J.Mol.Evol.8, 1-12 (1976)

Journal of Molecular Evolution

Simulation Studies on the Evolution of Amino Acid Sequences*

TOMOKO OHTA
National Institute of Genetics, Yata 1, 111, Mishima, Shizuoka-Ken, 411 Japan

Received September 24, 1975; January 13, 1976

Summary. A model of molecular evolution in which the parameter (intrinsic rate of amino acid substitution) fluctuates from time to time was investigated by simulating the process. It was found that the usual method of estimation such as Poisson fitting underestimates this variation of the parameter when remote comparisons are made. At the same time, four distance measures (minimum base difference, Poisson fitting, random nucleotide substitutions and negative binomial fitting) were tested for their accuracy. When the substitution rate is not uniform among the amino acid sites, the negative binomial fitting gives most satisfactory results, however, one needs to know the parameter beforehand in order to use this method. It was pointed out that the fluctuation of the evolutionary rate is expected if the nearly neutral but very slightly deleterious mutations play an important role on molecular evolution.

Key words: Amino Acid Substitution/Constancy of the Evolutionary Rate

INTRODUCTION

The question whether the rate of molecular evolution is constant or not is one of the most stimulating problems in recent study of evolution in conjunction with the big controversy between the "neutral" vs. "selection" hypotheses. According to Kimura (1968), the rate of amino acid substitution (k) is simply equal to the mutation rate per gamete (v) under his neutral mutation-random drift hypothesis. Here, if the "nearly neutral" mutant substitutions or, in particular, the very slightly deleterious mutations are prevalent at the molecular level, the above relationship (k = v) does not hold and the evolution becomes rapid in small populations such as the time

*Contribution No. 1087 from the National Institute of Genetics, Mishima, Shizuoka-ken, 411 Japan.

of speciation. In very large populations, the evolution stops or at least its rate slows down (Ohta, 1973, 1974). On the other hand, if Darwinian selection is the main cause of molecular evolution, one would expect parallel changes between the molecular and the phenotypic levels (Kimura, 1969). The purpose of the present study is to clarify this problem by simulating the evolution of amino acid sequences and to find out the true pattern of molecular evolution. Throughout this paper, by the term "the number of amino acid substitutions", I mean the cumulative number of amino acid substitutions that have taken place in the course of evolution and this is different from the observed amino acid differences between the two sequences.

MODEL

In this study, I shall introduce two types of non-randomness in the amino acid substitutions; over time and over amino acid sites in a sequence. We assume that the number of amino acid substitutions in a protein sequence in some period (to be called a leg) follows a Poisson distribution. This assumption holds when each amino acid site has a very small probability of substitution and there are many such sites in a sequence. If the parameter of the Poisson distribution is a constant (λ), the generating function of the number of amino acid substitution is,

$$(1) \qquad P(s) = e^{-(1-s)\lambda} \quad \text{for } |s| \leq 1$$

Consider the situation in which λ is not a constant but a random variable following certain probability distribution such as gamma function. This is considered to represent one case where the rate of molecular evolution is not strictly constant but is influenced by various environmental factors such as population size or severeness of the environment. Then λ in formula (1) follows its distribution function with mean $\bar{\lambda}$ and variance V_λ.

The mean (M) and the variance (V) of the number of amino acid substitutions may be obtained by differentiating the above generating function and taking the expectation.

$$(2) \qquad M = E\{P'(1)\} = \bar{\lambda} \qquad \text{and}$$

$$(3) \qquad V = E\{P''(1) + P'(1)\} - M^2$$
$$= V_\lambda + \bar{\lambda}$$

where E denotes the expectation with respect to λ. Thus the

variance of the number of substitutions is the sum of the mean and the variance of λ. Now, if $\bar{\lambda}$ and V_λ are constant when measured in a unit time, V is also a constant. I shall investigate the two cases: constant and variable λ. In the simulation experiment, I shall also introduce non-uniform mutation rates among the amino acid sites, since this is one of the important properties found in actual amino acid sequences.

