J Mol Evol (1981) 17:182–187

Journal of Molecular Evolution

Statistical Properties of the Jukes-Holmquist Method of Estimating the Number of Nucleotide Substitutions: Reply to Holmquist and Conroy's Criticism

Masatoshi Nei and Yoshio Tateno

Center for Demographic and Population Genetics, University of Texas, PO Box 20334, Houston, Texas 77025, USA

Summary. Conducting computer simulations, Nei and Tateno (1978) have shown that Jukes and Holmquist's (1972) method of estimating the number of nucleotide substitutions tends to give an overestimate and the estimate obtained has a large variance. Holmquist and Conroy (1980) repeated some parts of our simulation and claim that the overestimation of nucleotide substitutions in our paper occurred mainly because we used selected data. Examination of Holmquist and Conroy's simulation indicates that their results are essentially the same as ours when the Jukes-Holmquist method is used, but since they used a different method of computation their estimates of nucleotide substitutions differed substantially from ours. Another problem in Holmquist and Conroy's Letter is that they confused the expected number of nucleotide substitution with the number in a sample. This confusion has resulted in a number of unnecessary arguments. They also criticized our X_2 measure, but this criticism is apparently due to a misunderstanding of the assumptions of our method and a failure to use our method in the way we described. We believe that our earlier conclusions remain unchanged.

Key words: Molecular evolution – Nucleotide substitution

In response to Nei and Tateno's (1978) comments on Jukes and Holmquist's (1972) method (J/H method) of estimating the number of nucleotide substitutions, Holmquist (1978) amended some parts of the J/H method but maintained the view that the J/H method is essentially correct. Holmquist, aided by Conroy (Holmquist and Conroy 1981), now attacks our study and claims that our conclusions about the J/H method are based on a selected set of data and are therefore incorrect. We find his claim unfounded and believe that their criticisms are based on a misconception of the sampling properties of the J/H method and the meaning of mathematical models in data analysis. In the following we shall reply to each of his criticisms. All sections in this Letter will be the same as those of Holmquist and Conroy.

Before going into the detail, however, we should remind the reader that some of Holmquist's quotations from our paper are incorrect or incomplete. For example, we have not made such a flat statement as "the maximum parsimony estimate as made by the augmentation algorithm systematically err by overestimating the total number of nucleotide replacements." Our conclusions about Goodman et al.'s *augmentation* procedure (not the maximum parsimony method) and Jukes and Holmquist's REH method were more modest than Holmquist's quotations suggest. Indeed, for the former method we concluded that "these findings (overaugmentation and the large variance of estimate when the number of nucleotide substitutions is large) provide a warning against an uncritical use of the method in the study of the rate of nucleotide substitution in evolution" (Tateno and Nei 1978; Nei and Tateno 1979). On the other hand, for the latter we stated that "as long as there is a suggestion that the J/H method leads to an overestimate, uncritical use of the method should be avoided" (Nei and Tateno 1978). It should also be noted that we have never criticized Holmquist's (1972) method of estimating REH from the proportion of different amino acids between two homologous proteins. Because of these incorrect and incomplete quotations, we would like to urge the reader to read this Letter together with our earlier papers on this subject. Further-

Offprint requests to: M. Nei

0022–2844/81/0017/0182/$01.20

more, replies to some of Holmquist and Conroy's criticisms are already given in the previous papers. We shall write our reply as briefly as possible.

Nonuniform Amino Acid or Nucleotide Substitution and *On the Measure X_2*

When the amino acid sequences of two homologous proteins from different organisms are compared, the frequency of amino acid sites with 2 minimum base differences is generally higher than that expected under random nucleotide substitution in the entire gene. To account for this discrepancy, Jukes and Holmquist (1972) proposed a model in which amino acid sites of a protein are classified into two groups, i.e. variable sites and invariable sites, and in the group of variable sites the theory of random nucleotide substitution applies, all codons having the same expected rate of substitution. The J/H method of estimating nucleotide substitutions is based on this model and heavily depends on the ratio (r) of the frequency of amino acid sites with 2 and 3 minimum base differences to the frequency of amino acid sites with one minimum base difference. The number of nucleotide substitutions per gene estimated by this method is denoted by REH, whereas the number per codon by REHC.

