Biochem J. (1980) **191**, 349–354
Printed in Great Britain

The random character of protein evolution and its effect on the reliability of phylogenetic information deduced from amino acid sequences and compositions

Athel CORNISH-BOWDEN
Department of Biochemistry, University of Birmingham, P.O. Box 363, Birmingham B15 2TT, U.K.

(*Received 7 March 1980/Accepted 17 June 1980*)

Because evolution occurs by random events, the actual number of substitutions that occur in any period is not exactly equal to the number expected from the mean rate of substitution, but is statistically distributed about it. In consequence, even if rates of evolution are constant in different lineages, 'trees' deduced from descendant protein sequences contain random errors. When there are fewer than about eight differences between the sequences of the most distantly related pair from a set of proteins, this random effect is very large. It can then render trivial the statistical disadvantage inherent in using a crude measure of protein difference, such as amino acid composition or immunological cross-reactivity, in preference to a measure based on amino acid sequence. In some cases, such as classification of mammals on the basis of cytochrome *c* structure, it appears to make little difference to the reliability of the results whether the sequences of the proteins concerned are known or not. It may also be possible to obtain more reliable phylogenetic information from composition measurements on several kinds of protein than one could obtain from sequence measurements on a single kind of protein.

Despite the success of protein sequence studies in providing a detailed and largely self-consistent context for discussion of evolution, it is clear that protein sequences do not provide an exact record of evolution. There are discrepancies not only between molecular phylogenies and those derived by traditional methods, but also between the phylogenies derived from the sequences of different proteins (see, e.g., Goodman, 1976). Moreover, although phylogenies derived from crude measures of protein similarity, such as immunological cross-reactivity (Prager & Wilson, 1971) or amino acid compositions (Cornish-Bowden, 1979*a*), are in most respects inferior to those derived from sequences, the degree of inferiority is by no means as great as one might expect. Consider, for example, the two 'trees' in Fig. 1, which are derived from the α-chains of the haemoglobins of nine primates. The two classifications are rather similar and both are almost certainly correct in many respects. Where the sequences and compositions disagree, the classification according to sequence is usually the more reasonable: for example, the sequences in Fig. 1 show the α-chains (though not the $^3\alpha$-chains, which are minor variants) of all the Anthropoidiae as more similar to one another than any is to that of the loris, a prosimian, whereas the compositions interpose the loris between the Cercopithecidae and the other Anthropoidiae. Nonetheless, in some points of detail the compositions provide a more reasonable classification: for example, they group the three species of Cercopithecidae together, whereas the sequences suggest that the hanuman langur is more closely related to the human than to either of the other species of Cercopithecidae. Moreover, the sequences show the chimpanzee (as attested by its $^3\alpha$-chain) as only a distant relative of the human, whereas the compositions indicate a much closer relationship.

In this example no confusion is likely, as there is no reason to doubt that the $^3\alpha$-chains are paralogous to the normal α-chains, both because they are not major components but minor variants, and because the gorilla is represented in the data set by both types of chain. In less clear-cut circumstances the absence of definite evidence that a particular species contains a protein that is only distantly related, or not related at all, to the same protein in related species can cause considerable confusion. It was, for example, very difficult to account for the large differences between the lysozyme of the goose and those of the duck and chicken when these proteins

