Construction of Phylogenetic Trees By Pattern Recognition Procedures

Muhammad Abdallah Sharaf, Bruce R. Kowalski, and Boris Weinstein*
Department of Chemistry BG-10, University of Washington, Seattle, Washington 98195, USA

Z. Naturforsch. **35 c,** 508–513 (1980); received October 10/December 31, 1979

Cytochrome c, Pattern Recognition, Phylogenetic Tree, Protein Homology

The sequence of a protein can be graphed as a discrete function and a cross-correlation between any two such number sets produces a similarity score. The scores are used to prepare a phylogenetic tree involving hierarchical cluster analysis, non-linear mapping, and minimal spanning routines. Changes are suggested in the sequences of cytochrome c's from Mediterranean fruit fly, locust, and rattlesnake. The method is faster than existing procedures and does not require human intervention at any stage.

The study of molecular evolution *via* the changes of specific amino acid residues in homologous proteins was introduced in 1962 [1]. As a result, several mathematical models have been developed to trace phylogenetic relationships between similar and dissimilar proteins [2–13]. The cytochrome c family is a favorite test case in this area due to the wide distribution and availability of the compound in nature [14–20]. Based on various points of view, most existing procedures give evolutionary trees for this protein that are in agreement with those obtained by classical botanical and zoological studies. Without exception these techniques involve long computational time, arbitrary assumptions as to branch length, and human attention at one or more stages. Among other problems, it may be noted that the common ancestor route gives rise to degenerate node sequences, which increase as one encounters earlier evolutionary branches. Matrix methods have the added task of defining and estimating several statistical factors, including the variable genetic code for the amino acids, as well as evaluating the number of changes J (insertions or deletions) that brings any two proteins to the same length [3]. We have developed a simple and fast procedure to overcome these difficulties and the resulting data can be routinely subjected to pattern recognition methods involving cluster analysis and non-linear mapping [21–23].

If a score (1, 2, . . . 20) is assigned to each of the 20 common amino acids, then a protein can be represented as a discrete function. As an example, porcine glucagon, a peptide having 29 residues, is easily changed into a set of graphed points (Fig. 1). This "spectrum" enables a cross-correlation to be made between any two proteins in a facile fashion. For this purpose, a similarity score (SS) is formulated as follows:

$$SS_{i,j} = \left[\frac{N \cdot L'}{(B_{i,j} + 1)\, L^2}\right]^{1/2}$$

where

N = Number of corresponding common amino acids in the two proteins at a specific alignment,
L' = Length of overlapping portions of the two proteins,
B = Total number of minimum base changes required to change protein i to protein j,

and

L = i. Length of the longer protein where a comparison is made between *any* two proteins;
ii. Length of the shorter protein where it is part of the longer protein;
iii. Overlapping length of the two protein segments where a search is being made for local similarity.

The above function generates an "interferogram" in which the central spike corresponds to the displacement needed for an optimal alignment.

The similarity score (the height of the spike) is normalized to give 1 for a total match and 0 for a total dissimilarity. This similarity scoring function obviates the need to estimate any weighting factors and the data need not be transformed into a distance

Reprint requests to Prof. Dr. Boris Weinstein.
0341-0382/80/0500-0508 $ 01.00/0

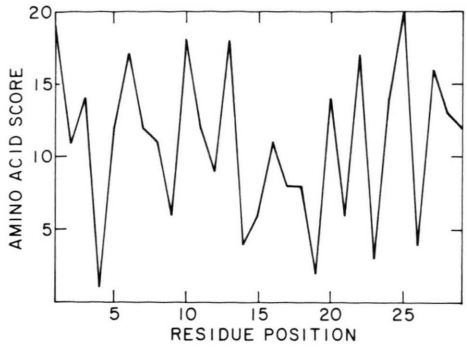

Fig. 1. Numerical representation of porcine glucagon.

measure, which is a long process and subject to error [18]. Also, proteins may be of varying length, so the number J is unnecessary. The SS values are actually calculated only after the best alignment has been determined for a family of homologous proteins. A routine can be written to introduce gaps at the proper sites and to help align regions within various proteins, but there was no need for it in the examples discussed here.

