J. Mol. Evol. 11, 67–73 (1978)

Journal of
Molecular Evolution

Goodman et al.'s Method for Augmenting the Number of Nucleotide Substitutions

Yoshio Tateno and Masatoshi Nei

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

Summary. Statistical properties of Goodman et al.'s (1974) method of compensating for undetected nucleotide substitutions in evolution are investigated by using computer simulation. It is found that the method tends to overcompensate when the stochastic error of the number of nucleotide substitutions is large. Furthermore, the estimate of the number of nucleotide substitutions obtained by this method has a large variance. However, in order to see whether this method gives overcompensation when applied together with the maximum parsimony method, a much larger scale of simulation seems to be necessary.

Key words: Nucleotide substitution – Molecular evolution

Moore et al. (1973) developed a method for estimating the number of nucleotide substitutions from the amino acid sequences of two homologous proteins, using information on the amino acid sequences from related species. This method (maximum parsimony method) is designed to detect hidden backward and parallel nucleotide substitutions. However, if the two organisms to be compared are distantly related and there are not many related species available, the estimate (parsimony distance) obtained by this method is expected to be an underestimate, since many backward and parallel substitutions would not be detected.

To overcome this difficulty, Goodman et al. (1974) introduced a method for augmenting the parsimony distance to compensate for undetected nucleotide substitutions. Their method is briefly as follows: first, using the maximum parsimony method, the topology of the evolutionary tree for a group of organisms is inferred, and the nucleotide sequences for all organisms and for all ancestral organisms at the branching points (nodes) in the tree are estimated. The number of nucleotide differences is then computed for every pair of nucleotide sequences estimated. This number is called the direct distance (DD). The DD value for a given link between two sequences may be augmented if there is a pair of sequences connected with two or more links but having the same DD value. Namely, if the sum of the DD values (or the augmented distances if the DD's for the links have already been augmented) of all links connecting this

0022-2844/78/0011/0067/$01.40

pair of sequences (path from one sequence to another) is larger than the DD value under augmentation, then it is used as an estimate of the true distance for this link and called the augmented distance (AD). The difference between AD and DD is called the augmentation factor (AF). This AF is added to all links with the same DD value. This is based on the implicit assumption that the rate of nucleotide substitution is the same for all evolutionary branches. This computation is started from the smallest DD value and proceeded for all DD's.

There are four rules in the augmentation procedure: (1) if there are many paths which have the same DD value, the one with the largest number of links is chosen for augmenting this DD value; (2) if there happen to be two or more such paths, the one which gives the smallest AF is chosen; (3) even if a path satisfies rules (1) and (2), the AD value for each link in the path must be smaller than the DD under augmentation - otherwise, this path is discarded and a new path is examined; (4) if the value of AF for a given value of DD, say X, is smaller than that for DD = X - 1, the AF for the latter value is used for augmenting this DD.

If nucleotide substitution occurs deterministically without stochastic errors, the above method seems to give valid corrections for undetected nucleotide substitutions. In practice, however, nucleotide substitution occurs stochastically, and this process may lead to an overcompensation, since the DD value for a relatively short link may by chance become equal to that for a pair of sequences with many small links but a longer total length (Nei and Chakraborty, 1976). In this note we present some of our computer simulations, which substantiate this possibility.

Before going into the details, however, we would like to emphasize that in this note we are concerned only with the augmentation procedure, though in practice the procedure is applied to the outcome of Moore et al.'s maximum parsimony method. The reason for this is that the results of the maximum parsimony method depend heavily on the tree structure used, and to get a proper estimate of ancestral sequences a tree with a large number of species (dense tree) must be used, which requires a large computer time. Furthermore, in the maximum parsimony method the topology and branch length are estimated simultaneously, and the topology estimated may vary from replication to replication even if the true tree is the same, so that comparison of the estimates of link lengths among replications becomes complicated. At any rate, in this note, we shall assume that the topology of the tree and the nucleotide sequences of all contemporary and ancestral organisms have been correctly estimated and restrict our attention solely to the augmentation procedure. For our purpose, the above assumption seems to be justifiable as long as the DD values are designed to be smaller than the true distances for long branches. This is because we are interested in knowing whether or not Goodman et al.'s (1974) algorithm leads to a proper augmentation *when the DD values for some links are smaller than the true values.* At any rate, the study of the augmentation procedure alone seems to bring some insight into Goodman et al.'s method of estimating the number of nucleotide substitutions, as will be discussed later.