0306-3275/80/110349-06$01.50/1

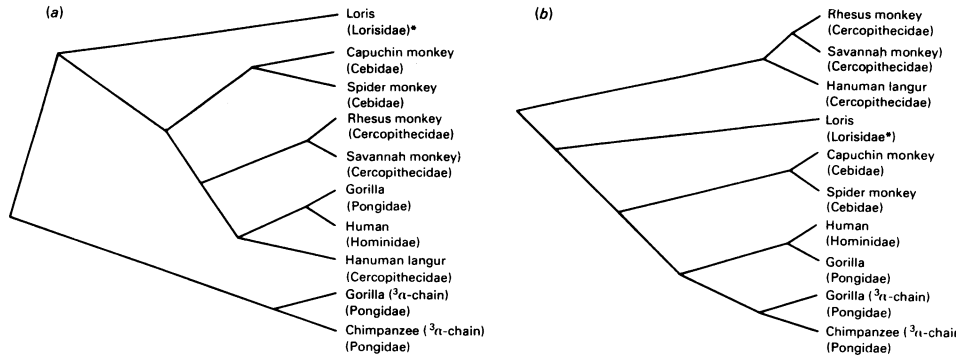

Fig. 1. *Classification of the α-chains of haemoglobin*

All of the primate α-chains listed in Alignment 37 of Hunt & Dayhoff (1976) are included, apart from the two chains from irus macaque, both of which lack residues 105–139. The $^3\alpha$-chains are minor variants observed in chimpanzee and gorilla. (The normal α-chain of chimpanzee is identical with that of human.) Families are shown in parentheses. All are in sub-order Anthropoidiae, apart from the family Lorisidae (*), which is in sub-order Prosimii. In (*a*) the clustering was done as described in the Methods section by using the amino acid sequences; in (*b*) the amino acid compositions were used.

were first studied. The anomaly was only resolved when the black swan was found to have two kinds of lysozyme, one like that of the goose and one like that of the duck (Arnheim & Steller, 1970).

In both of these examples from Fig. 1 the apparent superiority in some respects of the classification derived from the compositions is presumably due to chance rather than to any real superiority of the approach. Nonetheless, the fact that such anomalies occur at all emphasizes the importance of chance in influencing the results of a classification. It also raises the question of whether the enormously greater experimental effort needed to determine a protein sequence rather than just its composition or immunological properties is certain to be repaid in the form of more reliable phylogenetic information. I show in the present paper that this is by no means certain and that, on the contrary, the same experimental effort devoted to determining the amino acid compositions of several proteins in a group of species can lead to better phylogenetic information than would be available from the sequences of a single protein in the same species.

It has been recognized for many years that evolution must be a stochastic process, i.e. one that proceeds by the accumulation of random events, so that the 'molecular clock' must be stochastic and not metronomic, i.e. regular and non-random. Nonetheless, although Fitch (1976) comments that 'no-one expects evolutionary phenomena to be metronomic', there are numerous discussions of protein sequences in the literature that attach a significance to small numbers of differences that seems to imply that the authors are regarding the clock as metronomic. For example, the observation that the cytochrome *c* sequences of three Artiodactyla (cow, sheep and pig) are identical, whereas that of a fourth (hippopotamus) differs at three loci, has been taken as verification that evolutionary rates are not constant (Thompson *et al.*, 1978). Assuming, for the sake of illustration, that the proportions of the total length of the true evolutionary 'tree' relating the four species are 20% each for hippopotamus and pig, and 12.5% each for cow and sheep, the probability that three random substitutions will all occur on a limb unique to one species is $(0.2^3 + 0.2^3 + 0.125^3 + 0.125^3)$, or about 2%. This probability seems hardly small enough to disprove the hypothesis of constant rates, especially in view of the large number of cytochrome *c* sequences now known: as this is well over 50, one should not be surprised to find one with properties with only 2% likelihood if considered in isolation.

A major consequence of the stochastic character of evolution is that it can render the observed number of differences between two sequences a very unreliable guide to the expected number, and hence to the time since separation, even if the simplest possible statistical behaviour applies. This is especially true for closely similar sequences, and in the limit it makes a classification based on a highly conservative protein, such as histone H4, little or no improvement on classification at random. The purpose of the present paper is to study the severity of this effect and to determine how different the sequences of a set of proteins have to be for them to provide a reliable classification and, in particular, one that is more reliable than one could obtain from cruder information about the same proteins.

A preliminary account of some of this work was

presented at the 11th International Congress of Biochemistry (Cornish-Bowden, 1979*b*).

Methods

Estimation of amount of sequence difference from amino acid compositions

For any pair of proteins of equal length the number of sequence differences was estimated as the value of $S\Delta n$ defined as follows (Cornish-Bowden, 1977):

$$S\Delta n = \tfrac{1}{2}\sum_{i=1}^{18}(n_{iA}-n_{iB})^2$$

in which n_{iA} is the number of residues of the ith type in protein A and n_{iB} is the corresponding number in protein B. The summation is carried out over the 18 types of amino acid commonly distinguished in composition measurements, i.e. asparagine is counted with aspartate, and glutamine is counted with glutamate. This convention was also followed in counting the numbers of differences between known sequences: although this meant that a small proportion of the information contained in the sequences was discarded, it had only a slight and in most cases insignificant effect on the classifications produced. It also avoided the need for a complicated definition of the number of sequence differences in experimental circumstances where some of the amide assignments were known and some were not.

