A Model for Computer Identification of Micro-organisms

BY H. G. GYLLENBERG

Department of Microbiology, University of Helsinki, Finland

(Received 4 February 1965)

SUMMARY

Specific groups (or taxa) of micro-organisms can be defined as to their location and extension in a space which is produced when the recorded characters are imagined as dimensions. Since a culture to be identified can be imagined as a point in this space, identification can be achieved by examining which group or taxon occupies the region where the 'unknown' lies.

INTRODUCTION

The use of electronic computers for identification purposes does not constitute a particular problem from the point of view of programming computers. However, as compared with the large literature on numerical classification procedures, the interest concerning computer identification has remained rather scanty. In a recent review by Sneath (1964) only a few references on this topic were given. Nevertheless, in connexion with identification, the requirement for a rapid treatment of available information is often much more pronounced than in connexion with classification. Where this need for obtaining results quickly is combined with the necessity of handling large quantities of material, computers may be successfully applied. An automatic identification procedure based on the principles of numerical classification and description of micro-organisms is presented below.

RESULTS

Background

The present author (Gyllenberg, 1965) has used a geometrical model for the description of microbial populations in numerical terms. The basic idea for this model was adopted from Silvestri and his co-workers (Dr L. R. Hill, private communication; Silvestri, Turri, Hill & Gilardi, 1962). According to Silvestri et al. bacterial cultures under study can be imagined as points in a multi-dimensional space, where the dimensions are given by the two-state tests used to characterize the cultures. By application of factor analysis to the primary data, the number of 'dimensions' can be decreased to a more manageable quantity, and, finally, classification can be based on the clusters or heaps of points (cultures) recognized in the imagined space.

In population description (Gyllenberg, 1965), the population can be defined in a given space by its centre of gravity and its extension (radius). The estimate of the centre of gravity is given by the means of the co-ordinates (e.g. x, y, z, respectively, in a three-dimensional space) of the cultures used to represent the population. As
an estimate for the radius, the empirical formula $r = 2\sqrt[2]{(\Sigma d^2/n)}$ can be used. In this formula $d$ is the distance from the points representing the individual cultures to the centre of gravity, and $n$ is the number of cultures which make up the isolated sample of the population.

However, specific groups of cultures can be defined in a similar way. Automatic identification of an unknown culture is then possible by examining whether, when imagined as a point in a given space, it lies in the region occupied by one of the defined groups. If not, no one of the defined groups (i.e. defined taxa) shares the pattern of characters of the culture under examination; but, at any rate, the taxon to which it shows the greatest affinity can be recognized.

**General procedure**

*Reference information.* Identification along the above lines, like all identification procedures, requires a considerable amount of reference information. Accordingly, the reference information must first be collected according to the purpose of the identification procedure, processed and stored. The first step is to define a standard set of two-state tests for recording the corresponding characters on which the identification will be based. It lies outside the scope of this paper to discuss the principles for selecting the basic tests. It suffices to note that the particular identification purpose is of especial importance, but that yet other factors contribute to the information content of a given test.

The next step is to define the taxa or groups to be considered in identification and to select a suitable number of cultures as representatives of each taxon. These cultures can be referred to as reference cultures. The method of constructing or defining the reference taxa (i.e. whether monothetic or polythetic) does not directly influence the identification model (presuming that the taxa are really valid, which means that the representatives of each taxon show considerable mutual affinities). The characters of the reference cultures are then recorded by means of the selected series of basic two-state tests. The data thus obtained constitute the primary information, which defines the identification model.

*Treatment of the primary information.* The next step is to compute the correlation coefficients for each pair of characters from the primary reference information. To obtain significant coefficients, the total of reference cultures must be large: the number depends on the purpose of the identification and the kind and number of taxa to be considered, but a number greater than 100 can be given as a rough guide. Since two-state characters are used, the attribute $A$ can be numerically expressed as 1, and the attribute not-$A$ as 0, which makes it easy to obtain the correlation coefficients, even without computer aid. The correlation coefficients are arranged into a matrix, which provides the information needed for the principal component analysis.

The reference cultures when imagined as points are not evenly distributed in the multidimensional space which is defined by the basic characters (each character corresponding to a particular dimension), and accordingly their dispersion is hyper-ellipsoidal and not spherical. The principal component analysis (which is a particular kind of factor analysis; see Harman, 1960) determines the vectors or axes of the dispersion hyper-ellipsoid in decreasing order, namely, first the longest, then the
next longest orthogonal to the first, etc. To obtain a space with a manageable number of dimensions, the desired number of axes (or components) is selected in order of length. For population description, three dimensions have been considered sufficient (Gyllenberg, 1965). For identification purposes four or five may be needed, but for the sake of simplicity, only three-dimensional models are considered here. For the principal component analysis treatment by computer is required, but suitable routines may be available for most types of computer.

When the original multidimensional space has been reduced to a three-dimensional space, the final identification space is obtained. The next step, therefore, is to project the points which represent the reference cultures into the identification space. For this purpose a special computer routine is necessary. This ‘projection routine’ condenses the original information about each reference culture into three numbers (in the case of a three-dimensional identification space), namely, the three co-ordinates which indicate the location of the point in the identification space.

