J Mol Evol (1981) 17:121–122

Journal of Molecular Evolution

Doubt About Studies of Globin Evolution Based on Maximum Parsimony Codons and the Augmentation Procedure*

Motoo Kimura

National Institute of Genetics, Mishima 411, Japan

Summary. Both the maximum parsimony method of codon assignment and the augmentation procedure, as used by Goodman and his associates, are liable to serious errors and therefore should not be used for studying molecular evolution in general, and globin evolution in particular.

Key words: Globin evolution – Evolutionary distance estimation

In response to my criticism (Kimura, this issue) of the work of Goodman and his associates on globin evolution, Goodman (this issue) now says that the maximum parsimony codons which are presented extensively in Goodman et al. (1974) are "simply ambiguous rather than wrong." According to him, the letter U in the third position of their maximum parsimony codons stands for either U or C, and in many cases any of the four bases U, C, A or G. Similarly, G in the third position stands for either A or G, and in many cases any of the four bases. He claims that this is apparent as stated in the footnote to Table 1 of Goodman et al. (1974). The situation is not clear to me when a maximum parsimony codon ends with letter A. In the case of rabbit α globin, there are four maximum parsimony codons ending with A, and all four are in fact wrong.

On the whole, I found the footnote rather ambiguously written (i.e., not clear enough to indicate the ambiguous nature of Goodman et al.'s codons). I simply could not imagine that Goodman et al. (1974) would use the letters U, C, A and G instead of such letters as R for purine, Y for pyrimidine, and X or N for any of the four, when the actual bases are unknown or only known ambiguously. Now, I am puzzled as to what was the purpose of publishing enormous tables, filling page after page, to present maximum parsimony codons if Goodman (the first author) knew that these codons were so ambiguous.

Nevertheless, under this circumstance, my criticisms of the maximum parsimony codons lose some force. This does not mean, however, that Goodman's maximum parsimony operation is a valid scientific procedure. In fact, the fundamental fallacy of maximum parsimony codons has been revealed in a dramatic way by a recent study by Holmquist (1979).

Previously, Tateno and Nei (1978) performed Monte Carlo experiments of molecular evolution with a computer, simulating the process of divergence of nucleotide sequences by randomly accumulating mutational changes (each hypothetical sequence consisted of 300 nucleotides, corresponding to 100 codons, the first half of which were assumed to be variable and the remaining half invariable). Starting from a common ancestor, phylogenetic trees involving 21 contemporary sequences were produced. What is important is that, in these simulation experiments, unlike the evolutionary processes in nature, the exact topology, the numbers of accumulated nucleotide substitutions in each branch of the phylogenetic tree, and all the ancestral sequences at the branch points are known. Holmquist obtained the data of these simulation experiments from Nei and Tateno, and he sent the 21 contemporary sequences (after translating them into protein sequences) to Goodman, asking him to obtain the maximum parsimony solution including both the topology and reconstructed ancestral amino acid sequences. (This was called the "inferred topology" solution.)

Holmquist also asked Goodman to provide him with the ancestral amino acid sequence reconstructions, given

*Contribution No. 1351 from the National Institute of Genetics, Mishima, 411 Japan

0022–2844/81/0071/0121/$01.00

the correct topology. (The solution thus obtained was designated as the "known topology" solution.) These amino acid sequences obtained by Goodman were then compared with the correct ancestral sequences. It turned out that there are numerous errors in the amino acid sequences inferred by the method of maximum parsimony. Particularly noteworthy is the high rate at which errors accumulate as the distances between the nodal sequences and the contemporary sequences increase. For nodes 200 or more nucleotide substitutions apart, the error rate becomes more than 80%. Actually, errors are so numerous in remote ancestral reconstructions, even when the correct topology is supplied, that the reconstructed sequences are totally worthless (they even contain chain terminating codons!). It is clear now that any discussion which involves very early stages of globin evolution based on maximum parsimony reconstruction is meaningless.

