Microbiology (1998), **144**, 1157–1170 Printed in Great Britain

Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks

Royston Goodacre,[1] Éadaoin M. Timmins,[1] Rebecca Burton,[1] Naheed Kaderbhai,[1] Andrew M. Woodward,[1] Douglas B. Kell[1] and Paul J. Rooney[2]

Author for correspondence: Royston Goodacre. Tel: +44 1970 621947. Fax: +44 1970 622354. e-mail: rrg@aber.ac.uk

[1] Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion SY23 3DD, UK

[2] Bronglais General Hospital, Aberystwyth, Ceredigion SY23 1ER, UK

Three rapid spectroscopic approaches for whole-organism fingerprinting – pyrolysis mass spectrometry (PyMS), Fourier transform infra-red spectroscopy (FT-IR) and dispersive Raman microscopy – were used to analyse 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 methods to reduce the dimensionality of these hyperspectral data. 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 (ANN) approaches exploiting multi-layer perceptrons or 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. Comparable ANNs trained with Raman spectra correctly identified some 80% of the same test set. PyMS and FT-IR have often been exploited within microbial systematics, but these are believed to be the first published data showing the ability of dispersive Raman microscopy to discriminate clinically significant intact bacterial species. These results demonstrate that modern analytical spectroscopies of high intrinsic dimensionality can provide rapid accurate microbial characterization techniques, but only when combined with appropriate chemometrics.

Keywords: artificial neural networks, Fourier-transform infrared spectroscopy, pyrolysis mass spectrometry, Raman microscopy, urinary tract infection

INTRODUCTION

Urinary tract infections (UTIs) are a major clinical problem, especially among adult women. The family doctor consultation rate for this group for UTI is 63·5 consultations per 1000 women per year in the UK (Wilkie *et al*., 1992).

Abbreviations: ANN, artificial neural network; DF, discriminant function; DFA, discriminant function analysis; HCA, hierarchical cluster analysis; FT-IR, Fourier-transform infrared spectroscopy; MLP, multilayer perceptron; PC, principal component; PCA, principal-components analysis; PyMS, pyrolysis mass spectrometry; RBF, radial basis function; RMSEF, root mean squared error of formation.

The bacteria typically associated with UTI in hospitals are *Escherichia coli* (causative organism of 50% of the cases), *Klebsiella* species (14%), other coliforms (4%), staphylococci (6%), *Enterococcus faecalis* (10%) and *Pseudomonas aeruginosa* (3%) (Slack, 1995). Quantitative culture of urine is used to confirm the clinical diagnosis, and the finding of 10^5 c.f.u. per ml of urine is defined as 'significant bacteriuria' (Morgan & McKenzie, 1993). The empirical choice of an effective treatment is becoming more difficult as urinary pathogens are increasingly resistant to commonly used antibiotics (Gruneberg, 1994). Consequently it is necessary to perform antibiotic-sensitivity testing on significant isolates.

The degree to which a causative organism requires identification varies but identification is most useful in complex clinical cases; for example to distinguish relapse, to indicate failure of an antibiotic treatment, and to detect reinfection with a different organism in patients with recurrent infections (Lewis, 1989). Using conventional methods, laboratory examination of urine is expensive, time-consuming and labour-intensive: approximately 24 h incubation is required to obtain an accurate colony count. An additional 12–24 h is needed for organism identification and susceptibility testing, which may further delay administration of the most appropriate narrow-spectrum antibiotic (Casadevall, 1996; Pappas, 1991).

For routine purposes the ideal method for microbial characterization would require minimum sample preparation, would analyse samples directly (i.e. would not require reagents), would be rapid, automated and (at least relatively) inexpensive. With recent 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) (Helm *et al*., 1991; Naumann *et al*., 1991a, 1991b) 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; Schrader, 1995). Therefore they give quantitative information about the total biochemical composition of a sample. However, the interpretation of these multidimensional spectra, or what is known as hyperspectral data (Abousleman *et al*., 1994; Goetz *et al*., 1985; Wilson *et al*., 1995), has conventionally been by the application of 'unsupervised' pattern recognition methods such as principal components analysis (PCA), discriminant function analysis (DFA) and hierarchical cluster analysis (HCA). With 'unsupervised learning' methods of this sort the relevant multivariate algorithms seek 'clusters' in the data, thereby allowing the investigator to group objects 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 (Chun *et al*., 1993; Freeman *et al*., 1994; Goodacre *et al*., 1994b; 1996a, b; Sisson *et al*., 1995); arguably the most significant of these is the application of 'intelligent' systems based on artificial neural networks (ANNs) (Bishop, 1995; Wasserman, 1989).

We previously reported a study exploiting PyMS and ANNs to discriminate between susceptible and methicillin-resistant *Staphylococcus aureus*, illustrating that it is possible to detect very subtle physiological differences between strains of the same species of bacteria using these techniques (Goodacre *et al*., 1998). In the present study a group of 59 clinically significant urinary isolates of bacteria were collected from the local hospital. All isolates were typed by conventional biochemical tests to belong to *Escherichia coli*, *Proteus mirabilis*, *Klebsiella* species, *Pseudomonas aeruginosa*, and *Enterococcus* species. 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.

METHODS

Strains and cultivation. A group of 59 bacteria isolated from the urine of patients with urinary tract infection (UTI) were collected from Bronglais General Hospital, Aberystwyth. All isolates were typed by conventional biochemical tests to belong to *E. coli* (17, coded Ea–Eq), *Pr. mirabilis* (10, coded Pa–Pj), *Klebsiella* spp. (10, coded Ka–Kj), *Ps. aeruginosa* (10, coded Aa–Aj), and *Enterococcus* spp. (12, coded Ca–Cl). All strains were cultivated axenically and aerobically on LabM Malthus blood agar base (37 mg ml^{-1}) for 16 h at 37 °C. After subculturing 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) to approximately 40 mg ml^{-1}. The samples were then analysed by PyMS, FT-IR and dispersive Raman spectroscopies.

Pyrolysis mass spectrometry (PyMS). Five-microlitre volumes of the bacterial samples (approx. 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.

PyMS was then performed on a PyMS-200X instrument (Horizon Instruments). For full operational procedures see Goodacre & Kell (1996) and Goodacre *et al*. (1993, 1994a). Conditions used for each experiment involved heating the sample to 100 °C for 5 s followed by Curie-point pyrolysis at 530 °C for 3 s with a temperature rise time of 0·5 s.

PyMS data may be displayed as quantitative pyrolysis mass spectra (e.g. as in Fig. 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 most straightforward 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 1×10^6–3×10^6.

Diffuse reflectance-absorbance Fourier-transform infrared (FT-IR) spectroscopy. Ten microlitres of each bacterial sample was evenly applied onto a sand-blasted aluminium plate. Prior to analysis the samples were oven-dried at 50 °C for 30 min. Samples were run in triplicate. The instrument used was a Bruker IFS28 FT-IR spectrometer (Bruker Spectrospin) equipped with an MCT (mercury-cadmium-telluride) detector cooled with liquid N_2. The aluminium plate was then loaded

onto the motorized stage of a reflectance TLC accessory (Bouffard *et al*., 1994; Goodacre *et al*., 1996c; 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 Fig. 2.

ASCII data were exported from the OPUS software used to control the FT-IR instrument and imported into Matlab version 4.2c.1 (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 derivatives of these normalized and detrended spectra were calculated using the Savitzky–Golay algorithm (Savitzky & Golay, 1964) with 5-point smoothing.

