Distance as a Measure of Taxonomic Similarity[1]

ROBERT R. SOKAL [2]

Introduction

THE LAST FEW YEARS have seen the development of a number of methods for quantifying the classificatory process in systematics (Michener and Sokal, 1957; Sokal and Michener, 1958; Sokal, 1958; Sneath, 1957a; Sneath, 1957b; Cain and Harrison, 1958; Rogers and Tanimoto, 1960). These sudden developments, largely independently arrived at, appear to stem from a growing dissatisfaction with the arbitrariness and subjectivity of the customary taxonomic procedure and would also appear to be connected with the increasing availability of rapid methods of data processing and computation. In spite of certain differences among them, these methods show substantial philosophical and procedural agreements. They call first of all for the computation of a measure of resemblance or similarity between taxa. Resemblances are computed between every pair of taxa in a given study and these resemblances are arranged in the form of a so-called Q-type matrix, which is a matrix of similarities among taxa. These matrices are subsequently analyzed by various types of clus-

ter or factor analysis and in this manner the structure and degree of relationships among taxa are revealed. These methods have been called "quantitative systematics" in the past. Recently Sokal and Sneath have renamed them collectively as "numerical taxonomy" (Sokal, 1960).

The computation of taxonomic similarity depends on several assumptions. The most important of these are (1) that the evaluation of similarities among taxa is based on the observed values of characters and not on phylogenetic speculations and interpretations. Hence taxonomic affinities or similarities are *phenetic* rather than phyletic. (2) The similarities are evaluated on the basis of many characters and all characters are considered to be of equal taxonomic value; hence no one character is weighted more or less than any other character.

Three types of coefficients have been suggested to date as suitable measures of similarity between pairs of taxa. These measures can be conveniently classified as coefficients of association, correlation, and distance. Coefficients of association are generally restricted to characters subdivided into two states only. These are the so-called "present-absent" characters employed in various types of ecological association coefficients and first applied to taxonomic work by Sneath (1957b). Coefficients of association vary considerably in the details of their formulation, but are based on an arrangement of the data in the form of a 2 x 2 table. The coefficient of correlation which has been employed in numerical taxonomy (Sokal and Michener, 1958; Morishima and Oka, 1960) has been of the ordinary Pearson

product-moment type, permitting the characters to be subdivided into more than two states.

The third type of coefficient determines the similarities between two taxa as a function of their distance in an n-dimensional space whose coordinates are the characters. The present paper will elaborate on this particular approach.

The use of distance as a measure of taxonomic similarity is not new. The earliest application that I have been able to locate is by Heincke (1898) working on herrings. A similar method was employed by Anderson and Abbe (1934) in comparing specific and generic differences in the birches. Zarapkin's (1934) work is related indirectly to the concept of distance. Distance was extensively employed to summarize the resemblances between samples of anthropological material. This work resulted in the formulation by Karl Pearson (1926) of the coefficient of racial likeness from which has sprung the concept of the generalized distance suggested by Mahalanobis (1936) and further elaborated by Rao (1952). In a later section of this paper the relation of these statistics to the simple distance proposed here as a measure of taxonomic similarity will be discussed. In 1958 Cain and Harrison suggested a so-called "mean character difference" which was not interpreted as a distance by them but which is mathematically related to the concept of distance as will be shown below.

The Computation of Distance

Let us assume that we are studying two species, 1 and 2, which are described on the basis of two characters X and Y. Characters X and Y are measured, if continuous, and numerically coded in other cases. The two species may have any value for X and Y within the range of variation of these characters. Then we can plot our data as in Figure 1, where each of the two species assumes a fixed position in the two-dimensional space defined by the two axes X and Y. A measure of the similarity

between species 1 and 2 based on the two characters would be the distance in the two-dimensional space (i.e. on the plane) between the two species. Thus the greater the distance the less the two species would resemble each other. This distance $\delta_{1,2}$ can easily be computed by means of a well known formula of analytical geometry,

$$\delta_{1,2} = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2},$$

where X_1, Y_1 are the coordinates of species 1. This formula is often conveniently stated in its squared form in order to avoid the square root on the right side. Therefore,