Nei and Tateno (1978) claimed that the excess of amino acid sites with 2 minimum base differences is due not only to the variation of mutability among sites but also to the *nonrandom (not nonuniform) amino acid* substitution. It is well known that amino acid substitution in protein is not random and an amino acid is replaced more often by a similar amino acid than expected under random substitution. Using this data, they proposed a new method of estimating the number of nucleotide substitutions from the proportion (p_d) of different amino acids between the homologous proteins. In practice, their method is based on Dayhoff's (1972) mutation probability matrix obtained from empirical data and their estimate X_2 per 100 codons is approximately given by 1.7 PAM, where PAM is Dayhoff's "accepted point mutation per 100 codons." (Because of the nature of Dayhoff's empirical mutation probability matrix, Nei and Tateno's formulation should not be used when p_d is smaller than 0.02.) Conducting computer simulations, Nei and Tateno (1978) showed that the estimate of nucleotide substitutions obtained by the above method is always smaller than that obtained by the J/H method and pointed out the possibility that the J/H method gives an overestimate.

Holmquist and Conroy now attempt to compare the values of PAM, REH, and X_2 per 100 codons in their Table 1 and Fig. 1 and find that X_2 is always larger than either PAM or REH. Because of this finding, they claim that the assumption of random *nucleotide* substitution (or neglect of nonrandom substitution) in the J/H method does not lead to overestimation of nucleotide substitutions. However, it should be noted that the REH values in Table 1 and Fig. 1 are not those obtained by the J/H method but the expected values corresponding to the proportion of different amino acids between homologous proteins (p_d) when nucleotide substitution is assumed to occur at random over the entire region (Holmquist 1972). As mentioned earlier, what we have questioned is not this relationship but the J/H method in which REH is computed from the r value. Therefore, Holmquist and Conroy's Table 1 and Fig. 1 have nothing to do with our finding. They have also computed the estimate of nucleotide substitutions obtained by the Jukes-Cantor-Kimura-Ohta method (X_1). This method is based on the same principle as Holmquist's (1972) p_d method, so that X_1 should be close to REH, as his Table 1 shows.

In this connection it should be noted that PAM represents the number of amino acid substitutions rather than the number of nucleotide substitutions, so that it cannot be compared with REH, X_1, or X_2. On the other hand, X_2 is intended to be applied to actual amino acid sequence data where nonrandom substitution is common and mutability is not the same for all codon sites, whereas REH and X_1 in Holmquist and Conroy's Table 1 are based on the assumption that all codon sites are equally mutable and nucleotide substitution occurs at random. Therefore, X_2 is expected to be larger than REH and X_1 when the same p_d value is used. Nothing in his Table 1 is contradictory with our criticism on the J/H method.

Note also that Nei and Tateno have never stated that the assumption of random nucleotide substitution always leads to overestimation. What Nei and Tateno (1978) claimed is that nonrandom *amino acid* substitution may result in an unduly large value of r, i.e. the ratio of the frequency of amino acid sites with 2 and 3 minimum base differences to that of amino acid sites with one base difference, and this large value of r would give an overestimate of the number of nucleotide substitutions in the J/H method. In fact, we have shown by using Dayhoff's empirical probability matrix that the r value can be large even if the expected number of nucleotide substitutions is small. In this connection it should be noted that we have never stated that the *nonuniform* rates of codon substitutions at different sites produce an overestimate of REH. Therefore, their argument about nonuniform substitution is irrelevant to our paper.

Holmquist and Conroy attempt to show that the computation of X_2 in our computer simulation (Table 1 of Nei and Tateno 1978) is erroneous and that if REH and X_2 are computed in their own way X_2 is larger than REH. Their computations are, however, again based on an erroneous conception of the problem. The essence of Holmquist and Conroy's argument is as follows: A gene consists of variable and invariable codons. The distinction between these two groups is

not absolute, but to describe the evolutionary change of a gene the J/H model in which this property is incorporated is sufficiently accurate. In this model the number of variable codons can be estimated by Eq. (5) or Eq. (6) in their paper. Since Nei and Tateno, they think, assume that "the entire structural gene is free to accept mutations", X_2 should be computed by using the proportion of different amino acids for the part of variable codons. Namely, p_d in Nei and Tateno's (1978) formula (11) should be replaced by $T_2 p_d$ in their Eq. (10). When this method is applied to the data from our computer simulation in which Dayhoff's empirical probability matrix was used, they find that X_2 (T_2X_2 in their terminology) is considerably larger than REH. They then conclude that X_2 rather than REH gives an overestimate of the number of nucleotide substitutions.