Dispersive Raman microscopy. Spectra were collected using the Renishaw dispersive Raman spectroscope (Ramascope) (Williams *et al*., 1994a, b) with a low power (30 mW) near-infrared 780 nm diode laser with the power at the sampling point typically at 3 mW. The instrument was wavelength calibrated with a silicon wafer focused under the × 50 objective and collected as a static spectrum centred at 520 cm^{-1} for 10 s.

Samples were presented as 0·5 ml bacterial suspensions (40 mg ml^{-1} or $\sim 3 \times 10^9$ cells ml^{-1}) in 2 ml Supelco clear glass vials, covered with solid caps with aluminium liners. These glass vials were placed sequentially into the sample holder of a Renishaw Macropoint assembly. A 16 mm 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 s. 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. Stokes Raman spectra were collected over the wavenumber range 200 cm^{-1} to 2300 cm^{-1}. The spectral resolution used was $\sim$ 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 into Matlab version 4.2c.1. 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 detector) 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 (Alsberg *et al*., 1997) 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 wavenumber domain. Typical Stokes Raman spectra are shown in Fig. 3. Note that although the fluorescence is relatively low when cells are excited at 780 nm, the system can not discriminate

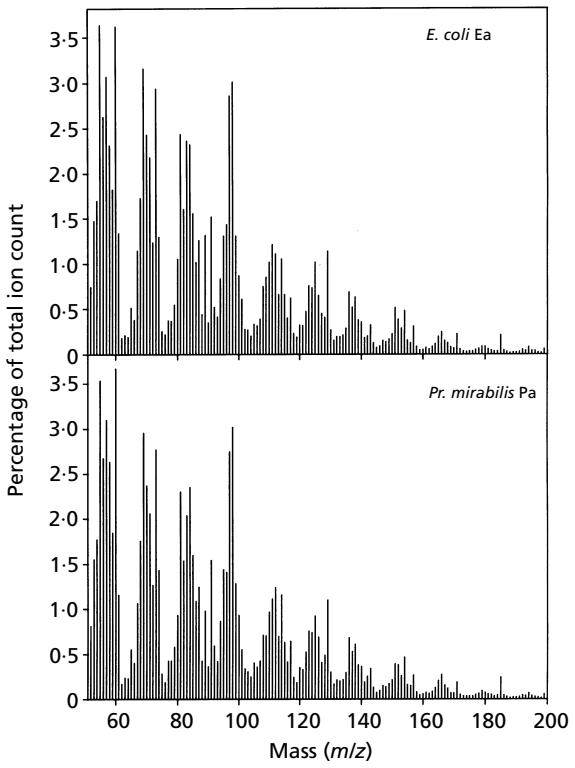

Fig. 1. Normalized Py-MS spectra of *E. coli* isolate Ea and *Pr. mirabilis* isolate Pa.

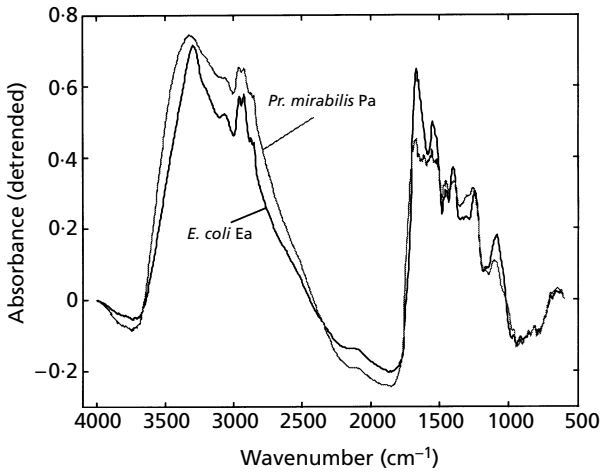

Fig. 2. FT-IR diffuse reflectance-absorbance spectra of *E. coli* isolate Ea and *Pr. mirabilis* isolate Pa.

whether individual photons arise by fluorescence or are scattered via the Raman effect.

Cluster analysis. The typical procedure for multivariate analysis is detailed in Fig. 4. 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); thus this process does not bias the analysis in any way. DFA was programmed according to Manly's principles (Manly, 1994).

DFA was not performed on the original feature space (spectra) because one can not feed collinear variables or too many variables into DFA. The starting point for DFA is the inverse of the pooled variance–covariance matrix within a priori groups. This inverse can only exist when the matrix is non-singular, i.e. its determinant is other than zero, which implies that it is of full rank (Dixon, 1975; MacFie *et al.*, 1978); i.e. generally if

$$(N_s - N_g - 1) > N_v \quad (1)$$

where N_s is the number of samples, N_g the number of groups, and N_v the number of inputs (variables; i.e. mass intensities, absorbances at particular wavenumbers, or photon counts at particular wavenumber shifts for PyMS, FT-IR and Raman respectively). For PyMS, FT-IR and Raman, the number of inputs is 150, 882 and 2283, respectively; this is far in excess of the number of samples (177) minus the number of groups (59), which comes to only 118. In addition, singularity can be caused by collinearity, and PCA removes collinearities whilst also reducing the number of inputs (so as to obey the above) to the DFA algorithm.

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 (MLP). All 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, Totton, Southampton, UK), 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.*, 1994a, 1995, 1996b).

The structure of the MLP used in this study to analyse the hyperspectral data consisted of three layers. The first layer contained either (a) the full spectra (made up of the 150 input nodes for PyMS, 882 for FT-IR, and 2283 for Raman; see Table 1 for details) or (b) the first few PC scores (see Table 1 for details), one 'hidden' layer, and five output nodes (encoded in binary fashion for the bacterial identities). These were binary encoded such that *E. coli* was coded as 10000, *Pr. mirabilis* as 01000, *Klebsiella* spp. as 00100, *Ps. 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 Fig. 5 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 (ϑ), which also has a modifiable weight. Nodes sum the signals feeding to them (Net):

$$Net = i_1w_1 + i_2w_2 + i_3w_3 + \ldots i_iw_i + \ldots i_nw_n = \sum_{i-1}^{n} i_1w_1 + \vartheta \quad (2)$$

The sum of the scaled inputs and the node's bias are then

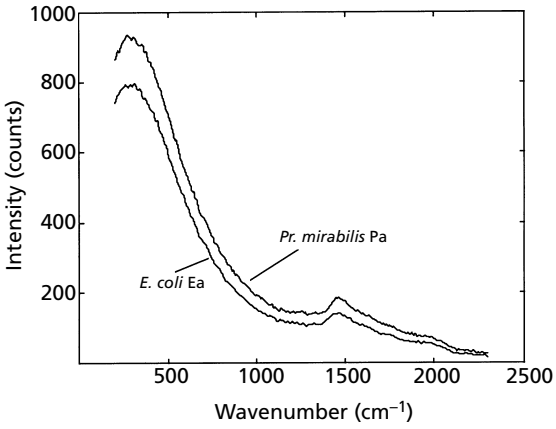

Fig. 3. Dispersive Raman spectra of *E. coli* isolate Ea and *Pr. mirabilis* isolate Pa.

