Rapid and Quantitative Analysis of the Pyrolysis Mass Spectra of Complex Binary and Tertiary Mixtures Using Multivariate Calibration and Artificial Neural Networks

Royston Goodacre,* Mark J. Neal, and Douglas B. Kell
Institute of Biological Sciences, University of Wales, Aberystwyth, Dyfed SY23 3DA, U.K.

Binary mixtures of the protein lysozyme with glycogen, of DNA or RNA in glycogen, and the tertiary mixture of cells of the bacteria *Bacillus subtilis, Escherichia coli,* and *Staphylococcus aureus* were subjected to pyrolysis mass spectrometry. To analyze the pyrolysis mass spectra so as to obtain quantitative information representative of the complex components of the mixtures, partial least-squares regression (PLS), principal components regression (PCR), and fully interconnected feedforward artificial neural networks (ANNs) were studied. In the latter case, the weights were modified using the standard back-propagation algorithm, and the nodes used a sigmoidal squashing function. It was found that each of the methods could be used to provide calibration models which gave excellent predictions for the concentrations of determinands in samples on which they had not been trained. Neural networks were found to provide the most accurate predictions. We also report that scaling the *individual* nodes on the input layer of ANNs significantly decreased the time taken for the ANNs to learn. Removing masses of low intensity, which perhaps mainly contributed noise to the pyrolysis mass spectra, had little effect on the accuracy of the ANN predictions though could dramatically speed up the learning process (by more than 100-fold) and slightly improved the accuracy of PLS calibrations.

Pyrolysis mass spectrometry (PyMS) has been widely applied to the characterization of microbial systems over a number of years (see refs 1–5 for reviews) and, because of its high discriminatory ability,[6] presents a powerful fingerprinting technique, which is applicable to any organic material. Only in the last decade, however, has the chemical basis for any such differences either been sought or found.

From the range of possible chemometric techniques,[7–10] Windig and Meuzelaar[11] successfully used factor and discriminant analyses[12,13] to uncover the concentration of components (expressed in the form of "variance diagrams") from various sets of simulated mixtures (biopolymers, lignites, and grass leaves). The same authors and their colleagues have also extended their factor analysis techniques to the analysis of jet fuels[14] and have implemented self-modeling curve resolution by factor analysis of a continuous series of pyrolysis mass spectra.[15] Factor analysis has also been applied to the extraction of pure component Fourier transform infrared (FTIR) spectra from the spectra of mixtures.[16] More recently, a method using principal components analysis (PCA) to detect pure variables (m/z for mass spectrometry) that have intensity contributions from only one component, called simple-to-use interactive self-modeling mixture analysis (SIMPLISMA) has been successfully applied to the deconvolution of mass spectral data.[17,18]

Kaltenbach and Small[19] successfully demonstrated that the interpretation of FTIR spectra by linear discriminant analysis (LDA) and piecewise linear discriminant analysis (PLDA)(19) was possible, while two other linear multivariate statistical techniques, partial least-squares (PLS) and principal components regression (PCR), have also been widely applied to the analysis of infrared (IR) spectra.[8,20–24] Finally, multiple least-squares regression techniques have been applied to mass spectral data derived from the pyrolysis of the simple biochemical mixture of glycogen, dextran, and serum albumin.[25]

During pyrolysis, intermolecular reactions can take place in the pyrolysate,[26,27] leading to a lack of superposition of the spectral components and to a possible dependence of the mass spectrum on sample size. In this sense it is perhaps surprising