DISTANCE MEASURES

Minimum Base Difference

I have adopted Fitch's minimum difference matrix, which he obtained directly from the code table (Fitch & Margoliash, 1967).

Poisson Fitting

Zuckerkandl & Pauling (1965), Margoliash & Smith (1965) and Kimura (1969) and others have estimated the number of multiple substitutions by fitting a Poisson distribution. If P_d is the fraction of amino acid sites at which two sequences being compared have different amino acids, the number of amino acid substitutions per site is estimated by the following formula;

$$(4) \qquad K_{aa} = -\log_e(1 - P_d)$$

The variance of this estimate becomes,

$$(5) \qquad \sigma^2_{K_{aa}} = \frac{P_d}{(1 - P_d)n_{aa}}$$

where n_{aa} is the number of amino acid sites being compared (Kimura, 1969). The assumption underlying this procedure is that all amino acid sites have an equal probability of substitution. When the amino acid sites have different probabilities of substitution, this method underestimates the true value somewhat; however, we shall show later that the bias is not large unless very remote comparisons are made.

Random Nucleotide Substitutions

Holmquist (1972a,b) developed a method to estimate the total number of base substitutions (including synonymous mutations) between the two sequences and he called it random evolutionary hits. Jukes & Cantor (1969) and Kimura & Ohta (1972) presented

the simpler method. I shall call it "random nucleotide substitutions" or RNS. Let P_d be the fraction of different amino acid sites as before and let D_E be the average number of base substitutions per codon. Then D_E is obtained from the following formula:

$$(6) \qquad D_E = -\frac{9}{4}\log_e\left(1 - \frac{4}{3}\gamma\right)$$

where γ is the fraction of nucleotide sites for which the two sequences differ from each other ($0 \leq \gamma \leq \frac{3}{4}$) and satisfy the following cubic equation,

$$(7) \qquad 1 - (1 - \gamma)^2\left(1 - \frac{1}{4}\gamma\right) = P_d$$

The variance of D_E becomes (Kimura & Ohta, 1972);

$$(8) \qquad \sigma^2_{D_E} = \frac{16P_d\,(1 - P_d)}{(1 - \gamma)^2\,(3 - \gamma)^2\left(1 - \frac{4}{3}\gamma\right)^2 n_{aa}}.$$

Underlying assumptions of this procedure are: base substitutions occur spatially at random and in uniform probability over the sequence, and at each site a given nucleotide mutates with equal probability to any one of the remaining three.

Negative Binomial Fitting

Uzzell & Corbin (1971) suggested that the negative binomial distribution gives a better estimation of the number of amino acid substitutions than the Poisson distribution, since the substitution rate is not uniform among the sites. This distribution is theoretically expected when the mutability follows the gamma distribution among the amino acid sites. The generating function of the negative binomial distribution takes the following form,

$$(9) \qquad P(s) = \left(\frac{p}{1 - qs}\right)^r \quad \text{for } |s| \leq 1$$

where $p = 1 - q$ and r are the parameters of the distribution. Here r reflects the non-randomness of mutability among the sites. For $r \to \infty$, the negative binomial distribution converges to the Poisson distribution. For smaller r, more non-randomness exists. Thus we need to know both parameters in order to fit this distribution. If the form of the distribution of the rate of substitution among the amino acid sites does not

change over time, the value of r may be estimated from the phyletically inferred number of substitution in each site. Uzzell & Corbin (1971) have estimated r ≃ 2 using cytochrome c data. Once r is known, p may be estimated from P_d (fraction of amino acid sites by which two sequences differ), and the average number of substitution per site (K_{NB}) may be obtained from the following formula:

$$(10) \qquad K_{NB} = rq/p$$

where $p = (1 - P_d)^{\frac{1}{r}}$ and $q = 1 - p$. The difficulties involved in this procedure are: (i) one needs to know r before-hand, and (ii) the variance of the above estimate is larger than that of the Poisson distribution.

There are several other distance measures, some of which are quite general and model non-specific (e.g. Grantham, 1974; Dayhoff, 1972; Beyer et al., 1974). It is impossible to examine all these distances in the present study.