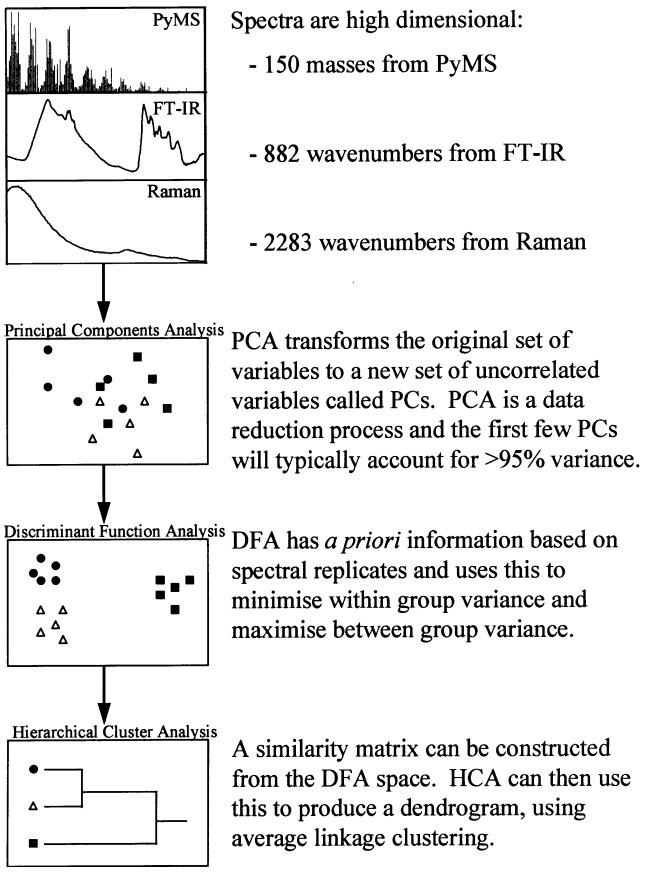

Fig. 4. Flowchart of unsupervised learning analysis used to cluster the PyMS, FT-IR and Raman spectra.

Table 1. Artificial neural network conditions used to identify the bacteriuria isolates

	Conditions used	MLPs*	PC-MLPs*	RBFs†
PyMS	Architecture	150-8-5	10-6-5	150-40-5
	Explained variance (%)	–	97·16	–
	No. of epochs	3×10^3	2×10^3	–
	Time (min)	4	0·5	0·2 (10 s)
FT-IR	Architecture	882-12-5	20-6-5	882-20-5
	Explained variance (%)	–	96·88	–
	No. of epochs	5×10^3	3×10^3	–
	Time (min)	60	1	0·2 (10 s)
Raman	Architecture	2283-12-5	5-3-5	2283-50-5
	Explained variance (%)	–	78·86	–
	No. of epochs	2×10^4	1×10^4	–
	Time (min)	1800 (30 h)	10	2

* The number of epochs and time taken to reach an RMSEF of 0·01 (1 %) were calculated from the mean of five re-trained models.

† The optimum number of kernel functions was found by calculating the minimum RMSEF for the training set.

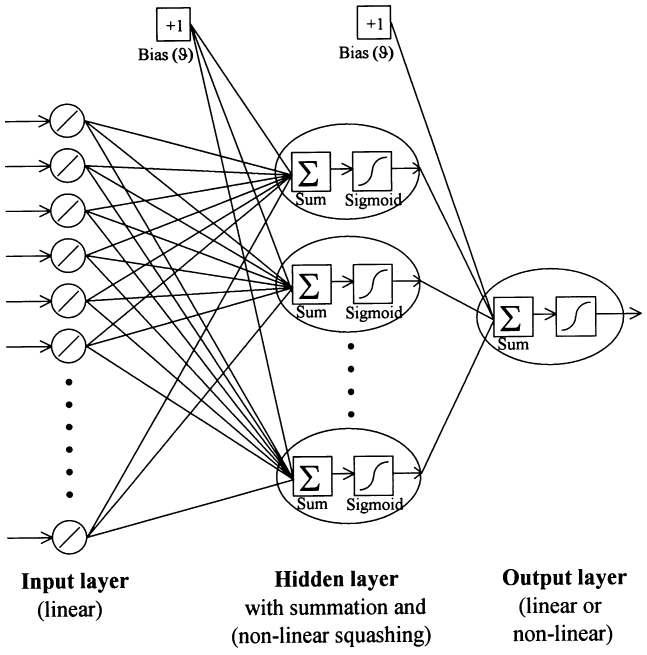

Fig. 5. An MLP neural network consisting of an input layer connected to a single node in the output layer by one hidden layer. In the architecture shown, adjacent layers of the network are fully interconnected although other architectures are possible. Nodes in the hidden and output layers consist of processing elements which sum the input applied to the node and scale the signal using a sigmoidal logistic squashing function.

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:

$$Out = \frac{1}{(1+\exp^{-Net})} \quad (3)$$

These signals (*Out*) are then passed to the output node, which sums them and the resulting values are 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 nodes were normalized between 0·1 and 0·9, whilst the output nodes were normalized between 0 and 1. 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 (Neal *et al*., 1994). 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 (Chauvin & Rumelhart, 1995; 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). An 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 0·2 and a momentum of 0·8 were used.

Each epoch represented the connection weight updatings and a recalculation of the root mean squared (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. Initially MLPs were trained until the RMSEF was 0·005 (0·5 %), and their ability to generalize was assessed on the test set. It was found that MLPs trained until the RMSEF was 0·01

(1%) were still able to generalize well, and since these MLPs obviously took less time to train and were less likely to overfit the input data (i.e. fitting to noise or the fitting of a model to outliers: Goodacre *et al*., 1996b; Kell & Sonnleitner, 1995), all MLPs were trained until the RMSEF was 0·01 (1%).

The error function on the output layer of these MLPs uses RMS error calculations, and cross-entropy may be a better choice of error function for some studies, since it allows one to assign Bayesian a posteriori probabilities (Richard & Lippmann, 1991). However, because it relies on probability density functions, it requires that the a priori population distributions be known, which is rarely the case (Bishop, 1995). We would add that for quantitative studies global functions such as RMSEF can be distorted by one error at the large end of the range much more than by a big error at the small end of the range. However, for qualitative identification studies this is rarely a problem.

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 (RBF) neural networks. All RBF analyses were also carried out with NeuFrame version 3.0.0.0 as detailed specifically by Saha & Keller (1990).

RBF networks are hybrid neural networks encompassing both unsupervised and supervised learning (Beale & Jackson, 1990; Bishop, 1995; Broomhead & Lowe, 1988; Hush & Horne, 1993; Moody & Darken, 1989; Park & Sandberg, 1991; Saha & Keller, 1990; Walczak & Massart, 1996; 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 (Fig. 6). 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 given in Fig. 7, 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 (Fig. 7).

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 normally 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 & Keller (1990). Briefly the training proceeds in two stages:

Stage 1. This involves unsupervised clustering of the input data (the input nodes were normalized between 0·1 and 0·9), typically using the *K*-means clustering algorithm (Duda & Hart, 1973; Everitt, 1993; Hush & 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 convenient because it positions the kernels relative to the density of the input data points. Next the receptive field is determined by the nearest-neighbour 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 jth RBF) and its nearest neighbour (k), and an overlap constant (*Overlap*) is used:

$$r_j = \text{Overlap} \times \min(\|w_j - w_k\|) \quad (4)$$

where $\| \ldots \|$ 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 neighbour; this optimum was also in agreement with the studies of Saha & Keller (1990).

The output from nodes in the hidden layer is dependent on the shape of the basis function and the one used was a Gaussian. Thus this value (R_j) for node j when given the ith input vector (i_i) can be calculated by:

$$\mathrm{R_j}(i_i) = \exp^{-(i_j^2/r_j^2)} \quad (5)$$

Stage 2. This involves supervised learning using simple linear regression. 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 five 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 training set (see Table 1 for details). 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.