Porcine glucagon and porcine secretin are gene duplication products and are related on a variety of grounds including 14 common amino acid residues [24]. To show this relationship through a similarity score, the sequence of the 29 amino acids in glucagon is held stationary and the sequence of the 27 amino acids in secretin is passed under it, residue by residue. A simple FORTRAN program, ALI BABA, is used for this purpose. In step one, the N-terminal histidyl residue of secretin falls beneath the C-terminal threonyl residue in glucagon and the SS calculation is therefore:

$$(SS) = \left[\frac{0 \cdot 1}{(2+1) \cdot 27^2}\right]^{1/2} = \left[\frac{0}{2187}\right]^{1/2} = 0 .$$

At step 27, the 27 residues of secretin have been placed under the initial 27 residues of glucagon. The common amino acid residues rise to 14 and the corresponding score (SS value) reaches a corresponding maximum of 0.161, which signifies an optimal match at this alignment. In step 28, the histidly residue at position one in secretin is now under the seryl residue at position two in glucagon and an immediate drop occurs in the SS value. This process

Table 1. SS Values for glucagon-secretin.

Odd Step	Number of Residues Compared (L′)	Common Amino Acids (N)	Minimum Base Changes (B)	SS L = Long (i)	SS L = Short (ii)	SS L = Overlap (iii)
1	1	0	2	0.000	0.000	0.000
3	3	0	5	0.000	0.000	0.000
5	5	0	9	0.000	0.000	0.000
7	7	0	10	0.000	0.000	0.000
9	9	0	16	0.000	0.000	0.000
11	11	0	18	0.000	0.000	0.000
13	13	1	20	0.003	0.003	0.061
15	15	1	22	0.028	0.030	0.054
17	17	0	27	0.000	0.000	0.000
19	19	2	26	0.041	0.044	0.062
21	21	2	28	0.041	0.045	0.057
23	23	3	32	0.050	0.054	0.063
25	25	1	36	0.028	0.030	0.033
27	27	14	19	0.150	0.161	0.161
29	27	2	42	0.039	0.042	0.042
31	25	4	32	0.060	0.064	0.070
33	23	4	30	0.059	0.064	0.075
35	21	1	32	0.028	0.030	0.038
37	19	0	32	0.000	0.000	0.000
39	17	1	23	0.029	0.031	0.050
41	15	1	20	0.029	0.031	0.056
43	13	2	20	0.038	0.041	0.086
45	11	1	19	0.026	0.027	0.067
47	9	0	14	0.000	0.000	0.000
49	7	1	13	0.024	0.026	0.101
51	5	1	9	0.024	0.026	0.141
53	3	0	7	0.000	0.000	0.000
55	1	0	2	0.000	0.000	0.000

continues until all possible alignments are made between glucagon and secretin (Table I). For M proteins, an $M \times M$ similarity matrix is generated and in the case of porcine glucagon and secretin, a standard 2×2 matrix gives the values 1.00 (complete identity) and 0.161 (optimal match).

The maximum SS value is now used to calculate a dissimilarity score. The dissimilarity function is defined as follows:

$$d_{i,j} = 1 - (SS)_{i,j} .$$

This dissimilarity function satisfies the usual conditions in that:

$$d_{i,j} = d_{j,i}$$
$$d_{i,j} > 0 \text{ for } i \neq j$$
$$d_{i,j} = 0 \text{ if and only if } i = j.$$

Although the dissimilarity function was not obtained by calculating Euclidean distances between points in an n-dimensional space, several pattern recognition methods are applicable to the data. For example, hierarchical cluster analysis (HIER) is ideally suited to our problem as it groups samples according to their distance to one another and produces a so-called dendrogram. HIER scans the distance matrix for the smallest distance, aligns the two corresponding points, and represents them as a two sample subset. The remaining items are processed until an optimum placement exists for all clusters. The resulting plot provides a visual display, since it gives the dissimilarity at which two samples are grouped together. This procedure has been discussed in some detail for related problems [25].

