Use of Canonical Variates Analysis in Differentiation of Bacteria by Pyrolysis Gas–Liquid Chromatography

By HALLIDAY J. H. MACFIE, COLIN S. GUTTERIDGE AND JOHN R. NORRIS

Agricultural Research Council, Meat Research Institute, Langford, Nr Bristol, Avon BS18 7DY

(Received 6 July 1977; revised 25 August 1977)

Low-resolution pyrolysis gas–liquid chromatography (p.g.l.c.) can differentiate genera of aerobic food spoilage bacteria. Multivariate statistical techniques were applied but neither principal components nor furthest neighbour cluster analysis produced a consistent differentiation although both confirmed the reproducibility of p.g.l.c. When the distance between genera was redefined in terms of Mahalanobis $D^2$ – a generalized concept taking into account scatter around the mean – good differentiation was observed and could be displayed graphically by plotting the genus group means relative to the first two canonical variate axes. The coefficients of the canonical variates provide a strategy for discriminating between the genus groups. Some practical problems in the identification of unknowns using this technique are discussed.

INTRODUCTION

Pyrolysis gas–liquid chromatography (p.g.l.c.), a process in which complex molecules are rapidly and reproducibly thermally degraded in an inert atmosphere and the products are separated and quantified, was originally proposed as a method for detecting the presence of bacteria by Wilson, Oyama & Vango (1962), who designed a miniaturized system for inclusion on the Surveyor series of spacecraft. The potential of p.g.l.c. for reproducibly fingerprinting micro-organisms was not recognized until Reiner (1965) described the pyrolysis of whole bacterial cells and showed that it was possible to distinguish between species of *Mycobacterium* and even serotypes of *Escherichia coli*. Since then the technique has been applied to numerous types of micro-organisms including *Salmonella* chemotypes and serotypes (Reiner et al., 1972); strains of *Vibrio cholerae* (Haddadin et al., 1973); and species and strains of *Aspergillus* (Vincent & Kulik, 1974). Furthermore, the coupling of pyrolysis to mass spectrometry has been successful in differentiating streptococci (Kistemaker, Meuzelaar & Posthumus, 1975); mycoplasmas (Quinn, 1976); and mycobacteria (Meuzelaar et al., 1976).

Pyrograms of a wide variety of micro-organisms may show characteristic qualitative differences but for most applications bacterial fingerprints have a general qualitative similarity (Quinn, 1974). Thus sophisticated data-processing techniques are required to highlight significant variations in the heights of specific peaks.

Menger et al. (1972) developed a computer program which matched unknowns against a library using both peak heights and retention times. Carmichael, Sekhon & Sigler (1973), using a Taxometric Map procedure (Carmichael & Sneath, 1969), were able to classify some dermatophytes which formed well-defined clusters, and noted a greater difference
between pyrograms of opposite mating types within a species than between pyrograms of different species. Eshuis, Kistemaker & Meuzelaar (1977) calculated numerical expressions for reproducibility, specificity and characteristicity. They investigated the effect upon the discrimination between pyrolysis mass spectra of two Listeria serotypes of using a modified Euclidean distance formula and two-dimensional non-linear mapping of distance matrices. Superior differentiation between serotypes was observed when weight vectors were introduced which were derived from the proposed expression for characteristicity (i.e. the observed ratio between inter-strain variance and intra-strain variance for each peak).

As many as 200 components have been resolved from pyrolysates of whole bacterial cells, but a high efficiency capillary column was required to achieve this (Quinn, 1974). This paper reports the application of multivariate analysis techniques, and in particular canonical variates analysis, to data generated by a low-resolution pyrolysis gas–liquid chromatography system which used short packed columns resolving less than 50 separate components.

**METHODS**

**Organisms and growth conditions.** The bacteria used (see Table 1) were all isolates from meat spoilage experiments with the exception of *Pseudomonas* sp. NCMB320, and one strain of *Microbacterium thermosphactum* which was received from Dr D. S. Clark, NRC, Canada. Strains designated *Microbacterium thermosphactum* were all Gram-positive rods in chains, breaking down to cocci in old cultures, catalase positive, growing at 4 °C but not at 37 °C and capable of growth on streptomycin sulphate/thallous acetate/actidione medium (Gardner, 1966). The *Lactobacillus, Pseudomonas, Micrococcus* and *Moraxella* strains were all identified using the schemes of Cowan & Steel (1974).