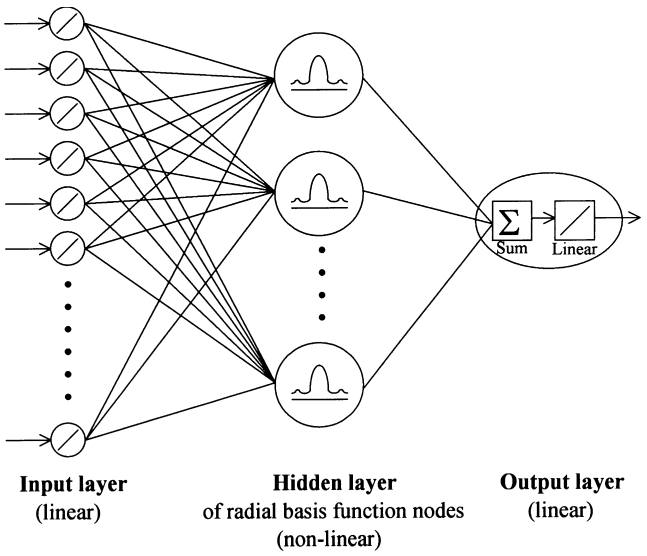

Fig. 6. RBF neural network consisting of an input layer connected to a single node in the output layer by 1 hidden layer. The hidden layer consists of radially symmetric Gaussian functions, although others exist (e.g. Mexican hat and thin plate splines).

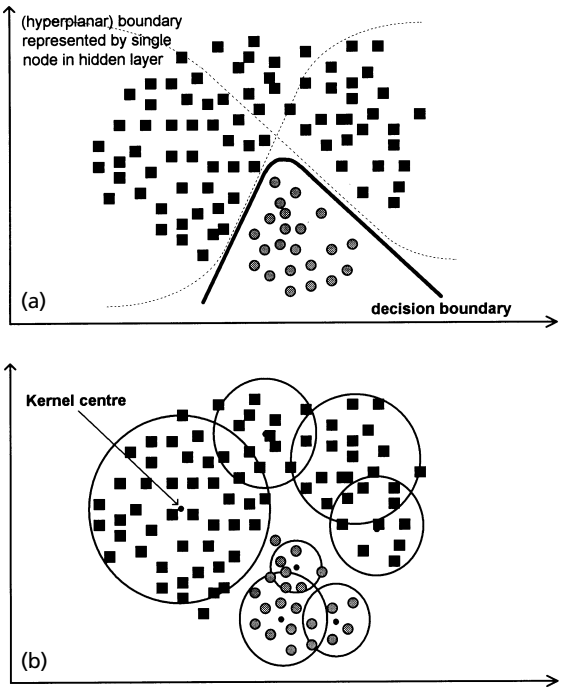

Fig. 7. (a) Typical decision boundary for a classification problem created between two data classes by an MLP with two nodes in the hidden layer, for two input nodes. Each hidden node represents a non-linear boundary and the nodes in the output layer interpolate this to form a decision boundary. (b) The same classification problem modelled by seven 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.

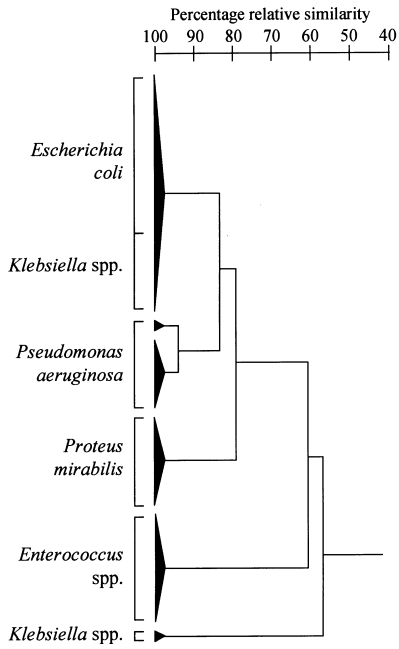

Fig. 8. Dendrogram based on PyMS data showing the relationship between the 59 bacterial isolates.

RESULTS

The raw spectra

Typical normalized PyMS spectra for *E. coli* isolate Ea and *Pr. mirabilis* isolate Pa are shown in Fig. 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 diffuse reflectance-absorbance FT-IR and dispersive Raman spectra of the same isolates are shown in Figs 2 and 3 respectively. These vibrational spectra and those from the other 57 bacteria all showed broad and complex contours; indeed for the Raman spectra it is 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 greatly reduced by the use of the near-infrared laser: Baraga *et al.*, 1992; Davey & Kell, 1996; Graselli & Bulkin, 1991).

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, essentially 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, was coded to give 59 individual groups, and analysed by DFA and HCA as detailed above. The resulting dendrogram from the analysis of the PyMS data is shown in Fig. 8, where it can be seen that five clusters are recovered. Although the *Ps. aeruginosa*, the *Pr. mirabilis* and the enterococcal strains form three well-defined clusters, the *Klebsiella* spp. do not form one group and some of them cluster with the 17 *E. coli* strains analysed. When the *Klebsiella* spp. were identified further by conventional means it was found that there were six *K. pneumoniae* and four *K. oxytoca* isolates. With respect to the clustering of these isolates in Fig. 8 it was found (identities not shown) that all six *K. pneumoniae* and two *K. oxytoca* isolates grouped with the *E. coli* strains, whilst the other two *K. oxytoca* isolates clustered separately. Therefore, the existence of two groups of *Klebsiella* spp. seen in the dendrogram (Fig. 8) was not due to two different species being isolated from the infected urine samples.

The analysis of the 59 bacterial isolates from their FT-IR data by DFA is depicted in Fig. 9 as a pseudo-3D ordination plot. In this figure (and in any view of this three-dimensional cube) it is also clear that the *Ps. aeruginosa* (A), the *Pr. mirabilis* (P) and the enterococcal

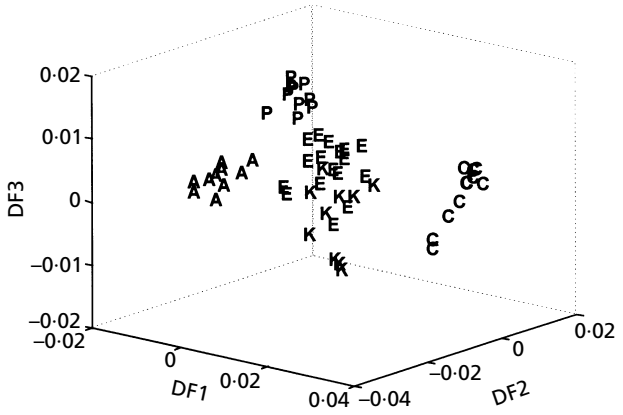

Fig. 9. Pseudo-3D DFA plot based on FT-IR data showing the relationship between the 59 bacterial isolates. The bacterial isolates are coded as follows; *E. coli* (E), *Pr. mirabilis* (P), *Klebsiella* spp. (K), *Ps. aeruginosa* (A), and *Enterococcus* spp. (C).

(C) strains form three distinct groups; however, the fourth, larger, cluster is a mixture of all the *E. coli* (E) and all the *Klebsiella* strains (K). This result again indicates that this unsupervised learning approach could not be used to give accurate identities of this group of clinical bacterial specimens.

Finally, DFA was used to analyse the Raman spectra; the results are shown in Fig. 10. Fig. 10(a) shows the analysis of all the strains, and the first discriminant function (DF 1) indicates that the majority of the variation was between the *Ps. aeruginosa* (A) strains and all the other isolates. This was possibly due to a small amount of fluorescence, since *Ps. aeruginosa* naturally fluoresces due to the production of pyocyanin (blue-green) and fluorescein (yellow) pigments (Pitt, 1990), 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 Fig. 10(b), 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.