In order to test these programs on a more interesting and complex situation, 73 cytochrome c sequences were collected from the literature [26–40]. The SS values for each two homologous proteins were obtained for all possible alignments and the highest scores were selected. The algorithm developed for this task was capable of performing about 320 calculations per second on the cytochrome c's and this phase of the work required 1200 seconds on a CDC-6400 computer. The final dendrogram or phylogenetic tree was generated in an additional 24 seconds (Fig. 2).

There is no inherent parameter in the SS function that indicates a goodness-of-fit measure, since only one diagram is produced when the matrix is scanned for hierarchical clustering. Any discrepancies found, especially when the final plot is compared to those produced by classical studies, can be attributed to the inadequacy of the function. In turn, this fault is related to the protein discriminating ability, the correctness of the amino acid sequences, and the inability to separate highly related species. Our representation is in excellent agreement with those produced by other procedures [13, 15], but several placements differ from those based on current evolutionary ideas or zoological grounds. For instance, the position found for Mediterranean fruit fly comes in part from differences existing at three amino acid residues (positions 54, 64, and 65) [31]. If these residues (asp, asn, and glu) are actually identical to those in fruit fly (asn, gln, and asp), then a common relationship to screwworm fly develops, which is more widely accepted. A similar situation exists with the sequence of locust [39] where our program aligns it to prawn. By allowing two of the residues (positions 4 and 64) to be changed to those found in all the flies and moths (a shift of gln and asp to ala and gln), then locust is seen as an early insect.

The most controversial case is rattlesnake where use of the originally assigned structure placed snake with starfish [33]. In another attempt, five residues (positions 86, 89, 93, 101, and 104) were changed (ser, lys, asn, lys, and ala to lys, ser, asp, ala, and lys) and snake appeared as an offshoot of the branch leading to frogs and birds [30]. We have found that the alignment of snake with the rest of the eukaryotes becomes perfect by making two more such changes (positions 11 and 12) in a similar fashion (thr and met to val and gln). It is realized that if no errors exist in the reported sequences of Mediterranean fruit fly, locust, and snake, then the dendrogram discrepancies may be due to mutations occurring after the divergence of these species.

Our tree clearly separates Gymnosperm and Angiosperm plants, as well as differentiating the dicots from the monocots. Unfortunately, no cleavage is obtained for those members belonging to different subgroups (Commenlinidae, Magnoliidae, Liliidae, Rosidae, Asteridae, Hamamelidae, and Dilleniidae). This result is not unexpected for these closely related plants have been scrambled in other studies [41–43]. Such behaviour could be attributed to a possible case of converging evolution in plants or to the use of just one insufficient variable, namely, cytochrome c, to describe divergence at the subfamily level. It is sobering to realize that this problem has been ignored by many of the published cytochrome c

