Distortions of Taxonomic Structure from Incomplete Data on a Restricted Set of Reference Strains

By P. H. A. Sneath

Department of Microbiology, Leicester University, Leicester LE1 7RH, U.K.

(Received 16 June 1982; revised 27 August 1982)

The paper examines how well taxonomic relationships can be estimated when the data are restricted to the similarities between each of the strains and a small subset of reference strains. Such data represent a strip from the similarity matrix rather than the complete matrix. The methods studied were: (a) minimum spanning trees, (b) the definition of one group at a time, and (c) the calculation of ‘derived matrices’. A derived matrix is a complete matrix obtained solely from the entries of the incomplete matrix, by treating these as quantitative character states. The data used were taxonomic distances based on morphological, biochemical and physiological results, and were selected from a previous study to provide good examples of salient patterns of taxonomic relationship. The results that were most similar to those from the complete data were given by derived matrices. Surprisingly little taxonomic distortion occurred, even if the reference strains were rather few, provided these were suitably chosen. Reference strains should be well dispersed, because distortion was considerable if all were very similar to one another. Ideally there should be a reference strain from each cluster, and aids to ensuring this are discussed.

The method has considerable potential for serological or nucleic acid pairing studies in which it is usually impracticable to obtain complete data on numerous strains.

INTRODUCTION

Much taxonomic work is based on nucleic acid pairing or serology. With such methods it is generally impracticable to obtain similarity values between every pair of strains if the strains are numerous. Instead, it is usual to measure the similarity of the entire set of strains to a restricted set of reference strains for which nucleic acid preparations or antisera are available. The worker then draws taxonomic conclusions from this limited array of similarity values. Many examples of such studies might be cited, but those of Gasser & Gasser (1971), London et al. (1975), Krych et al. (1980) and Hebert et al. (1980) are typical.

The question then arises, how accurate are these conclusions from limited data? This is seldom considered, though Cristofolini (1980) has discussed it briefly. The present paper is an attempt to assess the best way to use such incomplete data, and to investigate the safe limits of conclusions drawn from them. The example is from a phenetic study that illustrates various common taxonomic situations. It was not possible to find a serological study in the literature that adequately illustrated these, but the findings apply a fortiori to serology and similar work.

Similarity values between all pairs of operational taxonomic units (OTUs) can be arranged as a square matrix of order \( t \times t \) where there are \( t \) entities to be classified. In numerical taxonomy the matrix is usually symmetrical, so that only the lower left triangle is shown, but with studies on serology and nucleic acids this is not so. It is therefore convenient to consider all similarity matrices as square. With serology and nucleic acid studies the taxonomic process, therefore, does not start from an \( n \times t \) table of \( n \) characters versus \( t \) OTUs, as in most numerical

Abbreviation: OTU, Operational taxonomic unit.
taxonomies, from which a similarity matrix is calculated. Instead, it starts from the similarity matrix itself, wherein the serological or nucleic acid values are the analogues of computed similarity coefficients. The distinction is important and will be referred to repeatedly.

If the data consist of the similarities to a restricted set of reference OTUs, the matrix contains rows or columns that correspond to the reference OTUs. The problem can then be viewed as how one may best obtain from the incomplete matrix a taxonomic structure that is as close as possible to the structure obtainable from the complete matrix.

Three strategies have been commonly used: (a) to define one taxonomic group at a time; (b) to construct a network of closest neighbours; and (c) to derive a new matrix from the incomplete similarities and analyse this. These are not the only possible strategies (see Discussion).

In defining one group at a time the worker must choose a reference OTU from which to start. He must then choose a critical level of similarity which he considers will define an appropriate taxonomic category such as a species. A second reference OTU is then chosen from the remaining OTUs, and the process is repeated. This is continued until all, or almost all, OTUs have been grouped into species. It is convenient to think in terms of dissimilarities, and this implies the choice of a particular radius about each reference OTU which will circumscribe the members of only one species. It is clear that the choice of the correct radius will be critical to success. The concept of defining one group at a time was employed in the algorithm of Rogers & Tanimoto (1960); a bacterial application is that of Silvestri et al. (1962), although the method did not consider incomplete matrices.

The commonest network method is to construct a minimum spanning tree. This is made by joining only the closest pairs of OTUs until all OTUs are linked. The method is closely related to Single Linkage Clustering (Gower & Ross, 1969) and has been occasionally used in microbiology, by Kurylowicz et al. (1975), for example, and various modifications have been used (e.g. Gasser & Gasser, 1971; London et al., 1975) when the similarity matrix is incomplete.

Derived matrices were first developed in serology (Lee, 1968; Basford et al., 1968). Some mathematical aspects are discussed by Sibson (1972). The observed serological cross-reactions were treated as though they were themselves character states of a complex character that was determined by a particular reference antiserum. The concept has been discussed by Cristofolini (1980): it implies that the cross-reactions can be regarded as scores on a serological dimension defined by the antiserum. The way in which a complete matrix is derived is illustrated later but, once obtained, the taxonomic structure can be analysed by usual methods such as clustering or ordination. The process of deriving the matrix introduces distortion. In a preliminary study (Sneath, 1980) it was noted that c reference OTUs define a (c - 1)-dimensional space, with an extra dimension for added OTUs, but with the added factor of reflection into one half of the c-space.

It should be clear that incomplete similarity matrices do not permit reconstruction of all details of the full configuration. Some detail is irretrievably lost. A geographical example will illustrate this. If distances between North American cities are only measured to New York and San Francisco, the distances for Winnipeg in the north and Dallas in the south are almost identical: it will not be possible from this restricted information to say whether Winnipeg and Dallas are close or distant, and different analytical methods will imply different assumptions. Minimum spanning trees will imply they are distant. Derived matrices will imply they are close. The method of constructing one group at a time may imply either, depending on the critical radius adopted. Unlike geographical relationships (where the number of dimensions p is fixed, and where therefore p + 1 reference points will fix positions of all points), taxonomic relationships are usually of indefinitely high dimensionality. There cannot be, therefore, a completely satisfactory solution, but one may hope to find those methods that are best suited to any particular situation.

**METHODS**

*Data.* The data are the taxonomic distances between 36 strains of *Serratia* and allied bacteria (Tables 1 and 2), selected from the larger study of Grimont et al. (1977) so as to illustrate main taxonomic patterns. Eight strains
were chosen from each of four species of *Serratia*, and four isolated strains (singletons) were added. The distances were obtained from $S_{ij}$ values (based on 244 binary characters) as $d = \sqrt{(1 - S_{ij})}$; the same strains are used as an illustration in Sneath (1978). The matrix, $sr0$, of $d$ values (Table 2) is the starting point for the subsequent experiments and represents the reference configuration or 'true relationships' against which the other results are judged.

**Representation of taxonomic structure.** UPGMA phenograms (Sneath & Sokal, 1973) and Principal Coordinate ordinations (Gower, 1966) were selected as commonly used methods, together with shaded similarity matrices for certain cases. Minimum spanning trees were prepared for the cases where there were only a few reference OTUs by retaining in Table 2 only the rows and columns pertaining to those reference OTUs. The trees were prepared from these incomplete matrices as follows. Table 3 shows four OTUs of which two are reference OTUs. The closest pair is 3 and 4, joining at 0.08; this link is therefore made. The next closest pair is 1 and 2, at 0.10. The third is 2 and 3 at 0.20. These three links join all the OTUs, so the tree is as follows, with lengths of the links in parentheses: 1 (0.10), 2 (0.20), 3 (0.08), 4.

