J. theor. Biol. (1983) **101**, 317–319

LETTER TO THE EDITOR

Phenetic Methods of Classification Use Information that is Disregarded by Minimum-Length Methods

Penny (1982) has demonstrated that information is lost when phenetic methods are used for classification. These are methods in which the first stage is to combine the observations of all the characters studied into a single measure of "distance" between each pair of taxa. Phenetic methods may fail to choose between trees that are of different total length and thus readily distinguishable by minimum-length methods such as the tree-building method of Foulds, Hendy & Penny (1979). It does not, however, follow that phenetic methods are necessarily inferior to minimum-length methods, because the converse observation can also be made: minimum-length methods disregard some of the information that is used by phenetic methods.

Any character that is present in only one taxon contributes exactly +1 to the minimum length of any tree connecting a set of taxa. Consequently such characters do not contribute to discrimination between trees by the minimum-length criterion. They are not disregarded by phenetic methods, however, because a unique character contributes to the distance between the taxon that displays it and any other taxon, but not to the distances between taxa that do not display it. This point can be appreciated more clearly in relation to a real example.

The seven ribonuclease sequences tabulated by Barker & Dayhoff (1976) contain 129 loci. At 65 of these the same kind of amino acid residue is found in all seven sequences, and so these loci are not used by any classification method. There are 24 loci (1, 2, 3, 4, 6, 16, 25, 27, 36, 55, 58, 66, 72, 76, 77, 82, 91, 104, 107, 111, 116, 123, 128 and 129) at which one sequence differs from the other six; there are 12 loci (12, 21, 22, 26, 62, 64, 65, 90, 99, 106, 114 and 118) at which two sequences differ from one another and from the other five; there are two loci (9 and 42) at which three sequences differ from one another and from the other four; and there are two loci (34 and 105) at which four sequences differ from one another and from the other three. This leaves only 24 loci that can be used for distinguishing between trees by the minimum-length criterion.

These 24 loci also contribute to phenetic classification, but in addition the 40 loci with unique characters only are taken into account. Whether

0022-5193/83/060317+03 $03.00/0

these loci in fact contribute real information or simply noise can be tested by classifying the seven ribonucleases on the basis of them alone. This cannot be done by the minimum-length method, because all of the 945 possible trees have exactly the same length of 62 amino acid substitutions, but it can be done by a phenetic method. The distance matrix is shown in Table 1, together with the matrix for the complete data set, and the results

TABLE 1

Phenetic distance matrices† for ribonuclease sequences‡

Cattle	0 (0)						
Giraffe	6 (11)	0 (0)					
Red deer	7 (17)	6 (15)	0 (0)				
Cattle semen	11 (23)	11 (25)	12 (26)	0 (0)			
Pig	9 (26)	9 (28)	10 (25)	13 (26)	0 (0)		
Horse	15 (34)	15 (32)	15 (33)	19 (34)	17 (29)	0 (0)	
Rat	27 (42)	27 (45)	26 (44)	29 (42)	27 (44)	31 (42)	0 (0)
	Cattle	Giraffe	Red deer	Cattle semen	Pig	Horse	Rat

† The first value in each case is derived from the 40 loci that do not contribute to classification by the minimum-length method; the second value (in parenthesis) is derived from the complete data set.

‡ Except for the enzyme from cattle semen, all are ribonucleases from pancreas.

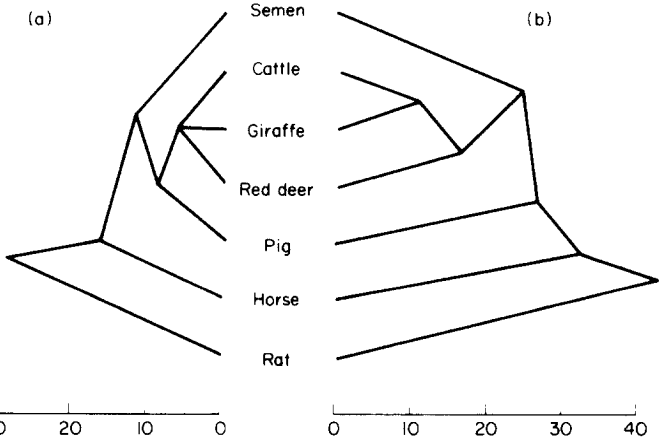

FIG. 1. Trees obtained by UPGMA from the data in Table 1, (a) using only the 40 loci disregarded by the minimum-length method, (b) using all the data.

of clustering both matrices by UPGMA (Sneath & Sokal, 1973) are shown in Fig. 1. The high degree of agreement between the two trees makes it clear that the 40 loci disregarded by the minimum-length method do contain real information.

It follows that although it is true that phenetic methods of classification disregard some of the information used by minimum-length methods (Penny, 1982), it is also true that minimum-length methods disregard some of the information used by phenetic methods. Consequently one cannot say without further study that one approach is inherently better than the other, and it is clearly desirable to search for new methods that make fuller use of the data than any of those in current use.

Department of Biochemistry,
University of Birmingham,
P.O. Box 363,
*Birmingham B*15 2*TT, England*

ATHEL CORNISH-BOWDEN

(*Received* 3 *August* 1982)

REFERENCES

BARKER, W. C. & DAYHOFF, M. O. (1976). *Atlas of Protein Sequence and Structure*, (Dayhoff, M. O, ed.). Vol. 5, suppl. 2, p. 78. Silver Spring: National Biomedical Research Foundation.

FOULDS, L. R., HENDY, M. D. & PENNY, D. (1979). *J. mol. Evol.* **13,** 127.

PENNY, D. (1982). *J. theor. Biol.* **96,** 129.

SNEATH, P. H. A. & SOKAL, R. R. (1973). *Numerical Taxonomy*. San Francisco: W. H. Freeman.