[1] Drucker, D. B. *Methods Microbiol.* **1976,** *9,* 51–125.
[2] Irwin, W. J. *Analytical Pyrolysis: A Comprehensive Guide*; Marcel Dekker: New York, 1982.
[3] Meuzelaar, H. L. C.; Haverkamp, J.; Hileman, F. D. *Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials*; Elsevier: Amsterdam, 1982.
[4] Gutteridge, C. S. *Methods Microbiol.* **1987,** *19,* 227–272.
[5] Berkeley, R. C. W.; Goodacre, R.; Helyer, R. J.; Kelley, T. *Lab. Pract.* **1990,** *39,* 81–83.
[6] Goodacre, R.; Berkeley, R. C. W. *FEMS Microbiol. Lett.* **1990,** *71,* 133–138.
[7] Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte. Y.; Kaufmann, L. *Chemometrics: A textbook*; Elsevier: Amsterdam, 1988.
[8] Martens H.; Næs, T. *Multivariate Calibration*; John Wiley and Sons: New York, 1989.
[9] Brereton, R. G. *Chemometrics: Applications of Mathematics and Statistics to Laboratory Systems*; Ellis Horwood: New York, 1990.
[10] Brown, S. D. *Anal. Chem.* **1992,** *64,* 22R–49R.
[11] Windig W.; Meuzelaar, H. L. C. *Anal. Chem.* **1984,** *54,* 2297–2303.
[12] Nie, N. H.; Hull, C. H. G.; Jenkins, J. G.; Steinbrenner, K.; Brent, W. H. *Statistical Package for the Social Sciences*; McGraw-Hill: New York, 1975.
[13] Windig, W.; Kistemaker, P. G.; Haverkamp, J. *J. Anal. Appl. Pyrol.* **1981,** *3,* 199–212.
[14] Windig, W.; McClennen, W. H.; Meuzelaar, H. L. C. *Chemom. Intell. Lab. Syst.* **1987,** *1,* 151–165.
[15] Windig, W.: Jakab, E.; Richards, J. M.; Meuzelaar, H. L. C. *Anal. Chem.* **1987,** *59,* 317–323.
[16] Gillette, P. C.; Lando, J. B.; Koenig, J. K. *Anal. Chem.* **1983,** *55,* 630–633.
[17] Windig, W. *Chemom. Intell. Lab. Syst.* **1992,** *16,* 1–16.
[18] Windig, W.; Heckler, C. E.; Agblevor, F. A.; Evans, R. J. *Chemom. Intell. Lab. Syst.* **1992,** *14,* 195–207.
[19] Kaltenbach, T. F.; Small, G. W. *Anal. Chem.* **1991,** *63,* 936–944.
[20] Heise, H. M.; Marbach, R; Janatsch, G.; Kruse-Jarres, J. D. *Anal. Chem.* **1989,** *61,* 2009–2015.
[21] Arnold, M. A.; Small, G. W. *Anal. Chem.* **1992,** *62,* 1457–1464.
[22] Small, G. W.; Kaltenbach, T. F.; Kroutil, R. T. *Trends Anal. Chem.* **1991,** *10,* 149–155.
[23] Marbach, R.; Heise, H. M. *Trends Anal. Chem.* **1992,** *11,* 270–275.
[24] Faix, O.; Bottcher, J. H. *Holzforschung* **1993,** *47,* 45–49.
[25] Vallis, L. V.; MacFie, H. J.; Gutteridge, C. S. *Anal. Chem.* **1985,** *57,* 704–709.
[26] Van de Meent, D.; de Leeuw, J. W.; Schenck, P. A.; Windig, W.; Haverkamp, J. *J. Anal. Appl. Pyrol.* **1982,** *4,* 133–142.
[27] Schulten, H-R.; Lattimer, R. P. *Mass Spectrosc. Rev.* **1984,** *3,* 231–315.

0003-2700/94/0366-1070$04.50/0

that quantitative analysis of mixtures by PyMS has been possible using *linear* discriminant analyses, such as factor analysis and methods based on principal component analysis. An alternative approach is the use of artificial neural networks (ANNs), which are, by now, a well-known means of uncovering complex, *nonlinear* relationships in multivariate data (vide infra).

To this end, our own aims have been to extend the PyMS technique to the *quantitative* analysis of the chemical constituents of microbial and other samples, and we have therefore sought to apply ANNs to the analysis and interpretation of pyrolysis mass spectra. Thus, we have been able to follow the production of indole in a number of strains of *E. coli* grown on media incorporating various amounts of tryptophan[28] and to estimate the amount of casamino acids in mixtures with glycogen.[29] With regard to classifications and discriminations, we have also exploited the combination of PyMS and ANNs for the rapid and accurate assessment of the presence of lower grade seed oils as adulterants in extra virgin olive oils[30,31] and for the identification of strains of *Propionibacterium* spp.[32]

