PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The Estimate of Total Nucleotide Substitutions from Pairwise Differences is Biased

W. M. Fitch

Phil. Trans. R. Soc. Lond. B 1986 **312**, 317-324
doi: 10.1098/rstb.1986.0010

Phil. Trans. R. Soc. Lond. B **312**, 317–324 (1986)
Printed in Great Britain

The estimate of total nucleotide substitutions from pairwise differences is biased

By W. M. Fitch

Department of Physiological Chemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706, *U.S.A.*

A nomographic method is presented that estimates the number of nucleotide substitutions since the common ancestor of two nucleotide sequences with no assumption about the proportion of transition and transversion substitutions except that it is constant over time. Of two previous methods of estimating this number, that of M. Kimura (*Proc. natn. Acad. Sci. U.S.A.* **78**, 454–458 (1981)) obtains the same result, and is thus confirmed by this work, while that of W. M. Brown, E. M. Prager, A. Wang & A. C. Wilson (*J. molec. Evol.* **18**, 225–239 (1982)) does not get the same result. The method presented here also obtains the fraction of all substitutions that are transitions. If one has three or more homologous sequences to compare, one can test the validity of the model by examining the constancy of the estimated proportion of substitutions that are transitions across the various pairs of sequences in a simple visual way. The method is general for any pair of mutually exclusive nucleotide substitutional categories, not just transitions and transversions. Mitochondrial data provide evidence that, for this and probably other current models correcting for superimposed substitutions, one or more of the underlying assumptions is incorrect. This is because there is some unknown systematic bias affecting this evolutionary process. It is suggested that at least part of the bias arises from incorrectly assuming that all sites are variable. In the absence of evidence that this bias is not present in other data, all estimates of the number of substitutions based upon pairs of sequences and current methods of estimating superimposed substitutions at a single site should be viewed as uncertain.

Introduction

It is a well recognized phenomenon that in certain circumstances, some classes of nucleotide substitutions are more frequent than others. For example, in the third coding position, transitions are more frequent than transversions, even in codons that are four-fold degenerate (Fitch 1980). This inequality is expected for the totality of third positions because only transitions are silent (that is, they do not change the encoded amino acid) for the two-fold degenerate codons. This inequality is not necessarily expected for the four-fold degenerate codons unless one invokes a specific mechanism of mispairing during replication, such as tautomerization, that specifically favours transitions (Topal & Fresco 1976). What is perhaps less generally appreciated, as Brown *et al.* (1982) have astutely pointed out, is that, as two sequences that are not otherwise constrained diverge, their observed *differences* in homologous positions will ultimately tend to show an excess of transversion differences over transition differences, even though most of the *substitutions* (mutations fixed) were transitions. This property of having the two classes of events appear, as time passes, as if all types of substitutions were more equally probable compared with opportunity, even though the underlying substitutional rates are not equal, will apply to any two mutually exclusive types of substitutional events, not just transitions and transversions. Thus most estimates of their

underlying relative frequency of occurrence from raw observational counts of the types of differences are systematically biased to underestimate that frequency as well as to underestimate the total number of substitutions. It is the purpose of this paper to provide a simple nomographic procedure to correct this bias. In this respect it accords with the purposes of work by Kimura (1981) and Brown *et al.* (1982). This study gives results in agreement with those of Kimura but not those of Brown *et al.*

In addition, however, the present study estimates the fraction of substitutions that are transitions. This fraction, in contrast to the reasonable expectation that it has some average rate over evolutionary time, is shown to have smaller and smaller estimates as evolutionary divergence increases, at least in mitochondrial DNA. This shows that a systematic bias remains in the data that is not 'accounted for' by any commonly used procedure for estimating rates of nucleotide substitutions by examining extant nucleotide sequences.

Methods

One wishes to estimate of the total number of nucleotide substitutions, and the fraction of them that are transitions, from observations on the number of transition and transversion differences between two sequences. The method used here is different in its derivation from any previously given, but the result provides values identical to those of Kimura (1981), thus confirming both methods, given the assumptions. For this reason, the method is relegated to the appendix.

