# INTELLIGENT SYSTEMS FOR THE CHARACTERIZATION OF MICROORGANISMS FROM HYPERSPECTRAL DATA

ROYSTON GOODACRE[1*], REBECCA BURTON[1], NAHEED KADERBHAI[1], ÉADAOIN M. TIMMINS[1], ANDREW WOODWARD[1], PAUL J. ROONEY[2] AND DOUGLAS B. KELL[1]
*[1]Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion, SY23 3DA, Wales, U.K. [2]Bronglais General Hospital, Aberystwyth, Ceredigion, SY23 1ER, Wales, U.K.*

## Summary

Three rapid spectroscopic approaches for whole-organism fingerprinting, *viz.* pyrolysis mass spectrometry (PyMS), Fourier transform infra-red spectroscopy (FT-IR) and dispersive Raman microscopy, were used to analyze a group of 59 clinical bacterial isolates associated with urinary tract infection.

Direct visual analysis of these spectra was not possible, highlighting the need to use multivariate methods to reduce the dimensionality of these hyperspectral data. First the unsupervised methods of discriminant function and hierarchical cluster analyses were employed to group these organisms based on their spectral fingerprints, but none produced wholly satisfactory groupings which were characteristic for each of the five bacterial types. In contrast, for PyMS and FT-IR, the artificial neural network-based approaches of multi-layer perceptrons and radial basis functions could be trained with representative spectra of the five bacterial groups so that isolates from clinical bacteriuria in an independent *unseen test set* could be correctly identified. ANNs trained with Raman spectra identified only 80% of the same test set.

These results demonstrate that modern analytical spectroscopies can provide rapid accurate microbial characterization techniques, but only when combined with intelligent systems.

## 1. Introduction

When a pathogen is isolated in a microbiology laboratory, the time taken for subsequent culture for identification and susceptibility testing may delay the administration of the most appropriate treatment.

For routine purposes the ideal method for microbial characterization would have minimum sample preparation, would analyze samples directly (i.e., would not require reagents), would be rapid, automated, and (at least relatively) inexpensive. With recent

*P.J. Stopa and M.A. Bartoszcze (eds.),*
***Rapid Methods for Analysis of Biological Materials in the Environment,*** 111–136.
© 2000 ***Kluwer Academic Publishers.***

developments in analytical instrumentation, these requirements are being fulfilled by physico-chemical spectroscopic methods, often referred to as 'whole-organism fingerprinting' (Magee, 1993). The most common such methods are pyrolysis mass spectrometry (PyMS) (Goodacre & Kell, 1996), Fourier transform infrared spectroscopy (FT-IR) (Naumann *et al.*, 1991) and UV resonance Raman spectroscopy (Nelson *et al.*, 1992).

PyMS, FT-IR and dispersive Raman microscopy are physico-chemical methods which measure predominantly the bond strengths of molecules (PyMS) and the vibrations of bonds within functional groups (FT-IR and Raman) (Colthup *et al.*, 1990; Ferraro & Nakamoto, 1994; Griffiths & de Haseth, 1986; Meuzelaar *et al.*, 1982). Therefore they give quantitative information about the total biochemical composition of a sample. However, the interpretation of these multidimensional spectra has conventionally been by the application of "unsupervised" pattern recognition methods such as principal components (PCA), discriminant function (DFA) and hierarchical cluster (HCA) analyses. With "unsupervised learning" methods of this sort the relevant multivariate algorithms seek "clusters" in the data, thereby allowing the investigator to group objects together on the basis of their perceived closeness (Everitt, 1993); this process is often subjective because it relies upon the interpretation of complicated scatter plots and dendrograms. More recently, various related but much more powerful methods, most often referred to within the framework of chemometrics, have been applied to the "supervised" analysis of these hyperspectral data; arguably the most significant of these is the application of intelligent systems based on artificial neural networks (ANNs) (Bishop, 1995; Wasserman, 1989).

Urinary tract infection (UTI) or bacteriuria remains a major clinical problem and bacterial resistance to antibiotics is increasing. Indeed, many consultations in general practice are because of bacteriuria (Sleigh & Timbury, 1986). The number of culturable organisms which is regarded as significant clinical bacteriuria is $>10^5$ cells.ml$^{-1}$, and the bacteria typically associated with UTI are *Escherichia coli* (causative organism in 60- 90% of bacteriurias), *Staphylococcus saprophyticus* (30% of women with bacteriuria aged <25 years), *Proteus mirabilis* (10%), *Klebsiella* spp. (which are often multiply resistant to antibiotics), *Staphylococcus aureus* and *Pseudomonas aeruginosa* (especially after catheterization).

In this study a group of 59 bacteria commonly associated with bacteriuria were collected from the local hospital. All isolates were typed by conventional biochemical tests to belong to *E. coli*, *Prot. mirabilis*, *Klebsiella*, *Pseud. aeruginosa*, and *Enterococcus*. The aim of this study was to compare the phenotypic differentiation of these 59 bacterial isolates by PyMS, FT-IR and Raman spectroscopies, and to use ANNs to identify the bacteria from these hyperspectral measurements.

## 2. Materials and Methods

***Strains and cultivation.***

A group of 59 bacteria commonly associated with urinary tract infection, or bacteriuria, were collected from the local hospital. All isolates were typed by conventional biochemical tests to belong to *Escherichia coli* (17), *Proteus mirabilis* (10), *Klebsiella* spp. (4 *K. oxytoca* and 6 *K. pneumoniae*), *Pseudomonas aeruginosa* (10), and *Enterococcus* spp (12). Details are given in Table 1. All strains were cultivated axenically and aerobically on LabM Malthus blood agar base (37 mg.ml$^{-1}$) for 16h at 37°C. After sub-culturing three times to ensure pure cultures, biomass was carefully collected using sterile plastic loops and suspended in 1 ml aliquots of sterile physiological saline (0.9% NaCl). The samples were then analyzed by PyMS, FT-IR and dispersive Raman spectroscopies.

***Pyrolysis mass spectrometry (PyMS).***

Five microlitres of the bacterial samples (ca. 40 mg ml$^{-1}$) were evenly applied to clean iron-nickel foils which had been partially inserted into clean pyrolysis tubes. Samples were run in triplicate. Prior to pyrolysis the samples were oven-dried at 50°C for 30 min and the foils were then pushed into the tubes using a stainless steel depth gauge so as to lie 10 mm from the mouth of the tube. Viton O-rings were next placed approximately 1 mm from the mouth of each tube.

Pyrolysis mass spectrometry was then performed on a Horizon Instrument PyMS-200X (Horizon Instruments Ltd., Heathfield, U.K.). For full operational procedures see (Goodacre and Kell, 1996; Goodacre *et al.*, 1993; Goodacre *et al.*, 1994). Conditions used for each experiment involved heating the sample to 100°C for 5s followed by Curie-point pyrolysis at 530°C for 3s with a temperature rise time of 0.5s.

PyMS data may be displayed as quantitative pyrolysis mass spectra (e.g., as in Figure 1). The abscissa represents the 150 *m/z* ratios, while the ordinate contains information on ion count for any particular *m/z* value ranging from 51 to 200. To remove the influence of sample size *per se* data were normalized as a percentage of the total ion count. Total ion counts were typically in the range 1 - 3.10$^{6}$.

***Diffuse reflectance-absorbance Fourier transform infrared (FT-IR) spectroscopy.***

Ten microlitres of the bacterial samples were evenly applied onto a sand-blasted aluminum plate. Prior to analysis the samples were oven-dried at 50°C for 30 min. Samples were run in triplicate. The FT-IR instrument used was the Broker IFS28 FT-IR spectrometer (Broker Spectrospin Ltd., Banner Lane, Coventry, UK) equipped with an MCT (mercury-cadmium-telluride) detector cooled with liquid $N_2$. The aluminum plate was then loaded onto the motorized stage of a reflectance TLC accessory (Bouffard *et al.*, 1994; Goodacre *et al.*, 1996; Winson *et al.*, 1997).

The IBM-compatible personal computer used to control the IFS28, was also programmed (using OPUS version 2.1 software running under IBM O/S2 Warp provided by the manufacturers) to collect spectra over the wavenumber range 4000 cm$^{-1}$ to 600 cm$^{-1}$. Spectra were acquired at a rate of 20 s$^{-1}$. The spectral resolution used was

4 cm$^{-1}$. To improve the signal-to-noise ratio, 256 spectra were co-added and averaged. Each sample was thus represented by a spectrum containing 882 points and spectra were displayed in terms of absorbance as calculated from the reflectance-absorbance spectra using the Opus software. Typical FT-IR spectra are shown in Figure 2.

ASCII data were exported from the Opus software used to control the FT-IR instrument and imported into Matlab version 4.2c. l (The MathWorks, Inc., 24 Prime Par Way, Natick, MA, USA), which runs under Microsoft Windows NT on an IBM-compatible personal computer. To minimize problems arising from baseline shifts the following procedure was implemented: (i) the spectra were first normalized so that the smallest absorbance was set to 0 and the highest to +1 for each spectrum, (ii) next these normalized spectra were detrended by subtracting a linearly increasing baseline from 4000 cm$^{-1}$ to 600 cm$^{-1}$, (iii) finally the smoothed first derivative of these normalized and detrended spectra were calculated using the Savitzky-Golay algorithm (Savitzky & Golay, 1964) with 5-point smoothing.