**Numerical analysis by derived matrices.** The mathematical conventions are based on those in Sneath (1980). The starting matrix $sr0$ was read by a computer program that allowed the user to specify a given subset of $c$ reference OTUs. The program then computed the new similarities, $S_g$, between all pairs of OTUs, using as the character states only those elements of $sr0$ that related to the reference OTUs. The program permitted a choice of coefficients for $S_g$; the most frequently used was

$$d_{g,k+1} = \left[ \frac{1}{c} \sum_{r=1}^{c} (d_{r,j} - d_{r,k})^2 \right]^{1/2}$$

where $d_{r,j}$ and $d_{r,k}$ are the distances between OTU $j$ and OTU $k$, respectively, and an OTU $r$ that is one of the $c$ reference OTUs. The subscript $g$ refers to the number of iterations of the process of obtaining $S_g$. In most experiments, $g = 0$, but in some there were repeated cycles of calculation.

An illustrative example is shown in Table 3, from which it will be seen that a complete similarity matrix can be obtained. It may be noted that although the derived matrix is symmetrical, the matrix from which the selected columns are available need not be, as in Table 3(a). The method is therefore available where the reaction of antigen $a$ with antiserum $b$ is not the same as that of antigen $b$ with antiserum $a$. There is no need to average the corresponding heterologous values. One can operate either with selected rows or selected columns, and thus obtain similarities between antisera or between antigens (a good example of each is given by Darbyshire et al., 1979). The same principle is true for DNA pairing. If the starting matrix is symmetric, or is made so by averaging, the resulting derived matrices are, of course, identical.

Other $S_g$ coefficients used were $d_{g,k+1}$, order to illustrate the behaviour with Simple Matching Coefficients, $S_{SS}$ (which are the complements of squared distances), and correlation coefficients (calculated from the usual formula over the $c$ reference OTUs, but with $c > 2$). Correlation coefficients are cosines of angles, $\cos \theta$, and to convert them to a form suitable for Principal Coordinates analysis they were transformed to their semichords $\sqrt{1 - \cos \theta}$. The semichords were also used for UPGMA and minimal spanning trees.

**Measures of clustering and dimensionality.** The tightness of the clusters was measured by the sums of squared distances of OTUs to the centroid of their own cluster, symbolized as $SS_d$. The four singletons were treated here as a fifth cluster, and the sum of the five $SS$ terms gave the within-cluster sum of squares, $SS_w$. The ratio

<table>
<thead>
<tr>
<th>Phenon* and species</th>
<th>Strain nos in present study</th>
<th>Respective nos in Grimont et al. (1977)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Serratia marcescens</td>
<td>1-8</td>
<td>5, 16, 293, 207, 60, 261, 57, 12.</td>
</tr>
<tr>
<td>B Serratia rubidaea†</td>
<td>25-32</td>
<td>22, 39, 37, 224, 231, 36, 513, 288.</td>
</tr>
<tr>
<td>C1 Serratia liquefaciens</td>
<td>9-16</td>
<td>9, 221, 319, 283, 239, 278, 507, 503.</td>
</tr>
<tr>
<td>Enterobacter cloacae</td>
<td>33</td>
<td>C1 = NCTC 10005‡</td>
</tr>
<tr>
<td>Enterobacter aerogenes</td>
<td>34</td>
<td>A1 = NCTC 10006‡</td>
</tr>
<tr>
<td>Enterobacter agglomerans</td>
<td>35</td>
<td>E20 = NCTC 9381‡</td>
</tr>
<tr>
<td>Erwinia carotovora subsp. atroseptica</td>
<td>36</td>
<td>E29 = NCPPB 549§</td>
</tr>
</tbody>
</table>

* Phenon letters as in Grimont et al. (1977).
† Listed in Grimont et al. (1977) and Sneath (1978) as *S. marinus* but according to the approved Lists (Skerman et al., 1980) this is a synonym of *S. rubidaea*.
‡ NCTC, National Collection of Type Cultures, London, U.K.
§ NCPPB, National Collection of Plant Pathogenic Bacteria, Harpenden, U.K.
Principal Coordinates analysis is possible with data that show some negative eigenvalues, although Gower (1978) has emphasized the need to watch this aspect critically. In the present study all the derived matrices using $d$ and semichord $r$ were Euclidean distances, so there were no negative eigenvalues larger than expected from rounding errors during computation. With $d^2$, however, substantial negative eigenvalues occurred, and this is referred to later.

The formulae for 'effective dimensionality', $n'$, are simplified from Sneath (1980).

$$n' = \frac{1}{\sum p_i^2}$$

where $p_i$ is $\lambda_i / \sum \lambda_i$, and the $\lambda_i$ values are the non-negative eigenvalues of the Principal Coordinates analysis. When the effective dimensionality $n'$ of a three-dimensional ordination is calculated, the summations are over only eigenvalues 1 to 3.

The concept of effective dimensionality is best appreciated from an ordination such as Fig. 9, in which the configuration can be seen to be almost one-dimensional, although three dimensions are shown. The quantity $n'$, however, can be obtained without an ordination from the $n \times n$ matrix $W$ of sums of squares and cross-products for quantitative or binary characters after subtracting from each character state the mean for that character. This matrix corresponds to the unrotated configuration in $n$-space. If $A$ is the diagonal matrix of eigenvalues of $W$, then the eigenvalues of $W^2$ are the squares of those of $W$, and their sum is the trace of $A^2$. Further, the trace of $W^2$ (because $W$ is real symmetric) equals the sum of the squares of the elements of $W$, which in turn equals $\text{tr}(A^2)$. Therefore $n'$ can be obtained as $[\text{tr}(W^2)] / \sum w_i$. The proof is derived from Searle (1966, pp. 177-178).

Choice of reference OTUs. The sets of reference OTUs which were chosen as of most interest, were as follows. Set a. All 36 strains. The derived matrix sra1 was operated on a second and a third time to give derived matrices sra2 and sra3. The use of all OTUs causes an even effect referred to as a uniform transformation. When, instead of taxonomic distances, correlations followed by semichords and squared taxonomic distances were used, derived matrices srl1 and srm1, respectively, were produced.

Set b. Strains 7, 12, 24 and 28, one from each species of *Serratia*. Within the species the strains were chosen at random. The derived matrix was srb1.
Taxonomy with incomplete data

Table 2. (continued)

|   | 0.000 | .288 | 0.000 | .257 | .214 | 0.000 | .335 | .281 | .308 | 0.000 | .359 | .336 | .316 | .377 | 0.000 | .374 | .385 | .286 | .369 | .370 | 0.000 | .495 | .465 | .482 | .520 | .474 | .517 | 0.000 | .487 | .481 | .479 | .508 | .466 | .505 | .145 | .000 | .490 | .449 | .468 | .506 | .460 | .512 | .192 | .170 | 0.000 | .490 | .449 | .468 | .514 | .469 | .512 | .192 | .158 | .158 | 0.000 | .502 | .469 | .487 | .516 | .482 | .521 | .202 | .158 | .170 | .192 | 0.000 | .481 | .449 | .468 | .481 | .460 | .520 | .241 | .232 | .202 | .202 | .230 | 0.000 | .511 | .469 | .498 | .526 | .496 | .537 | .272 | .241 | .212 | .192 | .241 | .249 | 0.000 | .514 | .471 | .497 | .480 | .472 | .523 | .346 | .316 | .359 | .341 | .354 | .346 | .348 | 0.000 | .540 | .544 | .555 | .508 | .571 | .522 | .594 | .585 | .575 | .582 | .583 | .575 | .587 | .560 | 0.000 | .546 | .577 | .576 | .585 | .570 | .568 | .523 | .535 | .530 | .530 | .534 | .527 | .552 | .452 | 0.000 | .569 | .573 | .587 | .543 | .606 | .529 | .610 | .597 | .598 | .598 | .603 | .592 | .603 | .588 | .519 | .569 | 0.000 | .649 | .636 | .646 | .595 | .657 | .576 | .700 | .692 | .695 | .701 | .703 | .695 | .711 | .650 | .532 | .655 | .520 | 0.000 |