MONTE CARLO EXPERIMENTS

The simulation experiment consists of the following steps: making the original sequence, mutation in the 2 descendants from the original sequence, and calculation of the distance between the 2 descendants. The original sequence was made by generating uniform random number between 0 ~ 1.0 (RAND 20 in TOSBAC 3400). First, the triplet code was made by asigning equal probability of $\frac{1}{4}$ to each of A, T, G and C. If any codon happened to be nonsense codon, it was discarded. A random sequence of a total of 100 amino acid codons was generated.

Mutation was carried out again by using random numbers. The number of mutants or the amino acid substitutions in a leg from the ancestral to the descendent sequence was determined by generating a Poisson random number (RAND 40 in TOSBAC 3400) with a specified parameter. If a synonymous mutation occurred, it was not counted but the base sequence was changed and the process was repeated. From an ancestral sequence, two descendants were made in order to compare each with the other and to calculate the distance in terms of amino acid substitutions. In one set of the experiments, the parameter of the Poisson distribution was a constant, whereas in another set, the parameter itself was a random variable following the gamma distribution, in which the mean and variance of the parameter were equal ($\bar{\lambda} = V_\lambda$). For this gamma distribution, the gamma random number (RAND 50 in TOSBAC 3400) was used.

The position of the mutation was determined again by generating a random number. Two different ways in asigning the

probability of mutation to each of 100 amino acid sites were employed. In one set, the uniform probability of mutation was asigned to each of the 100 sites by generating a uniform random number. In the other set, a gamma random number with mean one was used. Actually, if the gamma random number, $X_G = -100(\log_e X)/2$ falls in the interval i - 1 ~ i, the i-th site is chosen for mutation, where X is a uniform random number. Note that $-\log_e X$ is the gamma random number with both mean and variance one and that it is multiplied by 100 because there are 100 amino acid sites in a sequence. It is further divided by 2 for truncation. If X_G is larger than 100, it was discarded. By this asignment of mutability, I have empirically found that the parameter of the negative binomial distribution (r) may be 2, the value estimated from cytochrome c data (Uzzell & Corbin, 1971).

Thus the total of 4 sets of experiments designated I to IV were carried out as in the following table:

	Mutability of amino acid sites		
		Uniform	Gamma
Number of mutants in a leg	Poisson with constant parameter	I	II
	Poisson with variable parameter	III	IV

Each set of the experiments was conducted for 20 periods (i.e. 20 values of $\bar{\lambda}$) and for each period, 100 repeats were done, resulting in 2000 trials. In each period, the mean and the variance of each of the 4 distance measures were computed based on 100 repeats. Simultaneously, the expected variance was computed (formula 5).

For the negative binomial fitting, I put r = 2, the value which was found to fit to the cases of non-uniform mutability among the sites (II and IV). It seems impossible to estimate this parameter for each case in the present simulation, since there are only two legs to compare. When many sequences and more complicated phylogenetic tree are available, this parameter may be estimated as Uzzell & Corbin (1971) have done.

RESULTS

Figure 1 represents the mean of the estimated number of amino acid substitution between the two descendants by using the four methods: the minimum distance, Poisson fitting, random