Results

The following analysis assumes that one has compared two homologous sequences and observed d differences per site, of which d_v are transversions/site and d_s are transitions/site. Hence $d = d_v + d_s$. The task is to obtain r, the total rate of substitution per site and v and s, the fraction of r that are transversions and transitions, respectively. Hence $v + s = 1$. The computer program was run using values of $r = x^2$ where x ranges from 0.1 to 2.0 in increments of 0.1 (x is a dummy variable to get r to increase exponentially rather than linearly). Values of s ranged from 0.20 to 1.00 in increments of 0.05 plus the value of 0.99. The two pieces of information that one would normally possess, and that seem most useful in examining data nomographically, are the fraction d, of sites, that are different and the ratio of observed transversions to transitions, d_v/d_s. These constitute the axes of figure 1. Values of d and d_v/d_s were computed for the values of r and s noted, plotted on the figure, and lines drawn to connect points of constant r (horizontal curves) and constant s (vertical curves).

Given any pair of observed values of d and d_v/d_s, one finds the point on the figure and interpolates to get the values of r and s that would best account for the data. If for example, $d = 0.10$ and $d_v/d_s = 1.09$, then the best estimates of the underlying process are that there have been 0.11 substitutions per nucleotide (r) and 0.47 of all substitutions were transitions (s).

The method was applied to data on mitochondrial DNA, kindly supplied by Dr Ellen Prager, and the results are shown in figure 1 for all 21 pairwise comparisons.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

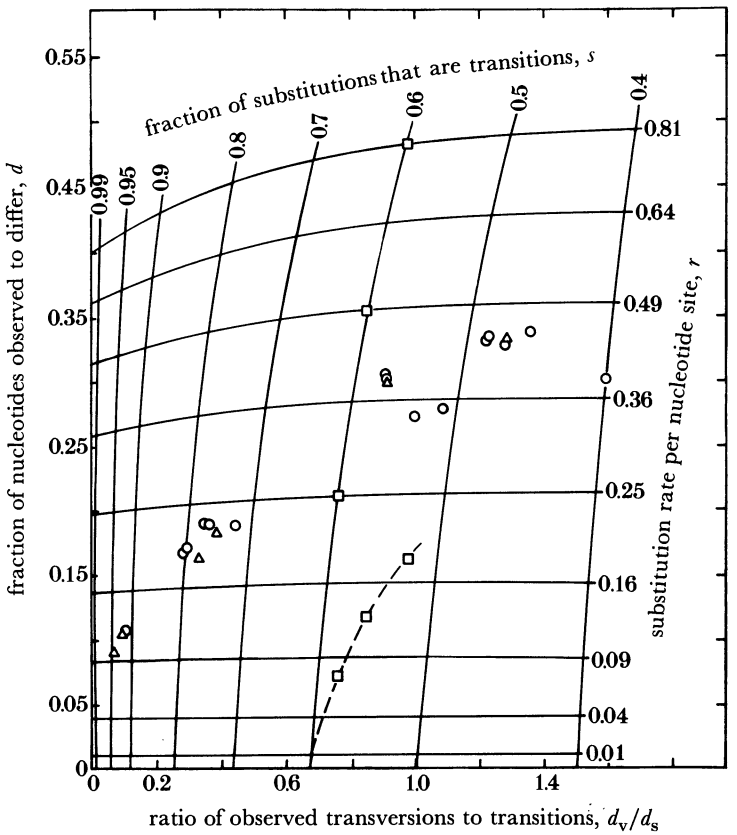

FIGURE 1. Real and hypothetical data plotted on the nomogram. The ordinate and abscissa are the variables d and d_v/d_s, directly observable in the data. These two values determine a point on the graph that relates to the lines that represent a second set of variables, the rate of substitution per nucleotide site (r) and the fraction of those substitutions that are transitions (s) and whose value may be estimated by interpolation. The squares represent points showing the effect of some sites being invariable (see text). The other 21 points are for pairs of taxa based on 896 nucleotides in mitochondrial DNA as provided by E. Prager. They appear from the lower left to the upper right in clusters of increasing size from 1 to 6. Each group includes one point, represented by a Δ, that is for the human and the taxon which is common to comparisons in that group. These taxa are, respectively, chimpanzee, gorilla, orang-utan, gibbon, mouse and cow. A group of n points is a comparison of the nth taxon in this sequence with all the taxa to its left plus humans.