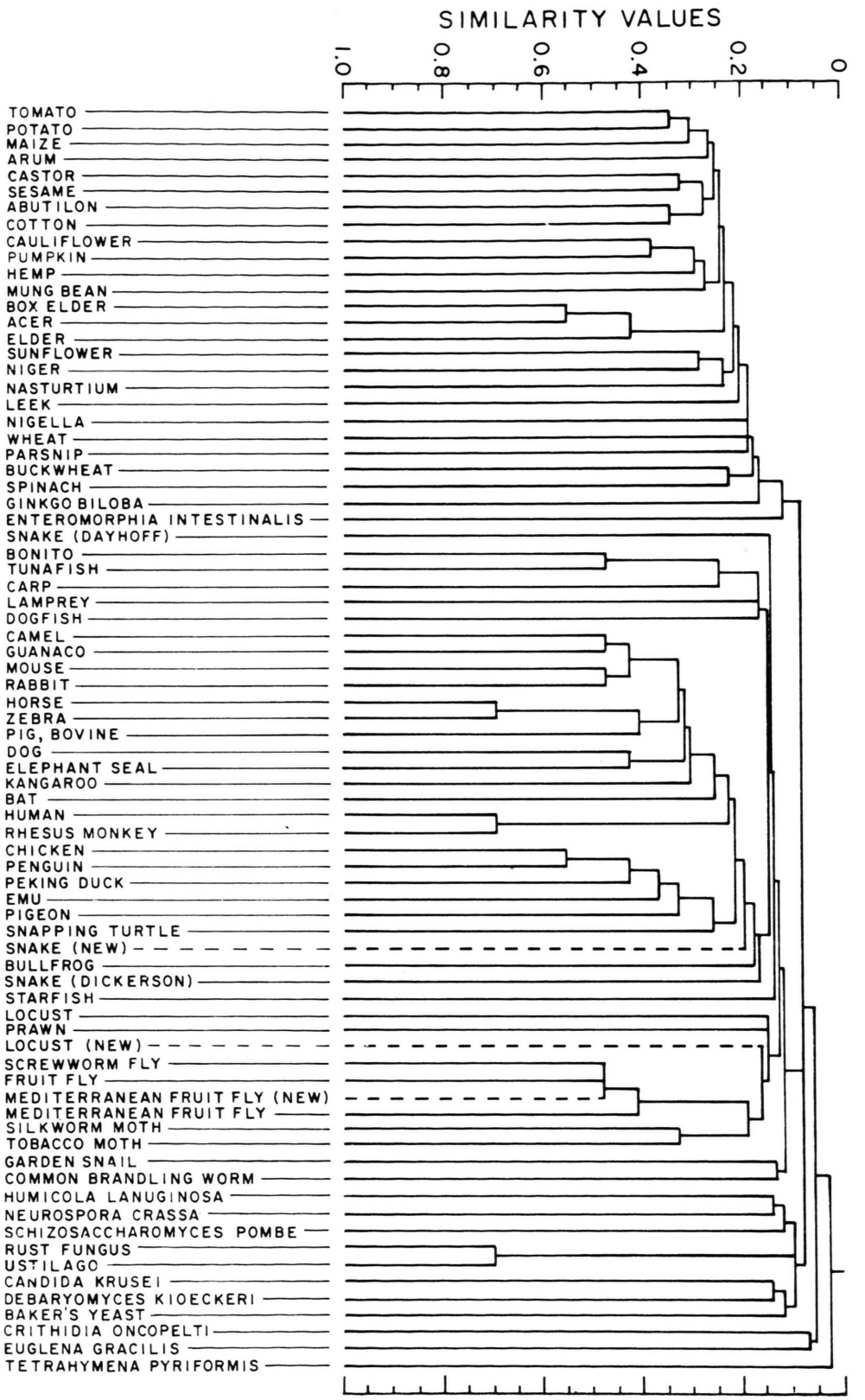

Fig. 2. Evolution of cytochrome c plotted by a hirarchical clustering technique.

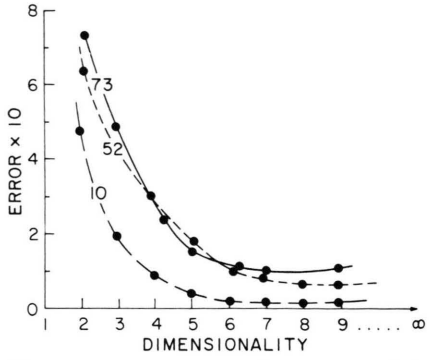

Fig. 3. Non-linear mapping of cytochrome c space for 10×10, 52×52, and 73×73 matrices.