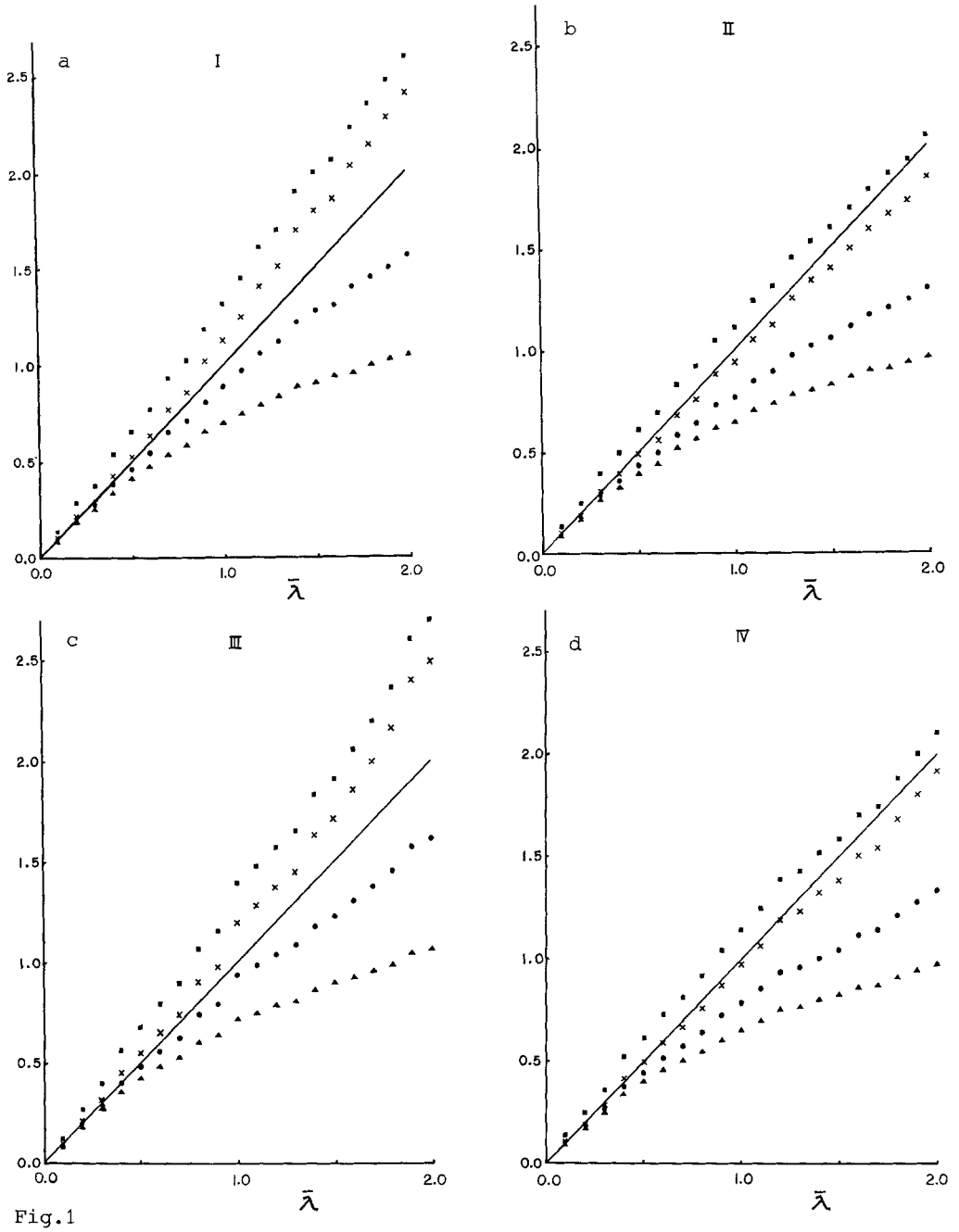

Fig.1

Results of simulation experiments I to IV. The ordinate is the average number of amino acid substitutions per site. The abscissa is the average of the parameter of the Poisson distribution, which determines the number of substitution in a leg, divided by the number of amino acid sites in a sequence ($\bar{\lambda}$). Thus, the ordinate and the abscissa have the same scale. In the figures, the straight line represents the expected actual number of substitutions; triangles, the observed minimum base difference; circles, the results of Poisson fitting; and crosses, the results of negative binomial fitting; and squares, the observed random nucleotide substitutions

nucleotide substitutions and negative binomial fitting. Figures 1a to 1d correspond to the sets I to IV respectively. The abscissa (λ) is measured by the expected number of amino acid substitutions and hence has the same scale as the ordinate. The expected actual number of substitution is shown by the straight line in the figure. In each figure, the triangles represent the results of minimum distance, the circles, those of Poisson fitting, the squares, those of random nucleotide substitutions and the crosses, those of negative binomial fitting.