$$\delta_{1,2}^2 = (X_1 - X_2)^2 + (Y_1 - Y_2)^2$$

which is the same as the formula given in Figure 1. When we wish to describe three characters for the two species we have to use a three-dimensional figure (a cube) in order to represent the exact position of the two species with respect to the three characters. This cube is shown in a two-dimensional projection in Figure 2. The computation of the distance between the two species, which are now suspended at fixed points in the three-dimensional space, is by an extension of the distance
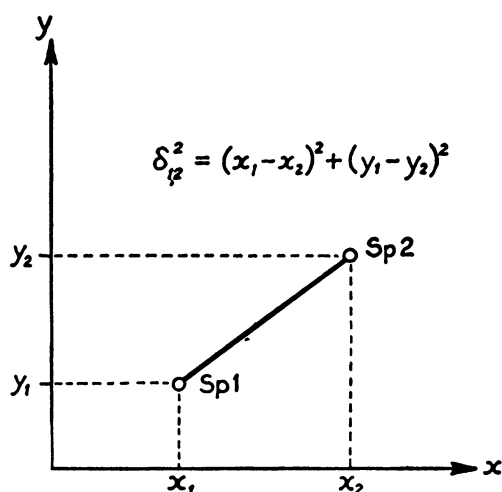

FIG. 1. Two-dimensional diagram showing the distance between two hypothetical species with reference to two characters, X and Y.

$$\delta_{12}^2 = (x_1-x_2)^2 + (y_1-y_2)^2 + (z_1-z_2)^2$$
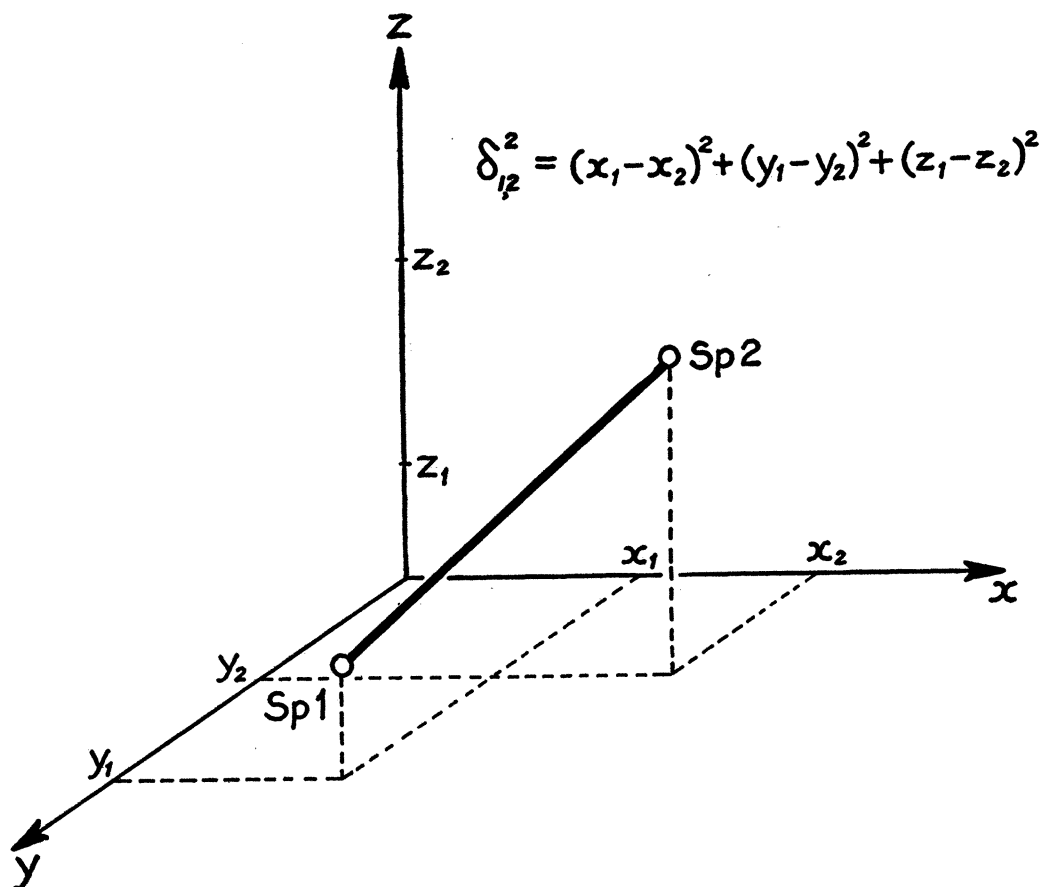
FIG. 2. Two-dimensional projection of the distance between two species with reference to three characters, X, Y, and Z.

