/\sqrt{N} = 11.04$, and only 1% would have standard deviations $\hat{\sigma}$ in excess of $\sigma + 2.326\sigma/\sqrt{(2N)} = 3.90$, if N = 50

[b]The values averaged for the protein sequence data are given in Table 4, and those for the corresponding mRNA sequence data are given in Table 5

[c]Number of replications in sample

[d]All 50 replications included

[e]REH calculated from REH = 1.5 × MBD (minimum base differences)

[f]REH calculated as in Steps 1 through 6 in Holmquist (1978b)

and 0.74, while n (1) has the expected values 6.62 and 11.81. The two ratios r are 0.030 and 0.063, respectively. These are sufficiently close to each other that considering the approximate expected errors in the two n(1) values of about $\sqrt{6.62} = 2.6$ and $\sqrt{11.81} = 3.4$, and in the two n(2) values of $\sqrt{0.20} = 0.45$ and $\sqrt{0.74} = 0.86$, an inversion of the magnitudes of the $\mu_2$ values for particular simulations of 2v = 10 and 2v = 20 might occur fairly often. The possibility of inverted $\mu_2$ values for protein data are most likely when the ratio r is near zero or unity, that is for $\mu_2 \sim 0$, or $\mu_2 \sim 6$. The vast majority of data for real protein sequences does not fall anywhere near these extremes (see Table 5 in Moore et al. 1976; and Tables 6, 8, and 10 in Holmquist et al. 1976). In fact for cytochrome *c*, α-hemoglobin, β-hemoglobin, myoglobin, and parvalbumin, the average $\mu_2$ and its population standard deviations are, respectively, 3.41 ± 1.39, N = 1,489; 2.01 ± 1.33, N = 1,081; 2.08 ± 0.95, N = 1,535; 0.88 ± 0.42, N = 173; and 3.11 ± 1.00, N = 76.

## Correlated Behavior Between the Fixation Intensity and the Number of Varions

In connection with the matters in the preceding section Nei and Tateno (1978, p 339) state "As theoretically expected there was a strong (but not perfect) negative correlation between $\hat{\mu}_2$ and $\hat{T}_2$." Just prior to their *Discussion* they report "In our earlier simulation on random nucleotide substitution we noted that the estimate of $\hat{\mu}_2$ for a shorter period of evolutionary time is often larger than for a longer period of time. The same pattern was observed in the present simulation. Namely, the average of $\hat{\mu}_2$ was 3.5 for PAM = 20 and 2.8 for PAM = 36, though the difference was not statistically significant. Thus, the larger value of REHC for PAM = 36 than that for PAM = 20 is caused by the increase in the estimate of the proportion of variable codons. The correlation between $\hat{\mu}_2$ and $\hat{T}_2$ was -0.475 for PAM = 36 and -0.536 for PAM = 20." They continue this line of reasoning in their *Discussion*: "We have seen that the increase in REHC with increasing evolutionary time is often caused by the increase in $\hat{T}_2$, while $\hat{\mu}_2$ stays more or less the same for a long period of time. This pattern is also observed in real data, as seen from Tables 4 and 5 in the J/H paper. Our simulation indicates that this is at least partially caused by a statistical artifact inherent in the J/H method, since it occurs even under random substitution. However, if a large part of this pattern is real, then the J/H model must be modified drastically. Namely, instead of fixing $T_2$ and allowing $\mu_2$ to increase with time as in the J/H model, we must fix $\mu_2$ and allow $T_2$ to increase. It is difficult for us to visualize how such an evolutionary change of protein occurs, unless it is assumed that once a codon site experiences a substitution, it becomes 'immune' from another substitution."

Nei and Tateno's numerical calculations are correct; their conclusions are not.

A negative correlation between $\mu_2$ and $T_2$ is *not* theoretically expected: In Equation 5, the denominator, 1–P (0), is the expected proportion of amino acid differences among the variable sites and is thus a monotonically increasing function of $\mu_2$. Therefore *if* the numerator in Equation 5, AAD, the total number of amino acid differences between the two homologous chains, remains constant, $T_2$ will decrease as $\mu_2$ increases. But the number of amino acid differences between the two chains need not remain constant. It can increase or decrease, independently of $\mu_2$, as well, the particular behavior actually observed being a matter of the end result of natural selection on protein and gene structure. Thus $T_2$ can either decrease, increase, or remain the same with a change in $\mu_2$. Nei and Tateno refer to the data of Jukes and Holmquist (1972) to support a negative correlation between $\mu_2$ and $T_2$: the Jukes and Holmquist article contains 33 $\{\mu_2, T_2\}$ points, mostly for cytochrome *c* and globin chains. Since then much