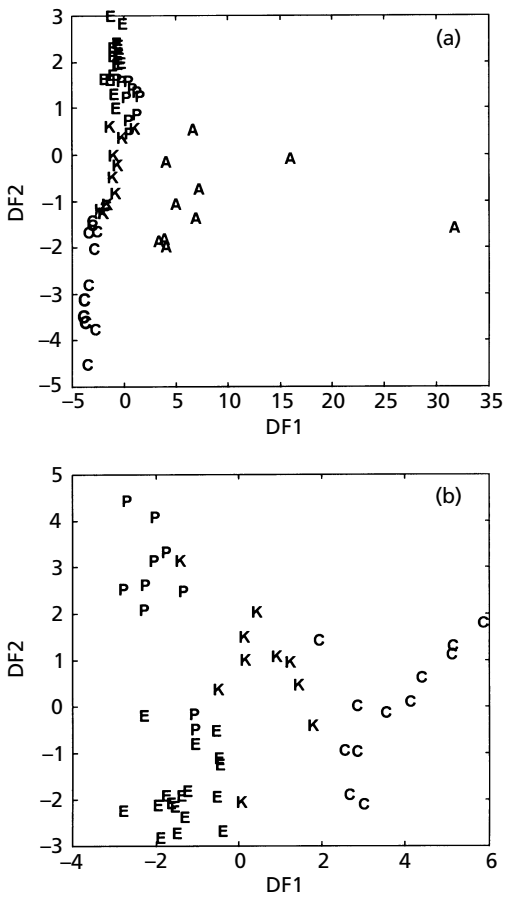

Fig. 10. DFA biplots based on Raman data showing the relationship between the 59 bacterial isolates (a) and the same after removal of the *Ps. aeruginosa* isolates (b). The bacterial isolates are coded as follows; *E. coli* (E), *Pr. mirabilis* (P), *Klebsiella* spp. (K), *Ps. aeruginosa* (A), and *Enterococcus* spp. (C).

Supervised analysis using ANNs

Since none of the spectroscopic data when analysed 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 ANN-based approaches of multilayer perceptrons (MLPs) and radial basis functions (RBFs).

As detailed above, the first five organisms in each of the five bacterial classes (a–e) were used to train the MLPs and RBFs (see Table 2). The input layers for the MLPs and RBFs were the full spectral data; therefore for PyMS these were 150 mass (m/z) intensities, for FT-IR these were normalized and detrended absorbances at 882 wavenumbers, and for Raman were the counts at 2283 wavenumber blocks. The outputs were always the same for both MLPs and RBFs and were binary encoded such that *E. coli* was coded as 10000, *Pr. mirabilis* as 01000, *Klebsiella* spp. as 00100, *Ps. aeruginosa* as 00010, and *Enterococcus* spp. as 00001.

The training set for the MLPs contained a relatively small number of spectra (75; five of each of the five bacterial classes in triplicate), and it is well known that if the number of parameters, or weights, in the calibration models is significantly higher than the number of exemplars in the training set then overfitting can more easily occur (Bishop, 1995; Chatfield, 1995; Seasholtz & Kowalski, 1993). Therefore, to help to obey the parsimony principle as described by Seasholtz & Kowalski (1993) the next stage was to reduce the number of inputs

Table 2. Identity of the bacteria used in the training set as judged by MLP analysis of their PyMS data

Bacterium	Bronglais identifier	Estimates from MLP				
		E. coli	*Pr. mirabilis*	*Klebsiella* spp.	*Ps. aeruginosa*	*Enterococcus* spp.
E. coli	Ea	**0·9**	0·0	0·1	0·0	−0·1
	Eb	**1·0**	0·0	−0·1	0·0	0·0
	Ec	**1·0**	0·0	−0·1	0·0	0·0
	Ed	**0·6**	0·0	0·4	0·0	0·0
	Ee	**1·0**	0·0	−0·1	0·0	0·0
Pr. mirabilis	Pa	0·0	**1·0**	0·1	0·0	0·0
	Pb	0·0	**1·0**	−0·1	0·0	0·0
	Pc	0·0	**0·9**	0·0	0·0	0·1
	Pd	−0·1	**1·0**	0·0	0·1	0·0
	Pe	0·0	**1·0**	0·0	−0·1	0·0
Klebsiella sp.	Ka	0·3	0·0	**0·7**	0·1	−0·1
	Kb	0·0	−0·1	**1·0**	0·0	0·0
	Kc	−0·1	0·1	**1·0**	0·0	0·0
	Kd	0·0	0·0	**1·0**	−0·1	0·0
	Ke	0·2	0·1	**0·7**	0·1	0·0
Ps. aeruginosa	Aa	0·2	−0·1	0·1	**0·8**	0·0
	Ab	0·0	−0·1	0·1	**1·0**	0·0
	Ac	−0·1	0·0	0·0	**1·1**	0·0
	Ad	0·0	0·1	−0·1	**1·0**	0·0
	Ae	0·0	0·0	0·0	**1·0**	0·0
Enterococcus sp.	Ca	0·0	0·0	0·0	0·1	**1·0**
	Cb	0·0	0·0	0·1	0·0	**1·0**
	Cc	0·0	0·0	0·0	0·0	**1·0**
	Cd	0·0	0·0	0·0	0·0	**1·0**
	Ce	0·0	0·0	0·0	0·0	**1·0**

to the MLPs. PCA is an excellent dimensionality-reduction technique, and the use of PC scores as inputs to MLPs, without deterioration of the calibration model, has previously been exploited in the analysis of UV–visible spectroscopic data (Blanco *et al*., 1995; Gemperline *et al*., 1991), for the identification of bacteria from their FT-IR spectra (Goodacre *et al*., 1996c), and for the quantification of biological systems from their PyMS spectra (Goodacre *et al*., 1997; Timmins & Goodacre, 1997). The optimal number of PCs as inputs to the PC-MLPs studied here was chosen based on the minimum needed whilst still being able to predict the test set correctly (as judged by the winning node in the output layer being taken as their identities); when too few PCs are used not enough information is present, and when more PCs are employed the later PCs contribute only noise to the model, thus increasing the probability of chance correlations between input and ouptut data. For PyMS the number of PCs used was the first 10 PCs (which accounted for 97·16% of the total variance); for FT-IR the first 20 PCs (96·88% of total variance), and for Raman the first 5 PCs (78·86% of total variance) were used. This PCA-based dimensionality-reduction process is of course not needed for RBFs since the first stage in this process involves the use of the unsupervised *K*-means clustering algorithm, and so bears similarity to the PC-MLP approach.

After training each of the three ANNs to an RMSEF of 0·01 in the training set, each calibrated system was challenged with both 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 unseen 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* Kg, which was identified as an *E. coli*. Exactly the same result was seen for the PC-MLPs (data not shown), but by contrast the full-spectral RBFs correctly identified all isolates (including *Klebsiella* strain Kg) in both the training and test sets.

