Journal of Applied Microbiology 1997, **83**, 208–218

Rapid quantitative analysis of binary mixtures of *Escherichia coli* strains using pyrolysis mass spectrometry with multivariate calibration and artificial neural networks

É.M. Timmins and R. Goodacre
Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion, UK

6006/11/96: received 28 November 1996, revised and accepted 5 December 1996

É.M. TIMMINS AND R. GOODACRE. 1997. Pyrolysis mass spectrometry (PyMS) and multivariate calibration were used to show the high degree of relatedness between *Escherichia coli* HB101 and *E. coli* UB5201. Next, binary mixtures of these two phenotypically closely related *E. coli* strains were prepared and subjected to PyMS. Fully interconnected feedforward artificial neural networks (ANNs) were used to analyse the pyrolysis mass spectra to obtain quantitative information representative of the level of *E. coli* UB5201 in *E. coli* HB101. The ANNs exploited were trained using the standard back propagation algorithm, and the nodes used sigmoidal squashing functions. Accurate quantitative information was obtained for mixtures with $>3\%$ *E. coli* UB5201 in *E. coli* HB101. To remove noise from the pyrolysis mass spectra and so lower the limit of detection, the spectra were reduced using principal components analysis (PCA) and the first 13 principal components used to train ANNs. These PCA-ANNs allowed accurate estimates at levels as low as 1% *E. coli* UB5201 in *E. coli* HB101 to be predicted. In terms of bacterial numbers, it was shown that the limit of detection for PyMS in conjunction with ANNs was 3×10^4 *E. coli* UB5201 cells in $1{\cdot}6 \times 10^7$ *E. coli* HB101 cells. It may be concluded that PyMS with ANNs provides a powerful and rapid method for the quantification of mixtures of closely related bacterial strains.

INTRODUCTION

In medicine and biotechnology there is a continuing need to find new pharmaceuticals, and hence for the development of rapid and efficient methods to screen large numbers of microbial cultures for the production of biologically active metabolites (e.g. Crueger and Crueger 1989; Tanaka and Omura 1993; Bevan *et al.* 1995). Such metabolites can additionally provide new structural templates for synthetic programs using rational methods of drug design through chemical synthesis, including combinatorial methods. It is imperative therefore that the concentration of the fermentation product (the determinand) is assessed accurately, such that the most high yielding strains are selected, and to assist the subsequent optimization of the bioprocess. In addition, it is also important that early detection of low-level microbial contamination of fermentations be assessed rapidly.

Pyrolysis mass spectrometry (PyMS) is a rapid, automated, instrument-based technique which permits the acquisition of spectroscopic data from 300 or more samples per working day. The method typically involves the thermal degradation of complex material in a vacuum by Curie-point pyrolysis; this causes molecules to cleave at their weakest points to produce smaller, volatile fragments called pyrolysate (Irwin 1982). A mass spectrometer can then be used to separate the components of the pyrolysate on the basis of their mass-to-charge ratio (m/z) to produce a pyrolysis mass spectrum, which can then be used as a chemical signature (fingerprint) of the complex material analysed (Meuzelaar *et al.* 1982).

PyMS is well established within (micro)biology for the characterization of bacterial systems and in particular the technique has been successful in the interstrain comparison

Correspondence to: Dr Royston Goodacre, Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion SY23 3DA, UK (e-mail: rrg@aber.ac.uk).

of a variety of medically important bacterial species and strains (see Gutteridge 1987; Magee 1993; Goodacre and Kell 1996 for reviews). Within our laboratory, this technique has been extended to the *quantitative* analysis of the chemical constituents of microbial and other samples, through the application of the *supervised* learning methods of artificial neural networks (ANNs) (for introductory surveys see Rumelhart *et al.* 1986; Wasserman 1989; Beale and Jackson 1990; Zupan and Gasteiger 1993; Bishop 1995) and the multivariate linear regression techniques of partial least squares regression (PLS) and principal components regression (PCR) (for texts see Haaland and Thomas 1988; Martens and Næs 1989; Brereton 1992). We have shown that using the combination of PyMS and ANNs it is possible to follow the production of indole in a number of strains of *Escherichia coli* grown on media incorporating various amounts of tryptophan (Goodacre and Kell 1993) to quantify the (bio)chemical constituents of complex binary mixtures of proteins and nucleic acids in glycogen (Goodacre *et al.* 1993, 1994b).

With regard to biotechnology, the combination of PyMS and ANNs can be exploited to quantify the amount of mammalian cytochrome b_5 expressed in *E. coli* (Goodacre *et al.* 1994a), and to measure the level of metabolites in fermentor broths (Goodacre *et al.* 1994d); samples from fermentations of a single organism in a complex production medium were analysed quantitatively for a drug of commercial interest, and the drug could be quantified in a variety of mutant-producing strains cultivated in the same medium, thus effecting a rapid screening for the high-level production of desired substances (Goodacre *et al.* 1994d). In related studies *Penicillium chrysogenum* fermentation broths were analysed quantitatively for penicillins using PyMS and ANNs (Goodacre *et al.* 1995), and to monitor *Gibberella fujikuroi* fermentations producing gibberellic acid (Goodacre and Kell 1996).

In this period we were also the first to show that PyMS with ANNs could be exploited to measure the concentrations of binary mixtures of *E. coli* and *Staphylococcus aureus* (Goodacre *et al.* 1996) and tertiary mixtures of cells of the bacteria *Bacillus subtilis*, *E. coli* and *Staph. aureus* (Goodacre *et al.* 1994b). It is likely that this approach could be exploited for determining microbial contamination in fermentations or bacterial levels in medically important biofluids such as urine.

In this study we firstly used PyMS with canonical variates analysis (CVA) to demonstrate the high degree of relatedness between *Escherichia coli* UB5201 and *E. coli* HB101. To mimic microbial contamination in a fermentation with an organism very closely related to the producer organism we then prepared various binary mixtures of these two *E. coli* strains. PyMS was exploited to analyse these mixtures and the pyrolysis mass spectra obtained was used to train ANNs so as to obtain quantitative information about the level of *E. coli* UB5201 in mixtures with *E. coli* HB101.