DISCUSSION

General properties

All curves of r and s are concave upward and to the left except for the strictly vertical line at $d_s = 0.\bar{3}$ ($d_v/d_s = 2.0$) and the x-axis at $d = 0$. Hence all values of d are underestimates of r except at $d = 0$ (at time 0!) and all observed fractions of differences that are transitions underestimate s except when $d_v/d_s \geqslant 2.0$. The latter tendency was noted by Brown *et al.* (1982). Moreover, all curves converge to $d = 0.75$, $d_v/d_s = 2$ as noted by Kimura (1981). That is, with the occurrence of a sufficient number of substitutions per site, two sequences will come to differ in three quarters of their sites and the ratio of transversion to transition differences will approach 2.00 (that is, they look like otherwise random sequences) regardless of the actual rate of transition substitutions relative to transversion substitutions. However, more complicated base substitutional schemes need not lead to a ratio of two (Holmquist 1982).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Comparison with other methods

Table 1 compares the results of this method with those of Kimura (1981) and of Brown *et al.* (1982). The method given here agrees exactly with that given by Kimura except for ± 1 in the second significant digit which represents the limits of interpolation on a larger, complete version of figure 1. Since the two methods were derived independently, and by rather dissimilar approaches (continuous compared with discrete), their agreement verifies the mathematical correctness of each.

TABLE 1. SUBSTITUTIONS PER NUCLEOTIDE BY THREE METHODS

	Brown *et al.*	Kimura	Fitch	s
chimpanzee (*Pan troglodytes*)	0.09	0.10	0.10	0.94
gorilla (*Gorilla gorilla*)	0.11	0.11	0.11	0.92
orang-utan (*Pongo pygmaeus*)	0.25	0.18	0.19	0.78
gibbon (*Hylobates lar*)	0.31	0.21	0.21	0.75
cow (not given)	0.88	0.38	0.39	0.57
mouse (not given)	1.24	0.44	0.44	0.47

Substitutions per nucleotide are for each of the taxa on the left compared with man. Sequences are the 896 nucleotides studied by Brown *et al.* (1982) and include the C-terminus of *URF4*, three t-RNAs and the N-terminus of *URF5*, all from mitochondrial DNA. The last column is the fraction of the substitutions calculated to be transitions. The values relating to cow and mouse were based on information kindly supplied by E. Prager (personal communication) and who calculated the values according to the method of Brown *et al.* (1982). The column labelled Kimura contains values obtained by his method (1981). The orang-utan, cow and mouse have, respectively, one less, two less and three extra nucleotides, compared with the others. These data appear in figure 1 as the points marked by a Δ.

The method of Brown *et al.* (1982) gives results that do not agree with those of Kimura, nor with mine. I can see no difference in our stated assumptions that would account for this discrepancy but, on the contrary, believe that all three parties are attempting to make the same correction. I believe that the difference lies in the attempt to estimate transition rates on the basis of transition differences independently of the transversions. The Jukes & Cantor formulation (1969), $\Lambda = -3/4 \ln (1-4\lambda/3)$, can easily be seen to be divisible into parts where λ_i are for different sites and the weighted average for the different Λ_i are used to give an overall Λ for all sites combined. It is not at all clear that it is legitimate to partition the transition and transversion processes at a single site (Λ_i and Λ_v), as Brown *et al.* (1982) do, to get an overall Λ for both processes combined. This doubt is applicable as well to the formulations of Perler *et al.* (1980) although the magnitude of the error may be less in their case to the extent that four-fold and two-fold degenerate codon positions persist in the degree of their degeneracy.