All three ANN-based methods correctly identified all isolates in the training and test sets from their FT-IR data (data not shown). The most notable feature of the various ANNs trained with the IR spectra was the time taken to train to an RMSEF of 0·01 (1%) (Table 1). Full-

Table 3. Identity of the bacteria used in the test set as judged by MLP analysis of their PyMS data

Bacterium	Bronglais identifier	Estimates from MLP				
		E. coli	*Pr. mirabilis*	*Klebsiella* spp.	*Ps. aeruginosa*	*Enterococcus* spp.
E. coli	Ef	**0·7**	0·0	0·3	0·0	0·0
	Eg	**1·0**	0·0	0·0	0·0	0·1
	Eh	**1·1**	0·1	−0·1	0·1	−0·1
	Ei	**1·1**	0·1	−0·1	0·0	−0·2
	Ej	**1·0**	0·1	0·0	0·1	−0·1
	Ek	**1·1**	0·0	−0·1	0·0	−0·1
	El	**0·9**	0·1	0·0	0·1	−0·1
	Em	**1·1**	0·1	0·0	0·1	−0·2
	En	**0·9**	0·0	0·2	0·0	0·0
	Eo	**1·0**	0·0	−0·1	0·0	0·0
	Ep	**1·0**	0·1	0·0	0·0	−0·1
	Eq	**1·2**	0·1	−0·1	0·0	−0·2
Pr. mirabilis	Pf	−0·1	**1·0**	0·1	0·0	0·0
	Pg	0·0	**1·1**	−0·1	0·0	0·1
	Ph	−0·1	**1·1**	−0·1	0·0	0·0
	Pi	0·1	**0·8**	0·2	0·1	−0·1
	Pj	0·1	**1·0**	0·0	0·0	−0·1
Klebsiella sp.	Kf	0·2	−0·1	**0·9**	0·0	0·0
	Kg	**0·7**	0·1	0·3	0·1	−0·2
	Kh	−0·1	0·0	**0·9**	0·1	0·1
	Ki	0·2	0·0	**0·7**	0·0	0·1
	Kj	0·1	0·0	**0·9**	0·0	−0·1
Ps. aeruginosa	Af	−0·2	0·0	0·3	**0·9**	0·0
	Ag	0·1	0·1	−0·3	**1·1**	0·0
	Ah	0·0	0·1	−0·1	**1·1**	0·0
	Ai	0·1	0·0	0·1	**0·8**	0·0
	Aj	0·0	0·0	0·0	**1·0**	0·0
Enterococcus sp.	Cf	0·0	0·0	0·0	0·0	**1·0**
	Cg	0·0	0·0	−0·1	0·0	**1·1**
	Ch	0·0	0·0	0·0	0·0	**1·0**
	Ci	0·0	0·0	−0·1	0·0	**1·0**
	Cj	0·0	0·1	0·1	0·0	**0·9**
	Ck	0·1	0·0	−0·2	0·0	**1·1**
	Cl	0·0	0·0	0·0	0·0	**1·1**

spectral MLPs trained with 882 IR absorbances as inputs took 5×10^3 epochs to train, which in 'real time' took 60 min on an IBM-compatible personal computer (dual P133 processor, 64 Mbytes RAM). The topology of these MLPs included 12 hidden nodes and five outputs nodes, and between the input and hidden layers and the hidden (including the single bias node) and output layers, these 882-2-5 MLPs contained 10649 weighted connections. When the first 20 PC scores were used as inputs, these 20-6-5 PC-MLPs took only 1 min (3×10^3 epochs), on the same personal computer, to reach the same RMSEF; this was hardly surprising since the number of weighted connections in these MLPs was only 155, i.e. was 68 times fewer parameters compared to the 882-2-5 MLPs. However, the RBFs were much the fastest to train and took just 10 s. The RBF is a hybrid ANN and involves first the unsupervised clustering of the IR spectra using *K*-means, followed by simple linear regression of the output from the Gaussians in the hidden layer on to the five output nodes. This means that this method is not computationally intensive, since unlike back-propagation-based MLPs they do not perform gradient descent (Walczak & Massart, 1996).

The results for the ANN analyses of the Raman spectra were slightly less successful. Whilst each method got 100% of the training set correct, full-spectral MLPs correctly identified 25 (74%), RBFs 26 (76%), and PC-MLPs 28 (82%) of the 34 isolates in the test set. Details of the results of interrogating the three ANN methods (test set only) are given in Table 4, where it can be seen that the *E. coli*, *Ps. aeruginosa* and *Enterococcus* spp. were nearly always identified but that the *Pr. mirabilis* and *Klebsiella* spp. isolates were mostly incorrectly

Table 4. Results from the analysis of Raman spectra in the test set using ANNs

Bacterium	No. in test set	No. (and %) correctly identified		
		MLPs	PC-MLPs	RBFs
E. coli	12	11 (92)	12 (100)	10 (83)
Pr. mirabilis	5	1 (20)	4 (80)	2 (40)
Klebsiella spp.	5	1 (20)	1 (20)	2 (40)
Ps. aeruginosa	5	5 (100)	5 (100)	5 (100)
Enterococcus spp.	7	7 (100)	6 (86)	7 (100)
Total	34	25 (74)	28 (82)	26 (76)

assigned by each of the methods. With one exception, the *Pr. mirabilis* isolates that were incorrectly identified were taken as belonging to the *Klebsiella* group (the other was wrongly identified as an enterococcus). The DFA of these and all the other isolates in Fig. 10(b) gives us some insight into this result in that the *Pr. mirabilis* isolates (P) grouped more closely with the *Klebsiella* isolates (K) than any of the other isolates. The eleven *Klebsiella* spp. which were wrongly identified were classified as belonging to *E. coli* (1), *Pr. mirabilis* (3) or enterococci (7); the DFA plot (Fig. 10a) suggests that the *Klebsiella* isolates are at the centre of a triangle where the three tips consisted of members of only *E. coli*, *Pr. mirabilis* or enterococcal isolates; this could explain why the *Klebsiella* strains were wrongly identified as belonging to others of those groups.

The time taken to reach a similar RMSEF level using the Raman data varies significantly (Table 1). The full-spectral MLPs took 30 h to train, compared to using PCA as a pre-processing stage to the MLPs, which trained in only 10 min. Finally, the full-spectral RBFs were fastest and took a mere 2 min to calibrate.

DISCUSSION

Three rapid spectroscopic approaches for 'whole-organism fingerprinting' – Curie-point PyMS, diffuse reflectance-absorbance FT-IR and dispersive Raman microscopy – were used to analyse 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 (Fig. 8) and ordination plots (Figs 9 and 10).

By contrast, for PyMS and FT-IR, the ANN-based approaches using MLPs or RBFs could be trained with small numbers of 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 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 aqueous slurries used for the Raman measurements was low; future studies will therefore concentrate on analysing the bacterial samples directly from colonies on agar plates or by drying them onto a metal surface and seeking to effect surface-enhanced Raman spectroscopy (SERS; Cotton *et al*., 1991; Nabiev *et al*., 1994; Nabiev & Manfait, 1993).

Whilst UV resonance Raman spectroscopy (Nelson & Sperry, 1991; Nelson *et al*., 1992) and FT-Raman microscopy (Puppels & Greve, 1993; Puppels *et al*., 1995) have been exploited for the discrimination of microbes, these are the first published data, as far as we are aware, showing the ability of dispersive Raman microscopy to discriminate clinically significant intact bacterial cells. Raman microscopy has the advantage over PyMS and FT-IR that it is possible to analyse single cells (Puppels & Greve, 1993), although to get a satisfactory signal-to-noise ratio this process is quite lengthy ($>$ 10 min) and the analysis of the hyperspectral data complex. This and the other features of the three spectroscopic methods are detailed in Table 5. The main advantage that PyMS conveys is that the multivariate analyses of these data are well developed and easily implemented (Goodacre & Kell, 1996; Goodacre *et al*., 1996b; Gutteridge, 1987; Magee, 1993), but this technique has the potential disadvantage that it destroys the sample. FT-IR has the advantage of speed and, particularly with our diffuse reflectance-absorbance approach (Goodacre *et al*., 1996c; Timmins *et al*., 1998; Winson *et al*., 1997), easily allows the acquisition of 400 samples per hour on a single 10×10 cm aluminium plate. Although the FT-IR spectra are of higher dimensionality than PyMS spectra, and the data analysis slightly more complex, it is fair to say that these slight disadvantages will diminish with computational advances (in both hardware and software).