MATERIALS AND METHODS

Organisms and cultivation

The strains to be quantified were *E. coli* HB101 (Maniatis *et al.* 1982) and *E. coli* UB5201 (de la Cruz and Grinsted 1982). To show the high degree of relatedness between these two strains, both isolates were analysed (by PyMS) together with several different bacteria representing other strains, species and genera (for details see Table 1). All strains were cultured aerobically on LabM blood agar base (37 mg ml^{-1}) for 16 h at 37°C. After growth, biomass was carefully collected using sterile plastic loops and suspended in 1 ml amounts of sterile physiological saline (0·9% NaCl).

Preparation of the binary *E. coli* mixtures for quantification by PyMS

Escherichia coli HB101 and *E. coli* UB5201 were grown separately in 4 l of liquid media (glucose [BDH], 10·0 g; peptone [Lab M], 5·0 g; beef extract [Lab M], 3·0 g; H_2O, 1 l) for 16 h at 37°C in a shaker incubator. After growth, the cells were harvested by centrifugation and washed in physiological saline. The dry weights of the cells were then estimated gravimetrically and used to adjust the weight of the final slurries with physiological saline to ≈40 mg ml^{-1}. The number of cells per ml of these slurries was determined using a plate count method (Collins and Lyne 1970). Three sets of binary mixtures were prepared (as detailed in Table 2). Set A consisted of mixtures containing between 0 and 100% *E. coli* UB5201 in *E. coli* HB101 (in 5% steps), set B consisted

Table 1 Designation of identifiers for multivariate data analysis on the PyMS spectra of different bacterial species and strains

Identifier for CVA plots	Bacterium
A	*Bacillus cereus*
B	*B. globigii* cbBg 1
C	*B. licheniformis*
D	*B. globigii* 1049
E	*B. megaterium*
F	*B. subtilis*
G	*Escherichia coli* UB5201
H	*Streptococcus faecalis*
I	*Escherichia coli* HB101
J	*E. coli* 045:K^{-}
K	*E. coli* 883
L	*E. coli* 861
M	*E. coli* 890
N	*Klebsiella pneumoniae*
O	*Kl. oxytoca*

CVA, Canonical variates analysis.

Table 2 Optimal neural network conditions used to quantify *Escherichia coli* UB5201 in binary mixtures of *E. coli* HB 101 and *E. coli* UB5201 from pyrolysis mass spectra of three sets of binary mixtures

Binary mixture set	A	B	C
Range	0–100% *E. coli* UB5201	0–20% *E. coli* UB5201	0–5% *E. coli* UB5201
Training data (% *E. coli* UB5201 in *E. coli* HB101)	0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100	0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20	0, 0·5, 1·0, 1·5, 2·0, 2·5, 3·0, 3·5, 4·0, 4·5, 5·0
Cross validation data (% *E. coli* UB5201 in *E. coli* HB101)	5, 15, 25, 35, 45, 55, 65, 75, 85, 95	1, 3, 5, 7, 9, 11, 13, 15, 17, 19	0·25, 0·75, 1·25, 1·75, 2·25, 2·75 3·25, 3·75, 4·25, 4·75
Architecture	150-8-1	150-8-1	150-8-1
Scaling on:			
Input layer*	0 to 5	0 to 6	0 to 6
Output layer†	−50 to 150	−10 to 30	−2·5 to 7·5
Spectral data averaged *before* training	No	No	Yes
Mean absolute error			
Training set	0·70 (σ 0·14)	0·33 (σ 0·01)	0·17 (σ 0·03)
Cross validation set	2·05 (σ 0·05)	0·95 (σ 0·09)	0·47 (σ 0·03)
Number of epochs	7×10^4	8×10^4	2×10^4

* Input layer was scaled across the whole mass range such that the minimum was set to 0 and the maximum to +1.

† Output layer was scaled with 50% headroom on the minimum and maximum values.

of mixtures with 0–20% *E. coli* UB5201 in *E. coli* HB101 (in 1% steps) whilst set C was made up of mixtures with 0–5% *E. coli* UB5201 in *E. coli* HB101 (in 0·25% steps).

Pyrolysis mass spectrometry

Five μl aliquots of these bacterial suspensions were evenly applied to clean iron–nickel foils which had been partially inserted into clean pyrolysis tubes. Samples were run in triplicate, except the samples of binary mixture set A which were run in quadruplicate. Prior to pyrolysis, the samples were oven-dried at 50°C for 20 min, 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 then placed $\approx$1 mm from the mouth of each tube.

Pyrolysis mass spectrometry was performed on a Horizon Instrument PyMS-200X (Horizon Instruments Ltd, Heathfield, UK). For full operational procedures, see Goodacre *et al.* (1994a,b,c). 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.

Data analysis

PyMS data may be displayed as quantitative pyrolysis mass spectra (Fig. 1). The abscissa represents the m/z ratio, while the ordinate contains information on ion count for any particular m/z value ranging from 51 to 200. Data were normalized as a percentage of the total ion count to remove the influence of sample size *per se*.

Multivariate data analysis to show the degree of relatedness between E. coli *UB5201 and* E. coli *HB101.* The normalized data of the 15 isolates (Table 1) were processed with the GENSTAT package (Nelder 1979) which runs under Microsoft DOS 6.2 on an IBM-compatible PC. This method has been previously described by Gutteridge *et al.* (1985). The initial stage in determining the between-group relatedness involved the reduction of the data by principal component analysis (PCA) (Chatfield and Collins 1980; Causton 1987; Flury and Riedwyl 1988; Everitt 1993); this is a well known technique for reducing the dimensionality of multivariate data whilst preserving most of the variance. Data were preserved by keeping only those principal components (PCs) whose eigenvalues accounted for more than 0·1% of the total variance. CVA then discriminated between groups on the basis of the retained PCs and the *a priori* knowledge of which spectra were replicates (MacFie *et al.* 1978; Windig *et al.* 1983). The next stage involved the construction of a percentage similarity matrix by transforming the Mahalanobis' distance between *a priori* groups in CVA with the Gower similarity coefficient S_G (Gower 1966).