Table 3. Example of calculation of a derived matrix

(a) Complete matrix of immunological distances

<table>
<thead>
<tr>
<th>Antiserum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antigen</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0-12</td>
<td>0-24</td>
<td>0-31</td>
</tr>
<tr>
<td>2</td>
<td>0-10</td>
<td>0</td>
<td>0-20</td>
<td>0-23</td>
</tr>
<tr>
<td>3</td>
<td>0-26</td>
<td>0-20</td>
<td>0</td>
<td>0-07</td>
</tr>
<tr>
<td>4</td>
<td>0-30</td>
<td>0-23</td>
<td>0-08</td>
<td>0</td>
</tr>
</tbody>
</table>

(b) Incomplete matrix from only two reference antisera, 1 and 3

<table>
<thead>
<tr>
<th>Antiserum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antigen</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>—</td>
<td>0-24</td>
<td>—</td>
</tr>
<tr>
<td>2</td>
<td>0-10</td>
<td>—</td>
<td>0-20</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>0-26</td>
<td>—</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>4</td>
<td>0-30</td>
<td>—</td>
<td>0-08</td>
<td>—</td>
</tr>
</tbody>
</table>

(c) Derived matrix: complete matrix of relationships between antigens, obtained by calculating taxonomic distances between rows in (b)

<table>
<thead>
<tr>
<th>Antigen</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antigen</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0-076</td>
<td>0-250</td>
<td>0-240</td>
</tr>
<tr>
<td>2</td>
<td>0-076</td>
<td>0</td>
<td>0-181</td>
<td>0-165</td>
</tr>
<tr>
<td>3</td>
<td>0-250</td>
<td>0-181</td>
<td>0</td>
<td>0-063</td>
</tr>
<tr>
<td>4</td>
<td>0-240</td>
<td>0-165</td>
<td>0-063</td>
<td>0</td>
</tr>
</tbody>
</table>
Fig. 1. Starting configuration: ordination. O, Phenon A, Serratia marcescens; ●, phenon B, S. rubidaea; ▲, phenon C1, S. liquefaciens; △, phenon C2, S. plymuthica; □, singletons. From matrix sr0. The baseplate is drawn at a height of −0.2 on axis III. The figure is another representation of that in Sneath (1978) from a different viewpoint.

Set c. Strains 7, 8, 12 and 15, two from Serratia marcescens and two from Serratia liquefaciens. Strains 8 and 15 were chosen at random from their clusters to add to 7 and 12. The derived matrix was srcl.
Set d. Strains 33, 34, 35 and 36, the four singletons, giving matrix srd1.
Set e. Strains 2, 4, 7 and 8, all from Serratia marcescens. Strains 2 and 4 were randomly chosen to add to strains 7 and 8 used in set c. The derived matrix was srel. Correlations with semichords gave matrix srgl.
Set f. Strain 2 from Serratia marcescens gave the matrix srf1.
Set g. Strain 20 from Serratia plymuthica was chosen because it was the strain closest to the centroid of all OTUs, and it gave matrix srgl.

RESULTS

Standards for comparison

The obvious standard for comparison is the starting configuration (Figs 1, 2). Recovery of this structure is the ideal. However, a uniform transformation (Fig. 3) is also acceptable and is easier to compare with the figures produced by varying the choice of reference OTUs. Figure 3 is also more appropriate when considering the Sums of Squares in Table 4. It should be noted that these are directly relevant to the phenograms, not necessarily to the ordinations.

Starting configuration

Figures 1 and 2 show the starting configuration as, respectively, the three-dimensional ordination and as the phenogram and minimum spanning tree. The four species clusters are well separated. The closest are S. liquefaciens (phenon C1, solid triangles) and S. plymuthicum (phenon C2, open triangles). They are also less compact than S. marcescens (phenon A, open circles) and S. rubidaea (phenon B, solid circles), and this can be also seen from Table 4 by their greater Sums of Squares. The total Sum of Squares (representing the scatter of all the strains) is 4.27, and the sum within the clusters is 1.53, so that the measure of clumping, $R$, is about 0.36 (Table 4).

The clusters appear looser in Fig. 2 than in Fig. 1. This is because the phenogram and minimum spanning tree use the distances in the full hyperspace, whereas the ordination uses
Fig. 2. Starting configuration: UPGMA phenogram (a) and minimum spanning tree (b). They are comparable to those in Sneath (1978) with transformation from $d^2$ to $d$. From matrix $sr_0$. ○, Phenon A, Serratia marcescens; ●, phenon B, S. rubidaea; ▲, phenon C1, S. liquefaciens; △, phenon C2, S. plymuthica; □, singletons.

part of the variation in such a way as to accentuate the inter-cluster distances over the distances within clusters. The UPGMA method gives a phenogram whose cross-bars are average distances between members of groups. The minimum spanning tree joins OTUs that are closest to one another, and thus the links between clusters make the clusters appear closer relative to the appearance in the phenogram. Further, the minimum spanning tree is an unfolded representation of a tree that is folded up within a hypercube whose sides measure 1. The angles
### Table 4. Parameters of the clusters in the configurations corresponding to the figures listed

<table>
<thead>
<tr>
<th>Matrix</th>
<th>No. of reference OTUs where applicable</th>
<th>Eigenvalues, $\lambda$, for principal axes (for comparison with ordinations)</th>
<th>Percentage variation in first three axes</th>
<th>Sums of Squares of:</th>
<th>Ratio $R$ of within to total</th>
<th>Effective dimensionality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>I</td>
<td>II</td>
<td>III</td>
<td>IV</td>
<td>Phenon A</td>
</tr>
<tr>
<td>sr0</td>
<td>1, 2, 14</td>
<td>1.084</td>
<td>0.831</td>
<td>0.508</td>
<td>0.420</td>
<td>56.7</td>
</tr>
<tr>
<td>sra1</td>
<td>3</td>
<td>0.233</td>
<td>0.138</td>
<td>0.118</td>
<td>0.052</td>
<td>77.4</td>
</tr>
<tr>
<td>sra2</td>
<td>4</td>
<td>0.069</td>
<td>0.035</td>
<td>0.022</td>
<td>0.010</td>
<td>86.9</td>
</tr>
<tr>
<td>sra3</td>
<td>5</td>
<td>0.026</td>
<td>0.012</td>
<td>0.004</td>
<td>0.003</td>
<td>89.4</td>
</tr>
<tr>
<td>srb1</td>
<td>6</td>
<td>0.307</td>
<td>0.158</td>
<td>0.122</td>
<td>0.081</td>
<td>87.9</td>
</tr>
<tr>
<td>src1</td>
<td>7</td>
<td>0.410</td>
<td>0.241</td>
<td>0.023</td>
<td>0.017</td>
<td>97.5</td>
</tr>
<tr>
<td>srd1</td>
<td>8</td>
<td>0.184</td>
<td>0.111</td>
<td>0.067</td>
<td>0.047</td>
<td>88.5</td>
</tr>
<tr>
<td>sre1</td>
<td>9, 15</td>
<td>0.725</td>
<td>0.037</td>
<td>0.030</td>
<td>0.004</td>
<td>99.5</td>
</tr>
<tr>
<td>srfl</td>
<td>10</td>
<td>0.910</td>
<td>0.040</td>
<td>0.000</td>
<td>0.000</td>
<td>100</td>
</tr>
<tr>
<td>srb1</td>
<td>11</td>
<td>2.735</td>
<td>1.904</td>
<td>0.879</td>
<td>0.856</td>
<td>69.2</td>
</tr>
<tr>
<td>srl1</td>
<td>12</td>
<td>5.511</td>
<td>2.253</td>
<td>0.888</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>srm1</td>
<td>13</td>
<td>0.016</td>
<td>0.011</td>
<td>0.005</td>
<td>0.001</td>
<td>106.7</td>
</tr>
</tbody>
</table>

* The values for $n'$ and $m'$ are the same because all the variation in these cases can be expressed in 3 dimensions or less.
† If calculated from the positive eigenvalues only, this becomes 95% (see text).
‡ Calculated from the positive eigenvalues only with $p_i = \lambda_i / \Sigma \lambda_i$ (+ve).
between the links are arbitrary, and chosen for convenience of display ('unfolded' is here used in the ordinary sense, not in that in the psychometric literature mentioned in the Discussion).