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 h. However, we have previously used PCA as a method for reducing the

Table 5. Features of the whole-cell fingerprinting methods studied

	Curie-point PyMS	Diffuse reflectance-absorbance FT-IR	Dispersive Raman
Destructive	Yes	No (although sample is dried)	No
Sample size	> 50 μg	> 50 μg on plates 5 μm diam. for microscope	Slurry in vials 1 μm diam. for microscope
Typical no. of cells	10^6–10^7	Plates: 10^6–10^7 Microscope: aggregates	Vials: 10^9 ml^{-1} Microscope: single cells
Typical speed	1·5–2 min	5–30 s	1–20 min
Automatable	Yes	Yes	Yes
Complex data capture	No	No	Fairly
Typical dimensionality	150	882	2283
Data analysis (1 = easy to 5 = hard)	1	2	4

number of inputs to ANNs (Goodacre *et al*., 1996c, 1997; Timmins & Goodacre, 1997) and in the present study using PC scores as inputs to MLPs reduced the training time to only 10 min, with a slight enhancement in the predictive ability of the PC-MLP. Finally the training time for the full spectral RBFs was very quick – only 2 min – with equivalent performance compared to the full-spectral MLPs.

At least for the present study the full-spectral RBF networks have the advantage of speed over full-spectral MLPs, and to a lesser degree over the PC-MLPs (Table 1). However, the results from the RBF's outputs are less quantized than those from the two MLP approaches; that is to say the outputs were not always very close to 0 or 1. For the FT-IR analysis, when the criterion for the identity of a bacterium was taken simply as the winning node in the output layer, all three ANNs predicted all 34 isolates in the test set correctly. However, if a more rigid criterion was used which stipulated that a correct identification was taken to be that the winning node must be > 0·75 and all other losing nodes < 0·25, then the full-spectral MLP incorrectly identified seven isolates, compared with the full-spectral RBF which now misidentified 13 of the 34 isolates. By contrast, the PC-MLP approach was best and only two isolates were wrongly assigned. When the speed of training and the more quantized predictions are considered, the PC-MLPs would appear to be the best ANN-based approach; moreover, this method was also best for predicting the identities of these bacteria from their Raman spectra (Table 4).

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

ACKNOWLEDGEMENTS

We are grateful to one of the reviewers for his/her useful comments on this paper. R. G. and E. M. T. are indebted to the Wellcome Trust for financial support (grant number 042615/Z/94/Z). N.K., A.M.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.

REFERENCES

Abousleman, G. P., Gifford, E. & Hunt, B. R. (1994). Enhancement and compression techniques for hyperspectral data. *Optic Eng* **33**, 2562–2571.

Alsberg, B. K., Woodward, A. M., Winson, M. K., Rowland, J. & Kell, D. B. (1997). Wavelet denoising of infrared spectra. *Analyst* **122**, 645–652.

Baraga, J. J., Feld, M. S. & Rava, R. P. (1992). Rapid near-infrared Raman spectroscopy of human tissue with a spectrograph and CCD detector. *Appl Spectrosc* **46**, 187–190.

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

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S. & Redon, M. (1995). Artificial neural networks for multicomponent kinetic determinations. *Anal Chem* **67**, 4477–4483.

Bouffard, S. P., Katon, J. E., Sommer, A. J. & Danielson, N. D. (1994). Development of microchannel thin layer chromatography with infrared microspectroscopic detection. *Anal Chem* **66**, 1937–1940.

Broomhead, D. S. & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Syst* **2**, 312–355.

Casadevall, A. (1996). Crisis in infectious diseases – time for a new paradigm. *Clin Infect Dis* **23**, 790–794.

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

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *J R Stat Soc Ser A* **158**, 419–466.

Chauvin, Y. & Rumelhart, D. E. (1995). *Backpropagation: Theory, Architectures, and Applications*. Hove, UK: Erlbaum.

Chun, J., Atalan, E., Ward, A. C. & Goodfellow, M. (1993). Artificial neural network analysis of pyrolysis mass spectrometric

data in the identification of *Streptomyces* strains. *FEMS Microbiol Lett* **107**, 321–325.

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

Cotton, T. M., Kim, J. H. & Chumanov, G. D. (1991). Application of surface enhanced Raman spectroscopy to biological systems. *J Raman Spectrosc* **22**, 729–742.

Davey, H. M. & Kell, D. B. (1996). Flow cytometry and cell sorting of heterogeneous microbial populations – the importance of single cell analyses. *Microbiol Rev* **60**, 641–696.

Dixon, W. J. (1975). *Biomedical Computer Programs*. Los Angeles: University of California Press.

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

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

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

Freeman, R., Goodacre, R., Sisson, P. R., Magee, J. G., Ward, A. C. & Lightfoot, N. F. (1994). Rapid identification of species within the *Mycobacterium tuberculosis* complex by artificial neural network analysis of pyrolysis mass spectra. *J Med Microbiol* **40**, 170–173.

Gemperline, P. J., Long, J. R. & Gregoriou, V. G. (1991). Nonlinear multivariate calibration using principal components regression and artificial neural networks. *Anal Chem* **63**, 2313–2323.

Goetz, A. F. H., Vane, G., Solomon, J. & Rock, B. N. (1985). Imaging spectrometry for earth remote sensing. *Science* **228**, 1147–1153.

Goodacre, R. & Kell, D. B. (1996). Pyrolysis mass spectrometry and its applications in biotechnology. *Curr Opin Biotechnol* **7**, 20–28.

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

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

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

Goodacre, R., Trew, S., Wrigley-Jones, C., Saunders, G., Neal, M. J., Porter, N. & 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. *Anal Chim Acta* **313**, 25–43.

Goodacre, R., Hiom, S. J., Cheeseman, S. L., Murdoch, D., Weightman, A. J. & Wade, W. G. (1996a). Identification and discrimination of oral asaccharolytic *Eubacterium* spp. using pyrolysis mass spectrometry and artificial neural networks. *Curr Microbiol* **32**, 77–84.

Goodacre, R., Neal, M. J. & Kell, D. B. (1996b). Quantitative analysis of multivariate data using artificial neural networks: a tutorial review and applications to the deconvolution of pyrolysis mass spectra. *Zentralbl Bakteriol – Int J Med Microbiol Virol Parasitol Infect Dis* **284**, 516–539.

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

Goodacre, R., Hammond, D. & Kell, D. B. (1997). Quantitative analysis of the adulteration of orange juice with sucrose using pyrolysis mass spectrometry and chemometrics. *J Anal Appl Pyrolysis* **40/41**, 135–158.

Goodacre, R., Rooney, P. J. & Kell, D. B. (1998). Discrimination between methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* using pyrolysis mass spectrometry and artificial neural networks. *J Antimicrob Chemother* **41**, 23–34.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325–338.

Graselli, J. G. & Bulkin, B. J. (1991). *Analytical Raman Spectroscopy*. New York: Wiley.

Griffiths, P. R. & de Haseth, J. A. (1986). *Fourier Transform Infrared Spectrometry*. New York: Wiley.