Holmquist and Conroy's computation and conclusion are of course irrelevant to the problem at issue, since T_2X_2 is their own invention and has nothing to do with the Nei and Tateno formulation. The crucial error in their computation is derived from their belief that in the Nei and Tateno formulation all codons in a gene are assumed to be equally mutable. (In their revised manuscript they state otherwise, but their computation is based on this misunderstanding.) We have never made such an assumption. Rather, we have assumed that the substitution of a codon is essentially controlled by the property of the amino acid coded for and the mutability varies from amino acid to amino acid. The mutabilities of amino acids we used are those obtained by Dayhoff from empirical data, and there is a 7-fold difference between the most mutable and least mutable amino acids (Dayhoff 1972, p. 91). Furthermore, in our formulation we used Dayhoff's empirical data on amino acid substitution, so that all sorts of nonrandom properties of amino acid substitution are taken into account. Because of these features, our method of computation of X_2 cannot be used in the way Holmquist and Conroy used. It should be applied to real data strictly in the way we described.

In this connection it should be noted that our model (N/T model) is not perfect and it is only an approximation to reality. In our model the mutability of a codon formally depends on the amino acid but not on the amino acid site in a protein. Strictly speaking, this is incorrect. However, as discussed in our earlier paper, there is a high correlation between site-dependency and amino-acid dependency. For example, many active sites or contact sites (highly immutable sites) in protein are occupied by cysteine, tyrosine, etc., which are highly immutable. Therefore, we have in effect considered both site-dependency and amino-acid dependency. We also note that a model based on site-dependency only, like the J/H model, is also incorrect, since the frequency of amino acid substitution is highly dependent on the biochemical property of the pair of amino acids that are involved in a substitution (Dayhoff 1972).

Holmquist and Conroy expend a considerable amount of space defending the J/H model, and try to justify it by using Fitch's (1972) concept of covarions. According to Fitch, the number of covarions increases as the diversity of the organisms studied increases. This certainly corresponds to the increase of the estimate of the number of variable codons (X_2) with increasing evolutionary time, but in our opinion it does not justify the J/H model at all. First of all, the concept of covarions is not the same as that of varions. The former is supposed to be a group of codons of which the substitutions are mutually dependent, whereas varions are merely a group of codons which are subject to independent substitution with the same probability. Second, the concept of covarions is not an established theory but a hypothesis still to be proven. Therefore, even if there are some agreements between the concepts of covarions and varions, it is not proof of the J/H model. The covarion hypothesis asserts that at any moment of time only a small proportion of codons are mutable. However, this assertion is based on a rough statistical analysis of the pattern of codon substitutions, and does not exclude the possibility that virtually all codons are mutable at any moment of time but the mutability greatly varies from codon to codon. The statistical method used by Fitch (1972) is not powerful enough to distinguish between these two possibilities. Third, despite Holmquist and Conroy's disclaimer, the J/H *model* has been shown to be valid only when the number of variable codons (T_2) is fixed throughout the evolutionary time. To our knowledge, no one has shown how randomness of codon substitution is assured when T_2 increases with time.

In their revised manuscript, which was prepared after seeing our reply, Holmquist and Conroy admit that in our model all codons in a gene are not assumed to be equally mutable. Yet they criticize that in the N/T model all codons are potentially variable and no distinction is made between invariable sites and variable sites. This criticism is valid, but do truly invariable sites really exist? If they do, what is their proportion in the entire gene? Fitch and Margoliash's (1967) statistical study certainly suggests the existence of such sites in cytochrome c but does not prove it. In hemoglobins at least more than 90 percent of amino acid sites seem to be potentially variable (Zuckerkandl and Pauling 1965). Therefore, it seems incorrect to say that our model is unrealistic at least for all practical purposes. Indeed, as shown in our earlier paper, the values of $P_e(0)$, $P_e(1)$, $P_e(2)$, and $P_e(3)$ obtained in our simulation are similar to the observed values in real data of hemoglobin and cytochrome c. They also argue that a two-parameter statistic is better than a one-parameter statistic. This is not true. It is well known in statistics

that a multi-parameter statistic could be worse than a single-parameter statistic. Indeed, REHC has a larger variance than X_1, a single-parameter statistic, as we have shown in our earlier paper.