Each strain was cultured without shaking for 48 h at 25 °C in a broth containing (g l⁻¹): tryptone, 0.5; yeast extract, 2.5; and glucose, 1; pH 7.2. Cells were harvested by centrifugation, washed three times with 0.85% (w/v) saline and lyophilized using an Exal freeze drier.

**Sample preparation for pyrolysis.** Lyophilized whole cells (2 mg) were suspended in 0.25 ml of distilled water and 10 µl of the suspension was layered directly on to the platinum coil of a Chemical Data Systems rgo Pyroprobe using a micro-syringe (Hamilton). Excess water was evaporated by repeatedly firing the pyrolyser at a setting of 50 °C. Refiring the probe did not produce any shadow peaks. The pyrolysate was swept on to the column by a stream of nitrogen maintained at 20 ml min⁻¹. The injector heater was set at 250 °C and the oven was temperature programmed to rise at 10 °C min⁻¹ from 75 to 200 °C after an initial 2 min hold, and was then maintained isothermally until no further products were eluted. The columns were then cleansed of high boiling point compounds by raising the oven temperature to 230 °C for 15 min. One complete analysis cycle took approximately 80 min. The sensitivity of the flame ionization amplifier was 5 x 10⁻¹⁸ amp for full-scale deflexion. Chart recorders (two in parallel) were set at 2 mV and 5 mV with a speed of 1 cm min⁻¹.

**Data collection.** Each of the 25 lyophilizates was analysed in triplicate. A baseline was set manually to every pyrogram taking into account the movement to and from zero of each cluster of peaks. The establishment of the baseline was an arbitrary procedure but once established it was set to each pyrogram in the same way. The heights of a standard set of 24 peaks, selected because of their reproducibility of resolution throughout the batch of analyses, were measured to the nearest mm. The data were standardized to a common total

<table>
<thead>
<tr>
<th>Group no. (for Figs 4, 5, 6)</th>
<th>Group name</th>
<th>No. of strains</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>Moraxella</em></td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td><em>Pseudomonas</em></td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td><em>Lactobacillus</em></td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td><em>Microbacterium thermosphactum</em></td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td><em>Micrococcus</em></td>
<td>5</td>
</tr>
</tbody>
</table>
peak height in order to remove variations attributable to sample size. The standardized data were punched on to computer cards for analyses using the ICL system 4/70 computer at Rothamsted Experimental Station.

Figure 1 shows the type of pyrogram obtained for *Microbacterium thermosphactum*, the baseline, the relationship between analysis time and oven temperature, and the subset of 24 peaks chosen for multivariate analysis. The pyrograms contained a large percentage of incompletely separated peaks but all the peaks included in the analytical subset could be identified on all the pyrograms using relative retention times. Comparison of Fig. 1 with Fig. 2, a pyrogram of a *Pseudomonas* strain, demonstrates the qualitative similarity of the traces. However, there are a number of obvious quantitative differences, particularly peaks 10, 15 and 19.

**Fig. 1.** Pyrogram of *Microbacterium thermosphactum* recorded at an amplifier attenuation of 32 and a chart recorder sensitivity of 2 mV.

**Fig. 2.** Pyrogram of a strain of *Pseudomonas*; attenuation and chart recorder sensitivity as Fig. 1.
Fig. 3. The three groups A, B and C have similar dispersion matrices. The Euclidean distance between the means of A and B, and B and C, is about the same. The generalized distance, on the other hand, takes account of the scatter of the variables and of the correlations among them. The generalized distance between A and B is much greater than that between B and C. [Reproduced with permission from Marriott (1974).]