formula previously discussed and the square of the distance is as shown in Figure 2. As soon as we describe species on the basis of more than three characters we are no longer able to depict their positions and relationships in a conventional diagram and we are forced to rely on algebraic manipulation of the data. However, the distance formula of analytical geometry is equally valid in an n-dimensional space (i.e., a hyperspace). The squared distance between two species in such a hyperspace of n characters can be computed very simply by summing the squares of the distances along each of the n axes (characters).

We may generalize the formula for the distance squared of two species and n characters as follows:

$$\delta_{1,2}^2 = \sum_{i=1}^{n} (X_{i1} - X_{i2})^2$$

where X_{i1} is the state code of species 1 for character i.

Such a distance formula can be applied to characters divided into two or more states. The number of states per character can vary from character to character. If the original data are set up in matrix form, as seems conventional in numerical taxonomy, with the columns of the matrices representing species (or other taxa) and the rows representing characters, then the distance squared is

simply the sum of the squares of the differences between any pair of columns. For computational purposes, either on a desk calculator or by means of electronic computation, it would probably be best to calculate the squared differences in the following form:

$$\delta_{1,2}^2 = \sum_{i=1}^{n} X_{i1}^2 + \sum_{i=1}^{n} X_{i2}^2 - 2\sum_{i=1}^{n} X_{i1} X_{i2}.$$

Such computations are extremely simple. For a reasonable number of taxa with simply coded character states they would not present a prohibitive amount of work, even if only a conventional desk calculator were available for processing the data. See Appendix for a computational example.

Modifications of the Procedure

The actual distance between taxa (or the square of the distance between them) is necessarily a function of the number of characters for which the taxa are compared. Other factors being equal, the greater the number of characters involved in the comparison, the greater the distance between taxa. While within a given study (i.e., between a number of taxa which are being compared simultaneously) the number of characters will remain constant and hence the distances will give a relative evaluation of taxonomic similarity, it would be preferable to use some generally acceptable standard by means of which these distance coefficients would be compared. It would therefore seen appropriate to divide the sum of the squared differences by n, the number of characters on which it is based, obtaining a mean squared difference,

$$\overline{\delta}_{1,2}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_{i1} - X_{i2})^2,$$

which seems appropriate as a general measure of the taxonomic distance. It is immediately apparent that this quantity is merely the variance of the differences between the character state codes of the two taxa being compared. If preferred, the square root of the above expression, representing the standard deviation of the differences may be employed.

In this connection it is desirable to discuss the relation of $\overline{\delta}_{1,2}^2$ to the mean character difference (M.C.D.) proposed as a measure of taxonomic similarity by Cain and Harrison (1958). The M.C.D. can be summarized by the following formula:

M.C.D. (between taxa 1 and 2)

$$= \frac{1}{n} \sum_{i=1}^{n} | X_{i1} - X_{i2}|.$$

Thus it is the average of the absolute differences between the character state codes of two taxa. It is therefore in the same relation to the distance coefficient proposed above as the average deviation, frequently used in the early days of statistics, is to the standard deviation. It shares some of the disadvantages of the earlier statistic. Thus it will always be less than the true distance and will not permit the partitioning which is possible in mean squares.