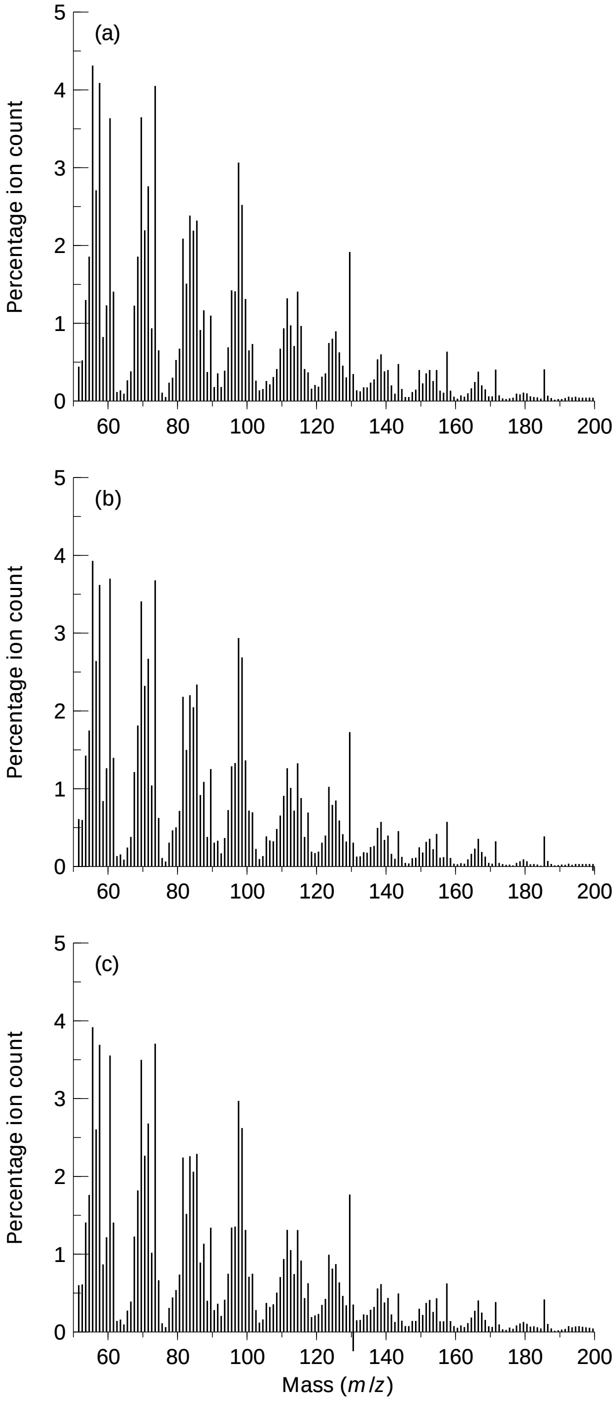

Fig. 1 Normalized pyrolysis mass spectra of (a) axenic *Escherichia coli* UB5201, (b) axenic *E. coli* HB101 and (c) 20:80 mixture of *E. coli* UB5201 and *E. coli* HB101

Artificial neural network analysis to quantify E. coli *UB5201 in the binary mixtures.* The normalized pyrolysis mass spectra for each set of binary mixtures were quantitatively analysed by ANNs using a user-friendly, neural network simulation program, NeuralDesk (Neural Computing Sciences, Southampton, UK), which runs under Microsoft Windows 3.11 on an IBM-compatible PC. Data were also processed prior to analysis using the Microsoft Excel 4.0 spreadsheet.

The algorithm used was standard back propagation (Rumelhart *et al.* 1986; Werbos 1994) which employs processing nodes (neurons or units), connected using abstract interconnections (connections or synapses). Connections each have an associated real value, termed the weight, that scales signals passing through them. Nodes sum the signals feeding to them and output this sum to each driven connection scaled by a 'squashing' function (*f*) with a sigmoidal shape, typically the function $f = 1/(1 + e^{-x})$, where $x = \Sigma$inputs.

For the training of the ANN, each input (i.e. normalized pyrolysis mass spectrum) is paired with a desired output (i.e. the amount of *E. coli* UB5201); together these are called a training pair (or training pattern). An ANN is trained over a number of training pairs; this group is called a training set. The input is applied to the network, which is allowed to run until an output is produced at each output node. The difference between the actual and the desired output, taken over the entire training set, is 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. Learning rate scales the magnitude of the step down the error surface taken after each complete calculation in the network (epoch), and momentum acts like a low-pass filter, smoothing out progress over small bumps in the error surface by remembering the previous weight change.

The structure of the ANNs used in this study to analyse pyrolysis mass spectra consisted of three layers containing 159 nodes made up of 150 input nodes (normalized pyrolysis mass spectra), one output (% *E. coli* UB5201) and one 'hidden' layer containing eight nodes (i.e. a 150-8-1 architecture). Each of the 150 nodes was connected to the eight nodes of the hidden layer which in turn were connected to the bias, making a total of 1217 connections, whose weights were altered during the training. Before training commenced the connection weights were set to small random values (Wasserman 1989).

RESULTS AND DISCUSSION

Degree of relatedness between the two *E. coli* strains

Typical PyMS spectra for 100% *E. coli* UB5201, 100% *E. coli* HB101 and a 20:80 mixture of the two strains are shown in Fig. 1. There was little qualitative difference between the spectra but, on closer inspection, quantitative differences may be observed. Such spectra readily illustrate the need to employ multivariate statistical techniques in the analysis of PyMS data.

After collection of the pyrolysis mass spectra from all 15 strains (Table 1) the first stage was to observe the relatedness between all these strains using CVA (Fig. 2a). It can be seen that all the *E. coli* strains and the two *Klebsiella* species (Enterobacteriaceae group) were recovered together into one tight cluster. The single *Streptococcus faecalis* strain clusters away from the other Gram-positive *Bacillus* species.

To observe finer discriminations, the relationship between the different Enterobacteriaceae was investigated further by performing multivariate analysis on these strains only; the resulting HCA can be seen in Fig. 2b. This dendrogram shows that there was clear differentiation between all the strains, that *E. coli* HB101 and *E. coli* UB5201 were very similar and were recovered together at a level of 90·8% relative similarity, and that this cluster showed <37% relative similarity to the other six isolates.