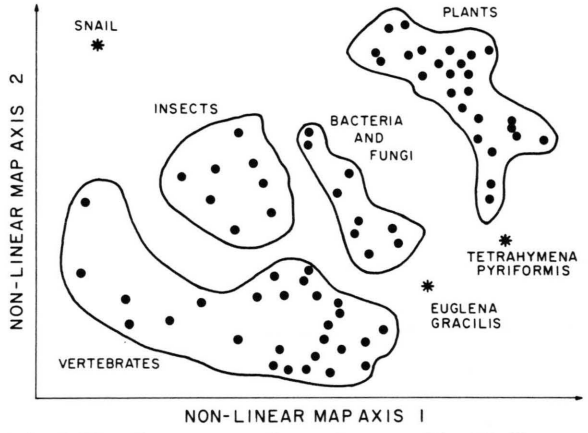

Fig. 4. Non-linear map of cytochrome c 73×73 distance matrix in two dimensions.

plots in the past, possibly due to small sample size or to biased human intervention. Further, one must not lose sight of the fact that the generalized cytochrome c phylogeny also disagrees with the traditional one in several more instances, including the following: chicken is placed next to penguin, not by ducks and pigeons; turtle appears more related to birds than to rattlesnake; and, man and monkey diverge from the mammals before marsupial kangaroo separates from the placental mammals.

We next made a non-linear mapping (NLM) of the dissimilarity matrix from its unknown dimensional space into two-, three-, or n-dimensional space by a conjugate gradient minimization technique [44]. The resulting transform provides an optimum representation of the original data. In brief, the data structure represented by the dissimilarity matrix is mapped to a lower dimensional space using mathematical criteria that minimize information loss (Fig. 3). No significant improvement appeared by going beyond $n =$ six and this behaviour was independent of the size of the dissimilarity matrix. Thus, it appears that the similarities between the 73 known cytochrome c sequences have an intrinsic dimensionality of six. Almost half of the information in the six-dimensional space distance matrix is represented by two principal components [21]. When these two components are plotted against each other, four separate clusters are observed, which include most of the cytochrome c's (Fig. 4). This result represents an approximate view into a cytochrome c space.

The groups detected in the NLM projection were verified using the minimal spanning tree method (TREE) of cluster analysis [45]. The outcome confirmed the empirical observation that the known cytochrome c sequences fall into five major families: bacteria, plants, insects, fish, and land animals.

In summary, the ALI BABA program for generating similarity scores between homologous proteins is efficient, quick, and does not require any manual steps. This procedure, however, is very sensitive to possible errors in the sequence of a protein and can provide controversial placements in a dendrogram. Taken in conjunction with HIER, NLM, and TREE pattern recognition techniques, a new and powerful computational procedure is now available for phylogenetic studies [46].

[1] E. Zuckerlandl and L. Pauling, Horizons in Biochemistry, (M. Kasha and B. Pullman, eds.), p. 189, Academic Press, New York 1962.
[2] C. R. Cantor and T. M. Jukes, Proc. Nat. Acad. Sci. USA **56,** 177 (1966).
[3] W. M. Fitch and E. Margoliash, Science **155,** 279 (1967).
[4] S. B. Needleman and C. D. Wunsch, J. Mol. Biol. **48,** 443 (1970).
[5] A. D. McLachlan, J. Mol. Biol. **61,** 409 (1971).
[6] T. M. Jukes and R. Holmquist, J. Mol. Biol. **64,** 163 (1972).
[7] T. M. Jukes and R. Holmquist, Science **177,** 530 (1972).
[8] R. Holmquist, J. Mol. Evol. **2,** 145 (1973).
[9] D. Sankoff and R. J. Cedergreen, J. Mol. Biol. **77,** 159, 164 (1973).
[10] G. W. Moore, M. Goodman, C. Callahan, R. Holmquist, and H. Moise, J. Mol. Biol. **105,** 15 (1976).
[11] M. S. Waterman, T. F. Smith, M. Singh, and W. A. Beyer, J. Theor. Biol. **64,** 199 (1977).
[12] T. H. Hseu, E. D. Jou, C. Wang, and C. C. Yang, J. Mol. Biol. **10,** 167 (1977).