more extensive compilations have been published: in particular we have published in Table 5 of Moore et al. (1976) 52 $\{\mu_2, T_2\}$points averaged from 655 individual pairs of cytochrome *c* sequences. This is a data base 60 times larger than in the J/H paper. A variety of behaviors is observed (Table 7): $T_2$ may remain constant with an increase or decrease in $\mu_2$ as among the mammalian divergences compared to the amphibian/teleost divergences; $\mu_2$ may remain constant with an increase or decrease in $T_2$ as among the mammalian/insect divergence compared to the lower-fish/Euglena-Crithidia divergences; both $\mu_2$ and $T_2$ may increase or decrease as among the mammalian divergences compared to the insect/plant divergence; or $\mu_2$ may increase (decrease) while $T_2$ decreases (increases) as among the lower-fish divergences compared to the amphibian/teleost divergence. Many further examples are found in the globin family of genes where 486 $\{\mu_2, T_2\}$points are tabulated (Holmquist et al. 1976). The globin data base there is 32 times larger than that considered by Nei and Tateno. In both papers it is emphasized that evolutionary rates are nonconstant and that "These rate differentials are resolved into two components (a) due to change in the number $[T_2]$ of codon sites free to fix mutations during the period of divergence of the species involved; (b) due to change in fixation intensity $[\mu_2]$ at each site. These two components also show [individually] nonuniformity along different lineages. Positive Darwinian natural selection can bring about an increase in either component, and negative or stabilizing selection in protein evolution can lead to decreases."

Table 7. Types of observed behaviors for the fixation intensity and number of varions[a]

| Type | Divergence | $\hat{\mu}_2$ | $\hat{T}_2/T$ | REHC |
|---|---|---|---|---|
| $T_2$ constant | mammals/mammals | 1.8 | 15 | 27 |
| $\mu_2$ increase | amphibians/teleosts | 4.9 | 15 | 74 |
| $\mu_2$ constant | mammals/insects | 3.6 | 25 | 90 |
| $T_2$ increase | lower fish/ Euglena-Critidia | 3.6 | 60 | 216 |
| $\mu_2$ increase | mammals/mammals | 1.8 | 15 | 27 |
| $T_2$ increase | plants/insects | 3.1 | 47 | 146 |
| $\mu_2$ increase | fungi/fungi | 3.7 | 36 | 133 |
| $T_2$ decrease | amphibians/insects | 5.6 | 26 | 146 |

[a]From Table 5, Moore et al. 1976

Many behaviors of $\mu_2$ relative to $T_2$ are thus observed in real data, and the restricted type of negative correlation reported by Nei and Tateno is due both to a lack of biological realism in their simulations as well as to the very few number of simulations and examples of real data they reported, not to a statistical artifact in the REH method.

These comments are not meant to imply that no correlation exists between $\mu_2$ and $T_2$ within a single family of given biological function. If the product $\mu_2 T_2$ = REH is plotted *vs* $T_2$ for the cytochrome *c* data, one obtains an approximately linear plot for which

$$\text{REH} = 4.2\, T_2 \quad . \tag{12}$$

The correlation coefficient was r = 0.89. The deviation of individual points from the line for a particular point $(\mu_2, T_2)$ can be large, the standard error of the estimate being 33. This would be anticipated from the wide range of behaviors observed in the data. From Eq. 12, on the average $\mu_2 \sim 4.2$ for cytochrome *c*. It needs emphasis that "on the average" is not the same as Nei and Tateno's statement that "$\hat{\mu}_2$ stays more or less the same for a long period of time." True constancy of a value should not be confused with its average, particularly when individual values, for reasons of natural selection, can vary from the expected value.

Equation 12 has a straightforward interpretation: on the average, once the cytochrome *c* gene has sustained between 4 and 5 fixed mutations, a new site, elsewhere in the molecule, previously not variable, becomes variable. This is the exact opposite of Nei and Tateno's statement "that once a codon site experiences substitution it becomes 'immune' from another substitution". The observed behavior summarized in Equation 12 is in complete accord with the covarion hypothesis of Fitch, and in fact, serves to illustrate it. Therefore, Nei and Tateno's statement that the Jukes/Holmquist model requires $T_2$ to remain fixed while $\mu_2$ must increase with time is not correct (Nei and Tateno 1978, p 345).

This section can be summarized as follows: For a given pair of sequences if $\mu_2$ is over (under) estimated, $T_2$ will be under (over) estimated. Nei and Tateno and we agree on this point. However, for two different pairs of sequences the correlation between the two $\mu_2$ and the two $T_2$ need not be negative as Table 7 in this letter illustrates. There is no contradiction between the last two sentences.

## Solution to a Paradox

Some time ago Zuckerkandl (1976) noted a paradox posed by the high turnover rate (0.75) reported by Fitch (1971) for the covarions in cytochrome *c*. That is, the probability of a covarion losing its variable status after the fixation of a single mutation elsewhere in the cytochrome *c* gene is 0.75. Zuckerkandl distinguished between two kinds of amino acid sites: *specific function* sites which must be occupied by a certain amino acid residue "with little or no leeway as to the residue and

very little or none as to the site", and *general function* sites in which any among several amino acid types will serve. The active site of an enzyme is an example of a specific function site, and those sites relating to solubility, for example, are general function sites. A particular locus may play a specific functional role as well as several functional roles.

The paradox noted by Zuckerkandl is that general function sites, for the time they remain so, should be permanent covarions. This contradicts the high turnover rate found by Fitch.