These results show that *E. coli* UB5201 and *E. coli* HB101 are phenotypically very closely related. The next stage was to ascertain whether PyMS with ANNs was able to quantify *E. coli* UB5201 in various mixtures with *E. coli* HB101.

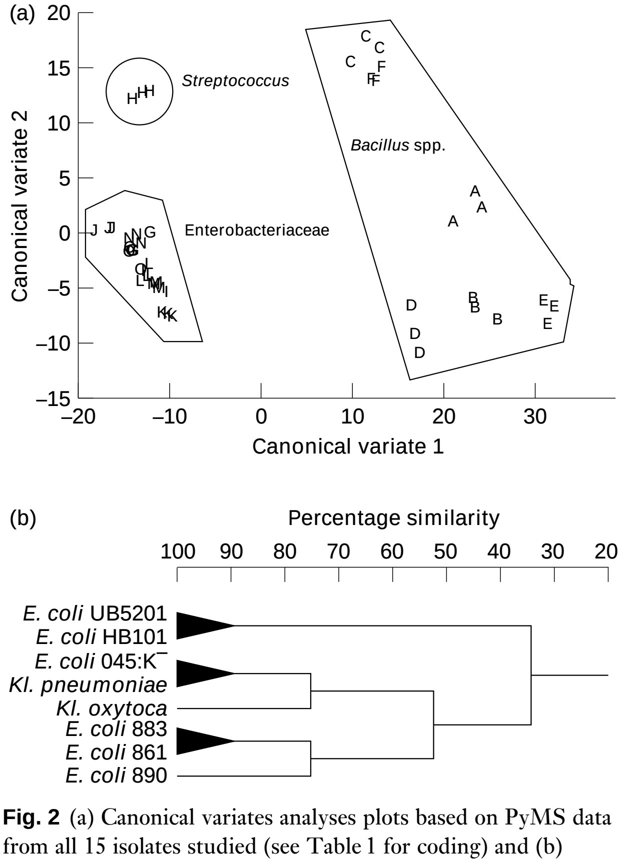

Fig. 2 (a) Canonical variates analyses plots based on PyMS data from all 15 isolates studied (see Table 1 for coding) and (b) dendrogram based on PyMS data from only *Escherichia coli* and *Klebsiella* strains

Quantification of *E. coli* UB5201 in *E. coli* HB101

Binary mixture set A. Table 2 shows that PyMS data from set A were split into two sets; the training set contained the normalized quadruplicate ion intensities from the pyrolysis mass spectra of 0, 10, . . . , 90 and 100% *E. coli* UB5201 in *E. coli* HB101, whilst the test set contained the normalized quadruplicate ion intensities from the 'unknown' pyrolysis mass spectra (5, 15, . . . , 85 and 95% *E. coli* UB5201 in *E. coli* HB101).

ANNs were then trained using the standard back propagation algorithm with the normalized quadruplicate PyMS data from the training set as the input and the percentage *E. coli* UB5201 mixed with *E. coli* HB101 as the output. As shown in Table 2, the output layer was scaled between −50 and 150. This range was chosen since it has been shown previously (Goodacre *et al.* 1993) that increasing the scaling on the output layer to quantify binary mixtures increases the accuracy of the network's predictions because it minimizes the influence of the sigmoidal activation function used to squash the signal passed through the output layer; thus a ± 50% headroom was used. It has also been shown (Goodacre *et al.* 1993) that altering the scaling on the input nodes has no effect on improving the ability of the network to generalize. Therefore as can be seen from Table 2, the input layer was scaled across the whole mass range such that the minimum was set to 0 and the maximum to +1. The effectiveness of training was expressed in terms of the plots of absolute error between the actual and the desired network outputs *vs* the number of epochs (interactions); this 'learning curve' is shown in Fig. 3. The learning curve of the test data (closed circles) is also shown in Fig. 3; it can be seen that whereas the training set error continues to decrease during training, the test set's learning curve initially decreases for $\approx 10^4$ epochs, then reaches a plateau which is followed by an increase at 8×10^4 epochs. This indicates that the ANN has been overtrained, and it is important not to overtrain ANNs since (by definition) the network will not generalize well (Goodacre and Kell 1993).

The network was therefore stopped at the optimum point (7×10^4 epochs) and interrogated with both the training and test sets. A plot of the network's estimate *vs* the true percentage of *E. coli* UB5201 (Fig. 4) showed that the network's estimates were accurate for each percentage, except for the estimates of 15% *E. coli* UB5201. PCA on these spectral data (Fig. 5) revealed that two of the four replicate 15% *E. coli* UB5201 samples were outliers. This indicates that rather than a failure of the ANNs, the inaccuracies in these two samples were due to experimental error in the preparation of the 15% *E. coli* UB5201 mixtures.

Five ANNs were then trained for 7×10^4 epochs and each ANN was interrogated with the training and test sets but with the two outlying 15% samples omitted. The mean network

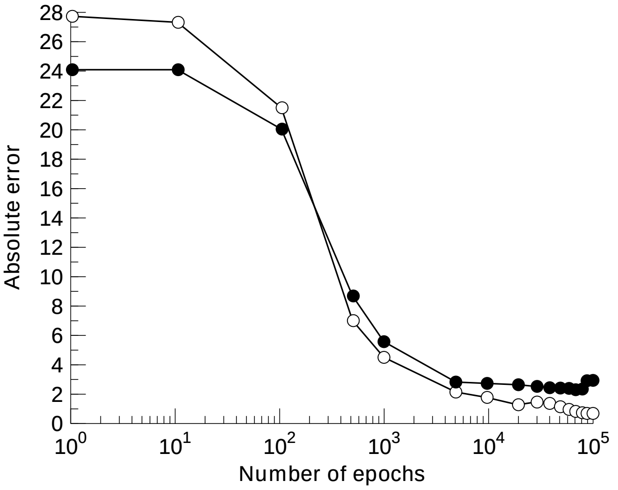

Fig. 3 Learning curves for the artificial neural networks (ANNs) trained to estimate the amount of *Escherichia coli* UB5201 in mixtures containing 0–100% *E. coli* UB5201 in 100–0% *E. coli* HB101. The standard back propagation algorithm was used with one hidden layer consisting of eight nodes. The open circles represent the absolute error of the data used to train the ANN while the closed circles represent the data from the test set. Training was taken to be finished at 7 × 10^4 epochs