As the taxa become more distantly related, the disparity between the Λ_i and Λ_v of Brown *et al.* becomes greater and so the method of averaging these two values could also contribute to the difference, especially since they seem to bracket the values obtained by Kimura and by this method. Their averaging method is shown by their equation (6) which is more complicated than necessary since it readily simplifies to

$$= [(1+2f_i)\,\Lambda_i + (1+2f_v)\,\Lambda_v]/4,$$

where f_i and f_v are the fraction of the differences that are transitions and transversions respectively (or d_s and d_v in this work). This is a peculiar weighting because in the case of recently diverged species such as human–chimpanzee with f_i close to one, it weights Λ_i nearly

three times as much as Λ_v, while in the limit where $f_v/f_i = 2$, it would weight Λ_i at only 5/7 of Λ_v. Moreover, if one computes Λ_i and Λ_v for the human–cow sequences (0.21 and 1.50, respectively), one would believe, in complete distinction to the human–chimpanzee data, that the rate of transversion substitutions was more than 7.5 times greater than that for transitions. If this were in fact true, why should one observe more transition than transversion differences? Under the circumstances, it seems that Λ_i and Λ_v are not properly conceived and hence there is no weighting scheme for Λ_i and Λv that in general is likely to yield the correct Λ.

Invariable nucleotide positions

Fitch & Margoliash (1967) demonstrated that nucleotide substitutions producing amino acid replacements in structural genes distribute themselves as if a portion of the codons were invariable with respect to such changes. But if that is true for codons, it must be true for nucleotides as well. How would one observe such an effect?

Consider a hypothetical human gene that is compared with the homologous gene in three successively more distantly related taxa. Suppose further that 0.6 of all substitutions were transitions and that there were 0.25, 0.49 and 0.81 substitutions per variable nucleotide site, respectively. The data, subject to stochastic variation, should in figure 1 fall on the solid line $s = 0.6$, at exactly the three points shown thereon by squares. This assumes that all of the nucleotide sites are indeed variable. But suppose that only a third of the sites are in fact variable. If every hundred variable sites for which this model is correct are accompanied by another 200 sites that are invariable, then all our observed values are still correct except that d will be low by that factor of three, and we will end up plotting the data on figure 1 as those squares connected by the dotted line. The key to recognizing the fact that d is improper because it includes invariable sites is the failure of the points all to lie on the same line of s value, the one indicating the fraction of all substitutions that are transitions.

Knowing the problem, the correction is easy. One simply determines by what number the d values of the data need be multiplied to cause them to lie on the same s line. This does, of course, presuppose that the fraction of all substitutions that are transitions, s, is relatively constant over time.

What do real data show? The data for the 896 mitochondrial nucleotides studied by Brown *et al.* (1982) are plotted in figure 1 along with values comparing the cow and the mouse to the primates. Clearly, they do not fall on a single s value. There are two readily available explanations: (i) the fraction of substitutions that are transitions is quite variable; (ii) the assumptions used in the model are not sufficiently comprehensive to reflect accurately the evolutionary processes in this part of the mitochondrial genome.

With respect to explanation (i), the predicted variability may well be correct, but it is not an adequate explanation because the data do not simply fall at random about some median s value. Rather, they fall in a very well-defined band, indicating that there is some undefined underlying systematic bias causing more distantly related sequences to appear to have had a lower fraction of transition substitutions in their past. This is true even after deliberately correcting for the problem recognized by Brown *et al.* (1982). It is difficult to believe that that bias is a change from predominantly transversion substitutions 100 million years ago to an overwhelming dominance of transitions in recent primate lineages. Explanation (ii) therefore appears the more reasonable.

The implications of these results are considerable, as they imply that the models upon which

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

substitution rates have been estimated from pairwise sequence differences may be inadequate for that purpose. Of course, the results may be peculiar to this segment of mitochondrial DNA, but there is, as yet, no evidence that the bias present here does not occur generally. It certainly has not been ruled out where such methods have been applied. Given that caveat with respect to generality, we may make several inferences.

The first inference is that, since none of the data points fall on the line $d_v/d_s = 2.0$, no method that ignores differences in transition and transversion rates, that is, that assumes the ratio is indeed 2.0, can be expected to estimate accurately the number of nucleotide substitutions separating two genes. This includes the methods of Jukes & Cantor (1979), Kimura & Ohta (1972) and Perler *et al.* (1980), among others.

The second inference is that, since none of the data points fall on any one line, even approximately, no method that ignores the problem of invariable sites (or, more generally, of sites with different rates) can be expected to estimate accurately the number of nucleotide substitutions separating two genes. This includes the three methods of the previous paragraph plus the methods of Kimura (1981), Brown *et al.* (1982) and this paper, among others.