ANNs can be considered as collections of very simple "computational units" which can take a numerical input and transform it (usually via summation) into an output (see refs 33–49 for excellent introductory surveys, and refs 28, 29, and 50–58 for applications in analytical chemistry). The relevant principle of supervised learning in ANNs is that the ANNs take numerical inputs (the training data) and transform them into "desired" (known, predetermined) outputs. The input and output nodes may be connected to the "external world" and to other nodes within the network. The way in which each node transforms its input depends on the so-called "connection weights" (or "connection strength") and "bias" of the node, which are modifiable. The output of each node to another node or the external world then depends on both its weight strength and bias and on the weighted sum of all its inputs, which are then transformed by a (normally nonlinear) weighting function referred to as its activation function. For present purposes, the great power of neural networks stems from the fact that it is possible to "train" them. Training is effected by continually presenting the networks with the "known" inputs and outputs and modifying the connection weights between the individual nodes and the biases, typically according to some kind of back-propagation algorithm,[33] until the output nodes of the network match the desired outputs to a stated degree of accuracy. The commonest architecture, as considered herein, is a fully interconnected feedforward network. The network, the effectiveness of whose training is usually determined in terms of the root-mean-square (rms) error between the actual and the desired outputs averaged over the training set, may then be exposed to "unknown" inputs and will then "immediately" output the globally optimal best fit to the outputs. If the outputs from the previously unknown inputs are accurate, the trained ANN is said to have generalized.

One reason that this general method is so attractive for the quantitative analysis of PyMS (or other multivariate spectral) data is that it has been shown mathematically [59–63] that a neural network consisting of only one hidden layer, with an arbitrarily large number of nodes, can learn any, arbitrary (and hence nonlinear) mapping of a continuous function to an arbitrary degree of accuracy. In addition, ANNs are widely considered to be relatively robust to noisy data, such as those which may be generated by PyMS.

To determine how well ANNs can be trained using back-propagation for the quantitative analysis of pyrolysis mass spectra, other chemometric methods need to be applied to the same data to give a benchmark. The two methods used for this purpose by McAvoy et al.[53] to deconvolute fluorescence spectra were principal components regression (PCR) and partial least squares (PLS). These authors and others studying multiple least-squares methods as well as the latent variable PCR and PLS methods,[64–66] have concluded that the best technique appears to be PLS; therefore PLS and PCR were used as a comparison in the present study.

PCR and PLS regression techniques are multivariate factor analysis methods[67–68] that are useful when the target matrix