In the present study we use a model tree which is similar to a part of Goodman et al.'s (1975) evolutionary tree of the globin family. It is given in Figure 1, where each of the contemporary and ancestral species (node points) is represented by a number, which is consistent with Goodman et al.'s computer algorithm. The total number of contemporary species used is 21.

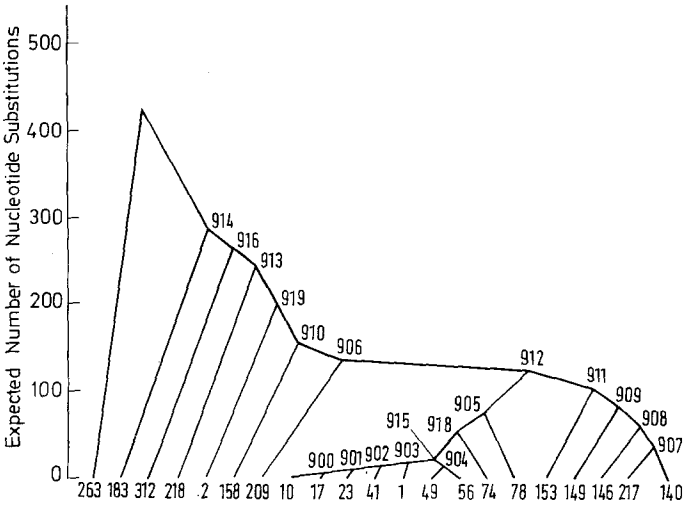

Fig. 1. Hypothetical evolutionary tree used for simulation studies. This tree mimics a part of Goodman et al.'s (1975) evolutionary tree of the globin family. The numbers given at the end points and nodal points of the tree refer to the contemporary and ancestral species

Using pseudorandom numbers, we produced a hypothetical cistron composed of a random sequence of 300 nucleotides (100 codons), excluding the nonsense codons. The first half of these nucleotides were subjected to random mutations, whereas the remaining half were assumed to be invariable. The number of mutations introduced for a given evolutionary interval (link) followed the Poisson distribution with a given expected number. This expected number is given in the ordinate of Figure 1. In this figure the length of each link has nothing to do with the expected number. To get the expected number one has to refer to the scale given at the left end. For example, the expected number of mutations for the link 906—912 is 10. In practice, the number of mutations for a given link (evolutionary time) was determined by a pseudorandom number following the Poisson distribution. Once the number for a link was determined, the mutations were distributed at random over the entire variable part of the cistron. When a particular nucleotide was hit by a mutation, it changed to one of the three remaining nucleotide types with probability 1/3. However, when a mutation resulted in a nonsense codon, the mutation was discarded and another mutation was generated at random at one of the 150 variable nucleotide sites. Thus, the total number of mutations for a given link was not affected by nonsense mutations.

We recorded the nucleotide sequences for all nodal points and contemporary "species", from which the DD values were computed. The actual number of nucleotide substitutions for each link, including synonymous mutations, was also recorded. For augmenting the DD values, we followed the Goodman et al. procedure. In practice, we used the computer program provided by Dr. M. Goodman. The augmented distance thus obtained was compared with the actual number of nucleotide substitutions (TD, true distance). This process was repeated four times.

The values of DD, AD, and TD for representative links are presented in Table 1. It is seen that when the true distance is 50 or less the AD value is close to it, indicating that the augmentation is sufficiently accurate. However, as the true distance increases, AD tends to be larger than TD. This tendency of overaugmentation is particularly high when TD is around 180 and the overaugmentation sometimes amounts to about 70 percent. When TD becomes extremely large, however, AD underestimates TD.

The above results indicate that, at least for certain values of TD, AD clearly overestimates the number of nucleotide substitutions. The reason for this can be seen by considering the process of augmentation for the link 910–158 in Replication 1. In this replication the DD for this link is 82 and the same as that of path 914–217. Since this path satisfies the augmentation rules, it is used for augmenting the DD for the link 910–158. The path includes 11 links and the sum of AD's for these links is 255. Therefore, the AD for the link 910–158 is 255, as shown in Table 1, and the AF is $255 - 82 = 173$. Note that this AF is already larger than the TD (144) for this link.