***Dispersive Raman spectroscopy.***
Spectra were collected using the Renishaw dispersive Raman spectroscope (Ramascope) (Williams *et al.*, 1994) with a low power (30mW) near infra-red 780nm diode laser with the power at the sampling point typically at 3mW. The instrument was wavelength calibrated with a silicon wafer focused under the x50 objective and collected as a static spectrum centered at 520 cm$^{-1}$ for 10 sec.

Samples were presented as 0.5ml bacterial suspensions in 2ml Supelco clear glass vials, covered with solid caps with aluminum liners. These glass vials were placed sequentially into the sample holder of a Renishaw Macropoint assembly. A 16mm focal length objective, fitted onto the objective system which fits into the standard microscope objective aperture and turns the beam through 90°, was then focused into the sample vial and the stage was locked. The spectrum was collected for 60 sec. The next sample was then placed into the sample holder and the spectral collection procedure was repeated.

The GRAMS WiRE software package running under Windows 95 was used for instrument control and data capture. Spectra were collected over the wavenumber range 200 cm$^{-1}$ to 2300 cm$^{-1}$. The spectral resolution used was ~0.92 cm$^{-1}$. Each sample was thus represented by a spectrum containing 2283 points and spectra were displayed in terms of the intensity of Raman scattering (counts).

ASCII data were exported from the GRAMS WiRE software used to control the Raman instrument and imported into Matlab version 4.2c. l. To minimize problems arising from cosmic rays and noise due to short sampling times the following procedure was implemented: (i) any cosmic rays (which excite the CCD camera) were removed using a median filter with a window of 9 data points; (ii) these spectra were then smoothed using a fast Fourier transform denoising routine which briefly removes the high frequency bins (bins 1-110 were kept) from the Fourier-domain spectra, since these contain predominantly noise. These Fourier-domain spectra were then inversely transformed back to the Raman domain. Typical Stokes Raman spectra are shown in

Figure 3. Note that although the fluorescence is relatively low when cells are excited at 780 nm, the system can not discriminate photons arising by fluorescence from those scattered via the Raman effect.

### *Cluster analysis.*

The initial stage involved the reduction of the dimensionality of the PyMS, FT-IR and Raman data by principal components analysis (PCA) (Causton, 1987; Jolliffe, 1986). PCA is a well known technique for reducing the dimensionality of multivariate data whilst preserving most of the variance, and Matlab was employed to perform PCA according to the NIPALS algorithm (Wold, 1966). Discriminant function analysis (DFA) then discriminated between groups on the basis of the retained principal components (PCs) and the *a priori* knowledge of which spectra were replicates (MacFie *et al.*, 1978; Windig *et al.*, 1983), and thus this process does not bias the analysis in any way. DFA was programmed according to Manly's principles (Manly, 1994). Finally, the Euclidean distance between *a priori* group centres in DFA space was used to construct a similarity measure, with the Gower similarity coefficient $S_G$ (Gower, 1966), and these distance measures were then processed by an agglomerative clustering algorithm to construct a dendrogram (Manly, 1994).

### *Multilayer perceptrons.*

All multilayer perceptrons (MLP) analyses (also known as back-propagation artificial neural networks (ANNs)) were carried out with a user-friendly, neural network simulation program, NeuFrame version 3,0,0,0 (Neural Computer Sciences, Lulworth Business Centre, Nutwood Way, Totton, Southampton, Hants), which runs under Microsoft Windows NT on an IBM-compatible personal computer. In-depth descriptions of the *modus operandi* of this type of MLP analysis are given elsewhere (Goodacre, et al., 1994; Goodacre *et al.*, 1994; Goodacre *et al.*, 1995).

The structure of the MLP used in this study to analyze the hyperspectral consisted of 3 layers, the first layer contains the spectra (made up of the 150 input nodes for PyMS, 882 for FT-IR, and 2283 for Raman), one "hidden" layer containing 8 (PyMS), 10 (FT-IR) or 12 (Raman) nodes, and five output nodes (encoded in binary fashion for the bacterial identities). These were binary encoded such that *E. coli* was coded as 10000, *Prot. mirabilis* as 01000, *Klebsiella* spp. as 00100, *Pseud. aeruginosa* as 00010, and *Enterococcus* spp. as 00001. Each of the input nodes were connected to the nodes of the hidden layer using abstract interconnections (connections or synapses) (see Figure 4 for a diagrammatic representation). Connections each have an associated real value, termed the weight ($w_i$), that scales the input ($i_i$) passing through them, this also includes the bias ($\vartheta$), which also has a modifiable weight. Nodes sum the signals feeding to them (*Net*):

$$Net \;= i_1 w_1 + i_2 w_2 + i_3 w_3 + \ldots + i_i w_i + \ldots + i_n w_n = \sum_{i=1}^{n} i_i w_i + \vartheta$$

The sum of the scaled inputs and the node's bias, are then scaled to lie between 0 and +1 by an activation function to give the nodes output (*Out*); this scaling is typically achieved using a logistic "squashing" (or sigmoidal) function:

$$Out = \frac{1}{(1+\exp^{-Net})}$$

These signals (*Out*) are then passed to the output node which sums them and in turn squashed by the above logistic sigmoidal activation function; the product of this node was then fed to the "outside world".

Before training commenced the values applied to the input and output nodes were normalized between 0.1 and 0.9. The scaling regime used for the input layer was to scale nodally, where the input nodes were scaled for *each* input node such that the lowest mass was set to 0.1 and the highest mass to 0.9. Finally, the connection weights were set to small random values (typically between -0.005 and +0.005).

The algorithm used to train the neural network was the standard back-propagation (BP) (Haykin, 1994; Rumelhart *et al.*, 1986; Wasserman, 1989; Werbos, 1994). For the training of the MLP each input (i.e. spectrum) is paired with a desired output (i.e., the identity of the bacteria); together these are called a training pair (or training pattern). A MLP is trained over a number of training pairs; this group is collectively called the training set, details of the training set are given in Table 2. The input is applied to the network, which is allowed to run until an output is produced at each output node. The differences between the actual and the desired output, taken over the entire training set are fed back through the network in the reverse direction to signal flow (hence back-propagation) modifying the weights as they go. This process is repeated until a suitable level of error is achieved. In the present work, a learning rate of 0.1 and a momentum of 0.9 were used.

Each epoch represented the connection weight updatings and a recalculation of the RMS error between the true and desired outputs over the entire training set (RMS error of formation; RMSEF). During training a plot of the error versus the number of epochs represents the "learning curve", and may be used to estimate the extent of training. All MLPs were trained until the RMSEF was 1.00.

Finally after training, all spectra collected from the bacterial isolates were used as the "unknown" inputs (test data); the network then calculated its estimate and for each sample the largest node in the output layer was taken as its identity.

***Radial basis function neural networks.***

All Radial basis function (RBF) analyses were also carried out with NeuFrame version 3,0,0,0 as detailed specifically by Saha and Keller (Saha & Keller, 1990).

RBFs are hybrid neural networks encompassing both unsupervised and supervised learning (Beale & Jackson, 1990; Bishop, 1995; Hush & Horne, 1993; Moody & Darken, 1989; Park & Sandberg, 1991; Saha and Keller, 1990; Wilkins *et al.*, 1994). RBFs are typically three-layer neural networks and in essence the sigmoidal squashing function is replaced by non-linear (often Gaussian or "Mexican hat") basis functions or kernels (Figure 5). The kernel is the function that determines the output of each node in the hidden layer when an input pattern is applied to it. This output is simply a function of the Euclidean distance from the kernel centre to the presented input pattern in the multi-dimensional space, and each node in the hidden layer only produces an output when the input applied is within its receptive field; if the input is beyond this receptive field the output is 0. This receptive field can be chosen and is radially symmetric around the kernel centre. Between them the receptive fields cover the entire region of the input space in which a multivariate input pattern may occur; a diagrammatic representation of this is shown in Figure 6, where a two dimensional input is mapped by seven radially-symmetric basis functions. This is a fundamentally different approach from the MLP, in which each hidden node represents a non-linear hyperplanar decision boundary bisecting the input space (Figure 4).

The outputs of the RBF nodes in the hidden layer are then fed forward via weighted connections to the nodes in the output layer in a similar fashion to the MLP, and each output node calculates a weighted sum of the outputs from the non-linear transfer from the kernels in the hidden layer. The only difference is that the output nodes of an RBF network are linear, whilst those of the MLP more typically employ a logistic (non-linear) squashing function.

The implementation of these RBF neural networks is exactly as described by Saha and Keller (Saha and Keller, 1990). Briefly the training proceeds in two stages:

*Stage 1* involves unsupervised clustering of the input data, typically using the K-means clustering algorithm (Duda & Hart, 1973; Everitt, 1993; Hush and Horne, 1993) to divide the high- dimensional input data into clusters. Next, kernel centres are placed at the mean of each cluster of data points. The use of K-means is particularly useful because it positions the kernels relative to the density of the input data points. Next the receptive field is determined by the nearest neighbor heuristic where $r_j$ (the radius of kernel $j$) is set to the Euclidean distance between $w_j$ (the vector determining the centre for the $j^{th}$ RBF) and its nearest neighbor ($k$), and an overlap constant (*Overlap*) is used:

$$r_j = \mathit{Overlap} \times \min\left(\| w_j - w_k \|\right)$$

where || ... || denotes a vector norm, or Euclidean distance.