 Certain strains stand out in all the representations. There is no markedly aberrant strain in phenon A, but in phenon B strain 32 is atypical, and is seen to be so in the ordination, phenogram and minimum spanning tree (where it is respectively the rightmost, the lowest, and the highest of its cluster). In phenon C1 strain 16 is atypical (respectively the highest, lowest and rightmost in the three representations). In phenon C2 strain 24 is atypical (respectively the highest, lowest and lowest). Of the singletons, strain 36 stands out as the most aberrant (rightmost, lowest and lowest respectively). These strains are convenient markers of the amount of within-cluster detail that is preserved in the later configurations.

*Uniform transformation: first cycle*

Figure 3 was obtained by employing all the OTUs as reference OTUs and calculating new Euclidian distances by formula 1 with $g = 0$. It can be seen that this produces an almost uniform
Fig. 4. Results of second cycle of uniform transformation with all 36 OTUs as reference OTUs: d,
ordination and UPGMA phenogram. Conventions as in Figs 1 and 2. From matrix sra2.

transformation of the starting configuration, in which no particular influence is exerted by one
OTU rather than another.

Comparison with Fig. 1 shows that the points have markedly condensed. The species clusters
are still well separated and C1 and C2 are still the closest. The OTUs that were furthest out in
Fig. 1 (particularly the singletons) have been relatively pulled inward. The outlying strains of the
clusters are still recognizable in Fig. 3 as being aberrant. The phenogram is similar to that in Fig.
2 except for the shrinkage, though the singletons are relatively closer and form a single group.
The minimal spanning tree is also similar to that in Fig. 2, though it appears to be somewhat
more linear in arrangement.

Although the total Sum of Squares has fallen from 4.27 to 0.63, the shrinkage has been
relatively greater within the clusters than between them, as is seen by the lower value of \( R =
0.18 \) (Table 4). The tightness of clustering has thus been somewhat increased.

Uniform transformation: repeated cycles

If the distances of the derived matrix used to prepare Fig. 3 are operated on for a second time
with formula 1 \((g = 1)\), with all 36 strains as reference OTUs, further shrinkage takes place (Fig.
4). The results of a third cycle are shown in Fig. 5. These figures are instructive in showing the
more extreme effects of the uniform transformation. The points shrink into a small region,
though the clusters remain separate, and indeed the tightness of clustering increases somewhat
\((R \text{ in Table 4 decreases})\).

The singletons move round to fill the gap between phenons A and B (though they link most
closely to C1 and C2). The shrinkage becomes extreme, and this is emphasized by retaining the
same scale of \( d \) in Figs 1–5 (the minimum spanning tree becomes too cramped to illustrate in
Figs 4 and 5). Nevertheless, the broad taxonomic structure is preserved, though detail is lost (in
Fig. 5 only the aberrant singleton 36 is still clearly recognizable by its extreme position).
Effect of different reference OTUs on derived configurations

Figures 6–10 show the effect of different selections of a few reference OTUs (marked with stars) on the derived configurations. The choice of reference OTUs was made to illustrate situations that might well occur in practice.

One reference OTU from each species. In Fig. 6 are shown the results of choosing one reference strain from each of the four species. The configuration has shrunk about as much as with the uniform transformation (Fig. 3), as shown by the Total Sums of Squares in Table 4. Each species is still broadly distinct, but there are pronounced distortions. Each species cluster appears to have expanded, but the ratio $R$ (Table 4) of 0.147 shows that they are on the whole about as tight relatively as in Fig. 3. The appearance of expansion is largely because the reference OTUs are relatively far away from other members of their own cluster and in each case they are peripheral. Each now appears to be an aberrant member of its own cluster; only strain 24 was aberrant in the starting configuration. This effect is sufficient to lead to misclustering of two reference OTUs in the phenogram. In contrast, the singletons have been compressed into a spurious cluster.

A major effect, therefore, is that strains that are not reference OTUs have been moved toward the centre. Some original detail is retained: strain 36 is still the most aberrant singleton (rightmost) and strain 16 is the most aberrant and rightmost member of cluster C1.

The minimum spanning tree also shows each species cluster as distinct, but the singletons are attached to two clusters. Three of the singletons might be mistaken as members of phenon C2. The bridging strains could be allocated to the wrong cluster. The most aberrant strains in this tree are not necessarily the most outlying in the original configuration, as this will depend on the situation of the reference OTU. Thus the strain at 9 o’clock in species A (strain 4) is not particularly aberrant in the original configuration.

Two reference OTUs from each of two species. Figure 7 shows the results of choosing two
Fig. 6. Effect of choosing four reference OTUs (marked with stars), one from each species: d, ordination, UPGMA phenogram (a) and minimum spanning tree (b) constructed from only those distances in the original configuration that relate to the reference OTUs. Conventions as in Figs 1 and 2. From matrix srbl.

reference OTUs from species A and two from species C1. It can be seen that these two species have expanded relative to Fig. 3, and this is not due simply to the peripheral position of the reference OTUs. The Sums of Squares (Table 4) for A and C1 are much larger than those for B or C2, but the clusters on the average are about as tightly clustered \( R = 0.14 \).

The reference OTUs appear as atypical members of their clusters, and the other members tend to lie in a line pointing toward the centre of the configuration. Opposite them, the other two clusters and the singletons are compressed together: this is also evident in the phenogram, in which several misclusterings are seen. The misclustered strains include the more aberrant members of clusters: thus the atypical strain 16 of phenon C1 appears within C2, and the atypical strain 32 of phenon B has been lost to a spurious cluster of singletons. The most distant singleton is still strain 36.
Fig. 7. Effect of choosing four reference OTUs, two from species A and two from species C1: d, ordination, UPGMA phenogram (a) and minimum spanning tree (b) constructed as described in Fig. 6. Conventions as in Figs 1, 2 and 6. From matrix sr1.

The minimum spanning tree shows confusion between clusters B and C2, with a few other misplaced strains. The singletons are attached to all the clusters.

Four reference OTUs, all singletons. Figure 8 shows the great compression of all strains other than reference OTUs, and R (Table 4) has increased, reflecting the poor distinctness of the species clusters. The phenogram shows the same features, with some miscustering. As with Fig. 7, the strains that have strayed into wrong clusters are mainly atypical cluster members (strains 16, 24 and 32 are among them).

The minimum spanning tree shows only two main groups, with species A and B in one, and C1 and C2 in the other as well as two misclusterings.

Four reference OTUs, all from one species. Figure 9 shows the results of choosing all four reference OTUs from species A. The configuration is somewhat less contracted than Figs 6–8
Fig. 8. Effect of choosing four reference OTUs, all singletons: (a) ordination, UPGMA phenogram and minimum spanning tree constructed as described in Fig. 6. Conventions as in Figs 1, 2 and 6. From matrix srd1.

(Total Sums of Squares = 0.80). The average tightness of clustering is good ($R = 0.13$), but this is partly due to the wide separation of phenon A from the rest.