Gruneberg, R. N. (1994). Changes in urinary pathogens and their antibiotic sensitivities 1971–1992. *J Antimicrob Chemother* **33**, **Suppl A**, 1–8.

Gutteridge, C. S. (1987). Characterization of microorganisms by pyrolysis mass spectrometry. *Methods Microbiol* **19**, 227–272.

Haykin, S. S. (1994). *Neural Networks: a Comprehensive Foundation*. New York: Macmillan.

Helm, D., Labischinski, H., Schallehn, G. & Naumann, D. (1991). Classification and identification of bacteria by Fourier transform infrared spectroscopy. *J Gen Microbiol* **137**, 69–79.

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

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

Kell, D. B. & Sonnleitner, B. (1995). GMP – Good Modelling Practice: an essential component of good manufacturing practice. *Trends Biotechnol* **13**, 481–492.

Lewis, D. A. (1989). Bacteriology of urine. In *Medical Bacteriology: a Practical Approach*, pp. 1–19. Edited by P. M. Hawkey & D. A. Lewis. Oxford: IRL Press.

MacFie, H. J. H., Gutteridge, C. S. & Norris, J. R. (1978). Use of canonical variates in differentiation of bacteria by pyrolysis gas–liquid chromatography. *J Gen Microbiol* **104**, 67–74.

Magee, J. T. (1993). Whole-organism fingerprinting. In *Handbook of New Bacterial Systematics*, pp. 383–427. Edited by M. Goodfellow & 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. & Hileman, F. D. (1982). *Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials*. Amsterdam: Elsevier.

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

Morgan, M. G. & McKenzie, H. (1993). Controversies in the laboratory diagnosis of community acquired urinary tract infection. *Eur J Clin Microbiol Infect Dis* **12**, 491–504.

Nabiev, I. & Manfait, M. (1993). Industrial applications of the surface enhanced Raman spectroscopy. *Rev Inst Français Petrole* **48**, 261–285.

Nabiev, I., Chourpa, I. & Manfait, M. (1994). Applications of Raman and surface enhanced Raman scattering spectroscopy in medicine. *J Raman Spectrosc* **25**, 13–23.

Naumann, D., Helm, D. & Labischinski, H. (1991a). Microbiological characterizations by FT-IR spectroscopy. *Nature* **351**, 81–82.

Naumann, D., Helm, D., Labischinski, H. & Giesbrecht, P. (1991b). 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.

Neal, M. J., Goodacre, R. & Kell, D. B. (1994). On the analysis of pyrolysis mass spectra using artificial neural networks. Individual input scaling leads to rapid learning. In *Proceedings of the World Congress on Neural Networks*, pp. I318–I323. San Diego: International Neural Network Society.

Nelson, W. H. & Sperry, J. F. (1991). UV resonance Raman spectroscopic detection and identification of bacteria and other microorganisms. In *Modern Techniques for Rapid Microbiological Analysis*, pp. 97–143. Edited by W. H. Nelson. New York: VCH.

Nelson, W. H., Manoharan, R. & Sperry, J. F. (1992). UV resonance Raman studies of bacteria. *Appl Spectrosc Rev* **27**, 67–124.

Pappas, P. G. (1991). Laboratory in the diagnosis and management of urinary tract infections. *Med Clin North Am* **75**, 313–325.

Park, J. & Sandberg, I. W. (1991). Universal approximation using radial basis function networks. *Neural Comput* **3**, 246–257.

Pitt, T. L. (1990). Pseudomonas. In *Topley and Wilson's Principles of Bacteriology, Virology and Immunity*, vol. 2, pp. 255–274. Edited by M. T. Parker & L. Collier. London: Edward Arnold.

Puppels, G. J. & Greve, J. (1993). Raman microspectroscopy of single whole cells. *Adv Spectrosc* **20A**, 231–265.

Puppels, G. J., Schut, T. C. B., Sijtsema, N. M., Grond, M., Maraboeuf, F., Degrauw, C. G., Figdor, C. G. & Greve, J. (1995). Development and application of Raman microspectroscopic and Raman imaging techniques for cell biological studies. *J Mol Struct* **347**, 477–483.

Richard, M. D. & Lippmann, R. P. (1991). Neural network classifiers estimate Bayesian *a posteriori* probabilities. *Neural Comput* **3**, 461–483.

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

Saha, A. & Keller, J. D. (1990). Algorithms for better representation and faster learning in radial basis functions. In *Advances in Neural Information Processing Sytems*, pp. 482–489. Edited by D. Touretzky. San Mateo, CA: Morgan Kaufmann Publishers.

Savitzky, A. & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem* **36**, 1627–1633.

Schrader, B. (1995). *Infrared and Raman Spectroscopy: Methods and Applications*. Weinheim: Verlag Chemie.

Seasholtz, M. B. & Kowalski, B. (1993). The parsimony principle applied to multivariate calibration. *Anal Chim Acta* **277**, 165–177.

Sisson, P. R., Freeman, R., Law, D., Ward, A. C. & Lightfoot, N. F. (1995). Rapid detection of verocytotoxin production status in *Escherichia coli* by artificial neural network analysis of pyrolysis mass spectra. *J Anal Appl Pyrolysis* **32**, 179–185.

Slack, R. C. B. (1995). Urinary infections. In *Antimicrobial Chemotherapy*, pp. 243–250. Edited by D. Greenwood. Oxford: Oxford University Press.

Timmins, É. M. & Goodacre, R. (1997). Rapid quantitative analysis of binary mixtures of *Escherichia coli* strains using pyrolysis mass spectrometry with multivariate calibration and artificial neural networks. *J Appl Microbiol* **83**, 208–218.

Timmins, É. M., Howell, S. A., Alsberg, B. K., Noble, W. C. & Goodacre, R. (1998). Rapid differentiation of closely related *Candida* species and strains by pyrolysis mass spectrometry and fourier transform infrared spectroscopy. *J Clin Microbiol* **36**, 367–374.

Walczak, B. & Massart, D. L. (1996). The radial basis functions – partial least squares approach as a flexible non-linear regression technique. *Anal Chim Acta* **331**, 177–185.

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: Wiley.

Wilkie, M. E., Almond, M. K. & Marsh, F. P. (1992). Diagnosis and management of urinary tract infection in adults. *Br Med J* **303**, 1137–1141.

Wilkins, M. F., Morris, C. W. & Boddy, L. (1994). A comparison of radial basis function and backpropagation neural networks for identification of marine phytoplankton from multivariate flow cytometry data. *Comput Appl Biosci* **10**, 285–294.

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

Williams, K. P. J., Pitt, G. D., Smith, B. J. E., Whitley, A., Batchelder, D. N. & Hayward, I. P. (1994b). Use of a rapid scanning stigmatic Raman imaging spectrograph in the industrial environment. *J Raman Spectrosc* **25**, 131–138.

Wilson, T. A., Rogers, S. K. & Myers, L. R. (1995). Perceptual-based hyperspectral image fusion using multiresolution analysis. *Optic Eng* **34**, 3154–3164.

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

Winson, M. K., Goodacre, R., Woodward, A. M., Timmins, É. M., Jones, A., Alsberg, B. K., Rowland, J. J. & Kell, D. B. (1997). Diffuse reflectance absorbance spectroscopy taking in chemometrics (DRASTIC). A hyperspectral FT-IR-based approach to rapid screening for metabolite overproduction. *Anal Chim Acta* **348**, 273–282.

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.

Received 7 October 1997; revised 15 January 1998; accepted 20 January 1998.