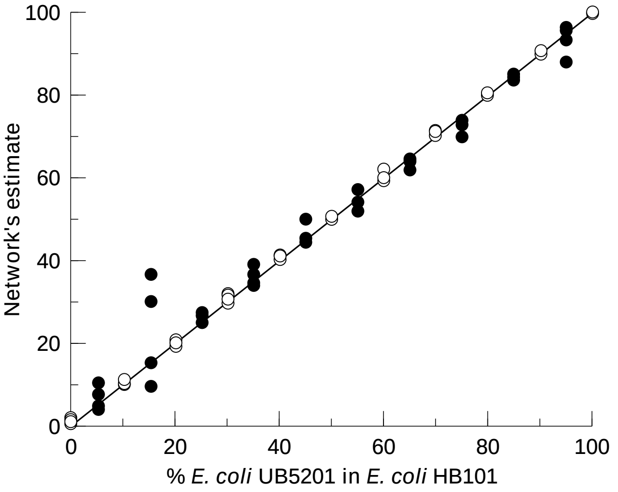

Fig. 4 Estimates of a trained 150-8-1 neural network against the true percentage of *Escherichia coli* UB5201 in mixtures of 0–100% *E. coli* UB5201 in *E. coli* HB101. Artificial neural networks (ANNs) were trained using the standard back propagation algorithm for 7 × 10^4 epochs. Data points from quadruplicate pyrolysis mass spectra are shown. The expected proportional fit is shown. ○, Results of seen data (training set); ●, results of unseen data (test set)

estimate *vs* the true percentage of *E. coli* UB5201 in *E. coli* HB101 was the same as that detailed in Fig. 4 (with the exception of the two outliers of course). Table 2 shows that the average absolute error for the seen and unseen data was 0·7 and 2·05, respectively, and the average standard deviations (σ) were 0·14 and 0·05, respectively. It was therefore evident that each of the five network's estimates of the percentage *E. coli* UB5201 was similar to the true quantity, both for spectra that were used for the training set and more importantly for the 'unknown' pyrolysis mass spectra. It is also clear that training was executed in a reproducible manner as is indicated by the very small error bars in Fig. 4.

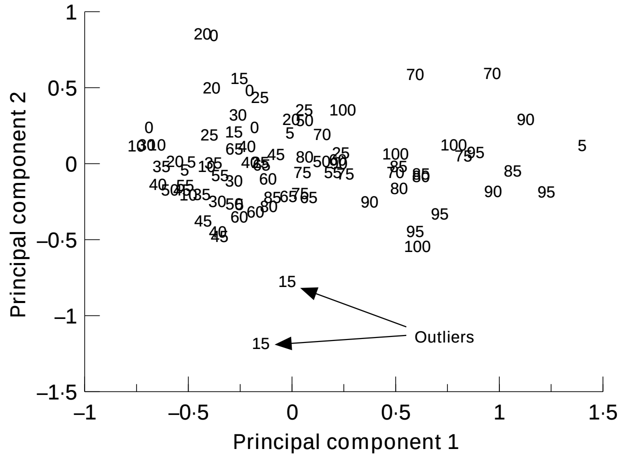

Fig. 5 Principal component analyses plot based on PyMS data of binary mixtures of 0–100% *Escherichia coli* UB5201 in 100–0% *E. coli* HB101. Two outliers have been highlighted which are of mixtures of 15% *E. coli* UB5201 in 85% *E. coli* HB101

Binary mixture set B. Mixtures were prepared over the range 0–20% *E. coli* UB5201 (Table 2). The training set contained the normalized triplicate ion intensities from the pyrolysis mass spectra of 0, 2, . . . , 18 and 20% *E. coli* UB5201 in *E. coli* HB101, and the test set contained the normalized triplicate ion intensities from the 'unknown' pyrolysis mass spectra (1, 3, . . . , 17 and 19% *E. coli* UB5201 in *E. coli* HB101).

As with set A, ANNs were trained using the standard back propagation algorithm with the normalized triplicate PyMS data as the input and the percentage *E. coli* UB5201 mixed with *E. coli* HB101 as the output. The input layer was again scaled across the whole mass range but this time the output layer was scaled between −10 and 30 (so as to give the same ±50% headroom). The 150-8-1 neural network was optimally trained five times (i.e. trained to give the best generalization as judged by test set cross validation) to an average value of 0·33 absolute error in the training set; training typically took 8 × 10^4 epochs.

After interrogating each of the five ANNs, a plot of the mean network estimate *vs* the true percentage of *E. coli* UB5201 (Fig. 6) revealed a proportional fit. Table 2 shows that the average absolute error for the unseen data was 0·95 (standard deviation was 0·09). It was therefore evident that

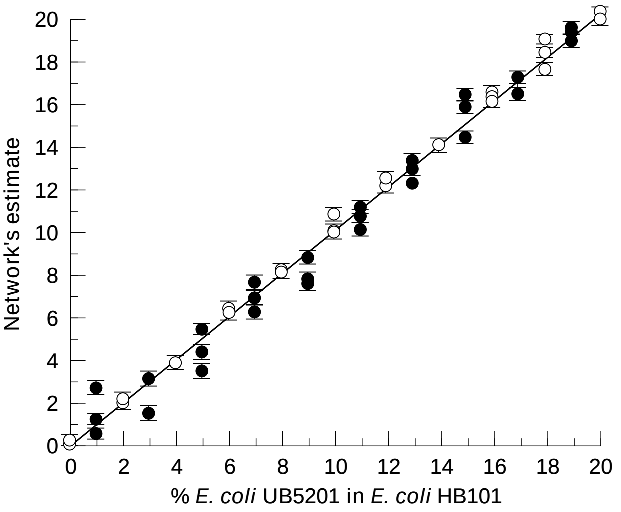

Fig. 6 Mean estimates of five trained 150-8-1 neural networks against the true percentage of *Escherichia coli* UB5201 in mixtures of 0–20% *E. coli* UB5201 in *E. coli* HB101. ANNs were trained using the standard back propagation algorithm for 8 × 10^4 epochs. Data points from triplicate pyrolysis mass spectra are shown. Error bars show the standard deviations over the five runs. The expected proportional fit is also shown. ○, Mean results of seen data (training set); ●, mean results of unseen data (test set)

