Correlative Similarity Coefficient: New Criterion for Forming Dendrograms

TAKICHI KANEKO

Institute of Physical and Chemical Research, Wako-shi, Saitama 351, Japan

Presumably, two strains that are similar to each other resemble each other with respect to similarity towards a third strain. An attempt to test this concept was made by a numerical analysis of a sample. The correlative similarity coefficient between two strains was defined as a correlative coefficient between their simple matching similarity coefficients to a third strain and was denoted by $S_C$. The two similarity coefficients were compared and employed for dendrogram preparation. As a result, it appeared that more reasonable clustering of strains was made and that the resemblance of each strain to all the others was well reflected with the use of $S_C$.

The increasing number of publications on computer analysis of taxonomic data does not necessarily mean that the methods presently used for this purpose are perfect. It would appear that additional basic approaches are still required.

When a similarity matrix is shaded, one’s attention is normally attracted to a series of triangles that form along the hypotenuse. However, this is not all that one can find in the matrix. The remaining area of the matrix is usually neither a uniform nor a random mosaic, but has more or less regular patterns. As the triangles indicate the close similarity of the strains involved, a stripe, or a relatively dark area surrounded by a light area or vice versa, suggests that neighboring strains resemble each other with respect to their similarity to a third strain.

Suppose that one may spatially locate the test strains as shown in Fig. 1. Strains 1, 2, and 3 resemble strain 1 nearly to the same degree, but they differ in their similarity to strain j. More generally, the closer strain 1 is to a randomly chosen strain k, the closer strain 2 also is to strain k. Thus, a positive correlation between the phenetic distance between strains 1 and k and that between 2 and k is suggested. In contrast, the resemblance between strains 2 and k is expected to be negatively correlated to that between 3 and k.

This report deals with an attempt to estimate and employ correlations like those referred to above for the preparation of a dendrogram. A new coefficient, to be defined below, is named “correlative similarity coefficient” and is denoted by “$S_C$.”

MATERIALS AND METHODS

Similarity data. For technical reasons, the data on 73 coryneform bacteria published by Davis and Newton (Fig. 1 of reference 3) were used as the sample to be analyzed. The data are given in terms of simple matching similarity coefficients ($SM$); the strain designations used here are the study numbers used by the above-mentioned authors. A complete list of the strains and the dendrograms prepared by simple and complete linkage principles are available in the original paper (3).

Definitions. Let $S(i,j)$ stand for a similarity coefficient between strains i and j, and let $n$ be the total number of strains employed. $S_C(i,j)$ is defined as a correlation coefficient between $SM(i,m)$ and $SM(j,m)$, where $m$ represents strains 1, 2, ..., and $n$.

Dendrograms. Dendrograms were prepared by the unweighted average linkage principle (5). Strain i was thus incorporated into a cluster comprising strains j and k, when the arithmetic mean of $S(i,j)$ and $S(i,k)$ was not less than a given value. Two clusters were linked together if the average of the S’s between each member of a cluster and each member of the other cluster reached a specified level.

Computer and programs. The FACOM 230-75 computer (Fujitsu Co., Tokyo) was used throughout this study. Programs were written in the Fujitsu dialect of Fortran.

RESULTS AND DISCUSSION

The supposition that two similar strains may have similar simple matching coefficients to a third strain was confirmed with some pairs simply by plotting the original data; a sample of this is illustrated in Fig. 2. In this case, the reported $SM$ was 72%, and the estimated $S_C$ was 0.79.

Values of $S_C$ calculated for all the combinations ranged from $-0.79$ (between strains 8 and 67) to 0.99 (e.g., between 63A and 64). This
suggests that some strains are situated together in a multidimensional space, whereas some are separated at almost the opposite ends of the range covered by the experiment. (Unlike $S_{SM}$, $S_C$ is variable and depends on what other strains are used for an analysis.)

Large and small $S_{SM}$'s may be expected to suggest positive and negative values of $S_C$, respectively. However, this held true only if an $S_{SM}$ was large enough or small enough. Figure 3 indicates that the observed deviation of $S_C$ was very large. In fact, even when simple matching

similarities for two pairs of strains were the same, the $S_C$ for one of them could be positive and the other might be significantly negative. This is consistent with the idea that two strains can equally resemble a third strain when directly compared or in terms of $S_{SM}$, but can differ in their positions relative to the others or in $S_C$ (Fig. 1).

Two dendrograms were then prepared with exactly the same program: one based on $S_{SM}$, and the other based on $S_C$. The results are shown in Fig. 4.

It is clear that the two different criteria of similarity produced very similar dendrograms. Actually, the strains fell into five major, very similar clusters in both dendrograms, and most of the strains could be placed in the same order in Fig. 4A and B. Thus, it can be concluded that the two dendrograms supported each other.

It is interesting to note that all of the strains joined in a single group only at a negative $S_C$. In agreement with this, the other dendrogram was completed at 54%, where a negative correlation is normally suggested (Fig. 3).

**Fig. 1.** Schematic similarity map. The irregular figure indicates the scope of an experiment, whereas figures and letters show individual strains.