The overlap that gave best results was found to be 2, which means that the edge of the radius of one kernel is at the centre of its nearest neighbor; this optimum was also in agreement with the studies of Saha and Keller (Saha and Keller, 1990).

The output from nodes in the hidden layer is dependent on the shape of the basis function and the one used was that of the Mexican hat. Thus this value ($R_j$) for node $j$ when given the $i^{th}$ input vector ($i_i$) can be calculated by:

$$R_j(i_i) = \exp^{-\left(\frac{\| w_i - i_i \|}{r_j}\right)^2}$$

*Stage 2* involves supervised learning in a single layer MLP. The inputs are the output values for all $n$ basis functions ($R_1$-$R_n$) for all the training input patterns to that layer ($i_1$-$i_n$), and the outputs are the bacterial identities binary encoded in 5 nodes as detailed above.

The output nodes are calibrated using simple linear regression. The optimum number of kernel functions was found by calculating the minimum error for the test set. Finally after training all spectra collected from the bacterial isolates were used as the "unknown" inputs (test data); the network then calculated its estimate and for each sample the winning node in the output layer was taken as its identity.

## 3. Results and Discussion

***The raw spectra.***

Typical normalized PyMS spectra for *Escherichia coli* isolate Ea and *Proteus mirabilis* isolate Pa are shown in Figure 1. These, and the spectra from all 59 bacteria, show an undulating, decaying feature with a periodicity of 14 atomic mass units; due to the loss of $CH_2$ units during pyrolysis (Meuzelaar, et al., 1982). The FT-IR diffuse reflectance-absorbance and Raman spectra of the same isolates are shown in Figures 2 and 3 respectively. These vibrational spectra and those from the other 59 bacteria all showed broad and complex contours; indeed for the Raman spectra it is very difficult to distinguish the Raman scattering from the background and/or any small levels of fluorescence by excitation using the 780 nm laser (although the contribution due to fluorescence should be reduced by the use of the near infra-red laser).

For all three spectral types there was very little qualitative difference between the spectra, although at least some complex quantitative differences between them were observed. Such spectra, uninterpretable by the naked eye, readily illustrate the need to employ multivariate statistical techniques for the analysis of PyMS, FT-IR and Raman data.

***Unsupervised cluster analysis.***

After collection of the three data types, each of the 59 strains, each represented by three replicate spectra, were coded to give 59 individual groups (see Table 1), and analyzed by DFA and HCA as detailed above: The resulting dendrogram from the analysis of the PyMS data is shown in Figure 7 where it can be seen that five clusters are recovered. Although the *Pseud. aeruginosa*, *Prot. mirabilis* and the enterococci strains form three well defined clusters, the *Klebsiella* spp. do not form one group and some of these cluster with the 17 *E. coli* (E) strains analyzed.

The analysis of the FT-IR data by DFA is depicted as a pseudo-3D ordination plot (Figure 8). In this figure it is again clear that the *Pseud. aeruginosa* (A), *Prot. mirabilis* (P) and the enterococci (C) strains form three distinct groups, however, the fourth cluster is a mixture of strains of the *E. coli* (E) and *Klebsiella* spp (O and K).

Finally, DFA was used to analyze the Raman spectra and the results are shown in Figure 9. Figure 9A shows the analysis of a'l the strains and the first discriminant function (DF1) indicates that the majority of the variation was between the *Pseud. aeruginosa* (A) strains and all the other isolates. This is likely to be due to small amount of fluorescence, since it is well known that *Pseud. aeruginosa* naturally fluoresce due to the production of pyocyanin (blue-green) and fluorescein (yellow) pigments (Sleigh and Timbury, 1986), and it is difficult to distinguish this electromagnetic radiation from Raman scattering as both are measured as a shift in wavelength from the 780 nm source laser. Therefore, these isolates were removed and the analysis rerun, the resultant DFA plot is shown in Figure 9B where it can be seen that the different isolates do not group together and only with *a priori* knowledge of the classes can any separation be inferred.

***Supervised analysis using artificial neural networks.***

Since none of the spectroscopies when analyzed by the various cluster analyses produced wholly satisfactory groupings which were characteristic for each of the five bacterial types, the next stage was to supervise the analysis using the artificial neural network-based approaches of multi-layer perceptrons (MLPs) and radial basis functions (RBFs).

As detailed above the first five organisms were used to train the MLPs and RBFs (see Table 2 for strain numbers). The input layers for the MLPs were either the full spectral data or the first few principal components; therefore for PyMS these were 150 m/z intensities or the first 10 PCs (which accounted for 97.16% of total variance), for FT-IR these were the absorbances at 882 wavenumbers or the first 20 PCs (which accounted for 96.88% of total variance), and for Raman were the counts at 2283 wavenumbers or the first 5 PCs (which accounted for 78.86% of total variance). Only the full spectra were used as inputs for the RBFs. The outputs were always the same for both MLPs, PC-MLPs or RBFs and were binary encoded such that *E. coli* was coded as 10000, *Prot. mirabilis* as 01000, *Klebsiella* spp. as 00100, *Pseud. aeruginosa* as 00010, and *Enterococcus* spp. as 00001.

After training each of the ANNs to a RMS error of 0.01 in the training set, each calibrated system was challenged with the training and test sets. For the PyMS data trained with a full spectral MLP the outputs for the training set are shown in Table 2 and the test set in Table 3. Using the criterion that the identity of an isolate from challenging a trained ANN is taken as the winning node (that is to say the largest value)

in the output layer, this PyMS-MLP correctly identified all 25 bacteria in the training set and 33 of the 34 isolates in the *unknown* (unseen) test set. The incorrectly assigned isolate was *Klebsiella* sp. JX58 and was identified as an *E. coli*; this was not surprising since in the dendrogram (Figure 7) this *Klebsiella* isolate and *E. coli* were recovered together. Exactly the same result was seen for the PC-MLPs, by contrast the full spectral RBFs correctly identified all isolates in both the training and test sets.
All three ANN-based methods correctly identified all isolates in the training and tests from their FT-IR data (data not shown).

Different results were seen for the ANN analyses of the Raman spectra. Whilst each method got 100% of the training set correct, each method only identified approximately 80% of the 34 isolates in the test sets correctly. MLPs correctly identified 25 (74%), RBFs 26 (76%), and PC-MLPs gave best prediction with 28 (82%). Generally the *E. coli, Pseud. aeruginosa* and *Enterococcus* spp. were always identified but approximately half of the *P. mirabilis* and *Klebsiella* isolates were incorrectly assigned.

## 4. Conclusions

Three rapid spectroscopic approaches for 'whole-organism fingerprinting' of PyMS, FT-IR and dispersive Raman microscopy were used to analyze a group of 59 clinical bacterial isolates associated with urinary tract infection.

Direct visual analysis of these spectra was not possible, highlighting the need to use multivariate methods to reduce the dimensionality of these hyperspectral data. Unsupervised learning methods of DFA and HCA were employed to group these organisms based on their spectral fingerprints, and although some groups were seen which were characteristic for each of the five bacterial types, wholly satisfactory clustering was not observed until *a priori* information was used in the interpretation of the complicated dendrograms (Figure 7) and ordination plots (Figure 8 and 9).

In contrast, for PyMS and FT-IR, the artificial neural network-based approaches of multi-layer perceptrons and radial basis functions could be trained with representative spectra of the five bacterial groups so that isolates from clinical bacteriuria in an independent *unseen test set* could be correctly identified. ANNs trained with Raman spectra identified only 80% of the same test set. It is likely that this was due to the sample presentation in that the concentration of cells in the slurries was low and future studies will therefore concentrate on analyzing the bacterial samples directly from colonies on petri dishes.

The ANNs for the very high dimensional Raman spectra (2283 wavenumbers) took a long time to train, and for the full-spectral MLPs this was 30 hours. However, we have previously used PCA as a method for reducing the inputs to ANNs (Goodacre *et al.*, 1997; Goodacre, et al., 1996; Timmins & Goodacre, 1997) and in the present study using principal components scores as inputs to MLPs reduced the training time to only 10 mins, with no degradation in the predictive ability of the PC-MLP. Finally the training time for the full spectral RBFs was very quick and only took 2 mins, with equivalent performance compared to the full spectral MLPs.

In conclusion, these results demonstrate that modern analytical spectroscopies, but only when combined with intelligent systems, can provide rapid accurate microbial characterization techniques.

## 5. Acknowledgments

R.G. and E.M.T. are indebted to the Wellcome Trust for financial support (grant number 042615/Z/94/Z). N.K., A.W. and D.B.K. thank the Chemicals and Pharmaceuticals Directorate of the UK BBSRC, Bruker Spectrospin, Glaxo Wellcome, Renishaw Transducer Systems and Zeneca Life Science Molecules for financial support.

