J. Mol. Evol. 13, 167–171 (1979)

Journal of
Molecular Evolution
© by Springer-Verlag · 1979

Augmentation Algorithm: A Reply to Holmquist

Masatoshi Nei, and Yoshio Tateno

Center for Demographic and Population Genetics, University of Texas at Houston, Houston, TX 77025, USA

Holmquist (1978a) recently criticized Tateno and Nei's (1978) Letter to the Editor on "Goodman et al.'s method for augmenting the number of nucleotide substitutions." He states that, unlike our conclusion, the results of the computer simulation presented in our paper proves the validity rather than the deficiency of Goodman et al.'s (1974) augmentation procedure. We believe that Holmquist's statement is based on his misunderstandings of our results. In the following we shall reply to each of his comments. Before going into detail, however, we would like to ask the reader to read this letter together with our earlier paper.

Does the Augmentation Procedure Assume Equal Rates of Nucleotide Substitution?

Holmquist claims that in Goodman et al.'s augmentation procedure the assumption of equal rates of nucleotide substitution in different evolutionary branches is not required. We disagree. In fact, we believe that if the rate of nucleotide substitution is allowed to vary with the evolutionary branch (link in Goodman et al.'s terminology), there is no way to correct for underestimate (or overestimation) of any link length. For example, an unduly small value of the number of nucleotide substitutions estimated by the maximum parsimony method for a link could be due either to a low rate of nucleotide substitution in this particular link or to simple underestimation. However, without the assurance that it is not due to a low rate of nucleotide substitution, how can we augment the link length? In practice, of course, nucleotide substitution is subject to a stochastic error whether it is aided by natural selection or not. Therefore, even if the (expected) rate of substitution is constant, the number of nucleotide substitutions for a given evolutionary period shows a large variation. As we have shown in our earlier paper, it is this stochastic variation that causes overaugmentation of link lengths in Goodman et al.'s method.

Does the Augmentation Procedure Overestimate the True Distance?

In our earlier paper we showed that with the evolutionary tree used in our computer simulation Goodman et al.'s augmentation method gives a good estimate of the number

0022–2844/79/0013/0167/$ 01.00

of nucleotide substitutions when the true number (or true distance: TD) is smaller than about 50, but tends to give an overestimate when TD is large. Particularly when TD was about 100 to 200 the augmented distance (AD) was considerably larger than TD. For the longest link with TD = about 500, however, the augmentation method gave a gross underestimate. (The reason for this has been discussed in our earlier paper.) Thus, we concluded that "Goodman et al.'s method introduces a systematic error, the difference between AD and TD depending on the value of TD (or direct distance: DD)."

Holmquist now compares the average values of AD and TD *over all evolutionary branches or links* in each of replications 1 and 2 of Czelusniak et al.'s (1978) computer simulation, which is essentially the same as ours, and finds that the averages are not significantly different. He then concludes that AD is not a seriously biased estimate of TD. We do not think that this conclusion is warranted. In the present case the average value of AD over all links is close to that of TD, because the overestimation of TD by AD for the range of TD = 60 ~ 200 is cancelled out by a serious underestimation for the case of TD = about 500. *However, no one would be interested in estimating the average value of TD over all links; it has no biological meaning.* Actually, what the augmentation method is intended to do is to estimate the TD value for *each link*. Therefore, if the means of AD and TD are to be compared, they should be computed over replications for each link, as we emphasized in our earlier paper. He has clearly misunderstood the probability space to be considered. The AD value for a particular link is subject to a large variation because nucleotide substitution is stochastic. Since AD is an extremely complicated function of DD values (link lengths before augmentation) as well as the tree used, it is virtually impossible to derive the mean and variance theoretically. Our computer simulation with four replications [Tateno and Nei's (1978) Table 1], however, has clearly shown that AD is an overestimate of TD when TD is about 60 to 200. Recently, we conducted a simulation of three more replications. The results obtained were essentially the same as those of the earlier four replications. We also note that essentially the same pattern is observed in Czelusniak et al.'s simulation, contrary to Holmquist's interpretation.

Holmquist (1978b) states that "Tateno and Nei (1978) report only 14 out of 39 augmented distances for their computer simulated network and discuss *selected* augmented distances among these 14 to force a case for a systematic bias in the augmentation procedure towards overestimation of the total nucleotide replacements". It is true that we presented augmented distances only for 14 links, but we included all links which had TD $\geqslant$ 58 in any one of the four replications. Since we claimed that a systematic overaugmentation occurred only when TD = 60 ~ 200, there was no need to present all the data. For this reason, distance data for the case of TD < 58 were presented only for a limited number of links just to show that "the augmentation is sufficiently accurate" in this case (Tateno and Nei 1978). We do not think that uninformative and unnecessary data should be published. Of course, we will be glad to supply our data to any reader who is interested in the detail.

Holmquist claims that the agreement between AD and the estimate of nucleotide substitutions obtained by Jukes and Holmquist's (1972) method in real data supports the validity of Goodman et al.'s augmentation method. This claim is obviously illogical, since both are estimates and the true number of nucleotide substitutions is not known in real data. Furthermore, Nei and Tateno (1979) have shown that Jukes and Holmquist's method too tends to give an overestimate.

Do the Augmented Distances have a Variance Larger than the Variance of the True Distances?