(28) Goodacre, R.; Kell, D. B. *Anal. Chim. Acta* **1993**, *279*, 17–26.
(29) Goodacre, R.; Edmonds A. N.; Kell, D. B. *J. Anal. Appl. Pyrol.* **1993**, *26*, 93–114.
(30) Goodacre, R.; Bianchi, G.; Kell, D. B. *Nature* **1992**, *359*, 594.
(31) Goodacre, R., Kell, D. B., Bianchi, G. *J. Sci. Food Agric.* **1993**, *63*, 297–307.
(32) Goodacre, R., Neal, M. J., Kell, D. B., Greenham, L. W., Noble, W. C.; Harvey, R.G. *J. Appl. Bacteriol.* **1994**, *76*, 124–134.
(33) Rumelhart, D. E.; McClelland, J. L.; and the PDP Research Group. *Parallel Distributed Processing. Experiments in the Microstructure of Cognition*; MIT Press: Cambridge, MA, 1986; Vols. I, II.
(34) Cowan, J. D.; Sharp, D. H. *Q. Rev. Biophys.* **1988**, *21*, 365–427.
(35) McClelland, J. L.; Rumelhart, D.E. *Explorations in Parallel Distributed Processing; A Handbook of Models, Programs and Exercises*; MIT Press: Cambridge, MA, 1988.
(36) Amit, D. J. *Modeling Brain Function; the World of Attractor Neural Networks*; Cambridge University Press: Cambridge, 1989.
(37) Kohonen, T. *Self-Organization and Associative Memory*; Springer: Heidelberg, 1989.
(38) Wasserman, P. D. *Neural Computing: Theory and Practice*; Van Nostrand Reinhold: New York, 1989.
(39) Wasserman, P. D.; Oetzel, R. M. *NeuralSource: the Bibliographic Guide to Artificial Neural Networks*; Van Nostrand Reinhold: New York, 1989.
(40) Aleksander, I.; Morton, H. *An Introduction to Neural Computing*; Chapman & Hall: London, 1990.
(41) Beale, R.; Jackson, T. *Neural Computing: An Introduction*; Adam Hilger: Bristol, 1990.
(42) Eberhart, R. C.; Dobbins, R. W. *Neural Network PC Tools*; Academic Press: London, 1990.
(43) Pao, Y.-H. *Adaptive Pattern Recognition and Neural Networks*; Addison-Wesley: Reading, MA, 1989.
(44) Simpson, P. K. *Artificial Neural Systems*; Pergamon Press: Oxford, 1990.
(45) Hecht-Nielsen, R. *Neurocomputing*; Massachusetts: Addison-Wesley, 1990.
(46) Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Addison-Wesley: Redwood City, 1991.
(47) Carpenter, G. A.; Grossberg, S. *Pattern Recognition by Self-Organizing Neural Networks*; MIT Press: Cambridge, MA, 1991.
(48) Peretto, P. *An Introduction to the Modelling of Neural Networks*; Cambridge University Press: Cambridge, 1992.
(49) Gallant, S. I. *Neural Network Learning*; MIT Press: Cambridge, MA, 1993.
(50) Curry, B.; Rumelhart, D. E. *Tetra. Comput. Methods* **1990**, *3*, 213–237.
(51) Long, J. R.; Mayfield, H. T.; Henley, M. V.; Kromann, P. R. *Anal. Chem.* **1991**, *63*, 1256–1261.
(52) Bos, A.; Bos, M.; van der Linden, W. E. *Anal. Chim. Acta* **1992**, *256*, 133–144.
(53) McAvoy,T.J.; Su, H. T.; Wang, N. S.; He, M. *Biotechnol. Bioeng.* **1992**, *40*, 53–62.
(54) Shadmehr, R.; Angell, D.; Chou, P. B.; Oehrlein, G. S.; Jaffe, R. S. *J. Electrochem. Soc.* **1992**, *139*, 907–914.
(55) Ball, J. W.; Jurs, P. C. *Anal. Chem.* **1993**, *65*, 505–512.
(56) Bruchmann, A.; Götze, H.-J.; Zinn, P. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 59–69.
(57) Richard, D.; Cachet, C.; Cabrol-Bass, D.; Forrest, T. P. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 202–210.
(58) Beebe, K. R.; Blaser, W. W.; Bredeweg, R. A.; Chauvel, J. P.; Harner, R. S.; LaPack, M.; Leugers, A.; Martin, D. P.; Wright, L. G.; Yalvac, E. D. *Anal. Chem.* **1993**, *65*, 199R–216R.
(59) Cybenko, G. *Math. Control Signals Syst.* **1989**, *2*, 303–314.
(60) Funahashi, K. *Neural Networks* **1989**, *2*, 183–192.
(61) Hornik, K.; Stinchcombe M.; White, H. *Neural Networks* **1989**, *2*, 359–368.
(62) Hornik, K.; Stinchcombe M.; White, H. *Neural Networks* **1990**, *3*, 551–560.
(63) White, H. *Neural Networks* **1990**, *3*, 535–549.
(64) Joreskog, K.; Wold, H. *Systems under indirect observation*; North Holland: Amsterdam, 1982
(65) Carey, W.; Beebe, K.; Sanchez, E.; Geladi, P.; Kowalski, B. *Sens. Actuators* **1986**, *9*, 223–234.
(66) Geladi, P.; Kowalski, B. *Anal. Chim. Acta* **1986**, *185*, 1–17.

(here equivalent to the output layer of ANNs) does not contain the full model representation,[69] i.e., there are more variables in the data matrix than in the target matrix. These methods are therefore ideally suited to quantify pyrolysis mass spectra since the number of variables in the data matrix is large, typically 150 (m/z 51–200). As with supervised learning in ANNs both approaches utilize a priori information about the samples.[70] The first stage in PCR is the decomposition of the data (**X**) matrix into latent variables by PCA; then each of the target (**Y**) variables are regressed onto this decomposed **X** matrix. PLS, however, performs a simultaneous and interdependent PCA decomposition in both **X** and **Y** matrices, in such a way that the information in the **Y** matrix is used *directly* as a guide for the optimal decomposition of the **X** matrix, and then performs regression of the latent variables on **Y**. It is considered that PLS usually handles several covarying **Y** variables better than does PCR and is superior for the simultaneous modelling of several *intercorrelated* target variables.[8,67]