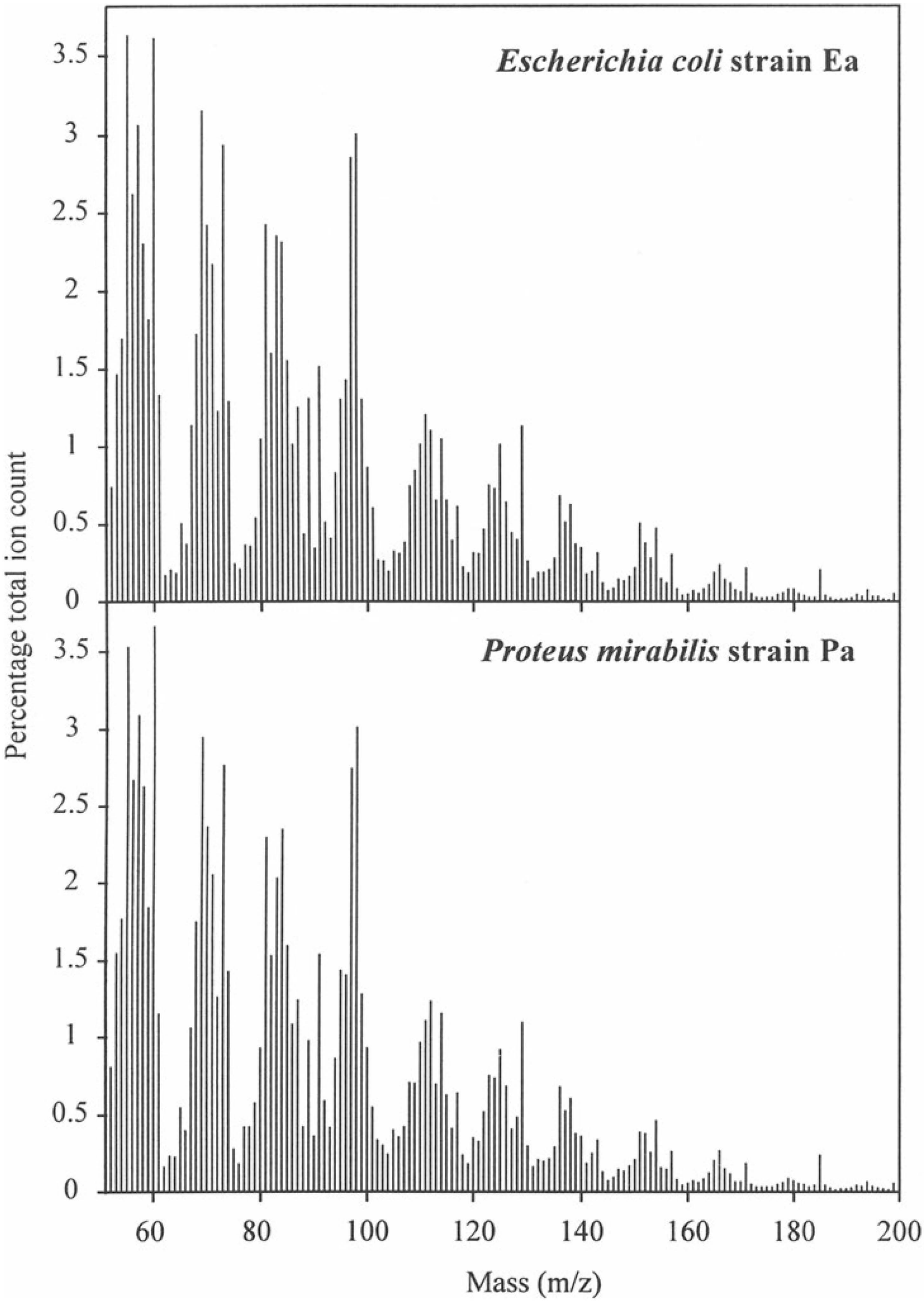

Fig. 1 Normalized pyrolysis mass spectra of *Escherichia coli* isolate Ea and *Proteus mirabilis* isolate Pa.

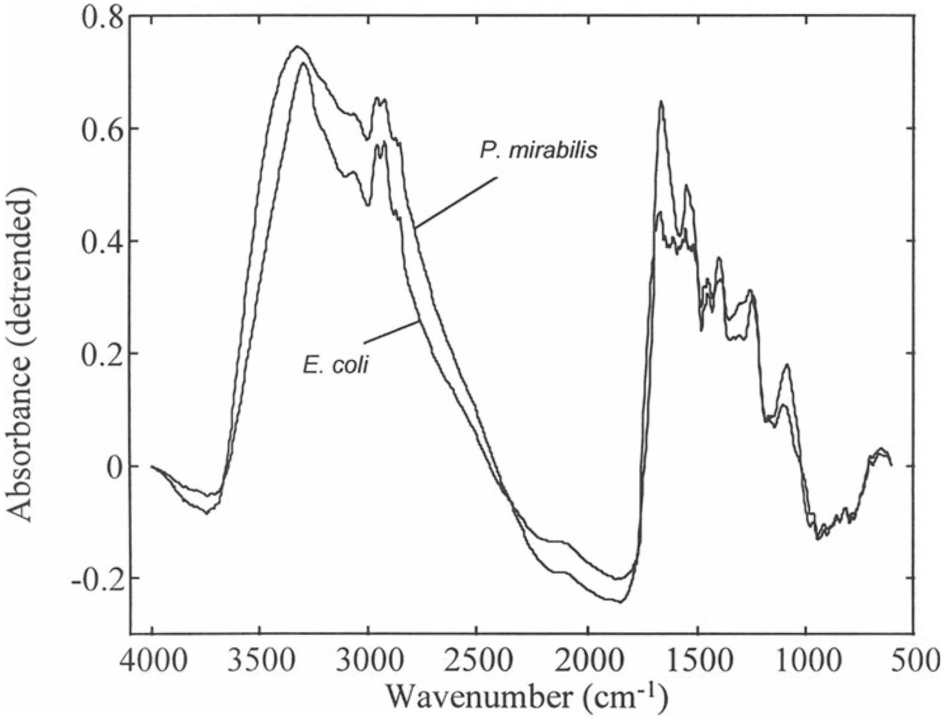

Fig. 2 FT-IR diffuse reflectance-absorbance spectra of *Escherichia coli* isolate Ea and *Proteus mirabilis* isolate Pa.

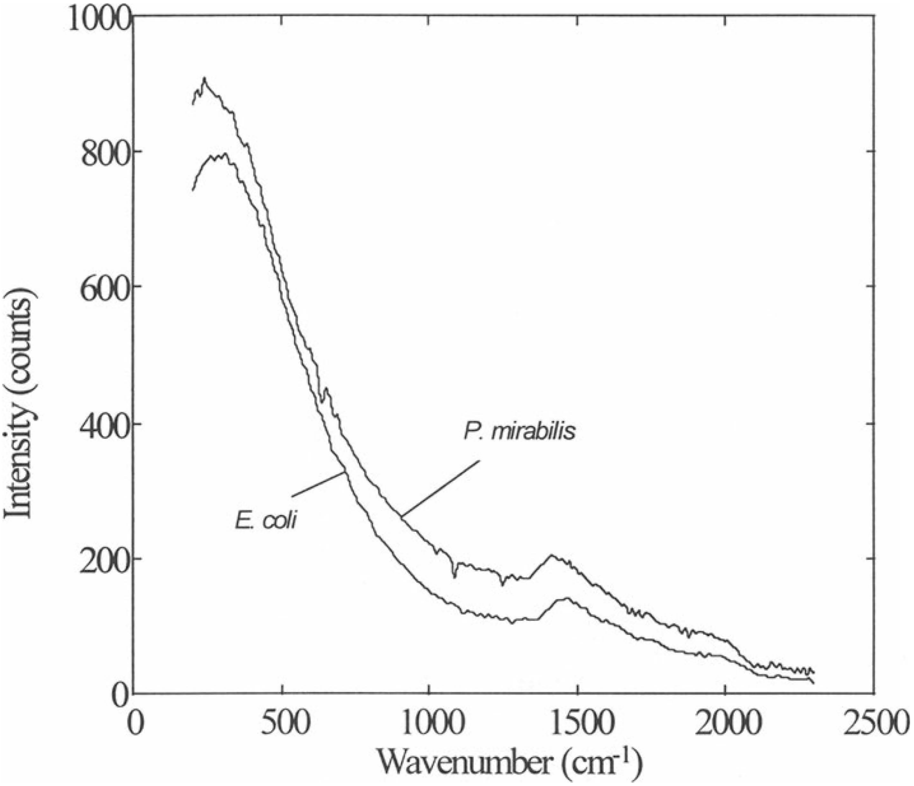

Fig. 3 Dispersive Raman spectra of *Escherichia coli* isolate Ea and *Proteus mirabilis* isolate Pa.

