J Mol Evol (1995) 41:254–255

JOURNAL OF MOLECULAR EVOLUTION

Point Counterpoint

On the General Reversible Markov Process Model of Nucleotide Substitution: A Reply to Saccone et al.

Ziheng Yang

Institute of Molecular Evolutionary Genetics, Pennsylvania State University, 328 Mueller Laboratory, University Park, PA 16802, USA

Before giving any comments on the model of Lanave et al. (1984), I would like to clarify a few relevant concepts—that is, the homogeneity, stationarity, and reversibility of a Markov process. A Markov process is said to be homogeneous if its rate matrix is independent of time; for phylogenetic analysis using DNA sequence data, this means that the pattern of nucleotide substitution remains the same along different parts of the tree. A homogeneous process has an equilibrium distribution, which is also the limiting distribution when time approaches infinity. Stationarity means that the process is at equilibrium—that is, nucleotide frequencies have remained more or less the same during the course of evolution. When base frequencies are quite different in different species, these assumptions are clearly violated. The two assumptions are made by virtually all methods or models currently used in phylogenetic analysis, more for mathematical tractability than for their accurate reflection of molecular sequence evolution. The rate matrix for a Markov process, $\mathbf{Q} = \{Q_{ij}\}$, where Q_{ij} is the rate of substitution from nucleotide i to j, is restricted by a mathematical requirement that the row sums are all zero (Grimmett and Stirzaker 1992, pp 239–246); this is analogous to the requirement that the row sums of the transition probability matrix over time t, $\mathbf{P}(t) = \exp(Qt)$, are all one. So a model which requires the assumptions of stationarity and homogeneity and which places no restriction on the structure of $\mathbf{Q}$ involves 12 (= 16 – 4) free parameters. [My counting of this number in Yang (1994) was 11, as only the relative rates were considered.]

Reversibility is an extra restriction placed on the structure of $\mathbf{Q}$—that is, $\pi_i Q_{ij} = \pi_j Q_{ji}$, or $\Pi\mathbf{Q} = \mathbf{Q}\Pi$, where π_i is the equilibrium frequency of nucleotide i and $\Pi = \text{diag}\{\pi_T, \pi_C, \pi_G\}$. This restriction reduces the number of free parameters by three and the model then has 9 parameters. In many cases reversibility leads to mathematical tractability, especially for the maximum-likelihood method which performs simultaneous comparison of all sequences (Felsenstein 1981, see also Keilson 1979). To the best of my knowledge, Tavare (1986) was the first to employ the general reversible process model (i.e., a model that places no other restrictions on the structure of $\mathbf{Q}$ except for reversibility, to nucleotide sequence analysis). Yang (1994, see also Zharkikh 1994) improved the notation to facilitate the implementation and interpretation of the model.

Lanave et al. (1984) unambiguously stated that their model did not make any restrictions about the structure of the rate matrix, and so their model was supposed to be a 12-parameter model, the same as the "unrestricted" model of Yang (1994) but different from the 9-parameter general reversible process model. The authors, however, made a transformation of the eigenvectors of the rate matrix (their equation 11), which, without any extra assumption, is unwarranted. This implicit assumption, which is both sufficient and necessary, is reversibility, and was pointed out by Gillespie (1986) and shown in more detail by Yang and Goldman (1994) and Zharkikh (1994). Lanave et al. (1984) ignored this fact, and the estimated rate matrices by their method (Tables 5 and 6 in Lanave et al. 1984) did not satisfy the mathematical requirement, mentioned above, that the sums of rows

(columns in their notation) in the rate matrix should all be zero, although the differences are small.

Therefore Saccone et al. (1995) seem to have justification in suggesting that Lanave et al. (1984), after the corrections made by Gillespie (1986), Yang and Goldman (1994), and Zharkikh (1994), was the first application of the general reversible process model to DNA sequence analysis. However, it does not seem to be a serious mistake to suggest that Lanave et al. (1984) attempted to use a more general model without the reversibility restriction, and their mathematical treatment involved inaccuracies. My discussion (Yang 1994) concerns the mathematical technicalities of these models, which are necessary for their correct implementation in the maximum-likelihood framework, where ambiguous formulations cannot be allowed. It does not concern the performance of the method of Lanave et al. (1994) in estimating sequence distances, for which the method was designed. In this regard, extensive simulations performed by Zharkikh (1994) suggest that Lanave et al.'s (1994) method is quite stable in estimating pairwise distances. As one might expect, the inaccuracies in their treatment have not led to serious bias in distance estimates. A more serious problem with their method might be that it is based on the assumption that substitution rates are constant across nucleotide sites, an assumption which is known to produce severe underestimation of distances when there exists rate variation at sites.

It may be worthwhile to mention that the reversible process, as a special class of Markov processes, is well studied in statistics. Keilson (1979) has discussed many nice mathematical properties of the reversible process and provided a proof that the rate matrix for a reversible process has only real eigenvalues and eigenvectors, as claimed but not proved by Yang (1994). The proof makes use of a matrix, $\Pi^{1/2}\mathbf{Q}\Pi^{-1/2}$, where $\Pi^{1/2} = \text{diag}\{\sqrt{\pi_1}, \sqrt{\pi_2}, \sqrt{\pi_3}, \sqrt{\pi_4}\}$, which is symmetrical under the reversibility assumption and thus has only real eigenvalues, and which has the same eigenvalues as **Q**. This proof also suggests that efficient algorithms can be used to calculate the eigenvalues and eigenvectors of the rate matrix for a general reversible process model.

I believe that my discussion (Yang 1994) of the model of Lanave et al. (1984) does not serve to undermine the interesting work of the authors, and hope that these exchanges of correspondence will help to promote the use of more realistic models in phylogenetic analysis.

Acknowledgments. I wish to thank Bill Bruno for his independent proof that the rate matrix for a general reversible Markov process has only real eigenvalues and eigenvectors, and Andrey Rzhetsky for the reference on reversible Markov processes (i.e., Keilson 1979).

References

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Gillespie JH (1986) Rates of molecular evolution. Ann Rev Ecol Syst 17:637–665

Grimmett GR, Stirzaker DR (1992) Probability and random processes. 2nd ed. Clarendon Press, Oxford

Keilson J (1979) Markov chain models: rarity and exponentiality. Springer-Verlag, New York

Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86–93

Saccone C, Lanave C, Pesole G, Preparata G (1995) The reversible stationary Markov process for estimating the pattern of nucleotide substitution: a response to Ziheng Yang. J Mol Evol 41:253

Tavare S (1986) Some probabilistic and statistical problems on the analysis of DNA sequences. In: Lectures in mathematics in the life sciences, vol 17, pp 57–86

Yang Z (1994) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111

Yang Z, Goldman N (1994) Evaluation and extension of Markov process models of nucleotide substitution. Acta Genetica Sinica 21: 17–23

Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. J Mol Evol 39:315–329