Simulation of evolutionary 'trees'

Methods were essentially as described by Dayhoff (1976). For each simulation a random ancestral sequence was generated with mean amino acid frequencies in accordance with the values compiled by Dayhoff & Hunt (1972). Random substitutions were introduced in accordance with the substitution-frequency data of Dayhoff *et al.* (1972).

Two different 'trees' were simulated. In the simpler of these the ancestral sequence gave rise to two descendants, each of which independently accumulated L substitutions on average before giving rise to two further descendants each, each of which then independently accumulated L more substitutions on average. (This is illustrated as an inset in Fig. 2.) In each leg of the 'tree' the actual number of substitutions was randomly distributed about the expected value L in accordance with a Poisson distribution. This 'tree' resembled the one simulated by Dayhoff (1976), except that her 'tree' was 'unrooted', i.e. it did not include a 'node' (branch point) for the common ancestor of all four descendants, and the number of substitutions between the two 'first-generation' descendants was L rather than $2L$, and her L values were exact rather than Poisson-distributed. The distribution of L was crucial in my work because of the very small values considered in some cases, but was of little importance in Dayhoff's (1976), because she was primarily concerned with the performance of different clustering methods at very high L values (up to 400 per 100 residues).

The other 'tree' was a more complex one with seven descendants, with topology and times of separation based on the evolution of seven hoofed mammals (sheep, goat, cow, llama, pig, horse and donkey) as given by Langley & Fitch (1974). It is shown as an inset in Fig. 3. The method of simulation was the same as for the simpler tree.

The use of Dayhoff's (1976) model of evolution requires some justification, because it almost certainly oversimplifies the true process. In particular, it assumes that the probability of substitution at any site is fully determined by the identity of the residue occupying the site and the substitution rate of the whole protein. It allows for no variation of substitution probabilities with time, between lineages, or according to the locations of sites within the whole structure. A consequence of this oversimplification is that the model is likely to generate fewer parallel substitutions than occur in reality. Nonetheless, for several reasons I have not attempted to set up a more complex and perhaps more realistic model. First, any such model could only be justified by a detailed and lengthy argument. Secondly, it would be difficult to make a meaningful comparison between new results and those in the literature. Thirdly, I have been concerned in the present paper to show that, even if the simplest assumptions about protein evolution apply, phylogenies deduced from amino acid sequences are subject to substantial random errors. Any additional complexities in the actual processes of evolution can only make this problem worse. Fourthly, the introduction of such complexities into the model would be unlikely to have a large effect on the relative reliability of phylogenies derived from sequence data compared with those derived from cruder data, because they would be expected to have parallel deleterious effects on all methods of analysis. Finally, the presence of some invariant loci, for example, would decrease the effective size of the protein considered, but it would have no other effect, because all methods of analysis referred to in the present paper are concerned with differences, not identities.

Method of clustering

After each 'tree' had been simulated, the four or seven descendant sequences were first adjusted to remove differences between aspartate and asparagine and between glutamate and glutamine. Compositions were then calculated from these adjusted sequences. For both sequences and compositions, difference matrices were compiled relating

the four or seven descendants. In the sequence case the differences were simple counts of the numbers of loci at which the residues were non-identical, i.e. the matrices were 'unitary matrices' in the terminology of Dayhoff (1976). More precise scoring is possible, but Dayhoff (1976) has shown that the differences between the results given by different kinds of sequence matrix are slight or even non-existent unless the period of evolution is very long. For the compositions the difference matrices consisted of $S\Delta n$ values as defined above.

Clustering was done by the UPGMA method (Sneath & Sokal, 1973) after addition to each value in the starting matrix of a random number from a Cauchy distribution with median 0 and median absolute value 10^{-5}, censored at ± 0.01 (i.e. random numbers numerically greater than 0.01 were discarded). In the absence of ties (equal values) in the starting matrix these random numbers were too small to affect the results, but if any ties were present in the data they caused them to be broken in an unbiased manner. A Cauchy distribution was used to ensure that the distribution was not altered by the averaging that occurs during the course of UPGMA clustering. [The distribution of the mean of a sample from a Cauchy distribution is the same as that of a member of the sample, i.e. the central limit theorem does not apply (Kendall & Stuart, 1969).]