This figure and the Sum of Squares for A shows that species A has greatly expanded by comparison with the uniform transformation (Fig. 3). The reference strains are again peripheral and widely separated from one another. A striking feature is the arrangement of all strains that do not belong to phenon A into almost a straight line leading away from the centre of phenon A. There is some misclustering of the other groups in the phenogram, but in the minimal spanning tree it is phenon A which shows most misclustering, or at least would be the most difficult phenon to interpret correctly.
Fig. 9. Effect of choosing four reference OTUs, all from one cluster: \( d \), ordination, UPGMA phenogram (a) and minimum spanning tree (b) constructed as described in Fig. 6. Conventions as in Figs 1, 2 and 6. From matrix srel.

Strain 32 is the rightmost member of phenon B in the ordination, and the lowest in the phenogram; strain 36 is still the most outlying singleton; but little fine detail has been preserved. As in earlier configurations the strains that are misclustered (e.g. 16 and 24) tend to be the aberrant members of their own clusters.

A single reference OTU. The extreme case of choosing a single reference OTU is shown in Fig. 10. This case might almost be said to be indeterminate. Strain 2 was chosen as reference OTU because it is the closest to the centroid of cluster A. The derived configuration is entirely linear, with all the variation in the first principal axis (\( \lambda_1 = \text{Total Sum of Squares, Table 4} \)). The
configuration is less contracted than the earlier derived configurations, but the low $R$ value is largely due to the gap between phenon A and the rest.

Phenon A is relatively much expanded (as shown by its Sum of Squares in Table 4). The reference OTU is at one end of the line of points and is very far out. The most typical member of species A is therefore, both in the ordination and the phenogram, made to seem the least typical. In contrast, the minimum spanning tree is a single fan, entirely different from the straight line of the ordination. Though in the minimum spanning tree the reference OTU is made to seem the most typical because it is in the centre, it should be noted that the same effect would be obtained whatever strain was used, even if it were the most atypical of all in the study, i.e. strain 36. In all

Fig. 10. Effect of choosing a single reference OTU: $d$, ordination, UPGMA phenogram (a) and minimum spanning tree (b) constructed as described in Fig. 6. Conventions as in Figs 1, 2 and 6. From matrix srf1.
Taxonomy with incomplete data

Another study was made with the single OTU 20 from phenon C2, which is the closest to the grand centroid of the original configuration. The results were similar to Fig. 10; the most typical strain in the study appeared the least typical in the ordination and phenogram; the other clusters were intermingled, and the singletons were at the end of the linear array opposite to the reference OTU. Cluster C2 was now expanded, but there was almost no gap between members of C2 and those of other clusters.

Derived matrices with other similarity coefficients

There are many possible similarity coefficients that could be used in deriving complete matrices, but they fall broadly into distance measures and angular measures. The correlation coefficient is the best known angular measure, and the experiments were repeated using this and transforming the correlations to semichords as described in Methods. The semichords, though they are Euclidean distances from the mathematical aspect, are approximately proportional to the angles for moderate angles, and thus reflect fairly closely the geometric properties expected from angular measures. In addition, some studies were made with the square of the distance in formula 1, which is equivalent to the complement of the Simple Matching Coefficient. Only a few examples are illustrated with figures. It should be noted that minimum spanning trees are either not applicable or are simple transformations of other minimum spanning trees; they are therefore not shown.

Correlation coefficient with all OTUs as reference OTUs. Figure 11 shows the equivalent of the uniform transformation using the correlation coefficient and semichord in place of \( d \). The space is measured in different units from those derived with \( d \), so one cannot say whether the
configuration has expanded or contracted, but the ratio $R$ shows that the clustering is somewhat tighter ($R = 0.22$ as against 0.36 for the original configuration) but about the same as the uniform transformation with $d$ ($R = 0.18$, see Table 4). The relative compactness of the species is much the same; thus the Sums of Squares show that C2 is still the most disperse of the four. Species C1 and C2 are the closest as before.

The taxonomic structure is well preserved, though the singletons now join species B before the other species join. The main difference from Fig. 3 (which was derived using $d$) is that the clusters, and also the singletons, are almost equidistant from the grand centroid. Most strains lie at a semichord distance of about 0.5; the coefficient of variation of the semichords is 0.12, which is smaller than the values for the distances of the strains from the grand centroid in the original configuration or the uniform transformation based on $d$ (coefficients of variation 0.17 and 0.22, respectively). The atypical strains 16 and 32 are seen as atypical in this configuration (the rightmost of their clusters in the ordination, and the lowest in the phenogram).

*Correlation coefficient with four reference OTUs from one species.* This analysis (Fig. 12) is equivalent to Fig. 9 with change of coefficient. On comparing it with that figure it can be seen that the ordination presents an entirely different appearance. In Fig. 12 there are three loose clusters, comprised predominantly of species C1, C2 and B, but each also contains a reference OTU from species A and either additional strains of species A or singletons. There is also one isolated reference OTU. Species A is obviously much expanded as shown by a large Sum of Squares. The phenogram shows the same main features. The appearance in Fig. 12 is very like the minimum spanning tree in Fig. 9: the species that provided the reference OTUs is disrupted and mingled with other species. The fine detail of the original configuration is lost, and none of the aberrant strains is readily recognizable as such.

The tendency for strains to be equidistant from the grand centroid (noted under Fig. 11) is very obvious in Fig. 12: it is indeed exaggerated by the dispersed appearance of the ordination,
Fig. 13. Uniform transformation using $d^2$ with all 36 OTUs: ordination and UPGMA phenogram. The baseplate lies at $-0.04$ on the third principal axis. Other conventions as in Figs 1, 2 and 6. From matrix srml.

because the coefficient of variation, 0-15, is not very low. Again, the average semichord distance from the centroid is close to 0-5, as in the last configuration.

Squared taxonomic distance with all OTUs as reference OTUs. Figure 13 shows the derived configuration obtained by substituting $d^2$ for $d$ in formula 1 and is comparable with Fig. 3. It is equivalent to using the Simple Matching Coefficient. Because of the change of metric the Sums of Squares are not directly comparable with those for Fig. 3. However, if one calculates $R$ from the square roots of the values in Table 4 (which provides some compensation for the change in metric) it is 0-452, indicating rather loose clustering as appears from Fig. 13.

In this configuration the shorter distances are proportionately smaller than in Fig. 3, and this has the effect of making the clustering in the phenogram tighter in Fig. 13, relative to the length of the stems that carry the clusters. The reason that the ordination shows looser clusters is because the metric is a squared Euclidean distance: this gives negative ('imaginary') eigenvalues and 'over' 100% of the variation is contained in the first three dimensions (Table 4). The negative eigenvalues commenced with the 18th, but were small except for the 35th and 36th. They raised the trace to 110-9%. The ordination is therefore not strictly comparable with that in Fig. 3, and comparison of the ordination and phenogram in Fig. 13 will show many discrepancies in detail which stem from this cause. The phenogram is the 'better' representation in the sense of considering all the variation: it shows that the taxonomic structure is broadly well preserved; atypical strains 16, 32 and 36 are the lowest of their clusters; detailed differences are because with the UPGMA method the averages of squared and unsquared distances are not always monotonic.

Other studies with correlations and squared distances. Other studies, corresponding to Figs. 6–9 but with correlations or squared distances, showed a few new points, which are briefly summarized as follows.

With correlations the selection of a reference OTU from each cluster gave a configuration
Fig. 14. Shaded similarity matrix of the original configuration. This is equivalent to that in Sneath (1978) with transformation from $d^2$ to $d$. From matrix sr0. Conventions as in Figs 1 and 2.

similar to Fig. 6 but with the singletons pulled in and clustered with species B and C2. The four species clusters were arranged at the apices of an almost regular tetrahedron of side (semichord) 0-8. The strains were all very close to being 0-5 (semichord) from the grand centroid (coefficient of variation 0-10).