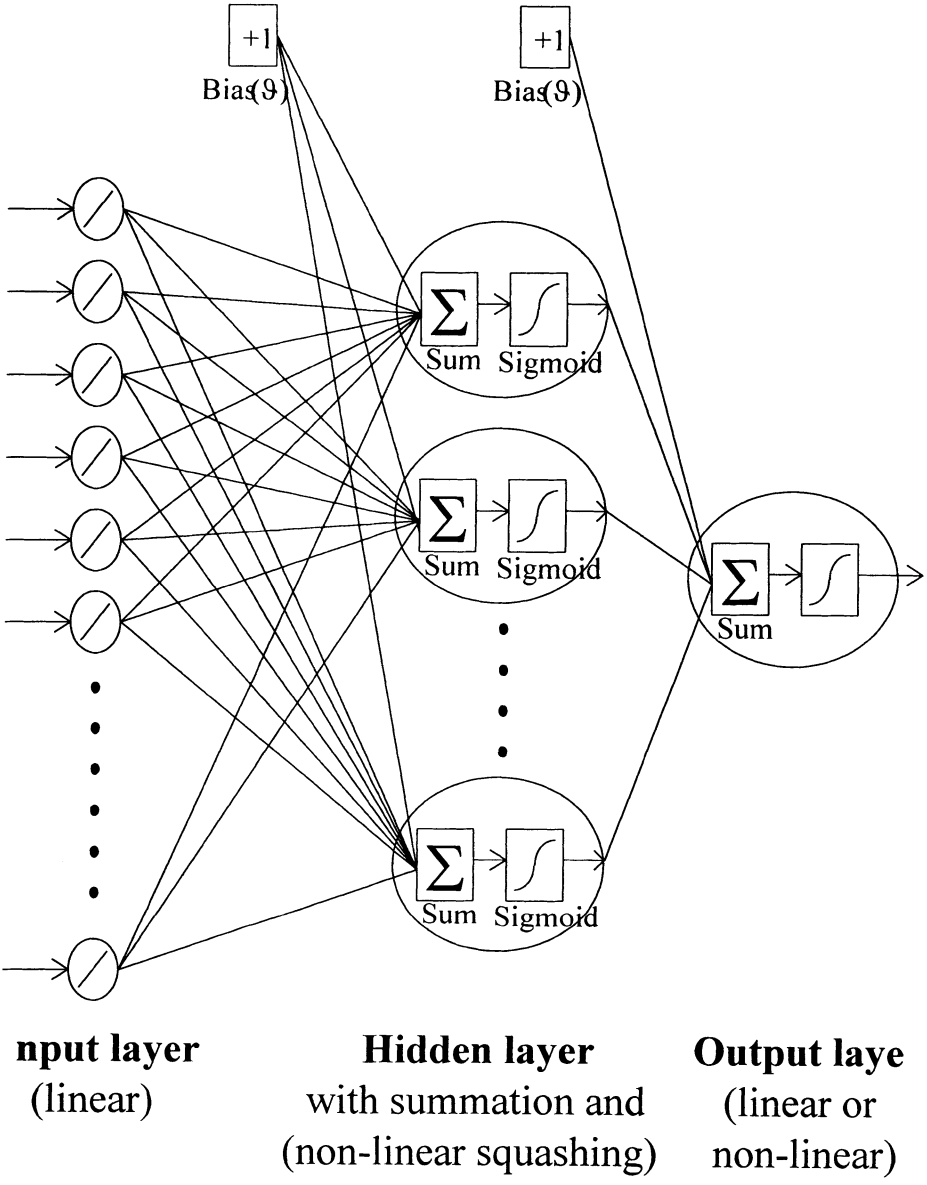

Fig. 4 A multilayer perceptron neural network consisting of an input layer connected to a single node in the output layer by 1 hidden layer. In the architecture shown, adjacent layers of the network are fully interconnected although other architectures are possible. Nodes in the hidden and output layers consist of processing elements which sum the input applied to it and scale the signal using a sigmoidal logistic squashing function.

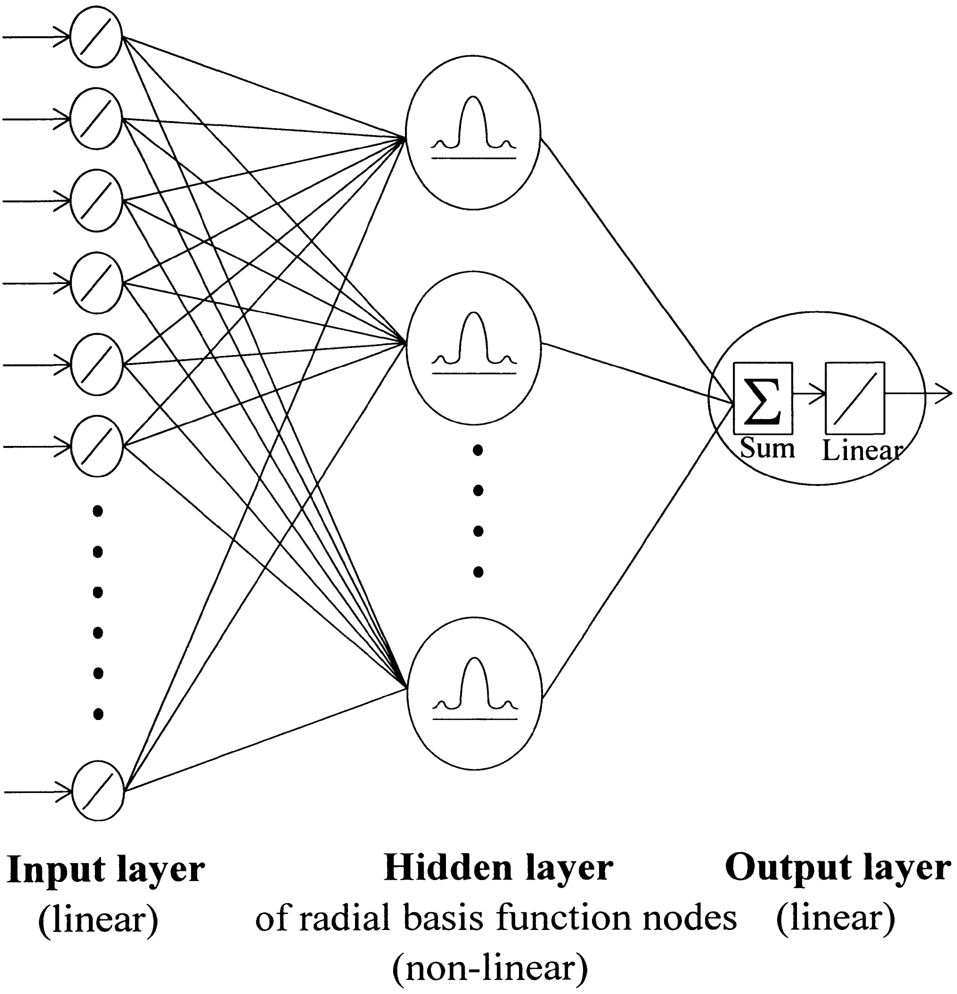

Fig. 5 Radial basis function neural net consisting of an input layer connected to a single node in the output layer by 1 hidden layer. The hidden layer consists of radially-symmetric gaussian functions, although others exist (e.g., Mexican hat and thin plate splines).