Scoring of clusters

For the simple 'tree' with four descendants, the resulting clusters were scored as correct if they had the correct topology (including the correct 'root') and incorrect otherwise. This simple binary scoring system seemed unreasonable for the more complex 'tree' with seven descendants, because it would take no account of considerable degrees of incorrectness that might be observed. For example, it would not differentiate between a 'tree' with llama and pig reversed but otherwise correct and one containing no correct clusters apart from the final one containing all seven species. Accordingly, each 'tree' was scored in the following way. For each pair of descendants the step of the UPGMA method in which they were clustered was compared with the step in which they ought to have been clustered. For example, in the true 'tree' the pig and cow are clustered in step 5, so a 'tree' clustering them at step 2 would have a difference of 3 for these two descendants. The whole 'tree' was given a score equal to the sum of squares of these differences, summed over all possible (21) pairs of descendants. With this scheme the best possible score is 0, for a 'tree' in which all of the correct clusters occur in the correct order. The worst possible score is 133, but this requires a deliberately perverse 'tree', such as one in which goat is clustered successively with horse, pig and llama, and then sheep is clustered successively with donkey, cow and the cluster containing goat. It is more useful to regard a typical bad score as 95, as this is the median score of 'trees' produced by clustering at random; 95% of such random 'trees' have scores of 52 or more. The maximum score for a 'tree' that contains the correct clusters formed in the wrong order is 18.

Results

Fig. 2 shows the probability of obtaining the correct topology for four 100-residue proteins that evolved according to the simple 'tree' defined above and illustrated in the inset. For values of L less than 2 there is little or no difference between the results obtained from the sequences and those from the compositions. Thus for these small values of L the statistical uncertainty, of the order of 40% (Cornish-Bowden, 1977, 1979*a*) inherent in estimating the amount of sequence difference from measurements of composition difference is trivial by comparison with the statistical uncertainty inherent in the stochastic nature of evolution. A value of $L = 2$ corresponds to about 8% difference between the most distantly related descendants, or a little less than the amount of difference between the cytochrome *c* sequences of the most distantly related mammals. Thus this simple model suggests that, if one used a protein such as cytochrome *c* for classifying mammals, it would make little difference whether one knew the sequences or only the compositions. For larger values of L the difference in performance

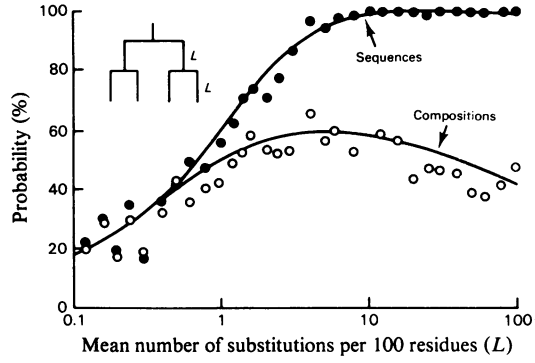

Fig. 2. *Probability of deducing the correct tree*
The simple tree shown as an inset was simulated as described in the Methods section. Each sequence contained 100 residues and each point represents the mean of 100 simulations. In each simulation the descendant sequences were clustered both according to sequences (●) and according to compositions (○). The resulting trees were considered correct if the descendants were paired correctly.

becomes progressively more marked: sequences give virtually perfect results if L is in the range 10–100, whereas compositions never offer much better than a 50% chance of obtaining a correct 'tree'.