We believe that Homlquist's answer to this question is also based on his misconception of the probability space to be considered. The variance of AD should be computed among replications for each link rather than among links in each replication. Tateno and Nei's (1978) Table 1 clearly shows that the variance of AD among replications is generally larger than that of TD. Namely, in 12 of the 14 links considered in this table, the variance of AD is larger than that of TD. For example, the variances of AD for links 911–153, 906–209, 910–158, and 919–2 are 371.6, 262.3, 2554.9, and 286.3, respectively, whereas the corresponding variances of TD are 26.9, 123.7, 32.9, and 78.3. Holmquist seems to be unhappy with these results and computes the standard deviations of AD and TD from two replications of Czelusniak et al.'s simulation. His results indicate that in 10 of the 14 links we have considered the standard deviation of AD is larger than that of TD, even though each standard deviation is based on only two observations in this case. He has also computed the standard deviation of the maximum parsimony augmentation distance (AD_{mp}). In our paper, however, we have not considered this quantity.

As mentioned earlier, we have recently added three more replications to our simulation study, so that we have results from seven replications. Using these results, we now computed the variances of AD and TD for each link. (Czelusniak et al.'s TD values in replications 1 and 2 are identical with those of our replications 1 and 2, respectively, so that their data were not included.) The results obtained indicated that in all of the 14 links the variance of AD was larger than that of TD. The average of the former variance over the 14 links was 533.6, whereas the average of the latter variance was 159.8. This clearly substantiates our earlier claim.

Quite independently of the above problem, he argues that an unbiased estimator is preferable to a biased one, whatever its variance is. We believe that this is a great challenge to the current view of statisticians, who generally use the mean squares error $[E(s - p)^2]$ from the population mean or parameter (p) as a criterion of accuracy of an estimator or statistic (s) whether s is unbiased or not (e.g., Cochran, 1963). Note that in the actual process of estimation of link length no replicate observations are obtainable for any link, unlike the case of computer simulation.

Do Topological Changes Affect the Augmented Distance?

As we predicted in our earlier paper, Holmquist finds substantial errors in the topology of the tree reconstructed by the maximum parsimony method. He then states that these topological changes did not alter the magnitude of the augmented distance nor did it change its variance. However, his comparison is again based on the mean of AD over all links rather than the mean over replications for each link. Therefore, it is not clear how the topological errors affected the AD values for individual links. This problem must be studied by using a large number of replications, as pointed out by Tateno and Nei (1978). It should also be noted that this problem is a minor part of our criticism on the augmentation method.

Does a Large Number of Nucleotide Substitutions Require a Different Model?

We have *not* suggested that the method developed by Jukes, Cantor, Kimura, and Ohta should be used *whenever* the number of nucleotide substitutions per codon is large. What we suggested is that in the particular case of Czelusniak et al.'s computer simulation the Jukes-Cantor-Kimura-Ohta method is better than the Poisson correction method. Note that in this case the model of random nucleotide substitution is used and the number of variable codons ($T_2 = 50$) is predetermined. All of Holmquist's computations in this section have nothing to do with our paper.

Are Ancestral Sequences Inferred by the Method of Parsimony Correct?

This question is irrelevant to our paper, but the answer is obvious without computation.

Conclusion

Holmquist states: "The statistical properties of the augmentation procedure are already clear from published theory, simulations, and analysis of real data. Tateno and Nei's critique to the contrary is a result of a biased sample of internodal link distances." We believe, however, it is clear from the above arguments that his criticism is based on his misinterpretation of the problems at issue and misunderstandings of our results. In our view Holmquist's paper has added no new finding about the properties of Goodman et al.'s augmentation method. Holmquist writes as though Moore (1977) provided proof of the nonoveraugmentation of Goodman et al.'s method. Actually, what Moore did is to restate the rules of Goodman et al.'s augmentation procedure in terms of mathematical terminologies and has nothing to do with the proof.

In this connection it should be noted that Tateno and Nei studied Goodman et al.'s augmentation method *separately* from their maximum parsimony algorithm and showed that overaugmentation may occur when TD is large. Czelusniak et al. (1978) do not question this finding but claim that it does not give an overestimate of nucleotide substitutions *when it is used together with their maximum parsimony algorithm*, since the latter method generally gives an underestimate. We are not sure about their claim, so we have suggested that a more careful study should be done before their method is widely used. Theoretically, there is no reason that the amount of overestimation introduced by the augmentation method does not exceed the amount of underestimation due to the maximum parsimony method. The logic of the former method is entirely different from that of the latter. It should be noted that in science logical consistency is more important than many examples of agreement between a theory and data without reason.

Holmquist has repeatedly criticized that we have published only selected data from our simulation. As mentioned earlier, however, we have presented all the data that are necessary for drawing an objective conclusion from our work. We do not think that publication of uninformative and redundant data has any scientific merit. Rather we believe that in scientific research or debate it is important to pay attention to the particular problem at issue and not to meddle with side problems. Holmquist, of course, has the right not to believe in our data or in our interpretation of data. In this case, however, he should repeat our simulation himself and check the validity of our data before he criticizes our work. Since we have given the detailed procedure of our simulation, it is easy to repeat it.

References

Cochran, W.G. (1963). Sampling Techniques, 2nd edition. New York: John Wiley
Czelusniak, J., Goodman, M., Moore, G.W. (1978). J. Mol. Evol. **11**, 75
Goodman, M., Moore, G.W., Barnabas, J., Matsuda, G. (1974). J. Mol. Evol. **3**, 1
Holmquist, R. (1978a). J. Mol. Evol. **12**, 17
Jukes, T., Holmquist, R. (1972). J. Mol. Biol. **64**, 163
Moore, G.W. (1977). J. Theoret. Biol. **66**, 95
Nei, M., Tateno, Y. (1978). J. Mol. Evol. **11**, 333–347
Tateno, Y., Nei, M. (1978). J. Mol. Evol. **11**, 67