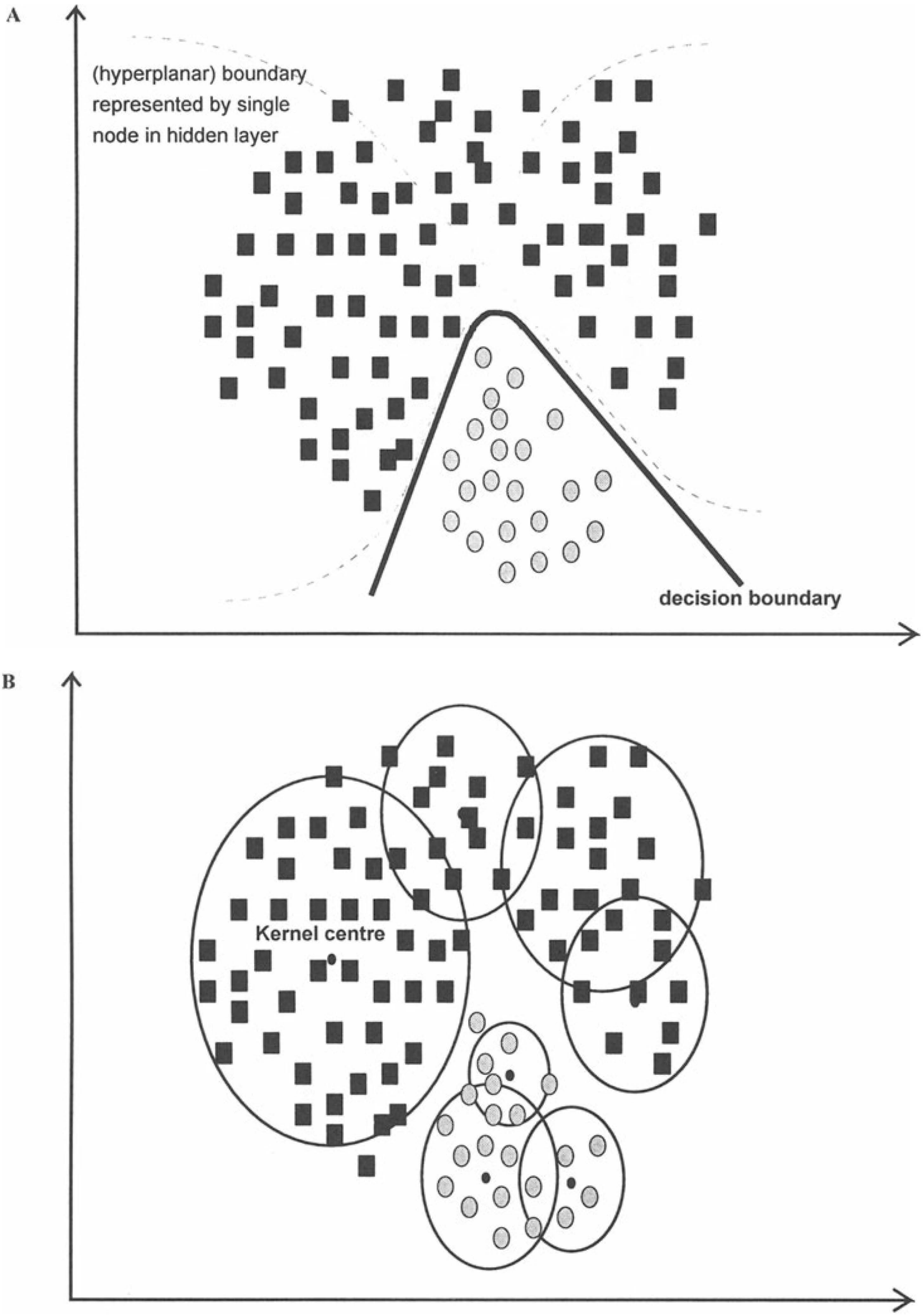

Fig. 6 (A) Typical decision boundary for a classification problem created between two data classes by a MLP with 2 nodes in the hidden layer, for 2 input nodes. Each hidden node represents a non-linear boundary and the nodes in the output layer interpolate this to form a decision boundary. (B) The same classification problem modeled by 7 radially-symmetric basis functions. The width of each kernel function (referred to as its receptive field) is determined by the local density distribution of training examples.

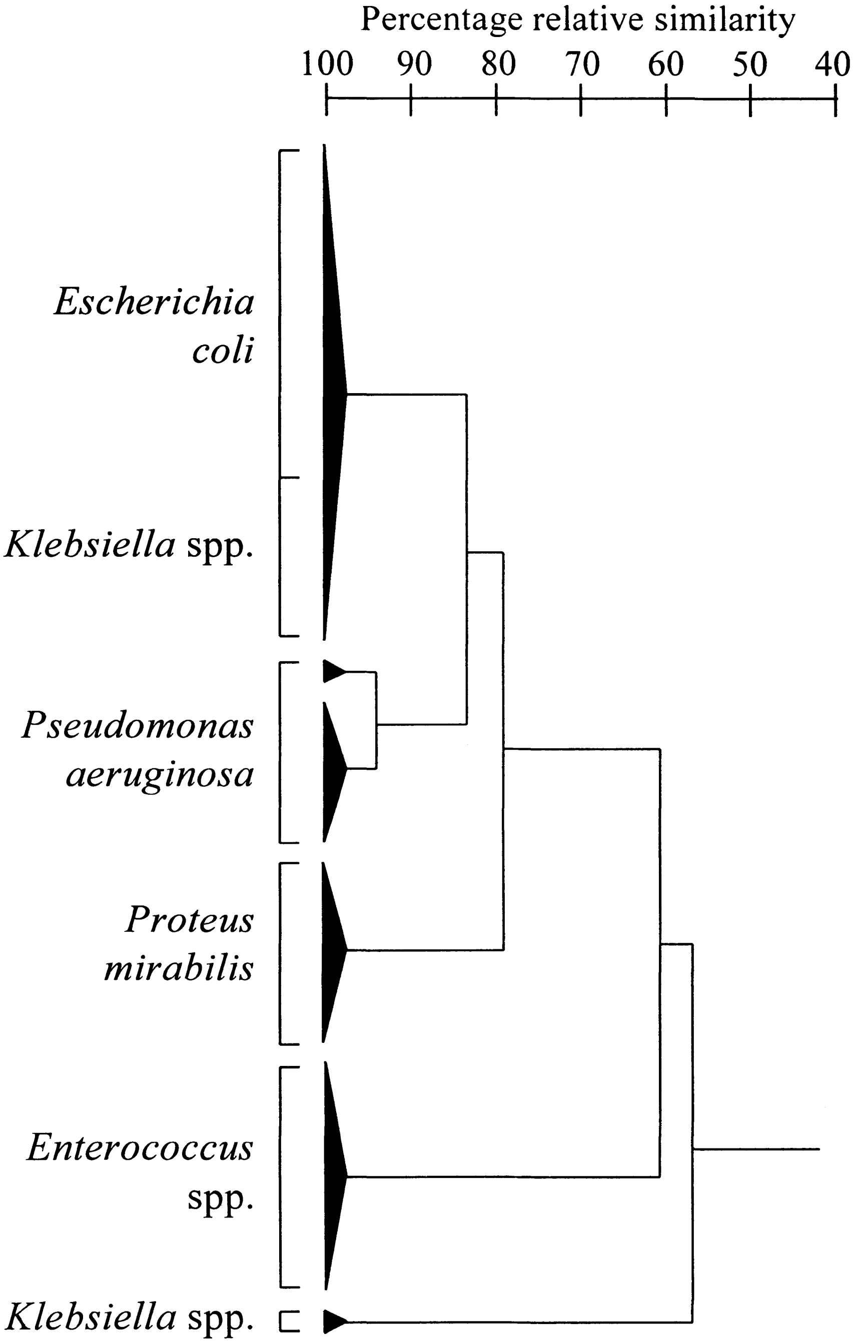

Fig. 7 Dendrogram based on PyMS data showing the relationship between the 59 bacterial isolates.

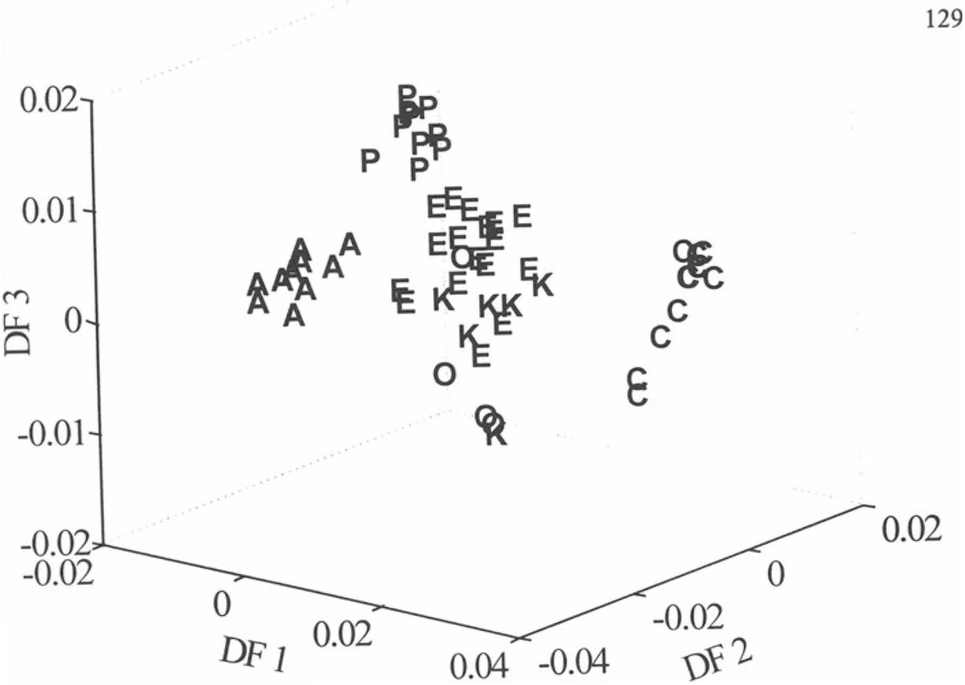

Fig. 8 Pseudo-3D discriminant analysis plot based on FT-IR data showing the relationship between the 59 bacterial isolates. The bacterial isolates are coded as follows; *E. coli* (E), *Prot. mirabilis* (P), *Klebsiella* spp (O and K), *Pseud. aeruginosa* (A), and *Enterococcus* spp. (C).