The amount of overaugmentation depends considerably on the number of nucleotide substitutions per site. In Table 1 we noted that an extreme overaugmentation occurs when TD is around 180. At this level DD does not increase very much even if TD increases, since it is close to the equilibrium level. In the present case the expected equilibrium value of DD is 112.5, since the theoretical equilibrium value of DD per nucleotide is approximately 3/4 and in our model 150 nucleotides are variable. When DD is close to the equilibrium value, a link may by chance have a DD that is identical with that of a path with many links, as in the above example. In this case a gross overaugmentation is expected to occur. Note, however, that overaugmentation occurs even when TD is about 60, though the extent is not large.

Table 1. Direct distances (DD), augmented distances (AD) and true distances (TD) for representative branches (links) of the evolutionary tree in Figure 1. All these distances include synonymous mutations. These results were obtained by computer simulation

Branch	Rep 1			Rep 2			Rep 3			Rep 4		
	DD	AD	TD	DD	AD	TD	DD	AD	TD	DD	AD	TD
919–910	33	42	40	21	23	35	31	40	31	36	44	46
907–217	37	46	44	32	43	38	26	31	32	30	35	39
918–74	36	45	48	33	44	39	26	31	42	46	66	57
908–146	38	52	49	43	62	54	43	64	52	43	62	55
905–78	51	82	71	54	77	63	40	58	55	41	52	52
909–149	55	96	81	53	76	67	45	66	67	54	81	70
911–153	63	129	93	68	117	97	61	100	87	58	85	86
906–209	65	131	133	70	149	107	76	154	117	71	119	113
910–158	82	255	144	78	158	152	81	177	140	85	253	151
919–2	90	263	192	86	225	171	88	239	186	87	255	182
913–218	92	265	270	99	248	233	95	277	221	102	292	259
916–312	98	271	247	116	281	269	101	287	233	96	267	244
914–183	88	261	252	86	225	279	87	205	262	108	298	286
914–263	105	278	522	104	269	558	93	244	551	110	300	535

The degree of augmentation also depends on the tree topology. In general, if a tree has many branches or links, the chance of overaugmentation is high. This is because in this case there are many paths having the same DD value as that of the link under augmentation.

Table 1 shows that the AD value for the link 914–263 is a serious underestimate of TD. This seems to be due to the fact that in all of the four replications the DD value for this link is not very high because of the saturation of mutations and by chance it did not happen to be the same as the DD value of any path that has many links. Theoretically it is possible for the AD value for the link 914–263 to be close to its expected value 560 or much larger than this. This would happen if the DD for this link is largest and equal to that of the path 183–914–912–10 or some other similar path with many links. In practice, the probability of occurrence of this event seems to be very small in the present tree.

It is now clear that the degree of augmentation depends considerably on chance. Even if the true distance is more or less the same for a group of links, the AD value can be drastically different. This is seen from the link 910–158 in Table 1. In all four replications the TD for this link is 140 to 152, but the AD varies from 158 to 255. This indicates that the estimate obtained by Goodman et al.'s augmentation method has a large variance.

The present study clearly shows that overaugmentation can occur in Goodman et al.'s algorithm. Namely, although the TD value increases linearly with evolutionary time (of course, with stochastic errors), the increase of AD is nonlinear. We are then interested in the question: does this sort of overaugmentation occur even when nucleotide sequences are estimated from amino acid sequences by using the maximum parsimony method? As mentioned earlier, the answer to this question depends considerably on the number of species used and the tree structure. With a small tree such as that used in our simulation, overaugmentation is unlikely to occur if the maximum parsimony method is used together with the augmentation procedure, since the tree is too sparse to give reasonably good estimates of nucleotide sequences and thus the DD values would be grossly underestimated when TD is large. In fact, Czelusniak et al. (1978) have shown that this is exactly the case. However, if the number of species used is large, *and the relative values of DD's (obtained by the maximum parsimony method) to TD's are more or less the same as those in our simulation,* then overaugmentation would occur.