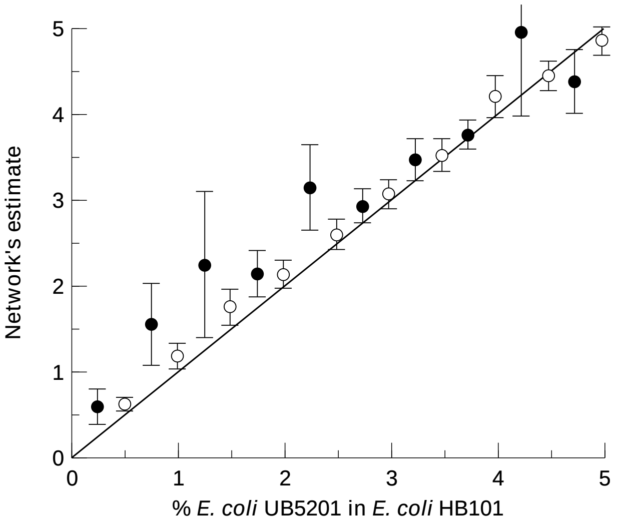

Fig. 7 Mean estimates of five trained 150-8-1 neural networks against the true percentage of *Escherichia coli* UB5201 in mixtures of 0–5% *E. coli* UB5201 in *E. coli* HB101. ANNs were trained using the standard back propagation algorithm for 2 × 10^4 epochs. Data points are the averages of triplicate estimates (triplicate data averaged after training five times). Error bars show the standard deviations over the five runs. The expected proportional fit is shown. ○, Mean results of seen data (training set); ●, mean results of unseen data (test set)

ANN analysis of PyMS data from mixtures containing between 0 and 20% *E. coli* UB5201 in 80 to 100% *E. coli* HB101 yielded accurate estimates of the amounts of *E. coli* UB5201.

Binary mixture set C. The training set of binary mixture set C contained the normalized triplicate ion intensities from the pyrolysis mass spectra of 0, 0·5, . . . , 4·5 and 5·0% *E. coli* UB5201 in *E. coli* HB101, while the test set contained the normalized triplicate ion intensities from the 'unknown' pyrolysis mass spectra (0·25, 0·75, 4·25 and 4·75% *E. coli* UB5201 in *E. coli* HB101). Five 150-8-1 ANNs were trained using test set cross validation to the optimal point; this took 2 × 10^4 epochs. The ANNs were then interrogated and the mean absolute error for the training and test sets were calculated. When the mean network estimates (with triplicate estimates averaged after training) were plotted against the actual percentage of *E. coli* UB5201 (Fig. 7) it was observed that the average estimates for the test set were rather inaccurate, although a general pattern of increase in estimation with increase in actual percentage did occur; this implies that at least some of the mass spectral features, attributing to the level of *E. coli* UB5201, were extracted. Larger error bars indicated that training of these ANNs was not as reproducible as those seen above for the analysis of sets A and B.

Table 3 shows that the absolute error for the training and test sets were 0·28 and 1·08, respectively. To attempt to reduce these error values, ANNs were trained using the *averaged* ion intensities (rather than the triplicates). Five ANNs were trained for 2 × 10^4 epochs and each ANN was then interrogated with the averaged training and test sets. The mean network estimates *vs* the true percentage of *E. coli* UB5201 in *E. coli* HB101 were plotted (data not shown). The network's estimates of percentages greater than 3% were similar to the true quantity, both for the test and training sets. Table 3 shows that using averaged ion intensities to train the neural network causes a reduction in the absolute error of the training and test sets from 0·28 and 1·08 to 0·17 and 0·47, respectively.

For percentages lower than 3%, the ANNs' estimations for the 'unknown' spectra were rather inaccurate. This result indicates that the limit of detection of the PyMS machine for *E. coli* UB5201 in binary mixtures with *E. coli* HB101 is 3%. That is to say, it is likely that below 3% the signal from *E. coli* UB5201 is at the same level as the noise associated with the mass spectra, and so difficult to extract.

To further reduce the effects of noise in the PyMS spectra, these data were reduced by PCA; PCA is an excellent method to reduce the dimensionality of the data, whilst preserving the variance. After the first few PCs, the axes generated will usually be due to random 'noise' in the data; these PCs can be ignored without reducing the amount of useful information

Table 3 Variation in the ability of artificial neural networks* (ANN) to estimate the percentage *Escherichia coli* UB5201 in mixtures containing 0–5% *E. coli* UB5201 and 95–100% *E. coli* HB101

Data type	Triplicate data averaged *before* training	ANN Architecture	Mean absolute training set error	Mean absolute test set error
Pyrolysis mass sepctra	No	150-8-1	0·28	1·08
Pyrolysis mass spectra	Yes	150-8-1	0·17	0·47
Principal components	No	13-3-1	0·19	0·65
Principal components	Yes	13-3-1	0·001	0·51

* Each ANN was trained for 2×10^4 epochs with the input layer scaled across the whole mass range and the output layer scaled from −2·5 to 7·5.

representing the data, since each PC is now independent of (uncorrelated with) any other PC. Therefore in other studies, ANNs were set up using PCs to represent the pyrolysis mass spectra of set C. The training set again contained triplicate PCs representing the pyrolysis mass spectra of 0, 0·5, . . . , 4·5 and 5·0% *E. coli* UB5201, while the test set contained triplicate PCs representing the 'unknown' pyrolysis mass spectra (0·25, 1·75, . . . , 4·25 and 4·75% *E. coli* UB5201 in *E. coli* HB101). To determine the optimum number of PCs which would represent a spectrum, a number of ANNs was trained using the standard back propagation algorithm with an increasing number of PCs (from 1 to 50) as the input layer and the percentage *E. coli* UB5201 mixed with *E. coli* HB101 as the output; furthermore all ANNs used three nodes in the hidden layer.