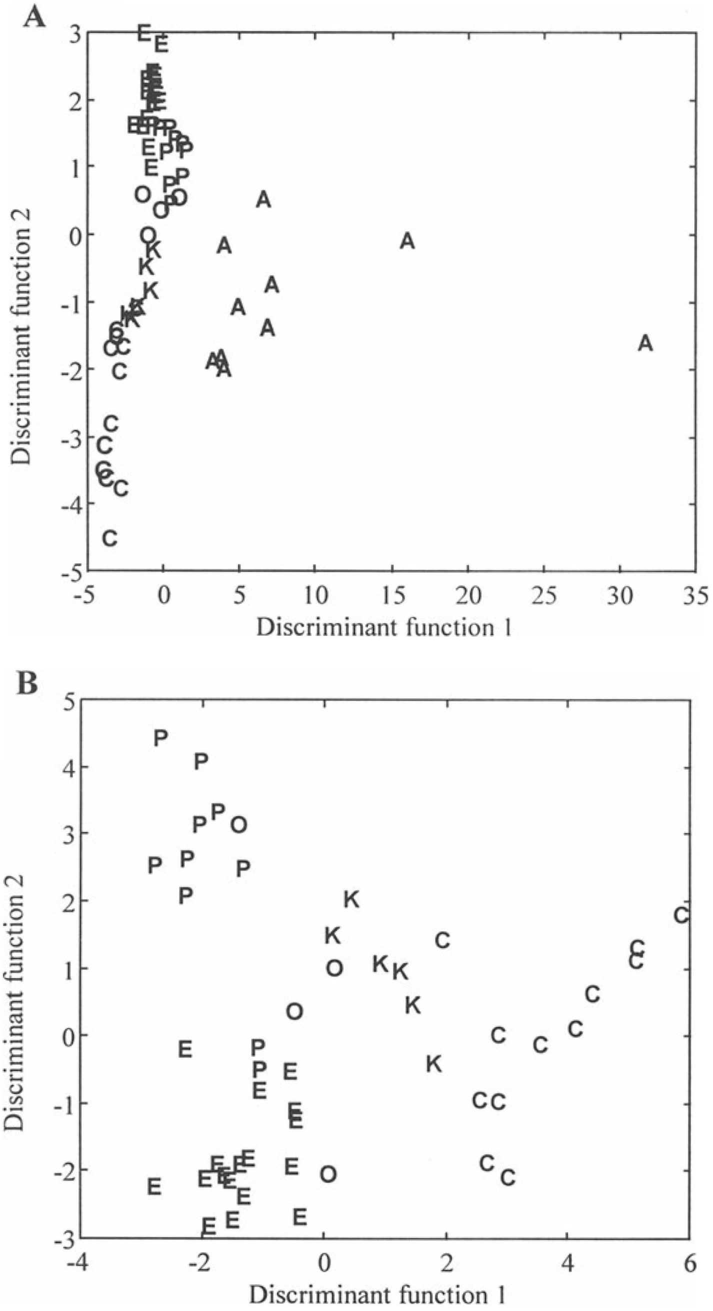

Fig. 9 Discriminant analysis biplots based on Raman data showing the relationship between the 59 bacterial isolates (A) and the same after removal of the *Pseudomonas aeruginosa* isolates (B). The bacterial isolates are coded as follows; *E. coli* (E), *Prot. mirabilis* (P), *Klebsiella* spp (O and K), *Pseud. aeruginosa* (A), and *Enterococcus* spp. (C).

TABLE 1. The 59 urinary isolates from clinical bacteriuria.

| Organism | Bronglais Hospital identifier | DFA identifier |
|---|---|---|
| *Eschericia coli* | KX 22, KX 3, QX 2, RB 1, JX 47, JX 44, KX 35, QX 54, EX 8, AX 48, QX 34, QX 27, QX 57, QX 32, GX 3, QX 7, JX 39 | Ea-Eq |
| *Proteus mirabilis* | EX 20, GX 13, JX 15, HX 23, KX 32, EX 25, HX 45, PX 11, HX 49, RX 44 | Pa-Pj |
| *Klebsiella oxytoca* | WX 12, RX 15, HX 2, JX 58 | Oa-Od |
| *Klebsiella pneumoniae* | VX 23, HX 45, DX 18, NX 1, RX 36, KX 44 | Ka-Kf |
| *Pseudomonas aeruginosa* | VX 54, WX 21, KX 13, PX 8, GX 5, QX 53, HX 36, BX 38, DX 35, HX 24 | Aa-Aj |
| *Enterococcus* spp. | TX 17, LX 44, CX 11, AX 44, QX 45, PX 6, JX 25, LX 9, VX 13, TX 28, WX 17, RX 51 | Ca-Cl |

TABLE 2. Identity of the bacteria used in the training set as judged by multilayer perceptron analysis of their PyMS data.

| Strain | Bronglais | Estimates from MLP | | | | |
|---|---|---|---|---|---|---|
| | Identifier | *E. coli* | *Prot. mirab.* | *Klebs.* spp. | *Pseud. aerug.* | *Enter.* spp. |
| *E. coli* | KX 22 | **0.9** | 0.0 | 0.1 | 0.0 | -0.1 |
| *E. coli* | KX 3 | **1.0** | 0.0 | -0.1 | 0.0 | 0.0 |
| *E. coli* | QX 2 | **1.0** | 0.0 | -0.1 | 0.0 | 0.0 |
| *E. coli* | RB 1 | **0.6** | 0.0 | 0.4 | 0.0 | 0.0 |
| *E. coli* | JX 47 | **1.0** | 0.0 | -0.1 | 0.0 | 0.0 |
| *Prot. mirab.* | EX 20 | 0.0 | **1.0** | 0.1 | 0.0 | 0.0 |
| *Prot. mirab.* | GX 13 | 0.0 | **1.0** | -0.1 | 0.0 | 0.0 |
| *Prot. mirab.* | JX 15 | 0.0 | **0.9** | 0.0 | 0.0 | 0.1 |
| *Prot. mirab.* | HX 23 | -0.1 | **1.0** | 0.0 | 0.1 | 0.0 |
| *Prot. mirab.* | KX 32 | 0.0 | **1.0** | 0.0 | -0.1 | 0.0 |
| *Klebs.* spp. | WX 12 | 0.3 | 0.0 | **0.7** | 0.1 | -0.1 |
| *Klebs.* spp. | RX 15 | 0.0 | -0.1 | **1.0** | 0.0 | 0.0 |
| *Klebs.* spp. | VX 23 | -0.1 | 0.1 | **1.0** | 0.0 | 0.0 |
| *Klebs.* spp. | HX 45 | 0.0 | 0.0 | **1.0** | -0.1 | 0.0 |
| *Klebs.* spp. | DX 18 | 0.2 | 0.1 | **0.7** | 0.1 | 0.0 |
| *Pseud. aerug.* | VX 54 | 0.2 | -0.1 | 0.1 | **0.8** | 0.0 |
| *Pseud. aerug.* | WX 21 | 0.0 | -0.1 | 0.1 | **1.0** | 0.0 |
| *Pseud. aerug.* | KX 13 | -0.1 | 0.0 | 0.0 | **1.1** | 0.0 |
| *Pseud. aerug.* | PX 8 | 0.0 | 0.1 | -0.1 | **1.0** | 0.0 |
| *Pseud. aerug.* | GX 5 | 0.0 | 0.0 | 0.0 | **1.0** | 0.0 |
| *Enter.* spp. | TX 17 | 0.0 | 0.0 | 0.0 | 0.1 | **1.0** |
| *Enter.* spp. | LX 44 | 0.0 | 0.0 | 0.1 | 0.0 | **1.0** |
| *Enter.* spp. | CX 11 | 0.0 | 0.0 | 0.0 | 0.0 | **1.0** |
| *Enter.* spp. | AX 44 | 0.0 | 0.0 | 0.0 | 0.0 | **1.0** |
| *Enter.* spp. | QX 45 | 0.0 | 0.0 | 0.0 | 0.0 | **1.0** |

TABLE 3. Identity of the bacteria used in the test set as judged by multilayer perceptron analysis of their PyMS data.

| Strain | Bronglais | Estimates from MLP | | | | |
|---|---|---|---|---|---|---|
| | Identifier | *E. coli* | *Prot. mirab.* | *Klebs.* spp. | *Pseud. aerug.* | *Enter.* spp. |
| *E. coli* | JX 44 | **0.7** | 0.0 | 0.3 | 0.0 | 0.0 |
| *E. coli* | KX 35 | **1.0** | 0.0 | 0.0 | 0.0 | 0.1 |
| *E. coli* | QX 54 | **1.1** | 0.1 | -0.1 | 0.1 | -0.1 |
| *E. coli* | EX 8 | **1.1** | 0.1 | -0.1 | 0.0 | -0.2 |
| *E. coli* | AX 48 | **1.0** | 0.1 | 0.0 | 0.1 | -0.1 |
| *E. coli* | QX 34 | **1.1** | 0.0 | -0.1 | 0.0 | -0.1 |
| *E. coli* | QX 27 | **0.9** | 0.1 | 0.0 | 0.1 | -0.1 |
| *E. coli* | QX 57 | **1.1** | 0.1 | 0.0 | 0.1 | -0.2 |
| *E. coli* | QX 32 | **0.9** | 0.0 | 0.2 | 0.0 | 0.0 |
| *E. coli* | GX 3 | **1.0** | 0.0 | -0.1 | 0.0 | 0.0 |
| *E. coli* | QX 7 | **1.0** | 0.1 | 0.0 | 0.0 | -0.1 |
| *E. coli* | JX 39 | **1.2** | 0.1 | -0.1 | 0.0 | -0.2 |
| *Prot. mirab.* | EX 25 | -0.1 | **1.0** | 0.1 | 0.0 | 0.0 |
| *Prot. mirab.* | HX 45 | 0.0 | **1.1** | -0.1 | 0.0 | 0.1 |
| *Prot. mirab.* | PX 11 | -0.1 | **1.1** | -0.1 | 0.0 | 0.0 |
| *Prot. mirab.* | HX 49 | 0.1 | **0.8** | 0.2 | 0.1 | -0.1 |
| *Prot. mirab.* | RX 44 | 0.1 | **1.0** | 0.0 | 0.0 | -0.1 |
| *Klebs.* spp. | HX 2 | 0.2 | -0.1 | **0.9** | 0.0 | 0.0 |
| *Klebs.* spp. | JX 58 | **0.7** | 0.1 | 0.3 | 0.1 | -0.2 |
| *Klebs.* spp. | NX 1 | -0.1 | 0.0 | **0.9** | 0.1 | 0.1 |
| *Klebs.* spp. | RX 36 | 0.2 | 0.0 | **0.7** | 0.0 | 0.1 |
| *Klebs.* spp. | KX 44 | 0.1 | 0.0 | **0.9** | 0.0 | -0.1 |
| *Pseud. aerug.* | QX 53 | -0.2 | 0.0 | 0.3 | **0.9** | 0.0 |
| *Pseud. aerug.* | HX 36 | 0.1 | 0.1 | -0.3 | **1.1** | 0.0 |
| *Pseud. aerug.* | BX 38 | 0.0 | 0.1 | -0.1 | **1.1** | 0.0 |
| *Pseud. aerug.* | DX 35 | 0.1 | 0.0 | 0.1 | **0.8** | 0.0 |
| *Pseud. aerug.* | HX 24 | 0.0 | 0.0 | 0.0 | **1.0** | 0.0 |
| *Enter.* spp. | PX 6 | 0.0 | 0.0 | 0.0 | 0.0 | **1.0** |
| *Enter.* spp. | JX 25 | 0.0 | 0.0 | -0.1 | 0.0 | **1.1** |
| *Enter.* spp. | LX 9 | 0.0 | 0.0 | 0.0 | 0.0 | **1.0** |
| *Enter.* spp. | VX 13 | 0.0 | 0.0 | -0.1 | 0.0 | **1.0** |
| *Enter.* spp. | TX 28 | 0.0 | 0.1 | 0.1 | 0.0 | **0.9** |
| *Enter.* spp. | WX 17 | 0.1 | 0.0 | -0.2 | 0.0 | **1.1** |
| *Enter.* spp. | RX 51 | 0.0 | 0.0 | 0.0 | 0.0 | **1.1** |