In our simulation we used the simplest form of nucleotide substitution, i.e. the Poisson process, where the variance of the number of substitutions is equal to the mean. In practice, however, nucleotide substitution is known to be typically nonrandom (e.g. Nei, 1975), and the variance of the rate of nucleotide substitutions seems to be about two times larger than the mean (Ohta and Kimura, 1971; Langley and Fitch, 1974). This is expected to increase the tendency of overaugmentation, since overaugmentation is caused by random fluctuation of the number of nucleotide substitutions. In this connection it should be noted that the tree estimated by the maximum parsimony method may be incorrect with a high probability if a large number of closely related species are used. It is then possible that this incorrect tree itself contributes

to overaugmentation to some extent, since it would inflate the variance of DD. However, in order to see whether overaugmentation really occurs when the maximum parsimony method is used together with the augmentation procedure, a large scale of simulation seems to be necessary, which we are not ready to undertake.

Nevertheless, two things are clear from our study. 1) Goodman et al.'s augmentation method introduces a systematic error, the difference between AD and TD depending on the value of DD. 2) The amount of augmentation is affected considerably by chance. These findings provide a warning against an uncritical use of their method in the study of the rate of nucleotide substitution in evolution. At the present time the general statistical properties of the augmented distance are quite unclear. How the mean and variance of this distance (for a particular link) is affected by tree structure and the matrix of DD values is virtually unknown. It is very important to clarify these properties before it is used extensively. They can be studied by computer simulation, preferably by using a more realistic model than ours.

In a realistic simulation we must follow both maximum parsimony method and augmentation method as they are actually used. Czelusniak et al.'s (1978) simulation is not complete, since they have not estimated the topology of the tree but used the known tree. If we estimate the topology from contemporary amino acid sequences, as is required in the actual maximum parsimony method, it would vary from replication to replication. In this case a particular link (for example, link 910–906 in Fig. 1) may not be represented in all replications. Nevertheless, it is possible to study the conditional mean and variance or even the conditional distribution of the distance of a particular link by eliminating replications in which this link is not represented. At any rate, it is important to realize that the distance of each link has its own distribution due to the stochastic nature of nucleotide substitution and the distribution is affected by the number of species used and the tree structure. Only through a large number of replications can the distribution of the distance of each link be clarified. Czelusniak et al.'s argument about the limiting case of the dense tree where the DD value of each link is 0 or 1 does not constitute any proof against the possibility of overaugmentation, since we claim that overaugmentation may occur only when the DD value is large. In fact, the limiting case would almost never occur in practice.

Czelusniak et al. seem to be happy when they discover that in their computer simulation the AD value is a little closer to the true value than the Poisson correction estimate. In our view, however, this does not support their method at all. When the number of nucleotide substitutions per site is large, "all the conditions of the Poisson model" are *not* met, unlike their statement. In this case they should have used the method developed by Jukes and Cantor (1969) and Kimura and Ohta (1972).

Acknowledgments. We thank Dr. Morris Goodman for sending us the computer program of their augmentation method and for his helpful discussions. This work was supported by the National Science Foundation Grant number DEB 76-06069 and the National Institute of Health Grant number GM 20293.

References

Czelusniak, J., Goodman, M., Moore, G.W. (1978). J. Mol. Evol.
Goodman, M., Moore, G.W., Barnabas, J., Matsuda, G. (1974). J. Mol. Evol. **3**, 1
Goodman, M., Moore, G.W., Matsuda, G. (1975). Nature **253**, 603
Jukes, T.H., Cantor, C.H. (1969). In: Mammalian protein metabolism, (H.N. Munro, ed.), pp. 21–123. New York: Academic Press
Kimura, M., Ohta, T. (1972). J. Mol. Evol. **2**, 87
Langley, C.H., Fitch, W.M. (1974). J. Mol. Evol. **3**, 161
Moore, G.W., Barnabas, J., Goodman, M. (1973). J. Theor. Biol. **38**, 459
Nei, M. (1975). Molecular population genetics and evolution. Amsterdam: North-Holland
Nei, M., Chakraborty, R. (1976). J. Mol. Evol. **7**, 313
Ohta, T., Kimura, M. (1971). J. Mol. Evol. **1**, 18

Received August 29, 1977; Revised January 3, 1978