The absolute error between the true amount of *E. coli* UB5201 and the predicted amount for both the training and the test set was calculated and plotted against the number of PCs used in the input layer (Fig. 8). It can be seen that there was a significant decrease in error in both the training and test set when greater than two PCs are used to represent a spectrum. Using two PCs (85·6% of the variance) to represent a spectrum produced an absolute test set error of 1·05 while using three PCs (92·4% of the variance) gave a test set error of 0·78. It can also be seen that the error in the training set continued to decrease with increasing number of PCs but that the lowest value of absolute error for the test set, indicating optimal calibration, was formed using 13 PCs. The error in the training set at this point was 0·20, whilst the test set was 0·68. Using more than 13 PCs causes overfitting, i.e. inaccurate predictions on the test data.

Five 13-3-1 ANNs were therefore trained using test set cross validation with the input layer (13 triplicate PCs) scaled across the whole mass range and the output layer scaled between −2·5 and 7·5; these ANNs were trained for ≈2×10^4 epochs. The ANNs were interrogated and the mean absolute error for the training and test sets calculated; 0·19 and 0·65, respectively. When the mean network estimates (with triplicate estimates averaged after training) were plotted against the actual percentage of *E. coli* UB5201 (data not shown) it was observed that the average estimates for the test set were not very accurate although again a general pattern of increase in estimation with increase in actual percentage did occur. Small error bars (data not shown) indicated that training of these ANNs was reproducible.

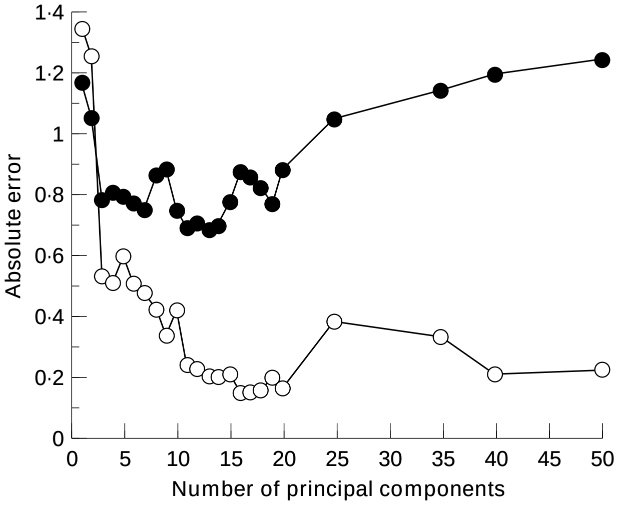

Fig. 8 Variation in the absolute error of the training (○) and test (●) sets for mixtures of 0–5% *Escherichia coli* UB5201 in *E. coli* HB101 as the number of PCs used to represent the spectral data was increased. At 13 PCs the absolute test set error was at its lowest, 0·68

A final study was performed in which five ANNs were trained for 2×10^4 epochs using 13 *averaged* PCs (rather than triplicates) to represent each spectrum. Each ANN was then interrogated with the averaged training and test sets and the mean network estimates *vs* the true percentage of *E. coli* UB5201 in *E. coli* HB101 were plotted (Fig. 9). It was observed that although the percentage *E. coli* UB5201 in the test set was not as accurately estimated for concentrations $>3\%$, compared with using raw mass spectra to train ANNs the limit of detection was significantly lower; 1% *E. coli* UB5201 in *E. coli* HB101. Table 3 shows that the absolute error for the training and the test sets were 0·001 and 0·51, respectively.

Plate counts on the original 40 mg ml^{-1} bacterial slurries showed that there were $3{\cdot}3 \times 10^9$ *E. coli* HB101 cells ml^{-1} while the *E. coli* UB5201 slurry contained 6×10^8 cells ml^{-1}. Therefore in terms of culturable bacterial numbers (since if present non-culturable cells will also contribute to the mass spectrum) the limit of detection for this experiment for binary strain mixtures was 3×10^4 *E. coli* UB5201 cells in $1{\cdot}6 \times 10^7$ *E. coli* HB101 cells (Table 4).

CONCLUSIONS

PyMS was used to assess the relatedness between 15 bacteria from four different genera. CVA showed that the Enterobacteriaceae grouped together and away from the five *Bacillus* spp. studied. On closer examination of *Kl. pneumoniae*, *Kl. oxytoca* and the six *E. coli* strains it was found that *E. coli* UB5201 and HB101 were very closely related.

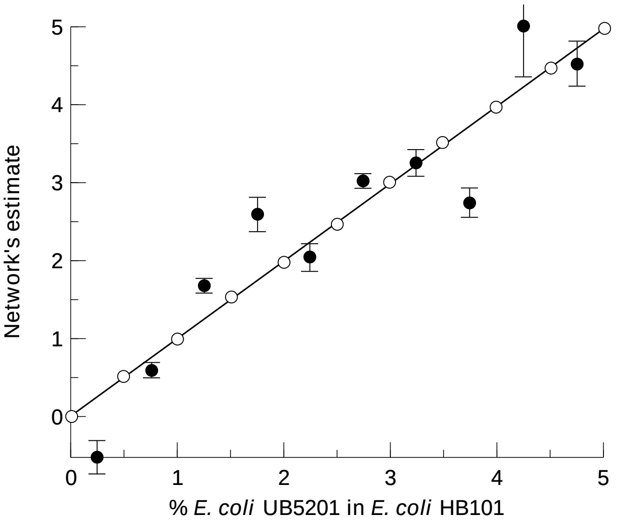

Fig. 9 Mean estimates of five trained 13-3-1 neural networks against the true percentage of *Escherichia coli* UB5201 in mixtures of 0–5% *E. coli* UB5201 in *E. coli* HB101. ANNs were trained using the standard back propagation algorithm for 2×10^4 epochs with the averages of 13 triplicate principal components which were used to represent the spectral data. Error bars show the standard deviations over the five runs. The expected proportional fit is also shown. ○, Mean results of seen data (training set); ●, mean results of unseen data (test set)

Table 4 Total number of bacterial cells per 5 μl samples with their corresponding percentages

% *Escherichia coli* HB101	% *E. coli* UB5201	No. of *E. coli* HB101 cells per 5 μl	No. of *E. coli* UB5201 cells per 5 μl
100	0	$1{\cdot}65 \times 10^7$	0
0	100	0	$3{\cdot}0 \times 10^6$
99	1	$1{\cdot}63 \times 10^7$	$3{\cdot}0 \times 10^4$

Initial experiments using ANNs to quantify mixtures from 0% to 100% *E. coli* UB5201 with HB101 were very successful with the exception of 15% UB5201 in 85% HB101. Rather than a failure of the ANNs, PCA showed that two of the 15% replicates had become contaminated and on PCA plots (Fig. 5) could be seen as outliers. This is significant because it highlights that ANNs, as well as other methods which exploit supervised learning algorithms, are sensitive to badly chosen data. Therefore outliers should be removed from both training and test sets before calibration commences.