Holmquist and Conroy state that the REHC value (REH/100 in his Table 1) obtained from p_d is in agreement with the REHC estimated by the J/H method. This statement is correct if amino acid substitution occurs completely at random and the number of amino acids examined is very large. In the presence of non-random amino acid substitution, however, the latter can be much larger than the former. For example, the latter value in replication 1 for PAM = 20 in Holmquist's Table 2 [Nei and Tateno's (1978) Table 1] is 0.74, whereas REHC obtained from the p_d value [0.214; see Nei and Tateno's (1978) Table 1] is 0.34. Similar comparisons can be made for all replications for PAM = 20 and PAM = 36, except in one case where the J/H method is not applicable. These comparisons indicate that the REHC obtained by the J/H method is about two times larger than the REHC from p_d. Holmquist and Conroy also state that "there is no large discrepancy between REH, X, and PAM". This statement is again incorrect when amino acid substitution is not random. Indeed, it is seen from Holmquist and Conroy's Table 2 that in the case of PAM = 20, REH (= $\mu_2 T_2$) ranges from 52 (= 2.29 × 22.6 in replication 2) to 100 (= 4.65 × 21.4 in replication 3), whereas in the case of PAM = 36 it ranges from 97 (= 1.99 × 48.9 in replication 5) to 142 (= 2.87 × 49.6 in replication 3) excluding the case where the J/H method is not applicable. Namely, REH is 2.5 ~ 5 times larger than PAM. Holmquist and Conroy's claim is again unsupported.

Holmquist and Conroy seem to believe that the agreement between REHC and T_2X_1 in his Table 2 is a support of the J/H theory. Actually, this is not true, since in the process of estimation the REHC value is adjusted so as to agree with T_2X_1. In his revised manuscript they deny this, but if one notes that both T_2 and $1 - P_e(0)$ are estimated by fitting the J/H model to the data, it is obvious that the T_2X_1 value computed from these T_2 and $1 - P_e(0)$ should become close to REHC. Their comparison of r_e and $\hat{r}_e$ also does not have much meaning, since $\hat{r}_e$ is not a parameter but just another estimate of parameter r, like r_e. Furthermore, there is no guarantee that $\hat{r}_e$ is always better than r_e, since the true expectations of these quantities are not known. Nei and Tateno (1978) and Holmquist (1978) worked out only the approximate expectations. We do not think that $(r_e - \hat{r}_e)/\hat{r}_e \equiv 1/n_e(1)$ gives any idea about $(r_e - r)/r$ or $(\hat{r}_e - r)/r$, i.e., the magnitude of deviation of r_e or $\hat{r}_e$ from the true value r. Indeed, $\hat{r}_e$ gives a very biased estimate of the expected number of nucleotide substitutions per variable codon (μ_2), as will be discussed later. They have also discussed the disagreement between the maximum parsimony estimates and X_2 values for hemoglobin α chain. See Nei and Tateno (1978) for our views on this problem.

Varions, Covarions and Convarions

Our comments on this issue are already presented in the foregoing section. The only thing we would like to add is that Fitch (1980) recently discussed this problem extensively, and his view is close to ours.

Computer Simulations

We believe that Holmquist and Conroy's comments in this section are based on their misconception about sample mean and population mean. In one of our computer simulations, in which a model gene with 100 codons (50 variable and 50 invariable codons) was used, we found the mean of the estimate ($\hat{\mu}_2$) of μ_2 from 5 replications to be 1.2 for $2v = 10$, considering only the cases where the J/H method could be used. They apparently misunderstand that this is the population mean, and compare the expected values of $n(i)$ for $\mu_2 = 1.2$ with those for $\mu_2 = 2v/50 = 0.20$ (the theoretical value in our simulation), where $n(i)$ is the number of amino acid sites with i minimum base differences. They then argue that the differences in the expected values of $n(i)$ between the two cases are so large, that Nei and Tateno must have made some error in their computation or selected atypical cases intentionally in the computation. This criticism is appalling. They do not appear to realize how large the deviation of $\hat{\mu}_2$ from μ_2 can be, though this was one of our major criticisms in our paper. In this connection it is worthwhile to note that if the number of variable codons (sample size) increases, the difference between $\hat{\mu}_2$ and μ_2 gradually diminishes.