Readers familiar with Pearson's coefficient of racial likeness (C.R.L.) will have noticed the similarity of the distance coefficient proposed here to the coefficient employed by Pearson (1926). The latter is conveniently formulated (in its squared form) as follows:

$$\text{C.R.L. (between taxa 1 and 2)} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{(m_{i1} - m_{i2})^2}{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}} \right\} - \frac{1}{n}$$

where n is the number of characters for which the two taxa are compared, m_{i1} is the mean for character i in taxon 1, s_{i1}^2 is the variance for character i in taxon 1 and n_1 is the number of individuals in taxon 1 from which the mean and variance have been computed.

As already mentioned above Pearson's formulations, which were employed to

measure the resemblances among anthropological material (mostly skulls), led to the development of more refined techniques, the generalized distance procedures of Mahalanobis (1936) and Rao (1952). Pearson's coefficient and the generalized distance procedures are not appropriate for the usual type of work in numerical taxonomy, since in the latter field we use a single code value to represent the state of a given character in a given taxon. We do not therefore consider variation for a given character within a taxon. On the other hand, the material with which Pearson and the Indian statisticians were dealing had characters which were largely continuous, varying with the population, and in their attempts to compute distance coefficients for such material they had to take into account the mean and variance of each of the characters considered. Since the material with which we are working consists to an appreciable proportion of discontinuous characters and since there is in many cases no variation of the particular state code within the taxon the problem of intrataxon variation in state codes does not usually arise in our data and we may assume their values to be fixed. Furthermore as a matter of simple practicality we cannot consider the methods of generalized distance in the study of numerical taxonomy, since these methods require inversions or similarly involved operations on matrices of the order of the number of characters considered. As we are often considering as many as a hundred or more characters the inversions of such matrices would be entirely impractical. No claim is made here that the characters which we are considering in numerical taxonomy are always invariable within taxa. On the contrary; in low ranking taxonomic units, such as subspecies, races, or varieties, as well as high ranking taxa, many character states are not constant for the entire taxon. How such data should be treated will be discussed in a forthcoming publication.

A further modification is desirable in the distance coefficient as proposed here. Since the coding of the various characters is arbitrary and since the range of state codes is therefore likely to vary from as little as one to as much as eight, nine, or even more, or since measurements can be in different units, the possible maximum distance between taxa on the various axes is likely to vary, and consequently unequal consideration of the various characters is likely to be introduced in the computation of a measure of taxonomic similarity. It would therefore be desirable to standardize the characters in some prearranged fashion in order to have the range of variation identical for all characters used in the study. Cain and Harrison (1958) dealt with this problem when they used the maximum expression of a character as 100% and arbitrarily graded all other expressions of the character as some portion of this total percentage. A difficulty attending this practice is that with the introduction of new taxa exhibiting more extreme expression of certain characters, the entire taxonomic material would have to be recoded and the study recomputed, since the new state codes would have to be a recalculated fraction of the new maximum expression of the character.

It would seem preferable to standardize the variation of characters in the customary manner by standardizing each row of the initial data matrix, i.e., all state codes for one character in a given study are transformed to possess a mean of zero and a variance of 1. This is easily accomplished by subtracting from each state code the mean state code for that character and dividing the difference by the standard deviation of the state codes for this character. This operation can be performed automatically when the data are fed into a digital computer or it can quite simply be performed on a desk calculator before the actual distances are computed. When the computations are carried out on a desk calculator it would be desirable to code the data in such a manner as to keep

them all positive in order to avoid having to use negative values (see Table 6, Appendix). It is, of course, true that by this procedure the inclusion of a new taxon exhibiting characters more extreme in their expression than those in the previous study will also provide some computational difficulty. However, if, as is likely, we can assume that the variance of the character is not too different, we can express it as a deviation from the previous mean of the character state codes using the previous standard deviation, unless we have too many new additions to the particular study.

The formula for distance squared of standardized state codes is:

$$\overline{\delta}_{1,2}^2 = \frac{1}{n} \sum_{i=1}^{n} (\xi_{i1} - \xi_{i2})^2,$$

where ξ_{i1} is the standardized state code of taxon 1 for character i. When we compare the above formula with the formula for Pearson's coefficient of racial likeness for standardized data (see Penrose, 1954),

$$\text{C.R.L. (between taxa 1 and 2)} = \frac{1}{n} \sum_{i=1}^{n} (\xi_{i1} - \xi_{i2})^2 - \frac{1}{n},$$

we note that the two are almost identical.