An apparent inconsistency between the results in Fig. 2 and those of Dayhoff (1976) requires comment. According to Fig. 2, the probability of obtaining the correct 'tree' from sequence data approaches about 20% as L approaches zero, whereas according to Dayhoff (1976) it is 100% for several methods of clustering for all values of L up to about 50. Although there are several differences between Dayhoff's (1976) simulations and those shown in Fig. 2 (for example, the total length of her 'trees' was $5L$ rather than $6L$, and her 'trees' were 'unrooted'), the main reason for the apparently different behaviour at $L = 0$ is likely to lie in the fact that Dayhoff (1976) was concerned with the behaviour of different clustering methods at very large L values and the smallest L value at which actual measurements were made was 25. Clearly, therefore, there is no real discrepancy, as extrapolation to zero of the sequence curve in Fig. 2 for the portion with L in the range 10–100 would suggest 100% success at $L = 0$.

Fig. 3 shows the scores obtained in simulations of the 'tree' with seven descendants. They confirm the qualitative impression given by the simpler example, i.e. they show that for very low rates of evolution sequences provide negligible phylogenetic information beyond what is contained in the corresponding compositions. The range of evolutionary rates covered by Fig. 3 includes the rates at which real proteins evolve, as indicated by the scale at the top of Fig. 3. For an extremely conservative protein, such as histone H4, neither sequences nor compositions provide any phylogenetic information; for cytochrome *c* there is only slight divergence between the sequence and composition curves; only with rapidly evolving proteins such as κ-casein do sequence data provide near-perfect results, and only then is there a substantial advantage in using sequence data rather than a cruder measure of structural difference.

As the failure of composition data to give perfect results is statistical in nature, one would expect to be able to improve the results by combining data for the

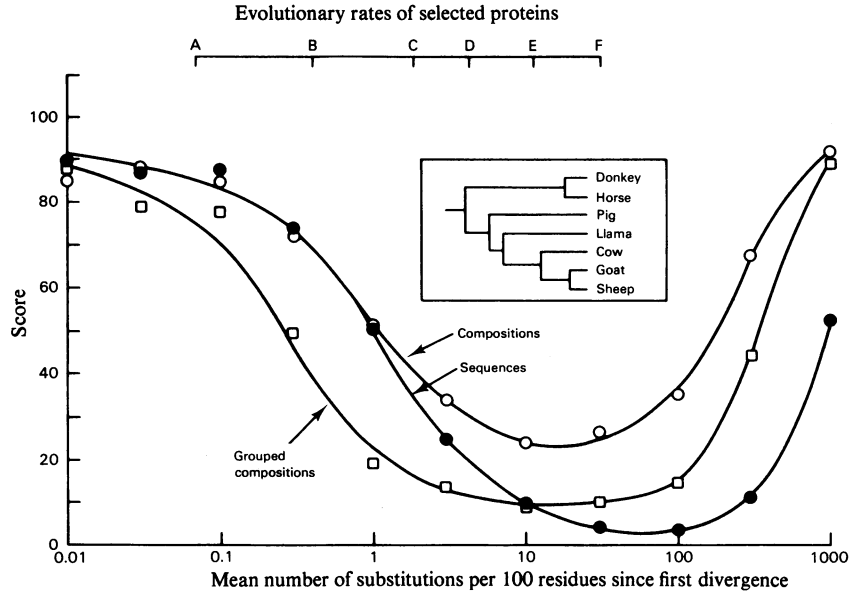

Fig. 3. *Scores obtained in simulation of a realistic tree*

The tree shown in the inset is based on information given by Langley & Fitch (1974). It was simulated, and the resulting sequences and compositions were clustered and scored, as described in the text. Results for sequences (●), compositions (○) and grouped compositions (□) are given. For the grouped compositions the simulations were grouped in sets of five, so that the difference matrix used for clustering was obtained by adding together five independent difference matrices. Each sequence had 100 residues, and 100 trees were simulated at each substitution rate. The rates of substitution of six real proteins are shown for comparison at the top of the Figure, as follows: A, histone H4; B, histone H2A; C, cytochrome *c*; D, trypsinogen; E, haemoglobin; F, κ-casein. In placing this reference scale along the abscissa the period since divergence of the hoofed mammals from a common ancestor was assumed to be 80 million years.

compositions of several proteins. The third curve in Fig. 3 confirms that this is so. Not only are the results given by combined data for the compositions of five proteins better at all rates of evolution than those given by the compositions of a single protein, they are also better than those obtained from the sequences of a single protein over much of the range of rates. Even at high rates of evolution one can do almost as well with the compositions of five proteins as one can with the sequences of a single protein. As Prager & Wilson (1978) have pointed out, combining results from several different proteins has the further advantage of decreasing the danger of invalid results from failure to recognize that particular proteins may be anomalous, e.g. goose lysozyme mentioned at the beginning of the present paper.