Table 1 gives the actual values of $n(i)$ for the first five replications of our simulation for the case of $2v = 10$ and the means over all 50 replications. It is clear that the sample values of $n(i)$ can be quite different from the expected values because of sampling error, but the mean values are very close to the latter values. Also note that $\hat{\mu}_2$ is very large compared with the theoretical values of $\mu_2 = 0.2$ when it is estimated by the J/H method. In our earlier version of this Letter we wrote "If Holmquist can not believe in these results, we would suggest that he conduct the simulation himself." He and Conroy then followed our advice. The results they obtained support our findings, though they claim otherwise (see below).

Their revised manuscript suggests that Holmquist now appreciates the large effect of sampling error on $\hat{\mu}_2$. However, instead of accepting our criticism, they attack us for not using the minimum base differences (MBD) method. Earlier Holmquist (1972) strongly criticized Fitch and Margoliash's (1967) MBD method, but when we (personal communication to Holmquist, 1975; Nei and Tateno 1978) indicated the inapplicability of the J/H method for the cases of $n(2) = n(3) = 0$ and $n(1) = 0$, Holmquist (1978) suggested that in these cases the MBD method be used. When $n(2) =$

Table 1. Numbers [$n(i)$] of amino acid sites with i minimum base differences in the group of (50) variable codons in Nei and Tateno's (1978) computer simulation for $2v = 10$ and the values of $\hat{\mu}_2$ and $\hat{T}_2$, and REH. (The remaining 50 codons were assumed to be invariable.) The result for the first five replications and the averages for 50 replications are given

Replication	$n(0)$	$n(1)$	$n(2)$	$n(3)$	$\hat{\mu}_2$	$\hat{T}_2$	REH
1	41	9	0	0	–[a]	–	–
2	41	8	1	0	0.75	21.95	16.5
3	40	8	2	0	1.50	15.96	23.9
4	45	5	0	0	–	–	–
5	41	9	0	0	–	–	–
Average over 50 replications	43.0	6.7	0.3	0	1.36[b]	14.6[b]	17.6[b]
Expected value	43.2	6.6	0.2	0	0.20	50.0	10.0

[a]Stands for the case where the J/H method could not be used

[b]Averages over all cases in which the J/H method could be used. Note that in Nei and Tateno's (1978) paper the averages for the first 15 replications are given. The theoretical value of REH is $2v = 10$ in this case

$n(3) = 0$ (but not when $n(1) = 0$), the MBD method is acceptable as an approximation, and we do not question this method. What Nei and Tateno (1978) questioned is the J/H method, as mentioned earlier. For this reason, we were particularly interested in the values of REHC, $\hat{\mu}_2$ and $\hat{T}_2$ (estimate of the number of variable codons) for the cases in which the J/H method was applicable. We therefore computed the averages of $\hat{\mu}_2$, T_2, and REHC for these cases. Holmquist and Conroy criticize that we selected the cases in which $\hat{\mu}_2$ and REHC were large. We do not understand why one can make such a criticism. Without separating the cases where the J/H method is applicable from the cases where it is not, how can one study the property of the J/H method?

The results of Holmquist and Conroy's simulation are essentially the same as ours and support our conclusion. For example, they could use the J/H method in 10 cases out of the 50 replications conducted, and the average of $\hat{\mu}$ for the case of $\mu_2 = 0.2$ becomes 1.01, which is close to our value (1.4) and five times higher than the expected value. Note that in the J/H method r_e is computed by equation (3) in Holmquist and Conroy's (1980) Letter rather than by (4). Other quantities are also similar to ours. However, Holmquist and Conroy computed REH by using the MBD method when the J/H method was not applicable, and averaged it over all 50 replications. Since the MBD method generally gives an underestimate and the J/H method gives an overestimate, the average of observed REH's may become close to the expected value. Indeed, in their simulation this was the case. This agreement of course does not support the validity of the J/H method, but rather supports our earlier conclusion. Note that in actual data analysis REH is estimated without replications and thus if the J/H method is used REH is overestimated, whereas if the MBD method is used, it is underestimated.