If all the characters chosen are functions of general size it may be that size will exert an undue influence on the determination of phenetic similarity. In such instances standardization of the columns in the original data matrix should follow the standardization of the rows and is likely to remove bias of this sort.

Evaluation

We do not as yet have comparative studies indicating the relative merits of the various types of coefficients in numerical taxonomy. It is obvious that the relation of the coefficient of distance to coefficients of correlation and association is a complementary one. The larger the distance, the smaller the degrees of association or correlation between taxa. In discussions with various taxonomists I have found that the concept of distance is one that appears more appealing to them than the concepts of correlation and association. It is also somewhat simpler to visualize; through the two- and three-dimensional model the idea of distance between taxa can be grasped easily. Distances are related to the type of result obtained in comparative serological studies. See for example Boyden (1958). From the point of view of computation the distance coefficient is not as easy to compute as an association coefficient, but is rather easier than a correlation coefficient, at least on a desk calculator. A digital computer will probably handle distances and correlations equally rapidly. At the University of Kansas we are at present preparing a program which will handle both correlations and distances simultaneously. The use of this program will permit us to evaluate the relative outcomes of the two methods. Persons interested in using such a program, which will be written for the IBM 650 digital computer with floating decimal arithmetic and index registers, are invited to contact the author.

In structuring the matrix of the coefficients of similarities based on distance we can use the customary methods of cluster analysis. Ordinary methods of factor analysis, which have proven useful in describing classifications and arranging the taxa in a hierarchic system, will not work on distance data as such. However, various methods of multivariate analysis should be able to partition the matrix of distances into its significant components.

Objections may be raised to using an orthogonal coordinate system to plot the distances between taxa, when we know that rows representing characters are correlated. This is a fundamental problem of all Q-type analysis in psychology as well as biology. At this time it is proposed to continue the use of characters as independent information. However, a

study of evaluating the effects of correlation of rows on structure in a Q-type study is planned.

Summary

A coefficient of distance between taxa is proposed as an alternative measure of resemblance or similarity in numerical taxonomy. This distance can be computed as the mean sum of squares of the differences between standardized character state codes of the taxa concerned.

REFERENCES

ANDERSON, E., and E. C. ABBE. 1934. A quantitative comparison of specific and generic differences in the Betulaceae. Jour. Arnold Arboretum, 15:43–49.

BOYDEN, A. 1958. Comparative serology; aims, methods and results. *In* Serological and Biochemical Comparison of Proteins, XIV Annual Protein Conference, W. Cole (ed.) Rutgers University Press, New Brunswick, pp. 3–24.

CAIN, A. J., and G. A. HARRISON. 1958. An analysis of the taxomist's judgment of affinity. Proc. Zool. Soc. London, 131:85–98.

HEINCKE, F. 1898. Naturgeschichte des Herings. I. Abhandl. Dtsch. Seefischerei-Vereins, 2.

MAHALANOBIS, P. C. 1936. On the generalized distance in statistics. Proc. Nation. Inst. Sci. India, 12:49–55.

MICHENER, C. D., and R. R. SOKAL. 1957. A quantitative approach to a problem in classification. Evolution, 11:130–162.

MORISHIMA, H., and H. OKA. 1960. The pattern of interspecific variation in the genus *Oryza;* its quantitative representation by statistical models. Evolution, 14:153–165.

PEARSON, K. 1926. On the coefficient of racial likeness. Biometrika, 18:105–117.

PENROSE, L. S. 1954. Distance, size and shape. Ann. Eugenics, 18:337–343.

RAO, C. R. 1952. Advanced statistical methods in biometric research. John Wiley and Sons, New York.

ROGERS, D. J., and T. T. TANIMOTO. 1960. A computer program for classifying plants. Science, 132:1115–1118.

SNEATH, P. H. A. 1957a. Some thoughts on bacterial classification. Jour. Gen. Microbiol., 17:184–200.
1957b. The application of computers to taxonomy. Jour. Gen. Microbiol., 17:201–226.