Definition of reference taxa in terms of identification space. Before the identification model can be used the reference taxa or groups to be considered in the identification procedure must be defined in terms of the identification space. For this purpose the centre of gravity and the radius of each taxon is computed according to the methods already described in the previous section (Background). Even with this model the borders between taxa which show clearly distinct centres of gravity may remain obscure, and some degree of overlap is perhaps impossible to avoid. However, as will be shown below, it is possible to overcome this confusion at least when the overlap is not too serious. At any rate, it is justifiable to examine the locations and extensions of the reference taxa in relation to each other. When considerable overlaps are found, e.g. when the centre of gravity of a given taxon is found to lie within the space occupied by another taxon, this may show that the reference material was inadequate and/or the set of basic tests was not properly selected, and that identification results may be of questionable value.

Identification procedure. For the ‘unknown’ organism to be identified the same characters have to be recorded as for the reference cultures. In this case too, this primary information is condensed according to the ‘projection routine’, and, in other words, the ‘unknown’ point can now be localized in the identification space. For this purpose the computer goes on to compare the ‘unknown’ with each one of the reference taxa. This is done by examining whether or not the distance from the ‘unknown’ point to the centre of gravity of the taxon considered is smaller than the corresponding radius. If not, the ‘unknown’ does not belong to this taxon or group. When the ‘unknown’ is found to belong to two groups (which is possible when the groups are mutually overlapping), the degrees of affinity of the ‘unknown’ to these groups can be evaluated from the relation \( a = r/d \), where \( a \) stands for affinity, \( r \) is the radius of the group considered, and \( d \) is the distance from the ‘unknown’ to the corresponding centre of gravity. The organism to be identified can then be placed in the group to which it shows the greatest affinity. If the ‘unknown’ does not belong to one of the reference taxa, the affinity-figure can be used to indicate to which taxon or group it is most closely related.

For the identification procedure a special routine is needed. It is of course possible to store additional information concerning the reference taxa in the computer (e.g. the main pattern of characters), and to instruct the computer to indicate
on which points an organism identified as belonging to a certain taxon differs from the main character pattern. It may also be convenient to store even such information which has not been used to derive the identification premises (e.g. which pattern of sensitivity to antibiotics is likely to be found, expected heat and disinfectant tolerances, significance as an indicator of some specific source of contamination, etc.); the computer then provides these data together with the result of identification.

Table 1. An illustration of the identification model

The unit is derived from the basic assumption: positive character = 1, negative character = 0, which means that each character can be described as the segment of a line of length 1; x, y, and z are the co-ordinates corresponding to the three axes of the identification space; d is the distance from the culture (point) to be identified to the centre of gravity under examination; a is the affinity of the culture (point) to the corresponding reference group.

<table>
<thead>
<tr>
<th>Isolate to be identified</th>
<th>Location in the identification space</th>
<th>Relation to the coliform group</th>
<th>Relation to the pseudomonas group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x</td>
<td>y</td>
<td>z</td>
</tr>
<tr>
<td>b06</td>
<td>3.53</td>
<td>0.51</td>
<td>0.26</td>
</tr>
<tr>
<td>b08</td>
<td>3.45</td>
<td>0.60</td>
<td>1.02</td>
</tr>
<tr>
<td>b20</td>
<td>3.36</td>
<td>1.68</td>
<td>0.35</td>
</tr>
<tr>
<td>c15</td>
<td>3.06</td>
<td>1.00</td>
<td>0.23</td>
</tr>
<tr>
<td>e45</td>
<td>3.25</td>
<td>0.43</td>
<td>0.12</td>
</tr>
</tbody>
</table>

An example of application. As a guide to the identification procedure presented above, the following simple example may be given. A model was prepared with the purpose of effecting identification of soil bacteria. Thirty-two simple tests were used to characterize the reference material which included representatives of several reference taxa. To recognize the most significant character correlations even the 'unknowns' to be identified (465 isolates in all) were used to derive the correlation coefficient matrix, and, accordingly, the identification space.

In this example coliform organisms and pseudomonads may be cited as reference groups. These groups were defined in terms of the identification space by the centre of gravity and the radius:

<table>
<thead>
<tr>
<th>Centre of gravity</th>
<th>Radius, r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coliform organisms</td>
<td>3.61 0.76 0.35 0.34 0.74</td>
</tr>
<tr>
<td>Pseudomonads</td>
<td>3.02 1.15 -0.53 0.84</td>
</tr>
</tbody>
</table>

The distance between two points in a three-dimensional space is given by

\[ d = \sqrt{((x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2)}. \]

When applied to the centres of gravity of the coliform organisms and the pseudomonads, respectively, this formula indicated a distance of 1.18, which showed that no overlap of these groups occurred, because 1.18 > 0.34 + 0.74.

For each 'unknown', the distances to the centres of gravity of the various reference groups were computed. For isolate b08 (see Table 1), the distance to the centre of gravity of the coliform group was

\[ d_e = \sqrt{((3.61 - 3.45)^2 + (0.76 - 0.60)^2 + (0.35 - 1.02)^2)} = 0.71. \]
The affinity of isolate b08 to the coliform group was

\[ a_e = \frac{0.34}{0.71} = 0.48. \]

The 'identification data' for five 'unknowns' are given in Table 1. Of them isolate b06 obviously belongs to the coliform group because \( a_e > 1 \) (\( r_e > d_e \)). The isolates b20 and c15, again, belong to the pseudomonas group. Isolate b08 shows a rather low affinity to both the reference groups considered here, and may belong elsewhere, whereas isolate e45 shows a considerable affinity to both groups but, nevertheless, lies outside their spaces.

REFERENCES