But we have shown at the beginning of this section how to allow for a fixed fraction of the sites being invariable. An important aspect of the Brown *et al.* (1982) data plus the additional data supplied by Prager is that no adjustment based upon d being underestimated will bring the mitochondrial data onto, or respectably near, a single s value line. Thus the third inference is that since no adjustment of d can cause the data points to fall on any one line, even approximately, no method currently available can be expected to estimate accurately the number of nucleotide substitutions separating two genes. These include all of the preceding methods plus the methods of Holmquist & Pearl (1980) and this paper.

Covariotides

After Fitch & Margoliash (1967) discovered that amino acid replacements in cytochrome *c* distributed themselves as if some positions might be invariable, Fitch & Markowitz (1970) then demonstrated that the estimates of the number of invariable positions increased as the range of taxa narrowed, suggesting that a considerable number of the positions may be invariable in any one taxon but that as replacements are fixed, those positions change. This leads naturally to the concept of concomitantly variable codons, or covarions for short. Fitch (1971) showed that only about two-thirds of the 60 plus invariable amino acid positions of cytochrome *c* estimated for fungi and metazoans could be common to both groups. This lent further support to the covarion concept. The interpretation in terms of the numbers of invariable positions may be overly narrow in that all that is really necessary is that there be significantly different rates of replacements at different sites. Nevertheless, the concept, that where the next nucleotide substitution or the next amino acid replacement may be fixed must be a function of what changes have preceded it, can hardly be denied in an evolutionary framework, and its applicability to nucleic acid data is surely as reasonable. Hence, covariotides or concomitantly variable nucleotides may be considered as a component to the perplexing problem of the location of the mitochondrial DNA points in figure 1.

The relatively horizontal march of the mitochondrial data across figure 1 would be expected under the following somewhat forced model. Let some fraction of the variable sites permit only transition substitutions and be able to evolve rapidly and thus be sooner saturated with changes. Let another fraction of the variable sites, considerably larger than the former, constitute the

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

remaining variable sites which evolve much more slowly but permit both transitions and transversions. These two fractions might be exemplified by the third codon position of two-fold degenerate amino acids and the first and second codon positions of all amino acids. Since about two-thirds of the mitochondrial data examined here are coding sequences, this model is at least possible, and it may account for the large proportion of transition differences in recently diverged sequences. It can certainly be tested, and it suggests that the cautionary remarks made about all current methods may prove to be unnecessary for non-coding sequences. Still, non-coding sequences may possess other problems to plague us. Caution about the robustness of our methods is seldom out of place for, while inconsistencies may demonstrate the invalidity of our assumptions, consistency cannot validate them because you cannot prove a null hypothesis, only reject it.

Thanks are expressed to Ron Niece, Ellen Prager and Allan Wilson for their assistance. This work was supported by NSF grant BSR-8400682.

References

Brown, W. M., Prager, E. M., Wang, A. & Wilson, A. C. 1982 Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. molec. Evol.* **18**, 225–239.

Cantor, C. R. & Jukes, T. H. 1969 Evolution of protein molecules. In *Mammalian protein metabolism* (ed. H. N. Munro), vol. III, pp. 21–132. New York: Academic Press.

Fitch, W. M. 1971 The non-identity of invariable positions in the cytochrome *c* of different species. *Biochem. Genet.* **5**, 231–241.

Fitch, W. M. 1980 Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three beta hemoglobin messenger RNAs. *J. molec. Evol.* **16**, 153–209.

Fitch, W. M. 1981 The old REH theory remains unsatisfactory and the new REH theory is problematical. *J. molec. Evol.* **18**, 60–67.

Fitch, W. M. & Margoliash, E. 1967 A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case. *Biochem. Genet.* **1**, 65–71.

Fitch, W. M. & Markowitz, E. 1970 An improved method for determining codon variability in a gene and its application to the rate of fixations of mutations in evolution. *Biochem. Genet.* **4**, 579–593.

Holmquist, R. 1983 Transitions and transversions in evolutionary descent. *J. molec. Evol.* **19**, 134–144.