In the figure, the symbols under the line indicate that they are underestimates. It is clear from the figure that, in all cases, the minimum distance (MBD) is a serious underestimate when the distance becomes large. The bias is pronounced when the substitution rate is not random among the amino acid sites (cases II and IV). For example, when MBD = 0.5, the bias is about 15% of the true value in cases II and IV but about 13% in cases I and III. When MBD = 1.0 it is the underestimation of almost 100% in the former cases.

The Poisson estimates (K_{aa}) also undervalue somewhat, particularly when the rate is non-random among the sites. For example when $K_{aa} = 0.5$, the bias is about 10% in cases II and IV but negligible in cases I and III. The bias gets nonnegligible for remote comparisons even for cases I and III. This is considered to be caused by the back mutation intrinsic to the genetic code (Farris, 1973). Actually, because of structural and functional constraints of protein molecules, the back mutation may occur much more frequently than random expectation. In practice, however, it can still be a good method of estimation unless very remote comparisons are made. In fact, Nei & Chakraborty (1976) found, using the empirical data, that the correlation is very high between the Poisson estimate and the maximum parsimony solution by Langley & Fitch (1974) and by Goodman et al. (1974).

Theoretically, random nucleotide substitutions (RNS) should be roughly 40% larger than the actual number of amino acid substitution (straight line). This relationship holds when the amino acid sites have uniform probability of substitution (cases I and III). However, when the rate is non-random among the sites (II and IV), RNS curves downwards as time gets larger. Since the linearity with respect to true distance is the most desirable property of a distance estimator, RNS is not much better than the Poisson estimate as long as there is variability in the substitution rate among the sites.

The negative binomial (K_{NB}) overestimates the number of substitution when the probability of substitution is uniform among the sites (I and III). In fact, it is clear that the model does not fit such cases. When the rate is gamma distributed (II and IV) among the sites, it fits better than any other distance

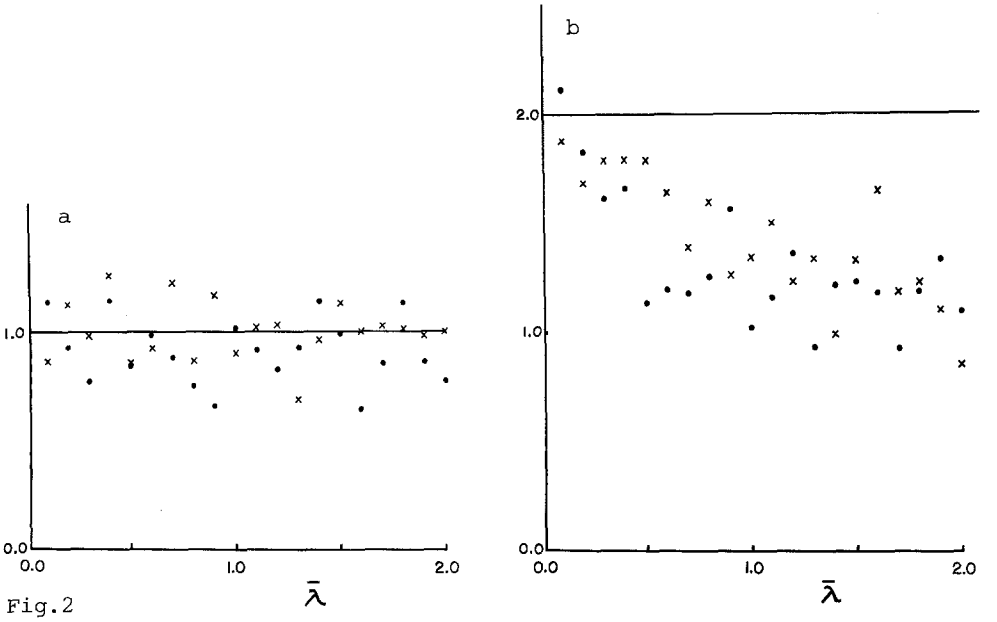

Fig.2

The variance of the rate of amino acid substitution as measured by the ratio of the observed to the expected variance (ordinate). Abscissa is the average parameter as in Figure 1. Figure 2a represents the results of the experiments I (×) and II (●), in which the parameter of the Poisson distribution is constant. Figure 2b represents the cases with a fluctuating parameter: experiments III (×) and IV (●). The straight lines in both figures are the theoretical expectation of the variance ratio

measures tried. The weakness of this measure as mentioned before, is that we need to know beforehand the parameter (r = 2 in the present case) of the distribution. Furthermore, the variance of the estimate is larger than that of the Poisson fitting.