The configuration from correlations corresponding to Fig. 8 (two reference OTUs from species A and two from C1) showed that species A was well-defined and far from the other strains. The other clusters were partly mingled with each other and with the singletons: all the strains not in cluster A formed a curved band opposite cluster A, centred on a point $P$ about $\frac{2}{3}$ of the way from the grand centroid to cluster A. All the strains were roughly 0-5 (semichord) from $P$. Clusters A and C1 were somewhat expanded relative to the others.

The use of correlations with the singletons (comparable with Fig. 8) gave poor taxonomic structure, as there was much misclustering: only species A retained its identity. The reference OTUs were to one side of the configuration, and the other strains were not as strongly compressed as in Fig. 8. The other strains formed a rough arc with a radius of about 0-5 (semichord) facing the singletons, which was centred at a point about halfway between strain 36 and most other strains.

No case with a single reference OTU can be obtained, because a minimum of three reference OTUs is required to give meaningful correlations. With four reference OTUs, also, all variation becomes expressible in three semichord dimensions. The main points about the configurations with correlations were: (a) all points tend to lie at about the same distance from some point $P$ in the ordination; (b) when correlation coefficients are used, clusters are more disperse than with distances and occupy broad regions of the arc; (c) reference OTUs are not markedly peripheral.

Analyses with $d^2$ corresponding to Figs 6–10 gave configurations very similar to those with $d$. The reference OTUs were much more peripheral, and there was greater expansion of clusters that contained reference OTUs than in those figures. Otherwise the differences were small. The negative eigenvalues were larger than for matrix srml. They commenced a little earlier (12th to
Taxonomy with incomplete data

14th eigenvalues), but were still predominantly concentrated in the last few (34th to 36th). The traces ranged from 137 to 177%. In such instances care must be taken in interpreting the percentage variation accounted for in the ordinations (see Table 4).

**Shaded similarity matrices**

Two shaded similarity matrices are shown, Fig. 14 from the starting configuration and Fig. 15 from the derived matrix with \( d \) and four reference OTUs from species A.

Figure 14 shows the usual features of such shaded matrices: the four clusters and the singletons are well displayed and the structure is recognizably similar to that in Figs 1 and 2. In contrast, Fig. 15 (which corresponds to Fig. 9) shows the confounding of clusters for which no reference OTU is available. Species A is reasonably clear, but even a different choice of shading levels does not display well the detail of the other groups. The very aberrant singleton, strain 36, is however recognizable as the lowest strain in the diagram.

**DISCUSSION**

**Previous examples of distortion**

Authors have seldom compared their earlier findings from a few reference OTUs with those obtained later using more of them. It seems that the problem of distortion is not readily recognized. Rohlf (in Sneath & Sokal, 1973, p. 299) noted that OTUs that were very dissimilar might appear spurious close in a derived matrix, and it may be that this referred to the genus *Dineutes* in the study of Basford et al. (1968) on which he was commenting. *Dineutes* was thought to be an outlying genus, but the derived matrix from serological data made it appear very close to certain other genera. As it was not a reference OTU, it may have been drawn inward in the manner illustrated above. Sommerville & Jones (1972) have noted that it may be difficult to obtain a consistent picture of the DNA–DNA relationships of the *Bacillus cereus* complex when
different reference strains were employed. This problem may be much greater than is generally realized.

The fullest discussion is by Cristofolini (1980). He pointed out that incomplete serological data made the position of an OTU uncertain, and that the uncertainty was inversely proportionate to the distance from a reference OTU. Cristofolini's concept of serological axes (each corresponding to one antiserum) differs from the model used here, but the implications are very similar. He illustrates several examples of distortion (Cristofolini, 1980, 1981). Thus Fig. 6 of the latter paper, derived from a single antiserum, shows both the expansion of a cluster of closely allied species and the misclustering of very different species. Several aspects of the problem are discussed by Hildebrand et al. (1982).

An example of extreme reduction in dimensionality is given in Fig. 6 of London et al. (1975), which illustrates the immunological distances of fructose diphosphate aldolase of various streptococci from one reference aldolase (Streptococcus faecalis ATCC 27792). The diagram is described as a phylogenetic map, but this is misleading because its relationships are phenetic and it is one-dimensional. It also gives the impression that very different species like Streptococcus pyogenes and Streptococcus salivarius are serologically very similar, whereas this can only be determined by comparing them directly. Such diagrams are useful if their limitations are realized. However, it is often not clear how they are constructed. Networks such as Figs 8–10 of Gasser & Gasser (1971) are not minimum spanning trees (they do not show full fan-shaped arrangements); they appear to be more similar to diagrams from derived matrices.

**Types of distortion with derived matrices**

Certain mathematical properties of derived matrices (Sneath, 1980) need mention, and can be illustrated by the present examples.

**Reduction in dimensionality.** With Euclidean distances (taxonomic distances and semichords are Euclidean) t points can be represented in t − 1 dimensions. The choice of c reference OTUs causes the derived configuration to be in c dimensions, i.e. c − 1 for the c reference OTUs plus an additional one for adding the non-reference OTUs. In addition, the points are reflected about the hyperplane defined by the c reference OTUs so that they lie in one half of the c-space. This can be seen from Table 4 for the configurations derived from four reference OTUs: all the variation is in four dimensions because it will be found that the sum of the first four eigenvalues equals the Total Sum of Squares. The extreme case is illustrated by Fig. 10, where one reference OTU gives a one-dimensional space with all points lying on the same side of the reference OTU.

This reduction in dimensionality itself causes difficulty in defining clusters. If clusters are to be recognized as distinct (whether by cluster analysis or ordination), they must not overlap heavily in space. Using simplifying assumptions (clusters are hyperspherical multivariate normal, are equal in dispersion, and are themselves distributed in the same fashion) it can be shown that the risk of confusion is related to the chi-square distribution (Sneath, 1980).

First, one makes an estimate of the ratio $\sigma = s_\text{cl}/s_\text{w}$, where $s_\text{cl}$ is the average standard deviation of the cluster centroids on each of the n dimensions, and $s_\text{w}$ is the average standard deviation within clusters for each dimension. These can often be estimated by the formula in Sneath (1974, 1979a), or obtained by inspection of ordination plots. Next, one requires a critical level of overlap such that this, or worse overlap, will be likely to cause clusters to be confused. For clusters of about equal size and density, the centroids must be separated by more than four average cluster standard deviations if they are to be seen as clearly distinct. In terms of the $W$ statistic (Sneath, 1977, 1979b), $W$ must be over 2.0. Third, look up in tables of chi-square the probability $P$ that corresponds to $W_{\text{crit}}^2 = (n - 1)$ for n degrees of freedom. With $W_{\text{crit}} = 2$ as advised, $P$ corresponding to chi-square of $8/\sigma^2$ is to be found. This is the probability that any pair of clusters taken at random overlaps to more than the critical value in n-space. If one wishes to be reasonably sure that no pair of clusters is likely to overlap, $P$ should be less than $2/q^2$ where q is the likely number of clusters in the study.

One may often assume with safety that all clusters are well separated in the full space on n characters. The reduction to a smaller number of dimensions, m, by ordination or by deriving new matrices, may cause undue overlap. If so, it is $P$ for m degrees of freedom that is of interest.
These probabilities increase greatly when \( m \) becomes small. Thus for two-dimensional ordination plots (\( m = 2 \)), clusters must be separated on the average by about 16 standard deviations if \( P \) is to be less than 1\% (Sneath, 1980; Table 1). These risks of not recognizing clusters as distinct (because of overlap due to reducing the dimensionality) apply not only to derivations with reference OTUs but also to ordinations as customarily employed in taxonomy.