Holmquist, R. & Pearl, D. 1980 Theoretical foundations for quantitative paleogenetics. Part III: the molecular divergence of nucleic acids and proteins for the case of genetic events of unequal probability. *J. molec. Evol.* **16**, 211–267.

Kimura, M. 1981 Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. natn. Acad. Sci. U.S.A.* **78**, 454–458.

Kimura, M. & Ohta, T. 1972 On the stochastic model for estimation of mutational distance between homologous proteins. *J. molec. Evol.* **2**, 87–90.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. 1980 The evolution of genes: the chicken preproinsulin gene. *Cell* **20**, 555–566.

Topal, M. D. & Fresco, J. R. 1976 Complementary base pairing and the origin of substitution mutations. *Nature, Lond.* **263**, 285–289.

Appendix

Let r be the rate of substitution per nucleotide over the interval since two homologous sequences being compared had a common ancestor, that interval being twice the amount of time since then. By using the usual Poisson formulation, the probability that any given nucleotide is unsubstituted in that interval is $P_0 = e^{-r}$ and the probability that it has suffered any particular number of substitutions, n, can be obtained by the recursion relation, $P_n = rP_{n-1}/n$. We wish to determine the value of r together with the fractions, s and v, of those

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

substitutions that are transitions and transversions, respectively. Note that $s+v=1$. This is to be done on the basis of the observations of a count of the differences between two sequences that are transitions and transversions. Let d_s and d_v equal the observed fraction of sites in the sequence that differ by transitions and transversions, and let $d=d_s+d_v$ equal the fraction of all sites that are observably different. The problem is to estimate, r, s and v from d_s and d_v. The approach is to use various values of r, s and v to calculate the expected values of d_s and d_v and plot them in a way convenient to get the former from the latter.

To get the expected values of d_s and d_v we need first to express P_n in a way that separates transitions and transversions and then add the separated values over all values of n until n is so large that P_n is negligible. For any value of n and r, $P_n=e^{-r}r^n/n!$ and the distribution of transition and transversion substitutions is obtained by multiplying the right side by the expansion of $(s+v)^n=1$, for which any given term is of the form $n!s^av^b/(a!b!)$ where $a+b=n$. Thus the computer program must loop over all combinations of a and b that add up to n and over all values of n for which P_n is not negligible. Negligibility in this work was achieved by examining all $n<20$.

All of the terms collectively give all possible ways $(a+b)$ that transitions and transversions can occur and they come with attached probabilities (P_n) of their occurrence. All that is needed now is a way of inferring how those combinations of transitions and transversions lead to observational differences in compared sequences. This proves to be relatively easy.

Consider first all terms for which all changes are transitions $(a=n, b=0)$. The first transition changes the nucleotide and a second necessarily returns the nucleotide to its original state and we start all over. Thus all values of $a=n$ lead to an observable transition difference if a is odd and to no observable difference if a is even.

If $b\neq 0 (a\neq n)$, then a becomes irrelevant and only the value of b matters, regardless of the order of the various transitions and transversions. This is because, if b is odd, the two nucleotides must necessarily differ by a transversion, and if b is even, they cannot differ by a transversion. If b is even $(b\neq 0)$ then the question arises whether the last transversion returned the nucleotide to its original state or to the one which was a transitional difference, and whether there was an odd or an even number of subsequent transitions. Since we can know neither of these, we simply assume that half of all cases where b is even will be observed as transition differences and half as no difference at all. Having examined the cases where $b=0$, b is even and b is odd over all values of b in which $a+b=n$ and over all values of n for which P_n is non-negligible, we have examined all pertinent cases. The sums of those cases where there is an expected observational difference are the values of d_s and d_v that one would expect to find upon examining two sequences evolving in the manner formulated. The computer program that calculated these values is available upon request.

The results of such computations are shown in figure 2 in the form of a nonogram where, given values of d and d_v/d_s, one can read off values of r and s. Comparable values could be obtained using Kimura's formulae (1981). The relation between his variables and mine are as follows:

$$d_s=P(T);\ d_v=Q(T);\ v=4\beta T;\ s=2\alpha T;\ r=2kT.$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES