ELSEVIER

FEMS Microbiology Letters 140 (1996) 233–239

FEMS MICROBIOLOGY LETTERS

Rapid identification of *Streptococcus* and *Enterococcus* species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks

Royston Goodacre [a,*], Éadaoin M. Timmins [a], Paul J. Rooney [b], Jem J. Rowland [c], Douglas B. Kell [a]

[a] *Institute of Biological Sciences, University of Wales, Aberystwyth, Dyfed SY23 3DA, UK*
[b] *Ysbyty Cyffredinol Bronglais (Bronglais General Hospital), Aberystwyth, Dyfed SY23 1ER, UK*
[c] *Department of Computer Sciences, University of Wales, Aberystwyth, Dyfed SY23 3DB, UK*

Received 19 April 1996; revised 6 May 1996; accepted 9 May 1996

Abstract

Diffuse reflectance-absorbance Fourier transform infrared spectroscopy (FT-IR) was used to analyse 19 hospital isolates which had been identified by conventional means to one of *Enterococcus faecalis*, *E. faecium*, *Streptococcus bovis*, *S. mitis*, *S. pneumoniae*, or *S. pyogenes*. Principal components analysis of the FT-IR spectra showed that this 'unsupervised' learning method failed to form six separable clusters (one for each species) and thus could not be used to identify these bacteria based on their FT-IR spectra. By contrast, artificial neural networks (ANNs) could be trained by 'supervised' learning (using the back-propagation algorithm) with the principal components scores of derivatised spectra to recognise the strains from their FT-IR spectra. These results demonstrate that the combination of FT-IR and ANNs provides a rapid, novel and accurate bacterial identification technique.

Keywords: Artificial neural network; Chemometrics; Fourier transform infrared spectroscopy (FT-IR); *Streptococcus*

1. Introduction

The ideal method for the examination of the relationships between bacterial strains would have minimum sample preparation, would analyse samples directly (i.e. would not require reagents), would be rapid, automated, non-invasive, quantitative and (at least relatively) inexpensive [1]. These requirements indicate a spectroscopic solution, the commonest such approach being pyrolysis mass spectrometry (PyMS) [2,3]. However, this is but one of several physico-chemical methods for microbial identification, often referred to as whole-organism fingerprinting and which include UV resonance Raman spectroscopy [4], and Fourier transform infrared spectroscopy (FT-IR) [5].

FT-IR allows the chemically-based discrimination of intact microbial cells, without their destruction, and produces complex biochemical fingerprints which are reproducible and distinct for different bacteria. Naumann and co-workers (e.g. [5,6]) have shown that FT-IR absorbance spectroscopy (in the

* Corresponding author. Tel: +44 (1970) 621 947; Fax: +44 (1970) 622 354; E-mail: rrg@aber.ac.uk

PII S0378-1097(96)00186-3

mid-IR range, usually defined as 4000–400 cm^{-1}) provides a powerful tool with sufficient resolving power to distinguish intact microbial cells at the strain level. However, like PyMS, the interpretation of the FT-IR spectra has conventionally been by the application of 'unsupervised' pattern recognition methods of correspondence analysis maps and cluster analysis [6]. With 'unsupervised learning' methods of this sort the relevant multivariate algorithms seek 'clusters' in the data [7], thereby allowing the investigator to group objects together on the basis of their perceived closeness; 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 PyMS data [1]. Arguably, the most significant of these is the application of (artificial) neural networks (ANNs). The first demonstration of the ability of ANNs to discriminate between biological samples from their pyrolysis mass spectra was by Goodacre et al. [8], who successfully used PyMS and ANNs for the assessment of the presence of lower-grade seed oils as adulterants in extra virgin olive oils. Several studies have now shown that the combination of PyMS and ANNs is very effective for the rapid identification of a variety of bacterial strains, as reviewed in [1,3].

Typically, the sample preparation for FT-IR absorbance measurements involves grinding the dried sample to a fine powder and mixing with KBr, although Naumann et al. [6] have replaced this slow and rather tedious method with the application of liquid samples (which are then dried) to one of 16 ZnSe windows on a rotating disc. However, we consider that a much more elegant approach, which is automated and can allow many more samples ($\gg$ 100) to be analysed in one data collection run, is to use reflectance methods. Diffuse reflectance-absorbance can be achieved by applying the sample onto a sand-blasted metal plate which can then be loaded onto a motorised stage of a reflectance TLC accessory [9]. It is noteworthy that such an approach also allows spectra to be obtained as a function of spatial location. Moreover, it has been shown that reflectance methods can be both more sensitive and discriminatory than absorbances [10].

The aim of this study was thus to use diffuse reflectance-absorbance FT-IR to examine a collection of streptococcal and enterococcal hospital isolates. Nineteen isolates were analysed which had been identified by conventional means to belong to one of *Enterococcus faecalis*, *E. faecium*, *Streptococcus bovis*, *S. mitis*, *S. pneumoniae*, or *S. pyogenes*. The FT-IR spectra were then analysed by principal components analysis (PCA) to observe any clusters and to elucidate if PCA could be used to identify these bacteria. Finally, we investigated the ability of artificial neural networks to identify these streptococci from their FT-IR spectra.

2. Materials and methods

2.1. Organisms and cultivation

Nineteen strains were isolated from Ysbyty Bronglais and classified as streptococci or enterococci by conventional means. These means were colonial and microscopic morphology, and biochemical characteristics; Lancefield groupings were elucidated using haemolysis with the Streptex kit (Oxoid). The API STREP kit (BioMérieux, Basingstoke, Hants., UK) was used to identify all of the isolates, with the exception of the pneumococci which were identified using Optochin disks. The 19 strains were thus identified to belong to the following six species (Ysbyty Bronglais identifier given in brackets): *Enterococcus faecalis* (3, 16, 17, 19), *Streptococcus pyogenes* (7, 14, 15, L19), *Streptococcus pneumoniae* (18, 20, H6, L18), *Streptococcus mitis* (5, 8, 11), *Streptococcus bovis* (2, 4), and *Enterococcus faecium* (10, 13).

To remove any effects of variable phenotype all strains were incubated aerobically with 7% CO_2 for 16 h at 37°C on a single batch of Lab M blood agar base, supplemented with 5% horse blood. After growth, biomass was carefully collected in physiological saline (0.9% NaCl) and frozen at −20°C.

2.2. Diffuse reflectance-absorbance FT-IR

Aliquots (20 μl) of the bacterial suspensions were evenly applied onto a sand-blasted aluminium plate (measuring 10 × 10 cm). To reduce the possible

effects on the FT-IR spectra of concentration-dependent phenomena all bacterial slurries were approximately 40 mg ml^{-1}. Prior to analysis the samples were oven-dried at 50°C for 30 min. Samples were run in triplicate. The FT-IR instrument used was the Bruker IFS28 FT-IR spectrometer (Bruker Spectrospin Ltd., Banner Lane, Coventry, UK) equipped with an MCT (mercury-cadmium-telluride) detector cooled with liquid N_2. The aluminium plate was then loaded onto the motorised stage of a reflectance TLC accessory [9,11].

The IBM-compatible PC 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 8 cm^{-1}, whilst the data point spacing in the Fourier transform of the interferogram (after using a zero-filling factor of 2) was 4 cm^{-1}. To improve the signal-to-noise ratio 256 spectra were co-added and averaged. Each sample was 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 and Kubelka-Munk theory [12].

2.3. Pre-processing and exploratory analysis

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 Park Way, Natick, MA), which runs under Microsoft Windows NT on an IBM-compatible PC. To minimize problems arising from baseline shifts, Matlab was used to take the smoothed second derivatives of the original FT-IR spectra using the Savitzky-Golay algorithm [13] using 5-point smoothing. Matlab was also employed to perform principal components analysis (PCA) so that exploratory data analysis could be conducted [14].

2.4. Artificial neural networks

All ANN analyses were carried out with a user-friendly, neural network simulation program,

Table 1
Identify of the bacteria used in the training and test sets as judged by artificial neural networks

Strain	Estimates from artificial neural networks (standard deviation) [a]					
	E. faecalis	*S. pyogenes*	*S. pneumoniae*	*S. mitis*	*S. bovis*	*E. faecium*
Training set:						
E. faecalis 3	**1.00**	0.00	0.00	0.00	0.00	0.00
E. faecalis 19	**0.99**	0.01	0.00	0.00	0.00	0.00
S. pyogenes 7	0.01	**1.00**	0.00	0.01	0.00	0.00
S. pyogenes 15	0.01	**0.98**	0.00	0.01	0.00	0.00
S. pneumoniae 18	0.00	0.00	**1.00**	0.00	0.00	0.00
S. pneumoniae L18	0.00	0.00	**0.98**	0.01	0.00	0.01
S. mitis 8	0.00	0.01	0.00	**0.99**	0.00	0.00
S. mitis 11	0.00	0.01	0.01	**0.97**	0.00	0.01
S. bovis 2	0.00	0.00	0.00	0.00	**0.99**	0.01
E. faecium 10	0.00	0.00	0.02	0.02	0.01	**0.98**
Test set:						
E. faecalis 16	**0.99**	0.01	0.00	0.00	0.00	0.00
E. faecalis 17	**0.87 (0.14)**	0.30 (0.29)	0.00	0.00	0.00	0.00
S. pyogenes 14	0.01	**1.00**	0.00	0.00	0.00	0.00
S. pyogenes L19	0.00	**0.99**	0.00	0.02	0.00	0.00
S. pneumoniae 20	0.00	0.00	**1.00**	0.00	0.00	0.00
S. pneumoniae H6	0.01	0.00	**0.99**	0.00	0.00	0.02
S. mitis 5	0.00	0.09 (0.12)	0.00	**0.99**	0.00	0.00
S. bovis 4	0.03	0.14 (0.16)	0.00	0.00	**0.82 (0.22)**	0.05 (0.06)
E. faecium 13	0.00	0.00	0.03	0.02	0.01	**0.96 (0.08)**

[a] Bold values indicate the winning node. The values given are the averages from training 10 ANNs with different random starting weights. The values in brackets are the standard deviations for the averages, and are only displayed if the standard deviation was greater than 0.05.

NeuFrame version 1,1,0,0 (Neural Computer Sciences, Totton, Hants., UK), which runs under Microsoft Windows NT on an IBM-compatible PC. In-depth descriptions of the modus operandi of this type of ANN analysis are given elsewhere [15–17].

For training the ANNs, each of the inputs were the first five principal components scores from the second derivative of the FT-IR reflectance-absorbance spectra derived from a total of 10 examples of the six different species (details are given in Table 1) and were paired with each of the desired outputs. These were binary encoded such that *E. faecalis* was coded as 100000, *S. pyogenes* as 010000, *S. pneumoniae* as 001000, *S. mitis* as 000100, *S. bovis* as 000010, and *E. faecium* as 000001. These training pairs collectively made up the training set. The structure of the ANN used in this study to analyse the FT-IR spectra therefore consisted of three layers: five input nodes, six output nodes (one for each strain), and one 'hidden' layer containing nine nodes (i.e., a 5-9-6 architecture, and see Fig. 3). For present purposes these 5-9-6 ANNs were trained to a RMS error of 0.01. This process was conducted 10 times so as (a) to observe whether training was reproducible; and (b) to use the 'committee' approach for prediction [18], where the outputs from the 10 5-9-6 ANNs were averaged.

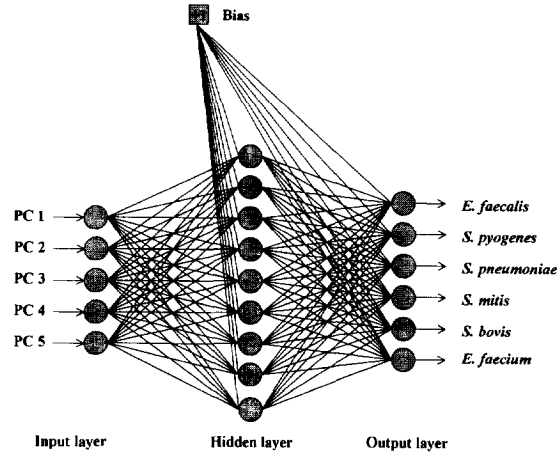

Fig. 1. An artificial neural network consisting of five inputs (one for each of the principal components) and six outputs (one for each type of four streptococci and two enterococci to be identified) connected to each other by one hidden layer consisting of nine nodes. In the architecture shown, adjacent layers of the network are fully interconnected although other architectures are possible.

3. Results and discussion

The three replicate FT-IR spectra for *S. pneumoniae* 18 and *S. pyogenes* 7 are shown in Fig. 1. These and the spectra from the other 17 bacteria all showed broad and complex contours. There was relatively little qualitative difference between all the spectra and such spectra readily illustrate the need to employ multivariate statistical techniques for the analysis of FT-IR data.

When collecting FT-IR spectra it is inevitable that baseline shifts will occur due to differences in the amount of sample which is interrogated by the spectrometer. The triplicate spectra of *S. pyogenes* 7 show this phenomenon where two of the spectra overlap but the third, although similar in form, clearly does not. Therefore to minimize the problems arising from unavoidable baseline shifts the smoothed second derivatives were calculated using the Savitzky-Golay algorithm [13] from the original FT-IR spectra. PCA was then performed on the transformed spectra and the resulting PCA plot, in which the first two principal components (PCs) accounted for 78.6% of the total variation in the data, is shown in Fig. 2.

Fig. 2 shows clearly that PCA cannot be used to cluster these bacteria based on their FT-IR spectra because the six different bacteria do not form distinct groups. Moreover, in Fig. 2 lines are drawn which

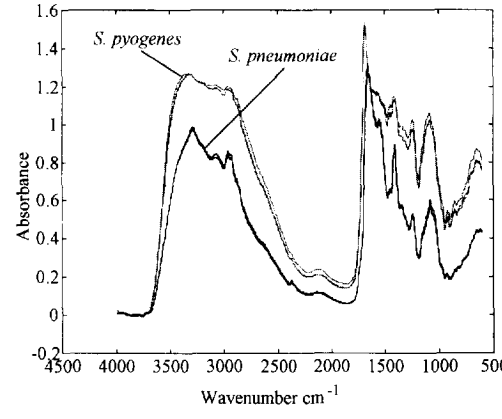

Fig. 2. Typical FT-IR diffuse reflectance-absorbance spectra of *Streptococcus pneumoniae* 18 and *Streptococcus pyogenes* 7. Spectra were obtained as described in the text.

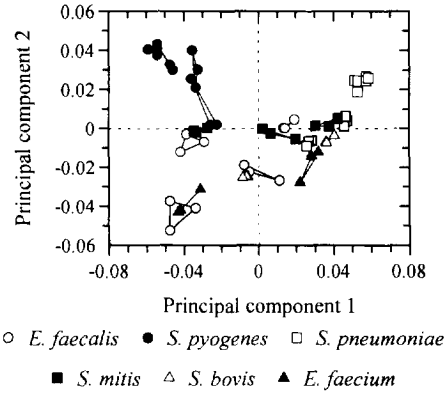

Fig. 3. Principal components plot based on the smoothed (five points) second derivative of the FT-IR data analysed by MATLAB showing the relationship between the 19 hospital isolates. The first two principal components are displayed and account for 55.8% and 22.8% (78.6% total) of the total variance respectively. The lines connect the triplicate spectra and thus form a triangle showing the spread over the replicates.

connect the triplicate spectra and a triangle is constructed which illustrates the reproducibility of the replicate FT-IR spectra. In some instances the replicates of bacteria from one species overlap with replicates from bacteria from different species. When other PCs were plotted (data not shown) these also failed to discriminate between these bacteria, and the replicates also overlapped.

To compare the effect on PCA of taking the second derivative spectrum, PCA was also carried out on the raw normalised FT-IR spectra. The first PC accounted for 91.9% of the total variation compared with 55.8% for the second derivative; it is likely that this difference of 36.1% in PC 1 was due largely to the removal of baseline variations and indeed highlights the point that PCA on raw data detected the overall absorbance as being a significant feature in all the spectra.

The most important conclusion to be drawn from the above PCA analyses on the FT-IR spectra is that this 'unsupervised' learning method failed to classify these bacteria and therefore could not be used to identify them. There is thus a requirement for numerical methods based on 'supervised learning' which allow easy direct interpretation of the identification of bacteria from their FT-IR spectra.

ANNs were trained with the raw normalised FT-IR spectra from the ten bacteria in the training set (see Table 1 for details). The six classes of species were binary encoded at the six output nodes as described above. The 882-10-6 ANN was trained using the standard back-propagation algorithm, and the effectiveness of training was expressed in terms of the average RMS error between the actual and the desired outputs. Training was stopped after the average error had reached 0.01; this typically took between 5×10^3–1×10^4 epochs and because of the large network topology took 5–6 h on a 486-based PC. The network was then interrogated using the FT-IR spectral data from the bacteria from both the training and test sets. Although the training set was correctly identified, these ANNs identified only the two *S. pyogenes* isolates in the test set (nine bacteria in total); it is likely that these ANNs failed because of the baseline shifts described above.

The next stage was to apply the second derivatives to the input nodes of the ANNs. The same training and test sets were employed as above and the 874-10-6 ANNs trained to 0.01 (because five smoothing points were used this meant that the first and last four wavenumber points could not be smoothed and so were removed). Training was also slow (5–6 h) and typically took between 5×10^3–1×10^4 epochs. However, these ANNs performed better and correctly identified four of the nine bacteria in the test set; both *E. faecalis*, one of the *S. pyogenes* and one of the *S. pneumoniae* were identified. Although using the second derivative had removed the baseline shifts it is obvious that this approach was not satisfactory. The training set for this ANN contained only 30 spectra (10 bacteria in triplicate), and it is well known that if the number of parameters, or weights, in the calibration model is significantly higher than the number of exemplars in the training set then over-fitting can more easily occur [18,19]. That the 874-10-6 ANNs employed here contained 8816 weights strongly suggests that this ANN had extracted features due to chance correlations or noise in the derivatized FT-IR spectra.

To obey the parsimony principle as described by Seasholtz and Kowalski [19] the next stage was therefore to reduce the number of inputs to the ANN. As detailed above, PCA is an excellent dimensionality reduction technique; thus the first five PC scores from the second derivatives were used as the input data, since these accounted for 91.3% of the total

variance. The use of PC scores as inputs to neural networks, without deterioration of the calibration model, has previously been applied to the analysis of UV/visible spectroscopic data [20]. The same training and test sets were employed as above and 5-9-6 ANNs trained until the average error had reached 0.01; training typically took between only 350 and 600 epochs, and the actual time taken to train was now only 2–3 min. The reason that more nodes were used in the hidden layer than in the input layer was because PCs are uncorrelated; thus to have enough degrees of freedom more nodes are required. Other ANNs were trained with fewer and more PC applied to the input nodes, whilst keeping nine nodes in the hidden layer failed to generalise sufficiently. 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 output data.