Let us turn our attention to the variance of the number of substitutions. We examine the results of the Poisson fitting in detail, since practically the same results were obtained for RNS. Figure 2 shows the results in terms of the ratio of the observed to the expected variance. The expected variance was calculated using formula (5) from the average P_d of 100 trials. Again, the expected *actual* ratio is shown by the straight lines. Figure 2a is the cases of constant parameter of the Poisson process among the legs (I and II). Crosses represent the results for case I and circles, those of case II. The observed values roughly agree with the expected values; however, there seems to be a slight tendency of underestimation when the substitution rate is not uniform among the sites (case II).

Figure 2b shows the results for the cases when the parameter of the Poisson distribution varies among the legs. Crosses represent the results of case III (uniform substitution among the sites) and circles, those of case IV (non-uniform substitution

among the sites). The expected actual ratio becomes two from formula (3) and by putting $V_\lambda = \bar{\lambda}$. It is interesting to find that the ratio decreases rapidly as the genetic distance gets larger. In other words, the observed variance is smaller than the true value and the deviation gets larger as the distance gets larger. For example, when the expected amino acid substitution is 0.3 per site ($\bar{\lambda} = 0.3$), the ratio is 1.8 in case III and 1.6 in case IV. It further decreases to almost 1.2 in case III and 1.0 in case IV when $\bar{\lambda} = 2.0$ (largest $\bar{\lambda}$ value in the figure). A more theoretical approach to this problem is under investigation. The fact that the observed variance underestimates the true value has an important bearing in assessing the true nature of molecular evolution and will be discussed later.

DISCUSSION

From the results of our Monte Carlo experiments, we can see that the non-randomness among the sites is reflected in the mean value of the distance whereas the non-randomness among the legs is reflected in the variance of the estimates. The interaction effects between the two types of non-randomness is rather small. Also we can say that the Poisson fitting is a good method as compared with the other more complicated approach not only because of its simplicity but also because of its reliability. As for the estimation of the mean, more assumptions are needed for more sophisticated analysis, and hence the reliability decreases as long as we are not sure of such assumptions. As for estimating the variance of the evolutionary rate, the Poisson fitting is considered to be satisfactory since the interaction effects of the two types of non-randomness mentioned is rather small.

The maximum parsimonious solution investigated by several investigators (e.g. Fitch, 1971; Goodman et al., 1974), is a kind of minimum distance and therefore it may be the underestimate when the number of branches (or, branchings) is small. However, in my Monte Carlo experiments, the true distance is not more than twice the minimum distance even at the maximum λ value studied. Thus the augmentation procedure by Goodman et al. (1974) which give 3 ~ 4 times of the minimum distance in some branches of the phylogenetic tree of the globins is questionable (Nei & Chakraborty, 1976).

Perhaps the most important finding of the present study is that the observed variance by Poisson fitting is the underestimate of the true value when the rate of evolution fluctuates. The bias is negligible for close comparisons but becomes large when the remote comparisons are made. This will make it difficult to find out the variation of the evolutionary

rate. Ohta & Kimura (1971) found by analysing several "semi-independent" comparisons of reported sequences from cytochrome c, hemoglobin α and hemoglobin β, that the variance of the evolutionary rate is roughly 2 ~ 3 times of the expected value from the strictly random process. Since the comparisons they used are not very remote, the bias would not be large. Essentially the same result has been obtained by Langley & Fitch (1974). In their analysis, the value of χ^2 becomes roughly twice the degrees of freedom for the maximum parsimonious solution by Fitch (1971). Therefore, these results, at least, fit to the present model of random fluctuation of the parameter among the legs. It has been noted by several investigators in cytochrome c and globins that the variance decreases as time gets larger (Ohta & Kimura, 1971; Romero-Herrera et al., 1973; Van Valen, 1974). This would be explained by the present result. Note here that the autonegative correlation as suggested by Van Valen (1974) is not necessarily needed to explain this fact.