With \( c \) reference OTUs it has been noted that the derived space has \( c \) dimensions. Due to the reflection the effective dimensionality is less, something like \( c \div \frac{1}{2} \). Thus, even if the choice of reference OTUs were such as to produce no distortion (in the sense of affecting OTUs differentially), there will be serious risks of not distinguishing clusters when very few reference OTUs are employed.

The risks are increased by differential distortion. An extreme example is shown by Fig. 9. All four reference OTUs are from one species; the derived configuration, however, is not effectively four-dimensional, so that \( m \) is not effectively \( c = 4 \), nor even 3:5; the effective dimensionality can be seen to be little greater than 1.0 (cf. Fig. 10). It is for this reason that the concept of 'effective dimensionality' is introduced, and shown by \( n' \) and \( m' \) in Table 4. It is at present not possible to give an analytic solution to how 'effective dimensionality' depends on the choice of reference OTUs, but some empirical observations are noted below.

The test for the risk of overlap may be illustrated by Fig. 8 where it is desired to know the approximate risk of serious overlap between pairs of clusters at random in the three-dimensional space, if one is given the configuration in four dimensions of the derived matrix. The singletons are excluded from consideration for simplicity. The principal coordinates permit calculation of the dispersion between and within clusters, and gives \( s^2 = 1.63 \times 10^{-3} \) and \( s^2 = 2 \times 10^{-4} \), so \( \sigma^2 \) is about 8-15. Then the \( P \) for chi-square of \( 8/\sigma^2 = 0.98 \) is about 0.08 for 4 degrees of freedom and about 0-19 for 3 d.f. The reduction of dimensions from three to four thus increases the risk of overlap, and 19\% is of the order of magnitude expected on considering the ordination in Fig. 8. The use of the effective dimensions for degrees of freedom may be justified as being statistically conservative: this somewhat increases the probabilities in this example.

**Effective dimensionality.** Table 4 shows the effective dimensionality of the various configurations. One may note that \( n' \) refers to the configuration itself and to the phenogram, whereas \( m' \) refers to the three-dimensional ordination shown in the corresponding figure. The latter is only included to give a visual impression of the effective dimensionality of the three-dimensional figures; it necessarily lies between 1 and 3. It is \( n' \) that is of most interest, though in extreme cases it may give misleading results. This quantity is a descriptive statistic, not a test of whether an eigenvalue is significantly greater than others. Such a test is that of Bartlett (1951; a worked example is given by Hope, 1968, p. 63), but this becomes indeterminate if any eigenvalue is zero; it is of course based on a specific model that is not readily applicable to the present situation.

The starting configuration does not have \( t - 1 = 35 \) effective dimensions, but only 7-5. This is because it is not hyperspherical: some axes are very small compared to others, so that they effectively contribute only a small fraction of a dimension. This may be appreciated by considering a straight line of points in a three-dimensional space; there is only one dimension of variation, and Fig. 10 (\( n' = 1.0 \)) illustrates this. In contrast, the ordination of Fig. 1 is almost spherical, so that \( m' \) is almost 3.

The effective dimensionality can be seen to reflect to some extent the adequacy of the reference OTUs. Figures 6 and 8 have widely spaced reference OTUs, and gave high \( n' \) and reasonable derived structures, although the inadequacy of choosing singletons as reference OTUs is not evident from \( n' \). At the other extreme is Fig. 10 (entirely one-dimensional) and Fig. 9 (almost one-dimensional, because all distances are measured from reference OTUs that are almost at the same point in the hyperspace). Figure 7, with two reference OTUs from each of two clusters is intermediate (\( n' = 2.1 \)), and it will be seen that this configuration is derived from distances that are measured from what are almost two points in hyperspace. A low value of \( n' \) therefore draws attention to the dangers of inadequate reference OTUs as evidenced by their positions as well as by their number. It helps to quantify the statements of Cristofolini (1980) and Hildebrand et al. (1982) that one should not choose reference antisera that are too much alike.
Alternative similarity coefficients and cluster methods for derived matrices

Similarity coefficients. There are two aspects to similarity coefficients. The first is the choice of scaling of the primary observations: this determines properties of the original configuration. Thus one might employ immunological distances or their logarithms, or convert DNA-DNA percent pairing values $D$ into $(100 - D)$ if these transformations gave better original configurations. Criteria for this need further study. In the same class are coefficients for determining similarity from the number of common antigens in gel diffusion tests: the Phi Coefficient has some advantages over the Jaccard Coefficient (Cristofolini & Feoli Chiapella, 1977; Cristofolini, 1980). The second aspect relates to the coefficients used in derivation of the complete matrix. The results given here suggest that there is not a lot to choose between $d$ and $d^2$, but that the correlation coefficient renders reference OTUs less peripheral in the derived configuration than $d$ or $d^2$, but also renders clusters more nearly equidistant. This is partly due to its mathematical properties: all points are equidistant 0.5 from some point in the full configuration when the semichord transformation is used, but this is not necessarily so in the ordinations. The evidence on whether correlations cause more misclustering is inconclusive, but the present study suggests that it does not recover the original configuration as well as $d$ does.

Clustering methods. UPGMA clustering was used as a standard method, and was superior to Single Linkage as judged by some preliminary studies. The relation of the latter to minimum spanning trees suggests that further work here is desirable, particularly on the influence of the numbers of OTUs in different clusters (Williams et al., 1971; Sneath, 1976). The difficulties in choosing appropriate levels for shading in shaded similarity matrices require emphasis: the usual levels may reveal little structure.
Taxonomy with incomplete data

Defining one group at a time

When defining one group at a time, it is clear that a decisive step is the choice of a critical distance, or rather radius, to determine group membership. Few workers explain how they arrive at this, and it is generally implied that experience allows a correct choice of a radius that defines, for example, a species. This is not the place to consider this in detail, but instead the dependence of the final results on the choice of reference OTUs will be briefly explored with the examples reported here.

The first point is that we lack information on the tightness and degree of separation of bacterial species when defined by serological or DNA-DNA measures. If they formed compact and widely separated clusters it would not matter what strains were chosen as reference OTUs, and critical radii would be quickly found from histograms of relationships.

If one attempts this approach it is necessary to know how frequently the choice of an inappropriate strain as reference OTU would lead to misgroupings. Table 5 shows the safe range for a critical radius to define the species in the original configuration. It can be seen that almost any choice of reference OTU will permit choice of an appropriate radius. Only with strain 17, and possibly with strain 21, will this be difficult. Nevertheless, these choices assume that one does already know what the species are. If one did not, the choices would be much more difficult. Some ranges are small. With some strains as reference OTUs, e.g. 13 and 15, it would be difficult to find a suitable radius from histograms of distances, and this would be scarcely possible if one started with a singleton. The critical radii, also, would differ with the species: it can be seen from Table 5 that no single value would be entirely satisfactory, though 0.39 or 0.40 would be acceptable in most cases. And it should be remembered that these clusters are relatively well-separated examples. In addition the order in which species are studied may make a difference to the results. It would be worthwhile to re-examine the method of Rogers & Tanimoto (1960) in the context of incomplete similarity matrices.

Comparison with the minimum spanning trees raises interesting points. The initial choice of one reference OTU yields a tree similar to that in Fig. 10. The choice of a second reference OTU that is clearly too far away to belong to species A would yield a tree with two fans, from which a third, and fourth reference OTU could be selected. This would ultimately give a minimum spanning tree similar to that in Fig. 6, which shows quite satisfactory structure. Nevertheless, any mistake in the choice could become misleading: if one started with singletons (Fig. 8, implying too big a radius), or accumulated several reference OTUs from one species (Fig. 6, implying too small a radius), the trees would be difficult to interpret and to improve. It would however be most instructive to apply the methods of derived matrices, minimum spanning trees, and defining one group at a time, to serological or DNA data whose detailed taxonomic structure was known, and also to DNA–RNA diagrams such as those of De Ley et al. (1978).