**Statistical methods**

The 24 heights which make up the description of a particular pyrogram can be thought of as the coordinates of a single point in 24-dimensional space, each peak representing one axis or dimension of the space. The 75 sets of peak heights (points), one for each replicate of the 25 organisms involved in the analysis will thus define a multi-dimensional scatter of which it is only possible to visualize two, or possibly three, dimensions. The aim of principal components analysis is to obtain successive axes that display the most variation among the points. Frequently only three or four of these new axes are needed to account for the majority of the variation present. In such a case, a plot of the points relative to the first two principal component axes, being that two-dimensional plot which displays the most variation, may allow clusters of points and any outliers to be identified visually. Using cluster analysis the distances between points are transformed to similarities and clusters of points are found by various algorithms (Sneath, 1972). In the present study a furthest neighbour hierarchical clustering algorithm was used. Using this technique, units form clusters, or clusters themselves combine, at the lowest similarity level between any two elements of the new cluster.

Canonical variates analysis. For each genus, the vectors of the mean heights of each peak give the coordinates of the genus mean in the 24-dimensional space already defined. In testing the discrimination between genera, the ordinary Euclidean distance between these means is of little use as it does not take into account the observed scatter of samples around their respective means. Consider the hypothetical example given in Fig. 3, where the two-dimensional scatter of three populations A, B and C is shown. The population means A and C are almost equidistant from B but clearly there is a sense in which population B is closer to C than A. To overcome this, a generalized distance known as Mahalanobis $D^2$ (Mahalanobis, 1936) is defined which is independent of the scale of peak heights and takes account of the correlations between them. This implicitly defines a new set of co-ordinates in which A is much further from B and C. The aim of canonical variates analysis is to recover this set of co-ordinates relative to a set of canonical variate axes which are expressed as linear combinations of the axes of the original 24-dimensional space. Plotting the means relative to the first two canonical axes displays as much of the generalized inter-distances as is possible in two dimensions. When the same scale is used on each axis, circular confidence regions may be placed around each mean (Seal, 1966). Further details of these techniques may be found in Marriott (1974) or Blackith & Reyment (1971).

**RESULTS**

**Principal components analysis**

A total of 42% of the variation displayed by the 75 pyrograms in 24-dimensional space was accounted for by the first two principal components. The plot is shown in Fig. 4. With some exceptions (15, 21, 31 and 41) the sample replicates clustered together quite well. There is, however, no indication of strains clustering naturally into generic groups.
Canonical variates in pyrolysis g.l.c.

Table 2. Mean percentage similarities between and within groups (75 samples)

<table>
<thead>
<tr>
<th>Group no.</th>
<th>Group name</th>
<th>Group no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Moraxella</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Pseudomonas</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Lactobacillus</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>Microbacterium thermosphactum</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>Micrococcus</td>
<td></td>
</tr>
</tbody>
</table>

Cluster analysis

The dendrogram representation of furthest neighbour cluster analysis of the corresponding 75 x 75 similarity matrix produced a similar picture of good overall replication but poor discrimination between groups. The mean between-group and within-group similarity matrix appears in Table 2. As most within-group similarities are larger than the between-group similarities, the strains of each group must, on average, be closer to one another than to the rest of the population. The contrast between group 4 (Microbacterium thermosphactum) and the rest is particularly evident.

Canonical variates analysis

A plot of the five group means relative to the first two canonical variate axes (Fig. 5) displayed 89% of the total generalized variation among group means. The 95% confidence region for group 4 is distinct from the rest and we conclude that we may be able to discriminate between strains of this genus and strains of the other four genera.

Discrimination between groups 1, 2 and 3 is very poor and the data were examined to see whether this was due to variation among replicates. The statistical analysis was repeated on the 25 replicate means; however, to avoid singularity of the within-group covariance matrix (Marriott, 1974) it was necessary to reduce the number of peaks to 20. In the absence of a suitable method of elimination the last four peaks of the subset were removed.

Table 3 gives the lower half of the matrix of Mahalanobis $D^2$ distances between the groups.
Table 3. Matrix of Mahalanobis $D^2$ among the groups (25 replicate means)

Note that there are no units as this is a generalized distance and is independent of scale.

<table>
<thead>
<tr>
<th>Group no.</th>
<th>Group name</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Moraxella</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Pseudomonas</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Lactobacillus</td>
<td>12</td>
<td>15</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Microbacterium thermosphae tum</td>
<td>50</td>
<td>50</td>
<td>58</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Micrococcus</td>
<td>18</td>
<td>24</td>
<td>17</td>
<td>53</td>
<td>0</td>
</tr>
</tbody>
</table>

It is similar in form to Table 2 but, being a distance matrix, only the off-diagonal terms are of interest. The corresponding plot of means relative to canonical variate axes appears in Fig. 6. (In comparing Figs 5 and 6, it will be seen that an apparent reflexion has occurred. This is due to different signs in the canonical coefficients and is unimportant. It is the relative positions of the means that are of interest.)