## 6. Literature Cited

Beale, R. and Jackson, T. (1990). *Neural Computing: An Introduction*. Bristol: Adam Hilger.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Bouffard, S. P., Katon, J. E., Sommer, A. J. and Danielson, N. D. (1994). Development of microchannel thin layer chromatography with infrared microspectroscopic detection. *Analytical Chemistry* **66**, 1937-1940.

Causton, D. R. (1987). *A Biologist's Advanced Mathematics*. London: Allen and Unwin.

Colthup, N. B., Daly, L. H. and Wiberly, S. E. (1990). *Introduction to infrared and Raman spectroscopy*. New York: Academic Press.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York: Wiley.

Everitt, B. S. (1993). *Cluster Analysis*. London: Edward Arnold.

Ferraro, J. R. and Nakamoto, K. (1994). *Introductory Raman Spectroscopy*. London: Academic Press.

Goodacre, R., Hammond, D. and Kell, D. B. (1997). Quantitative analysis of the adulteration of orange juice with sucrose using pyrolysis mass spectrometry and chemometrics. *Journal of Analytical and Applied Pyrolysis* **40/41**, 135-158.

Goodacre, R. and Kell, D. B. (1996). Pyrolysis mass spectrometry and its applications in biotechnology. *Current Opinion in Biotechnology* **7**, 20-28.

Goodacre, R., Kell, D. B. and Bianchi, G. (1993). Rapid assessment of the adulteration of virgin olive oils by other seed oils using pyrolysis mass spectrometry and artificial neural networks. *Journal of the Science of Food and Agriculture* **63**, 297-307.

Goodacre, R., Neal, M. J. and Kell, D. B. (1994). Rapid and quantitative analysis of the pyrolysis mass spectra of complex binary and tertiary mixtures using multivariate calibration and artificial neural networks. *Analytical Chemistry* **66**, 1070-1085.

Goodacre, R., Neal, M. J., Kell, D. B., Greenham, L. W., Noble, W. C. and Harvey, R. G. (1994). Rapid identification using pyrolysis mass spectrometry and artificial neural networks of *Propionibacterium acnes* isolated from dogs. *Journal of Applied Bacteriology* **76**, 124-134.

Goodacre, R., Timmins, É. M., Rooney, P. J., Rowland, J. J. and Kell, D. B. (1996). Rapid identification of *Streptococcus* and *Enterococcus* species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks. *FEMS Microbiology Letters* **140**, 233-239.

Goodacre, R., Trew, S., Wrigley-Jones, C., Saunders, G., Neal, M. J., Porter, N. and Kell, D. B. (1995). Rapid and quantitative analysis of metabolites in fermentor broths using pyrolysis mass spectrometry with supervised learning: application to the screening of *Penicillium chrysogenum* fermentations for the overproduction of penicillins. *Analytica Chimica Acta* **313**, 25-43.
Gower, J. C. (1966). Some Distance Properties of Latent Root and Vector Methods used in Multivariate Analysis. *Biometrika* **53**, 325-338.

Griffiths, P. R. and de Haseth, J. A. (1986). *Fourier transform infrared spectrometry*. New York: John Wiley.

Haykin, S. S. (1994). *Neural networks :a comprehensive foundation*. New York: Macmillan.

Hush, D. R. and Horne, B. G. (1993). Progress in supervised neural networks - what's new since Lippmann. *IEEE Signal Processing Magazine* **10**, 8-39.

Jolliffe, I. T. (1986). *Principal Component Analysis*. New York: Springer-Verlag.

MacFie, H. J. H., Gutteridge, C. S. and Norris, J. R. (1978). Use of canonical variates in differentiation of bacteria by pyrolysis gas-liquid chromatography. *Journal of General Microbiology* **104**, 67-74.

Magee, J. T. (1993). Whole-organism fingerprinting. In *Handbook of New Bacterial Systematics*, pp. 383-427. Edited by M. Goodfellow and A. G. O'Donnell. London: Academic Press.

Manly, B. F. J. (1994). *Multivariate Statistical Methods : A Primer*. London: Chapman & Hall.

Meuzelaar, H. L. C., Haverkamp, J. and Hileman, F. D. (1982). *Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials*. Amsterdam: Elsevier.

Moody, J. and Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation* **1**, 281-294.

Naumann, D., Helm, D., Labischinski, H. and Giesbrecht, P. (1991). The characterization of microorganisms by Fourier-transform infrared spectroscopy (FT-IR). In *Modern techniques for rapid microbiological analysis*, pp. 43-96. Edited by W. H. Nelson. New York: VCH Publishers.

Nelson, W. H., Manoharan, R. and Sperry, J. F. (1992). UV resonance Raman studies of bacteria. *Applied Spectroscopy Reviews* **27**, 67-124.

Park, J. and Sandberg, I. W. (1991). Universal approximation using radial basis function networks. *Neural Computation* **3**, 246-257.

Rumelhart, D. E., McClelland, J. L. and The PDP Research Group (1986). *Parallel Distributed Processing, Experiments in the Microstructure of Cognition*. Cambridge, Mass.: MIT Press.

Saha, A. and Keller, J. D. (1990). Algorithms for better representation and faster learning in radial basis functions. In *Advances in Neural Information Processing Systems*, pp. 482-. Edited by D. Touretzky: Morgan Kaufmann Publishers.

Savitzky, A. and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* **36**, 1627-1633.

Sleigh, J. D. and Timbury, M. C. (1986). *Notes on Medical Bacteriology*. London: Churchill Livingstone.

Timmins, É. M. and Goodacre, R. (1997). Rapid quantitative analysis of binary mixtures of *Escherichia coli* strains using pyrolysis mass spectrometry with multivariate calibration and artificial neural networks. *Journal of Applied Microbiology* **83**, 208-218.

Wasserman, P. D. (1989). *Neural Computing: Theory and Practice*. New York: Van Nostrand Reinhold.

Werbos, P. J. (1994). *The roots of back-propagation: from ordered derivatives to neural networks and political forecasting*. Chichester: John Wiley.

Wilkins, M. F., Morris, C. W. and Boddy, L. (1994). A comparison of radial basis function and backpropagation neural networks for identification of marine phytoplankton from multivariate flow cytometry data. *Computer Applications In the Biosciences* **10**, 285-294.

Williams, K. P. J., Pitt, G. D., Batchelder, D. N. and Kip, B. J. (1994). Confocal Raman microspectroscopy using a stigmatic spectrograph and CCD detector. *Applied spectroscopy* **48**, 232-235.

Windig, W., Haverkamp, J. and Kistemaker, P. G. (1983). Interpretation of sets of pyrolysis mass spectra by discriminant analysis and graphical rotation. *Analytical Chemistry* **55**, 81-88.

Winson, M. K., Goodacre, R., Woodward, A. M., Timmins, É. M., Jones, A., Alsberg, B. K., Rowland, J. J. and Kell, D. B. (1997). Diffuse reflectance absorbance spectroscopy taking in chemometrics (DRASTIC). A hyperspectral FT-IR-based approach to rapid screening for metabolite overproduction. *Analytica Chimica Acta* , in press.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, pp. 391-420. Edited by K. R. Krishnaiah. New York: Academic Press.