When training had ceased, the network was interrogated and the estimated output for each sample was calculated. As expected, the network's estimate of the bacterial identity of the training set was the same as the known identities (Table 1). The results of the network's final analysis of the unknown test set (given as the average of the outputs for the ten training runs) are shown in Table 1, where it is clear that all nine streptococci were correctly and unequivocally identified. For *E. faecalis* 7 the node for *E. faecalis* scored 0.87 whilst the node for *S. pyogenes* scored 0.3. This strain was identified as *E. faecalis* because the node's activation for *E. faecalis* was significantly greater than for any of the others; moreover, although the *S. pyogenes* node scored 0.3 the standard deviation on the 10 different ANNs was 0.29. Likewise *S. bovis* 4 was also correctly identified because the node for *S. bovis* scored 0.82, standard deviation 0.22; and 0.14, standard deviation 0.16, on the *S. pyogenes* node. This highlights the benefit of training several ANNs and using the committee approach for prediction as discussed by Bishop [18].

This study clearly showed that diffuse reflectance-absorbance FT-IR spectroscopy can be used to obtain biochemical fingerprints from intact microbial cells. PCA could not be used to identify hospital isolates belonging to one of four streptococcal or two enterococcal groups. However, back-propagation neural networks were also trained with the first five PCs from the smoothed second derivative FT-IR spectra to identify these strains successfully. This is the first example of the application of ANNs to the analysis of FT-IR spectra for bacterial identification. We therefore conclude that the combination of FT-IR and ANNs provides an objective, rapid and accurate discriminatory technique.

Acknowledgements

We would like to thank Drs. Bjørn Alsberg, Paul Turner and Andy Woodward for useful discussions. R.G. and E.M.T. thank the Wellcome Trust for financial support (grant number 042615/Z/94/Z). D.B.K. is indebted to the Chemicals and Pharmaceuticals Directorate of the UK BBSRC for financial support.

References

[1] Goodacre, R. and Kell, D.B. (1996) Pyrolysis mass spectrometry and its applications in biotechnology. Cur. Opin. Biotechnol. 7, 20–28.

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

[3] Goodacre, R. (1994) Characterisation and quantification of microbial systems using pyrolysis mass spectrometry: Introducing neural networks to analytical pyrolysis. Microbiol. Eur. 2, 16–22.

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

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

[6] Naumann, D., Helm, D., Labischinski, H. and Giesbrecht, P. (1991) The characterization of microorganisms by Fourier-transform infrared spectroscopy (FT-IR). In: Modern Techniques for Rapid Microbiological Analysis (Nelson, W.H., Ed.), pp. 43–96, VCH Publishers, New York.

[7] Everitt, B.S. (1993) Cluster Analysis. Edward Arnold, London.

[8] Goodacre, R., Kell, D.B. and Bianchi, G. (1992) Neural networks and olive oil. Nature 359, 594–594.

[9] Glauninger, G., Kovar, K.A. and Hoffmann, V. (1990) Possibilities and limits of an online coupling of thin-layer chro-

matography and FTIR spectroscopy. Fresenius J. Anal. Chem. 338, 710–716.

[10] Mitchell, M.B. (1993) Fundamentals and applications of diffuse reflectance infrared fourier transform (DRIFT) spectroscopy. Adv. Chem. Ser. 236, 351–375.

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

[12] Griffiths, P.R. and de Haseth, J.A. (1986) Fourier Transform Infrared Spectrometry. John Wiley, New York.

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

[14] Jolliffe, I.T. (1986) Principal Component Analysis. Springer-Verlag, New York.

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

[16] 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. Anal. Chim. Acta 313, 25–43.

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

[18] Bishop, C.M. (1995) Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

[19] Seasholtz, M.B. and Kowalski, B. (1993) The parsimony principle applied to multivariate calibration. Anal. Chim. Acta 277, 165–177.

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