Now when a covarion in the cytochrome *c* gene loses its variable status, Fitch's turnover rate would indicate that 1/0.75 or 1.3 mutations had been fixed somewhere in the gene. But from Equation 11, the estimate 1.3 is too low. If it is assumed that when a new covarion appears in the gene, an old one drops out (this is necessary to keep the size of the covarion set roughly constant), the proper turnover rate is about $1/4.2 \sim 0.24$. The actual turnover rate would thus appear to be about 1/3 that estimated by Fitch.

## Conclusions

The simulation data confirm the correctness of REH theory. A comparison of this data for proteins and for mRNAs demonstrates that amino acid sequence data is inherently not capable of providing accurate estimates of evolutionary parameters. A knowledge of the pattern of base replacements at the third nucleotide position within codons is a prerequisite for accurate assessment of the number of superimposed mutations at the three positions within codons, of the number of codon sites able to fix mutations, and of the fixation intensity $\mu_2$. This information can only come from experimentally sequenced mRNA or DNA sequences. Adequate theory for analysis of these nucleic acid sequences exists (Holmquist and Pearl 1980).

Nei and Tateno (1979) recently stated in regard to our criticism of their work that they "have presented all the data that are necessary for drawing an objective conclusion from our work. We do not think redundant and uninformative data has any scientific merit... Holmquist should repeat our simulation himself and check the validity of our data before he criticizes our work." It is clear from Table 4 and 5 of this letter that Nei and Tateno did not present the data necessary for drawing an objective conclusion about the study. The data they label redundant and uninformative is central to the issues they discuss. The data they reported is not representative of the process they described.

*Acknowledgments:* This work was supported by grant PCM76–18627 from the National Science Foundation and NASA grant "The Chemistry of Living Systems" NGR 05–003–460.

## References

Czelusniak J, Goodman M, Moore GW (1978) J Mol Evol 11: 67–74
Dayhoff MO, Eck RV, Park CM (1972) Atlas of protein sequence and structure 5:89–99
Dayhoff MO, McLaughlin PJ (1972) Atlas of protein sequence and structure 5:111–118
Dickerson R (1971) J Mol Evol 1:26–45
Fitch W, Markowitz E (1970) Biochem Genet 4:579–593
Fitch W (1971) J Mol Evol 1:84–96
Fitch W (1972) Hematologie und Bluttransfusion 10:177–215
Fitch W (1976) J Mol Evol 8:13–40
Hewett-Emmett D, Cook CN, Barnicot NA (1976) Old world monkey hemoglobins: Deciphering phylogeny from complex patterns of molecular evolution. In: Goodman M, Tashian R, Tashian J (eds) Molecular anthropology. John Wiley, New York, p 257
Holmquist R (1972 a) J Mol Evol 1:211–222
Holmquist R (1972 b) J Mol Evol 1:115–133
Holmquist R (1973) J Mol Evol 2:145–148
Holmquist R (1976 a) J Mol Evol 8:337–349
Holmquist R (1976 b) Random and nonrandom processes in the molecular evolution of higher organisms. In: Goodman M, Tashian RE, Tashian JH (eds) Molecular anthropology. Plenum Press, New York, p 89
Holmquist R (1978 a) J Mol Evol 12:17–24
Holmquist R (1978 b) J Mol Evol 11:361–374
Holmquist R (1978 c) J Mol Evol 11:349–360
Holmquist R (1979 a) J Mol Evol 13:173–178
Holmquist R (1979 b) Science 203:1012–1014
Holmquist R (1980) J Mol Evol 15:149–159
Holmquist R, Pearl D (1980) J Mol Evol 16:211–267
Holmquist R, Jukes TH, Moise H, Goodman M, Moore GW (1976) J Mol Biol 105:39–74
Holmquist R, Cantor C, Jukes TH (1972) J Mol Biol 64:145–161
Jukes TH, Cantor C (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian Protein Metabolism III. Academic Press, New York, p 21
Jukes TH, Holmquist R (1972) J Mol Biol 64:163–179
Jukes TH, Holmquist R, Moise H (1975) Science 189:50–51
Karon J (1979) J Mol Evol 12:197–218
Kimura M (1977) Nature 267:275–276
Kimura M, Ohta T (1972) J Mol Evol 2:87–90
Moore GW, Goodman M, Callahan C, Holmquist R, Moise H (1976) J Mol Biol 105:15–37
Moore GW (1977) J Theoret Biol 66:95–106
Nei M, Chakraborty R (1976) J Mol Evol 7:313–323
Nei M, Tateno Y (1978) J Mol Evol 11:333–347
Nei M, Tateno Y (1979) J Mol Evol 13:167–171
Noguchi T (1977) A hybrid model for molecular evolution and an evolutionary clock. In: Matsubara H, Yamanaka T (eds) Evolution of Protein Molecules. The Scientific Societies Press, Tokyo, p 61
Noguchi T (1978) Origins of Life: 489–494
Ratner VA (1978) Mathematical Biology and Medicine 1:210–257
Tateno Y, Nei M (1978) J Mol Evol 11:67–73
Wilson AC, Carlson S, White TJ (1977) Ann Rev Biochem 46: 573–639
Zuckerkandl E (1976) J Mol Evol 7:167–183

Received February 1, 1980