SOKAL, R. R. 1958. Quantification of systematic relationships and of phylogenetic trends. Proc. Tenth Internation. Congr. Entomology, 1:409–415.
1960. Die Grundlagen der numerischen Taxonomie. Proc. Eleventh Internation. Congr. Entomology (in press).

SOKAL, R. R., and C. D. MICHENER. 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull., 38: 1409–1438.

ZARAPKIN, S. R. 1934. Zur Phänoanalyse von geographischen Rassen und Arten. Arch. Naturgesch., new ser., 3:161-186.

ROBERT R. SOKAL is Associate Professor in the Department of Entomology at the University of Kansas, Lawrence.

Appendix

Some readers may find a computational example helpful. I have chosen the small, fictitious example employed by Cain and Harrison (1958) to illustrate the computations for the mean character differences. Table 1 gives the value of the character

TABLE 1—TAXA

	A	B	C	D	E	$\sum_{j=1}^{t} X_{ij}$	$\overline{X}$
Characters 1	100	90	95	5	10	300	60
2	100	85	80	5	5	275	55
3	75	100	95	20	15	305	61
4	5	5	0	80	100	190	38
5	90	100	—	20	25	235	58.75
6	100	50	25	—	—	175	58.33
7	90	95	100	0	0	285	57
$\sum_{i=1}^{n} X_{ij}$	560	525	395	130	155		
$\sum_{i=1}^{n} X_{ij}^2$	51850	46875	35075	7250	10975		

Character state codes for seven (*n*) characters of five (*t*) hypothetical species (taxa) from data by Cain and Harrison (1958). Maximum expression of a character is represented by 100, its absence by 0. Missing information is symbolized by a hyphen.

states for seven characters and for five species (taken from Table 1 of Cain and Harrison).

Table 2 gives differences and differences squared between taxa A and B. The sum of the squared differences, 3575, can also be obtained from the formula

$$\delta_{A,B}^2 = \sum_{i=1}^{n} X_{iA}^2 + \sum_{i=1}^{n} X_{iB}^2 - 2\sum_{i=1}^{n} X_{iA}X_{iB}$$

$\sum_{i=1}^{n} X_{iA}^2$ is obtained by squaring and summing column A of Table 1, $\sum_{i=1}^{n} X_{iB}^2$ by similar treatment of column B (Table 1). $\sum_{i=1}^{n} X_{iA}X_{iB}$ is the sum of the products of columns A and B. (Table 2). Hence $\delta_{A,B}^2$ will be $51{,}850 + 46{,}875 - (2)47{,}575 = 3575$. The mean squared differences $\overline{\delta_{A,B}^2}$, is obtained by dividing $\delta_{A,B}^2$ by n, the number of characters. Therefore $\overline{\delta_{A,B}^2} = 3575/7 = 510.7$. The theoretical maximum of $\overline{\delta^2}$ for any pair of taxa in this instance is clearly 10,000. Thus taxa A and B are quite closely related. Where there are missing values for certain character states, as in taxa C, D, and E, the deviations are taken only where applicable and are divided by the number of differences employed. Thus $\overline{\delta_{C,D'}^2} = 35{,}750/5 = 7{,}150$, as can be seen from Table 2.

If we wish to standardize the state codes for each character (i.e., each row of the original data matrix) we proceed as follows. We first subtract the mean of each row (shown in Table 1) from each character state code in that row as shown

in Table 3. These new rows will consequently add up to zero (as shown near the right hand margin of Table 3). Then we calculate the sum of squares of each coded character. Since the mean is zero, the sum of squares of character i is simply $\sum_{j=1}^{t} X_{ij}^2$, and the variance is that quantity divided by the number of taxa, t, over which this row has been summed. The standard deviation and its reciprocal value are calculated next. Table 4 shows each coded character state value (from Table 3) divided by the standard deviation (or multiplied by its reciprocal). These are the standardized state codes ξ. From them we can calculate $\overline{\delta_{j,k}^2}$ as before. The matrix of distance coefficients is given in Table 5. The expected value for the average distance squared between two randomly picked taxa is $(1.12)^2 = 1.2544$. Thus we find that some coefficients e.g., $\overline{\delta_{A,B}^2}$, are much less than expected, i.e., the taxa are very closely related, while others, e.g., $\overline{\delta_{B,E}^2}$, are much farther apart than expected, i.e., the taxa are not closely related.