[13] E. Margoliash, W. M. Fitch, and R. E. Dickerson, Brookhaven Symposia in Biology, No. 21, p. 259, Brookhaven National Laboratory, Brookhaven, New York 1968.
[14] E. M. Prager and A. C. Wilson, J. Mol. Evol. **11,** 129 (1978).
[15] M. O. Dayhoff (ed.), Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington, D. C. 1972.
[16] R. E. Dickerson, Sci. Am. **226,** 58 (1972).
[17] P. J. McLaughlin and M. O. Dayhoff, J. Mol. Evol. **2,** 99 (1973).
[18] W. A. Beyer, M. L. Stein, T. F. Smith, and S. M. Ulam, Math. Biosci. **19,** 9 (1974).
[19] S. S. Carlson, Ph. D. Dissertation, University of California, Berkeley 1975; Diss. Abstr. **37 B,** 187 (1976).
[20] W. Fitch, J. Mol. Evol. **8,** 13 (1976).
[21] B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc. **94,** 5632 (1972).
[22] B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc. **95,** 686 (1973).
[23] B. R. Kowalski, Anal. Chem. **47,** 1152 A (1975).
[24] B. Weinstein, Experientia **24,** 406 (1968).
[25] B. Everitt, Cluster Analysis, Heinemann Educational Books Ltd., London 1974.
[26] F. Lederer, Centre National de la Recherche Scientifique, Gif-Sur-Yvette, France, personal communication.
[27] L. R. Croft, Handbook of Protein Sequences, Joynson-Bruvuers, Oxford 1973.
[28] G. Martinez, H. Rochat, and G. Ducet, FEBS Lett. **47,** 212 (1974).
[29] B. T. Meatyard and D. Boulter, Phytochemistry **13,** 2777 (1974).
[30] R. E. Dickerson and R. Timkovich, The Enzymes, (P. O. Boyer, ed.), p. 397, Academic Press, New York 1975.
[31] J. M. Fernandez-Sousa, J. G. Gavilanes, A. M. Municio, J. A. Paredes, A. Perez-Aranda, and R. Rodriguez, Biochim. Biophys. Acta **393,** 358 (1975).
[32] G. W. Pettigrew, Biochem. J. **147,** 291 (1975).
[33] M. O. Dayoff, Atlas of Protein Sequence and Structures, **Vol. 5,** Supplement 2, National Biomedical Research Foundation, Washington, D. C. 1976.
[34] A. Lydiatt and D. Boulter, Comp. Bioch. Phys. **53 B,** 337 (1976).
[35] A. Lydiatt and D. Boulter, FEBS Lett. **62,** 85 (1976).
[36] A. Lydiatt and D. Boulter, FEBS Lett. **67,** 331 (1976).
[37] G. E. Tarr and W. M. Fitch, Biochem. J. **159,** 193 (1976).
[38] S. S. Carlson, G. A. Mross, A. C. Wilson, R. T. Mead, L. D. Wolin, S. F. Bowers, N. T. Foley, A. O. Muijsers, and E. Margoliash, Biochemistry **16,** 1437 (1977).
[39] A. Lydiatt and D. Boulter, Biochem. J. **163,** 333 (1977).
[40] R. L. Niece, E. Margoliash, and W. M. Fitch, Biochemistry **16,** 68 (1977).
[41] D. Boulter, J. A. M. Ramshaw, E. W. Thompson, M. Richardson, and R. H. Brown, Proc. R. Soc. London, Ser. B, **181,** 441 (1972).
[42] D. Boulter, Pure Appl. Chem. **34,** 539 (1973).
[43] D. Boulter, Syst. Zool. **22,** 549 (1974).
[44] J. W. Sammon, Jr., I. E. E. E. Trans. Comput. **C-20,** 68 (1971).
[45] C. T. Zahn, I. E. E. E. Trans. Comput. **C-20,** 68 (1971).
[46] Programs discussed here are available from Infometrix, P. O. Box 25888, Seattle, Washington 98125, USA.