Using pyrolysis mass spectra as the input to ANNs showed that the limit of detection was 3% *E. coli* UB5201 in *E. coli* HB101. In an attempt to reduce the influence of random noise in these spectra which hides the small signal due to *E. coli* UB5201, PCA was used to reduce the random noise by only using the first few PCs as inputs to other ANNs; the optimum number of PCs was found to be 13. ANNs were trained with the first 13 PCs after averaging and this increased the resolution so that the limit of detection was now only 1%; this equates to 3×10^4 *E. coli* UB5201 cells in $1{\cdot}6 \times 10^7$ *E. coli* HB101 cells.

It is plausible that the approach detailed here for quantifying low levels of bacteria in mixed populations could be used to detect the contamination of fermentor broths with small numbers of bacteria or fungi. Whilst we appreciate that such contamination may not be with a single micro-organism as demonstrated here, we have previously shown that it is possible to measure the concentrations of tertiary mixtures of cells of the bacteria *Bacillus subtilis*, *E. coli* and *Staph. aureus* (Goodacre *et al.* 1994b). With regard to the clinical laboratory we feel that this technique could be exploited for determining microbial contamination in fermentations or bacterial levels in medically important biofluids such as urine.

We conclude that the combination of PyMS and ANNs constitutes a powerful, accurate and precise methodology for the rapid quantification of mixtures containing very closely related strains of bacteria and would be applicable to assessing bacterial concentrations in medical and biotechnological samples.

ACKNOWLEDGEMENT

The authors are indebted to the Wellcome Trust for financial support (grant number 042615/Z/94/Z).

REFERENCES

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

Bevan, P., Ryder, H. and Shaw, A. (1995) Identifying small-molecule lead compounds – the screening approach to drug discovery. *Trends in Biotechnology* **13**, 115–121.

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

Brereton, R.G. (1992) *Multivariate Pattern Recognition in Chemometrics*. Amsterdam: Elsevier.

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

Chatfield, C. and Collins, A.J. (1980) *Introduction to Multivariate Analysis*. London: Chapman & Hall.

Collins, C.H., Lyne, P.M. and Grange, J.M. (1970) Counting micro-organisms. In *Microbial Methods 6th edn* ed. Collins, C.H., Lyne, P.M. and Grange, J.M. pp. 127–140. London: Butterworths University Park Press.

Crueger, W. and Crueger, A. (1989) *Biotechnology: A Textbook of Industrial Microbiology*. Sunderland, MA: Sinauer Associates Inc.

de la Cruz, F. and Grinsted, J. (1982) Genetic and molecular characterization of Tn*21*, a multiple resistant transposon from R100.1. *Journal of Bacteriology* **151**, 222–228.

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

Flury, B. and Riedwyl, H. (1988) *Multivariate Statistics: A Practical Approach*. London: Chapman & Hall.

Goodacre, R. and Kell, D.B. (1993) Rapid and quantitative analysis of bioprocesses using pyrolysis mass spectrometry and neural networks – application to indole production. *Analytica Chimica Acta* **279**, 17–26.

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

Goodacre, R., Edmonds, A.N. and Kell, D.B. (1993) Quantitative analysis of the pyrolysis mass spectra of complex mixtures using artificial neural networks – application to amino acids in glycogen. *Journal of Analytical and Applied Pyrolysis* **26**, 93–114.

Goodacre, R., Karim, A., Kaderbhai, M.A. and Kell, D.B. (1994a) Rapid and quantitative analysis of recombinant protein expression using pyrolysis mass spectrometry and artificial neural networks – application to mammalian cytochrome B5 in *Escherichia coli*. *Journal of Biotechnology* **34**, 185–193.

Goodacre, R., Neal, M.J. and Kell, D.B. (1994b) 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. (1994c) 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., Trew, S., Wrigley-Jones, C., Neal, M.J., Maddock, J., Ottley, T.W. *et al.* (1994d) Rapid screening for metabolite overproduction in fermentor broths using pyrolysis mass spectrometry with multivariate calibration and artificial neural networks. *Biotechnology and Bioengineering* **44**, 1205–1216.

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.

Goodacre, R., Neal, M.J. and Kell, D.B. (1996) Quantitative analysis of multivariate data using artificial neural networks: a tutorial review and applications to the deconvolution of pyrolysis mass spectra. *Zentralblatt für Bakteriologie* **284**, 516–539.

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

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

Gutteridge, C.S., Vallis, L. and MacFie, H.J.H. (1985) Numerical methods in the classification of microorganisms by pyrolysis mass spectrometry. In *Computer-Assisted Bacterial Systematics* ed. Goodfellow, M., Jones, D. and Priest, F. pp. 369–401. London: Academic Press.

Haaland, D.M. and Thomas, E.V. (1988) Partial least squares methods for spectral analyses.1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry* **60**, 1193–1202.

Irwin, W.J. (1982) *Analytical Pyrolysis: A Comprehensive Guide*. New York: Marcel Dekker.

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*. London: Academic Press.

Maniatis, T., Fritsch, F. and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Martens, H. and Næs, T. (1989) *Multivariate Calibration*. Chichester: John Wiley.

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

Nelder, J.A. (1979) *Genstat Reference Manual*. University of Edinburgh: Scientific and Social Service Program Library.

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

Tanaka, Y. and Omura, S. (1993) Agroactive compounds of microbial origin. *Annual Reviews of Microbiology* **47**, 57–87.

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.

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.

Zupan, J. and Gasteiger, J. (1993) *Neural Networks for Chemists: An Introduction*. Weinheim: VCH Verlagsgeesellschaft.