The necessarily negative values of some standardized character state codes may lead to computational errors when the work is carried out on desk calculators. Coding the data by adding a constant 5 to all values in Table 4 would remove this difficulty (see Table 6). The reader can easily convince himself that the values of $\overline{\delta_{j,k}^2}$ calculated from Table 6 are identical with those calculated from Table 4.

TABLE 2

		$(X_{1A}-X_{1B})$	$(X_{1A}-X_{1B})^2$	$X_{1A}X_{1B}$	$(X_{1C}-X_{1D})$	$(X_{1C}-X_{1D})^2$
Characters	1	10	100	9000	90	8100
	2	15	225	8500	75	5625
	3	−25	625	7500	75	5625
	4	0	0	25	−80	6400
	5	−10	100	9000	—	—
	6	50	2500	5000	—	—
	7	−5	25	8550	100	10000
	$\sum^{n}$		3575	47575		35750

Some intermediate computational steps for obtaining $\overline{\delta_{A,B}^2}$ and $\overline{\delta_{C,D}^2}$ as explained in the text.

TABLE 3—TAXA

Characters	A	B	C	D	E	$\sum_{j=1}^{t} X_{ij}$	$\sum_{j=1}^{t} X_{ij}^2$	s_x^2	s_x	$1/s_x$
1	40	30	35	−55	−50	0	9250	1850	43.01	.02325
2	45	30	25	−50	−50	0	8550	1710	41.35	.02418
3	14	39	34	−41	−46	0	6670	1334	36.52	.02738
4	−33	−33	−38	42	62	0	9230	1846	42.97	.02327
5	31.25	41.25	—	−38.75	−33.75	0.00	5318.7500	1329.69	36.46	.02743
6	41.67	−8.33	−33.33	—	—	0.01	2916.6667	972.22	31.18	.03207
7	33	38	43	−57	−57	0	10880	2176	46.65	.02144

Character state codes from Table 1 with the row mean subtracted from each code. Computational steps for standardizing rows.

TABLE 4—TAXA

		A	B	C	D	E	$\sum_{J=1}^{t} X_{1j}$
Characters	1	.9300	.6975	.8138	—1.2788	—1.1625	0
	2	1.0883	.7255	.6046	—1.2092	—1.2092	0
	3	.3834	1.0679	.9310	—1.1227	—1.2596	0
	4	—.7680	—.7680	—.8843	.9774	1.4429	0
	5	.8571	1.1314	—	—1.0628	—.9257	0
	6	1.3364	—.2672	—1.0690	—	—	0.0002
	7	.7074	.8146	.9218	—1.2219	—1.2219	0

Character state codes of Table 1 standardized by rows.

TABLE 5

	A	B	C	D	E
A	—	.4732	1.0655	2.7979	4.0242
B	.4732	—	.1191	4.0760	4.3145
C	1.0655	.1191	—	3.9896	4.4011
D	2.7979	4.0760	3.9896	—	.0446
E	4.0242	4.3145	4.4011	.0446	—

Matrix of squared distance coefficients between taxa of Table 4.

TABLE 6—TAXA

		A	B	C	D	E
Characters	1	5.9300	5.6975	5.8138	3.7212	3.8375
	2	6.0883	5.7255	5.6046	3.7908	3.7908
	3	5.3834	6.0679	5.9310	3.8773	3.7404
	'4	4.2320	4.2320	4.1157	5.9774	6.4429
	5	5.8571	6.1314	—	3.9372	4.0743
	6	6.3364	4.7328	3.9310	—	—
	7	5.7074	5.8146	5.9218	3.7781	3.7781

Standardized character state code of Table 4 coded by addition of 5 to avoid negative values.