Analogous situations

There are few parallels to the problem studied here. When one considers that most taxonomy is based on very incomplete data it is surprising how little attention has been given to the matter. It is known from experience that moderate percentages of missing values in \( n \times t \) tables do not greatly affect numerical taxonomies (Sneath & Sokal, 1973, p. 181), if these gaps are haphazard, or at least do not affect a majority of characters of any OTU. Similarly, methods for non-metric scaling and ordination may be adapted for missing similarity values, and the work of Gleason & Staelin (1975) suggests that a substantial proportion may be missing without serious effects provided there is nothing systematic in the pattern of missing values. The pattern is, however, very systematic in the present problem, so their findings scarcely apply. The process of deriving a new matrix has analogies to the iterative weighting of characters described by Hogeweg (1976) where the transformation is determined by the choice of characters rather than OTUs. A uniform transformation with the correlation coefficient was proposed by Kaneko (1979). Figure 11 is closely analogous to the results of his algorithm. It seems an open question whether Kaneko’s procedure is an improvement: one must balance tighter clustering against the risk of misclustering. Rather similar is the suggestion by Baum (1977) that one should power a
Table 5. Dependence of a species radius on the choice of reference OTU

The first value in the safe range is the distance from the strain to the farthest strain of its own species. The second value is the distance to the nearest strain (shown in parentheses) that belongs to another species (or several strains if distances are equal). The distances are from the original configuration (Table 2).

<table>
<thead>
<tr>
<th>Species A</th>
<th>Species B</th>
<th>Species C1</th>
<th>Species C2</th>
<th>Singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain no.</td>
<td>Safe range and no. of nearest strain outside species</td>
<td>Strain no.</td>
<td>Safe range and no. of nearest strain outside species</td>
<td>Strain no.</td>
</tr>
<tr>
<td>1</td>
<td>0.32-0.44 (10)</td>
<td>25</td>
<td>0.35-0.46 (20)</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>0.32-0.45 (9)</td>
<td>26</td>
<td>0.32-0.47 (23)</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>0.32-0.45 (10, 12)</td>
<td>27</td>
<td>0.32-0.45 (20)</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>0.33-0.46 (9)</td>
<td>28</td>
<td>0.34-0.45 (20)</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>0.33-0.44 (10)</td>
<td>29</td>
<td>0.35-0.47 (20)</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>0.31-0.45 (13)</td>
<td>30</td>
<td>0.35-0.45 (20)</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>0.31-0.44 (10)</td>
<td>31</td>
<td>0.35-0.47 (20)</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>0.32-0.43 (9, 10, 12)</td>
<td>32</td>
<td>0.35-0.47 (20)</td>
<td>16</td>
</tr>
</tbody>
</table>

NA, Not applicable, because there are no other strains of these species.

* The second value is less than the first, so there is no safe range; the distance to strains 14 and 15 of another species is less than 0.40, and these strains would be misclassified. In addition strains 9, 10 and 13 are only marginally over 0.40 from strain 17.

† The range is very small (0.386-0.392).
similarity matrix in the hope of finding clearer taxonomic structure, though Gower (1978) has drawn attention to certain dangers in this. A related problem of distortion is that which occurs with ordination of frequency data (see Williamson, 1978). The converse of the present problem has been studied in psychometrics under the name of multidimensional unfolding; this is to reconstruct a complete similarity matrix from a rectangle of entries (e.g. Gold, 1973; Davison, 1976).

Effect of experimental error

The model has assumed that serological or other distances are known exactly. They are of course, like phenetic distances, subject to some degree of uncertainty. Barely detectable cross-reactions correspond to indefinitely large distances, and some figure less than infinity must be allocated to absence of a cross-reaction. It is notable how little detail is usually given on the reproducibility or accuracy of serological and DNA values. This is particularly important for weak reactions, because small differences in weak reactions imply much larger differences in the apparent positions of OTUs than do differences in strong reactions. Some perspective to the example used here may be obtained from the formulae in Sneath & Johnson (1972): using these, the distances in Table 2 and Table 5 have a standard deviation of about 0.03.

Requirements for recovery of taxonomic structure

The method of derived matrices can give good results with appropriate choice of OTUs. Thus Sgorbati (1979) gave a phenogram of 20 bifidobacteria from results with three reference antisera. The matrix was derived using \( d \) from logarithms of indices of serological dissimilarity. The phenogram showed seven clusters. The same 20 strains were later tested with five additional antisera, and a new derived matrix and phenogram was prepared (Sgorbati & London, 1982). The phenograms were very similar. The same seven clusters were seen in the second with only minor rearrangements of two clusters. The three reference OTUs were chosen in the first study so as to be widely spread among the bifidobacteria, and the additional reference OTUs were chosen where possible from clusters that previously had none.

One may conclude that no cluster in a phenogram from a derived configuration can be relied upon unless it was built around a reference OTU. In consequence, the current trend in serology is to construct as complete a matrix of cross-reactions as possible (Cristofolini, 1980) and to associate additional OTUs with the complete structure (e.g. Gorman et al., 1980).

One can be sure with Euclidean spaces that if a pair of OTUs \( j \) and \( k \) (that are not reference OTUs) are shown as distant in a derived configuration then they are indeed distant. One cannot be sure they are indeed close if they appear to be close; they may be separated by a distance of up to the minimum of \( d_{jr} \) or \( d_{kr} \) for all reference OTUs \( r \). This principle has been exploited in strategies used in cladistics (Goodman & Moore, 1971) and could be considered for the present problem. It would be interesting to study various estimates of a probable distance \( d_{jr} \) by using formulae such as the minimum of \( \sqrt{(d_{jr}^2 + d_{kr}^2)} \) for any \( r \). Other strategies for this problem would be worth exploring. J. C. Gower (personal communication) notes that one can add a new point to a Principal Coordinate analysis (Gower, 1968; a taxonomic example is given by Wilkinson, 1970). It would therefore be possible to base the analysis on the \( c \) reference OTUs and add new OTUs to the space thus defined. This could then be followed by clustering or further ordination.

In conclusion, clusters may not be recognizable by any of the methods discussed here unless they are represented by reference OTUs. This therefore places the taxonomist in difficulty: if he has already correctly recognized the clusters he will obtain results that confirm them, but if he has not he may fail to find the clusters at all.

In practice the outlook is less gloomy. The method of defining one group at a time is well known and can be very successful if care is taken. Minimum spanning trees are probably less useful but could perhaps be put to better use than in the past. Derived matrices utilize all the available similarity values and, of the three methods, give the best taxonomic structure with proper choice of reference OTUs. They may even be applicable in a manner that at first sight would appear illogical. Much effort has gone into preparing pure antigens and DNA. It would seem retrograde to make deliberate mixtures of pure preparations. Yet the purification is
normally to separate unwanted material of the same organism. It would seem quite logical to mix pure antigens or pure DNA preparations from different organisms, or antisera to such preparations, in the hope that the resulting polyvalent mixtures would behave as if they were located at the cluster centres. This could offer interesting opportunities for replacing single reference OTUs by composites that better represent the clusters.

Grateful thanks are due to Barbara Sgorbati who drew attention to this problem, and to P. A. D. Grimont for the data here. D. Jones, J. C. Gower, M. J. Sackin, M. E. Rhodes-Roberts and E. M. H. Wellington and a number of other colleagues, are thanked for discussion and advice. The work was supported by a grant from the Medical Research Council.

REFERENCES