Random fluctuation of the evolutionary rate is expected if the *nearly neutral* mutant substitutions are numerous for molecular change. In particular, the behavior of very slightly deleterious mutations, which I postulated as important for molecular changes (Ohta, 1973, 1974), is greatly influenced by the population size. When the population is small, they behave as neutral mutants and can spread in the population. But when the population is large, they are effectively selected against and are eliminated from the population. The severity of the environment must also affect their behavior. Such factors may be regarded as random variables and not systematic factors such as progressive evolutionary force. In fact, if the progressive evolutionary force or the adaptive natural selection is the primary factor of molecular evolution, one should find much stronger correlation between the evolution at the phenotypic level and the molecular evolution than actually found (Kimura, 1969; King & Wilson, 1975).

The models used in the present study may be criticized as unrealistic, in particular, I did not incorporate the chemical nature of 20 kinds of amino acids, whereas it is now well recognized that it has an important bearing on amino acid substitutions (e.g. Dayhoff, 1972). Instead, I concentrated on the random nature of mutant substitutions. Since some distance measures used are based on the assumption of the random nature of amino acid substitutions, some of the agreements between the actual and the estimated distances may be circular. However, as a first step, it is necessary to test these distance measures under the assumption on which they are derived. The next step is to incorporate the chemical nature of amino acids into the model and to find out the aggregate effects of various types of non-randomness in the amino acid substitutions.

Acknowledgement. I thank Dr. Motoo Kimura for stimulating discussions and encouragement throughout the course of this work. Thanks are also due to Drs. Jack L. King and Masatoshi Nei for going over the manuscript and giving many valuable suggestions to improve it.

REFERENCES

Beyer, W.A., Stein, M.L., Smith, T.F., Ulam, S.M. (1974). Math.Biosci. 19, 9
Dayhoff, M.O.(1972). Atlas of protein sequence and structure, Vol. 5. Washington: National Biomedical Research Foundation
Farris, J.S. (1973). Am.Nat. 107, 531
Fitch, W.M. (1971). Syst.Zool. 20, 406
Fitch, W.M., Margoliash, E. (1967). Science 155, 279
Grantham, R. (1974). Science 185, 862
Goodman, M., Moore, G.W., Matsuda, G. (1975). Nature 253, 603
Holmquist, R. (1972a). J.Mol.Evol. 1, 115
Holmquist, R. (1972b). J.Mol.Evol. 1, 134
Jukes, T.H., Cantor, C.R. (1969). In: Mammalian protein metabolism, H.N. Munro, ed., pp. 21-132. New York: Academic Press
Kimura, M. (1968). Nature 217, 624
Kimura, M. (1969). Proc.Natl.Acad.Sci. 63, 1181
Kimura, M., Ohta, T. (1972). J.Mol.Evol. 2, 87
King, M.C., Wilson, A.C. (1975). Science 188, 107
Langley, C.H., Fitch, W.M. (1974). J.Mol.Evol. 3, 161
Margoliash, E., Smitch, E.L. (1965). In: Evolving genes and proteins, V. Bryson, H.J. Vogel, eds., pp. 221-242. New York: Academic Press
Nei, M., Chakraborty, R. (1976). J.Mol.Evol. (in press)
Ohta, T. (1973). Nature 246, 96
Ohta, T. (1974). Nature 252, 351
Ohta, T., Kimura, M. (1971). J.Mol.Evol. 1, 18
Romero-Herrera, A.E., Lehmann, H., Joysey, K.A., Friday, A.E. (1973). Nature 246, 389
Uzzell, T., Corbin, K.W. (1971). Science 172, 1089
Van Valen, L. (1974). J.Mol.Evol. 3, 89
Zuckerkandl, E., Pauling, L. (1965). In: Evolving genes and proteins, V. Bryson, H.J. Vogel, eds., pp. 97-166. New York: Academic Press