It might be argued that the numbers of nucleotide substitutions involved between various globin sequences studied by Goodman and his associates are in most cases much less than 200 per 50 variable codons, and therefore, the errors can be much less. I still think, however, that possible errors are too numerous to place much reliance on maximum parsimony plus augmentation operations when a reasonably remote ancestral sequence is involved. For example, according to Holmquist and Pearl (1980), 618 nucleotide substitutions separate the α and β globins of the rabbit, and the estimated number of "varions" is 111. This corresponds to about 278 nucleotide substitutions per 50 varions. This means that the common ancestral sequence of α and β globins corresponds roughly to point F or G in the phylogenetic tree illustrated in Fig. 1 of Holmquist (1979). From Table 2 of the same paper, we find that the errror amounts to more than 60%; the reconstructed parsimonious amino acid sequence of the common ancestor of α and β globins is certainly meaningless. More generally, according to Holmquist (1979, see Eq. 13), the percent error in the ancestral amino acid sequence is equal to 0.48 times TD, where TD is the true distance from the present in nucleotide replacements.

I am also doubtful about the validity of the augmentation procedure that has been used extensively by Goodman and his associates. To show my point, I shall take, as an example, the augmented distance from the common ancestor of α and β globins down to the contemporary human α globin. If we trace the relevant paths in Fig. 1 of Goodman et al. (1975), this number is 156. Their phylogenetic tree was based on 55 contemporary globins. Goodman now presents a part of the genealogical tree (see Fig. 1 of Goodman, this issue) constructed by using 159 globin sequences. The corresponding distance from this new Fig. 1 is 29 + 123 + 67 or 219. This new estimate is some 40% higher than the previous estimate of 156. On the other hand, the augmented total distance from the common ancestor of lamprey globin and α-β hemoglobins down to the contemporary lamprey globin was 145 when estimated by Goodman et al. (1975), but, the corresponding new estimate is 78 + 95 or 173. In this case, the new estimate is higher by only 19%. The purpose of the augmentation is to make corrections for hidden mutant substitutions in sparsely populated branches. The above results suggest, however, that it is not serving very well the purpose for which it was originary devised, for the line leading to the human must represent one of the most densely populated branches, while the line leading to lamprey must be one of the least densely populated. Criticisms of the augmentation procedure have already been made by Tateno and Nei (1978). It seems to me that this procedure brings a certain systematic bias into the estimation of the number of mutant substitutions, and therefore it is not a very reliable method. I am also quite skeptical of Goodman et al.'s (1974, 1975) facile explanations of various evolutionary amino acid substitutions exclusively in adaptive ("selectionist") terms. Although plausible, these explanations lack quantitative analysis based on population genetics theory, and we must take such explanations with a grain of salt.

I conclude that the work of Goodman and his associates based on the maximum parsimony method and augmentation procedure, although often accepted as a most thorough study of globin evolution, actually contains so many problems that its validity has to be seriously questioned.

Acknowledgement. I thank Drs. J.F. Crow, Y. Tateno, N. Takahata and K. Aoki for stimulating discussions and helpful suggestions in composing this manuscript.

References

Goodman M (1981) Globin evolution was apparently very rapid in early vertebrates: A reasonable case against the rate-constancy hypothesis. J Mol Evol (this issue)

Goodman M, Moore GW, Barnabas J, Matsuda G (1974) J Mol Evol 3:1–48

Goodman M, Moore GW, Matsuda G (1975) Nature 253:603–608

Holmquist R (1979) J Mol Biol 135:939–958

Holmquist R, Pearl D (1980) J Mol Evol 16:211–267

Kimura M (1981) Was globin evolution very rapid in its early stages?: A dubious case against the rate-constancy hypothesis. J Mol Evol (this issue)

Tateno Y, Nei M (1978) J Mol Evol 11:67–73

Received, January 5, 1980