Discussion

The main disadvantage of using amino acid sequences of proteins for classifying organisms is that the determinations are laborious and expensive to obtain. As a result, the number of interesting questions of evolution or classification that have been illuminated by comparing proteins is a meagre fraction of the number of such questions that can be asked. There can be little doubt that sequences provide more detailed and more accurate information about evolution than is available from cruder measures of protein structure, such as composition, immunological cross-reactivity or electrophoretic mobility. Equally, however, there can be no doubt that these cruder alternatives are much cheaper than the determination of sequences and, at least in the case of compositions, the literature already contains an enormous body of largely uninterpreted data. For purposes of classification, therefore, one can regard the choice of structural measure as one of cost-effectiveness. The data in the present paper indicate that, when a group of proteins to be classified differs at only a few sites, there is negligible advantage to be gained from measuring the sequences if the compositions are known. For proteins that evolve more rapidly, one may still do as well or better by using composition data for several different proteins rather than sequence data for a single protein.

I have been mainly concerned with amino acid compositions in the present paper, but most of the conclusions would apply with similar force to the immunological techniques that have been used for classification, which are of the same order of precision as composition measurements (Nei, 1977; Cornish-Bowden, 1979*a*). Nonetheless, composition measurements enjoy two additional advantages. First, compositional differences can be expressed on the same scale as sequence differences by means of the index $S\Delta n$ (Cornish-Bowden, 1977), so that it is simple and convenient to replace crude estimates derived from compositions by more precise ones derived from sequences if these become available. Similarly, there are no scaling difficulties to prevent mixing of data of the two different kinds. Secondly, amino acid compositions are frequently obtained independently of any investigations of evolution or classification for other (usually obscure) reasons. Consequently the cost of obtaining composition data for purposes of classification is often zero.

I am grateful to Dr. A. J. G. Moir for constructive criticism.

References

Arnheim, N. & Steller, R. (1970) *Arch. Biochem. Biophys.* **141**, 656–661

Cornish-Bowden, A. (1977) *J. Theor. Biol.* **65**, 735–742

Cornish-Bowden, A. (1979*a*) *J. Theor. Biol.* **76**, 369–386

Cornish-Bowden, A. (1979*b*) *Abstr. Int. Congr. Biochem. 11th* 226

Dayhoff, M. O. (1976) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **35**, 2132–2138

Dayhoff, M. O. & Hunt, L. T. (1972) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, pp. D355–D359, National Biomedical Research Foundation, Silver Spring, MD

Dayhoff, M. O., Eck, R. V. & Park, C. M. (1972) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, pp. 89–99, National Biomedical Research Foundation, Silver Spring, MD

Fitch, W. M. (1976) in *Molecular Evolution* (Ayala, F. J., ed.), pp. 160–178, Sinauer Associates, Sunderland, MA

Goodman, M. (1976) in *Molecular Evolution* (Ayala, F. J., ed.), pp. 141–159, Sinauer Associates, Sunderland, MA

Hunt, L. T. & Dayhoff, M. O. (1976) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 2, pp. 191–223, National Biomedical Research Foundation, Silver Spring, MD

Kendall, M. G. & Stuart, A. (1969) *The Advanced Theory of Statistics*, 3rd edn., vol. 1, pp. 248–249, Griffin, London

Langley, C. H. & Fitch, W. M. (1974) *J. Mol. Evol.* **3**, 161–177

Nei, M. (1977) *J. Mol. Evol.* **9**, 203–211

Prager, E. M. & Wilson, A. C. (1971) *J. Biol. Chem.* **246**, 5978–5989

Prager, E. M. & Wilson, A. C. (1978) *J. Mol. Evol.* **11**, 129–142

Sneath, P. H. A. & Sokal, R. R. (1973) *Numerical Taxonomy*, pp. 230–234, W. H. Freeman, San Francisco

Thompson, R. B., Borden, D., Tarr, G. E. & Margoliash, E. (1978) *J. Biol. Chem.* **253**, 8957–8961