It is clear from Figs 5 and 6 that eliminating the variation among replicates produces a remarkable improvement in discrimination. None of the 95% confidence regions overlap, implying that unknown strains from any group may be identified by plotting their position on the graph. Since 91% of the total generalized variation is displayed by the first two canonical variates, a satisfactory level of discrimination is achieved without the need to consider the influence of the third or fourth canonical variate axes.

DISCUSSION

Roy & Szinai (1977) have commented that 'despite a plethora of variables, p.g.l.c. can yield a series of peaks with the same relative retention times and the same relative peak heights, time and time again'. Furthermore, a large proportion of the data available in such pyrograms is redundant. Given these two constraints, the application of p.g.l.c. to the identification of bacteria must involve analytical systems, both chemical and statistical, which are demonstrably reproducible, discriminate clearly between groups of organisms, and allow the allocation of fresh data.

The results of the analyses described here indicate the power of principal components analysis as an exploratory technique. The method is particularly useful for detecting outliers or misrecorded data. Cluster analysis provides a useful back up when data cannot be satisfactorily resolved in two or three dimensions.

For discrimination, the Mahalanobis $D^2$ is a particularly powerful concept because it is invariant under any scaling of the original variables. Thus, given that the grouping of the samples is known, the method proposed by Eshuis et al. (1977) of scaling the variables according to reproducibility, specificity or characteristicity becomes redundant. Although the canonical variates analysis was able to reproduce the groupings achieved by conventional bacteriological methods this was not necessarily the best grouping of the pyrograms. To obtain this, other multivariate analyses (e.g. Friedman & Rubin, 1967) may be required and these are being studied.

Given that the canonical variates analysis can discriminate between low-resolution pyrograms of different genera, the next step is to assess its potential in identifying an unknown isolate. In concept this step does not require any further use of a computer because the same subset of peak heights can be multiplied by the canonical coefficients for both the first and second variates to produce the co-ordinates of the unknown sample relative to the original canonical variate axes. However, before low-resolution p.g.l.c. can be used for identification based on canonical variates analysis the following conditions must be
Canonical variates in pyrolysis g.l.c.

Fig. 5. Plot of the genus means of the 75 samples relative to the first two canonical variate axes: 89% of the scatter among the means is expressed in this plot. As both axes are plotted to the same scale the 95% confidence regions for each population are circular with radius equal to the square root of the chi square on two degrees of freedom ($\sqrt{5.99} = 2.45$).

Fig. 6. Plot of the genus means relative to the first two canonical variate axes of the 25 replicate means: 91% of the $D^2$ distances among means are accounted for. The 95% confidence regions are plotted in the same way as in Fig. 5 ($r = 2.45$).

fulfilled: (i) The chromatographic system must employ columns with long analytical lifetimes and good reproducibility. (ii) The axes must be established from data which are fully representative of the types of micro-organism within a particular sampling habitat. (iii) The accuracy of identification must be verified empirically. (iv) A suitable method needs to be defined for removing variables that do not contribute to the discrimination.

Further studies on chromatography column technology are required before low-resolution p.g.l.c. can be employed as a routine analytical system, but canonical variates analysis is ideally suited to both high-resolution p.g.l.c. using capillary columns and pyrolysis mass spectrometry, where reproducibility and repeatability are considerably improved. Furthermore, the rapid analysis times of automated pyrolysis mass spectrometry (Meuzelaar et al., 1976) mean that data libraries can be established in weeks rather than months and canonical variates analysis may well be the key to the rapid identification of samples using minimum calculating facilities.
We are grateful to Dr B. G. Shaw for provision of cultures and to Dr D. S. Mottram for many helpful discussions concerning gas–liquid chromatography. One of us (C.S.G.) thanks the Agricultural Research Council for receipt of a studentship during which this work was undertaken.

REFERENCES