**Fig. 2.** An example of correlation between the simple matching coefficients of two strains to a third strain. Each circle represents a single strain, but not all are shown because of overlapping.
Figure 3 shows that SC did not vary in parallel with SSM but that their correlation was sigmoidal so that similarity and dissimilarity were emphasized when expressed in SC. It was probably for this reason that the dendrogram and the similarity matrix appeared more clear-cut when SC was used.

Figure 5 shows part of a triangle matrix in which the strains were arranged as suggested by the SC dendrogram and similarity was indicated in SSM. It again suggests that the dendrogram was reasonable even from a different viewpoint.

Although the dendrograms resemble each other, their differences are of taxonomic interest. Numerical taxonomy is generally considered to be useful, especially for the classification of rather unusual organisms, for typical organisms are normally classified without much difficulty. In reality, however, atypical strains often fall into different clusters in dendrograms that are prepared by different principles.

For a comparison of dendrograms, a group of five strains (strains 94, 104, 105, 108, and 109) is used as an example. Davis and Newton constructed two SSM dendrograms: one on the basis of simple linkage and the other on complete linkage methods (3). Their results and the presently derived data may be summarized as shown in Table 1.

When an analysis was carried out with SSM according to the average linkage principle, strain 109 was separated from the other strains because of its low similarity to them and was incorporated into cluster III. Strain 108 was first linked to 105 because the similarity between them happened to be high, whereas strains 94 and 104 were linked together. The four eventually merged into major cluster IV (Fig. 4A, Table 1).

Analysis of the same strains on the basis of SC resulted in a remarkably different clustering (Fig. 4B, Table 1). Strain 108 had low similarity to the other strains of the group and was excluded from them. Figure 5 very clearly demonstrates that this exclusion was reasonable. It also suggests that the other four were rather uniform in their SSM to cluster III and that they had fairly high SSM's to cluster IV. However, they were distinct from cluster IV in their similarity to clusters III and, though not shown, V.
Only clusters I and IV are covered. Strains are labeled as indicated in the attached figure. Similarity table.

FIG. 5. Triangle similarity matrix of some strains. Only clusters III and IV are covered. Strains are shown in the same order as in Fig. 4B. Similarity levels are symbolized as indicated in the attached table.

![Triangle similarity matrix of some strains](image)

**Table 1. Distribution of some strains in different dendrograms**

<table>
<thead>
<tr>
<th>Strains linked by dendrogram preparation method:</th>
<th>( S_{SM} )</th>
<th>( S_C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster*</td>
<td>( S_{SM} )</td>
<td>( S_C )</td>
</tr>
<tr>
<td>Single linkage*</td>
<td>Complete linkage*</td>
<td>Avg. linkage</td>
</tr>
<tr>
<td>III</td>
<td>109</td>
<td>94,104</td>
</tr>
<tr>
<td>IV</td>
<td>94,104</td>
<td>105,108</td>
</tr>
<tr>
<td>Others</td>
<td>109</td>
<td>94,104</td>
</tr>
</tbody>
</table>

* Symbol \( S_{SM} \) indicates the similarity coefficient.

**Conclusions.** An obvious advantage of \( S_C \) over \( S_{SM} \) is that more data are employed for a judgment. In fact, in the course of preparing traditional dendrograms, only single \( S_{SM} \) values are considered for the first clusterings, and all of the others are neglected as negative controls, at least in a simple linkage analysis. Thus, a considerable portion of the extensive data can be left unused or, even in a complete linkage analysis, not fully utilized. It is possible that very important information may be ignored altogether. On the other hand, a single \( S_C \) is estimated from \( 2 \times n \) values when \( n \) strains are analyzed.

Although \( S_{SM} \) is the most popular of the similarity coefficients presently in use, a number of others have been proposed (4, 5) and evaluated (1). However, I should like to emphasize that they all fall into the same category in that they involve a direct comparison of each strain with each of the other strains. Thus, they are direct but can be too microscopic. On the other hand, \( S_C \) is indirect but more macroscopic in nature. In fact, it appears that an \( S_C \) dendrogram could better reflect the resemblance of each strain to all of the strains used for a comparison.

Incidentally, the confidence limits of the estimated correlation coefficients have been extensively calculated (2). However, the following, one of the popular formulae, will be satisfactory for our routine considerations:

\[ S_{SM} \]
Confidence limit = \( t(1 - r^2)^{0.5}/(n - 2)^{0.5} \),
where \( r \) and \( n \) are an estimated coefficient and the size of a sample, respectively, while \( t \) is that of Student with \( n - 2 \) degrees of freedom. Thus, the errors are maximal at \( S_C = 0 \) and vary with the absolute value of \( S_C \). The variance of \( S_{SM} \), on the other hand, has its maximum at \( S_{SM} = 0.5 \) and varies with the absolute value of \( S_{SM} - 0.5 \) (4). Thus, they are similar to each other in statistical properties. The two dendrograms shown in Fig. 4 appear to have roughly equivalent statistical errors per unit length of the horizontal line.

ACKNOWLEDGMENTS
I wish to thank Dr. T. Ando for his suggestions and encouragement. Also, I am grateful to Mr. K. Suzuki for his discussions.

LITERATURE CITED