In this connection it is interesting to note that Holmquist's (1978) unbiased estimate ($\hat{r}_e$) of r still gives an overestimate of μ_2. In their Table 4 Holmquist and Conroy give an average of the estimates of μ_2 equal to 0.84, which is four times larger than the expected value ($\mu_2 = 0.2$). Clearly, $\hat{r}_e$ is not an unbiased estimate.

All his remaining comments are again based on his misconception about the sample estimate and population parameter, so that we shall not further dwell on them. See also Nei and Tateno (1978) for replies to some of his comments. In their revised manuscript they have included estimation of REH from mRNA sequences. This problem is, however, irrelevant to our criticism, so that no comment will be made.

Correlated Behavior Between the Fixation Intensity and the Number of Varions

Here again Holmquist and Conroy confuse the sample property with the population property. Although we studied the correlation between sample estimates $\hat{\mu}_2$ and $\hat{T}_2$ for a given value of μ_2 or for a set of μ_2's, they discuss the correlation between population parameters μ_2 and T_2 without specifying the probability space to be considered, and criticize that the negative correlation observed in our simulation is "due to a lack of biological realism in our simulations." In practice, however, we find the same negative correlation in their simulation results. Furthermore, in an earlier section (*On the Measure* X_2) they had stated that "whenever μ_2 is overestimated, T_2 is underestimated so that their product REH is relatively constant." Here they clearly admit the negative correlation. They also accept our earlier finding that "the increase in REH with increasing time is often caused by the increase in $\hat{T}_2$."

Holmquist and Conroy seem to misinterpret our criticism on the J/H *model*. They write: "Nei and Tateno's statement that the J/H *model* requires T_2 to remain fixed while μ_2 must increase with time is not correct." Actually we have never made such a statement. What we have stated is that if T_2 increases with time but μ_2 stays more or less constant, as suggested by our computer simulation and Jukes and Holmquist's analysis

of actual data, the J/H *model* must be modified by taking into account this property. Holmquist and Conroy claim that the J/H *model* is valid whether T_2 changes with time or not. However, they have not proved it mathematically. Actually, it is not easy to develop a mathematical model in which T_2 varies with time but in the group of variable codons random nucleotide substitution is assured with the same substitution rate for all condon sites, as we emphasized in our earlier paper. To develop such a model, one must assume that T_2 is a random variable increasing with time. Since μ_2 remains constant, substitution must occur mostly in the sites which have recently become a variable-codon site. Then, how is the randomness of nucleotide substitution over all variable codon sites assured? Of course, even if a model approximating the above process is developed, it does not refute our criticism on the J/H *method*. This is because our main criticism is that the *r* value can be large under nonrandom amino acid substitution even if the average number of nucleotide substitutions per codon is relatively small.

Solution to a Paradox

This section is irrelevant to our paper.

Conclusions

It is now clear that Holmquist and Conroy's criticism on Nei and Tateno's (1978) paper is based on a misunderstanding of the latter paper. Nei and Tateno (1978) criticized Jukes and Holmquist's (1972) method of estimating the number of nucleotide substitutions. Holmquist and Conroy tried to defend this method by using other methods of estimating the number of nucleotide substitutions. Therefore, their arguments and computations are largely irrelevant to our earlier paper, and our criticism remains valid. However, this subject is now becoming obsolete because of the recent advance in DNA and RNA sequencing.

References

Dayhoff MO (1972) Atlas of Protein Sequence and Structure, v 5, Natl Biomed Res Found, Washington
Fitch WF (1972) Brookhaven Symp Biol 23:186
Fitch WF (1980) J Mol Evol 16:1–57
Fitch WF, Margoliash E (1967) Biochemical Genetics 1:65
Holmquist R (1972) J Mol Evol 1:211
Holmquist R (1978) J Mol Evol 11:361
Holmquist R, Conroy T (1981) J Mol Evol 17:167–181
Jukes TH, Holmquist R (1972) J Mol Biol 64:163
Nei M, Tateno Y (1978) J Mol Evol 11:333
Nei M, Tateno Y (1979) J Mol Evol 13:167
Tateno Y, Nei M (1978) J Mol Evol 11:67
Zuckerkandl E, Pauling L (1965) In: Bryson V, Vogel HJ (eds) Evolving Genes and Proteins. Academic Press, New York, p 97

Received August 14, 1980