In this study, using PyMS, we therefore analyzed mixtures of lysozyme in glycogen, as representative of complex proteins and carbohydrates and the nucleic acids DNA or RNA in glycogen, and exploited ANNs, PLS, and PCR to estimate the amount of the determinands lysozyme, DNA or RNA in "unknown" (i.e., unseen) spectra. Finally, the three approaches were used to estimate the amounts of different bacteria in a tertiary mixture of *Bacillus subtilis*, *Escherichia coli*, and *Staphylococcus aureus*. We conclude that accurate quantitative analysis of the pyrolysis mass spectra of binary and tertiary mixtures is easily possible using all three numerical techniques but that ANNs give more accurate predictions than do either PLS or PCR. We believe that this is the first study to compare directly neural networks and the linear regression techniques of PCR and PLS for the quantification of the pyrolysis mass spectra of complex mixtures. We also found that scaling the *individual* nodes on the input layer of ANNs significantly decreased the time taken for the ANNs to learn quantitatively to analyse the mass spectra from the binary mixtures. Removing masses of low intensity, which perhaps mainly contributed noise to the pyrolysis mass spectra, had little effect on the accuracy of the ANN predictions, though could dramatically speed up the learning process (by more than 100-fold) and slightly improved the accuracy of PLS calibrations.

EXPERIMENTAL SECTION

Preparation of the Binary Mixtures. Binary mixtures were prepared such that 5 μL of a solution contained 0–100 μg of either the determinand lysozyme (from chicken egg white, Sigma), ribonucleic acid (diethylaminoethanol salt type IX from torula yeast, Sigma), or deoxyribonucleic acid (sodium salt from salmon testes, Sigma; in steps of 5 μg) in 20 μg of glycogen (oyster type II, Sigma).

Preparation of the Tertiary Mixtures of Bacteria. The bacteria used were *Staphylococcus aureus* NCTC6571, *Bacillus subtilis* DSM10, and *Escherichia coli* W3110. All strains were grown in 4 L of liquid media (glucose (BDH), 10.0 g; peptone (LabM), 5.0 g; beef extract (LabM), 3.0 g; H_2O, 1 L) for 16 h at 37 °C in a shaker. After growth the cultures were harvested by centrifugation and washed in phosphate buffered saline (PBS). The dry weights of the cells were estimated gravimetrically and used to adjust the weight of the final slurries using PBS to approximately 40 mg/mL. Two sets of mixtures were then made. The training set consisted of x% *S. aureus*, y% *B. subtilis*, and z% *E. coli*, where x, y, and z were varied in units of 10; this set comprised 66 samples (Figure 9A). The second, "unknown" test set consisted of 16 samples whose quantities are shown in Table 1 and Figure 9A.

Table 1. Percentage Amounts of Bacteria Used in Preparing the Unknown Tertiary Mixture Set

sample no.	% *S. aureus*	% *B. subtilis*	% *E. coli*	sample no.	% *S. aureus*	% *B. subtilis*	% *E. coli*
1	25	75	0	9	25	50	25
2	25	0	75	10	50	25	25
3	0	25	75	11	66	17	17
4	75	25	0	12	17	66	17
5	75	0	25	13	17	17	66
6	0	75	25	14	15	42.5	42.5
7	33	33	33	15	42.5	15	42.5
8	25	25	50	16	42.5	42.5	15

Sample Preparation for Pyrolysis Mass Spectrometry. Clean iron–nickel foils (Horizon Instruments Ltd., Ghyll Industrial Estate, Heathfield, E. Sussex, TN21 8BR, U.K.) were inserted, using clean forceps, into clean pyrolysis tubes (Horizon Instruments), so that 6 mm was protruding from the mouth of the tube. Aliquots (5 μL) of the mixtures were evenly applied to the protruding foils. The samples were oven dried at 50 °C for 30 min, then the foils were pushed into the tube using a stainless steel depth gauge so as to lie 10 mm from the mouth of the tube. Finally, viton O-rings (Horizon Instruments) were placed on the tubes. Samples for the binary mixture analysis were replicated four times, samples for tertiary mixtures were replicated three times.

Pyrolysis Mass Spectrometry. The pyrolysis mass spectrometer used in this study was the Horizon Instruments PYMS-200X, as initially described by Aries et al.[71] The sample tube carrying the foil was heated, prior to pyrolysis, at 100 °C for 5 s. Curie-point pyrolysis was at 530 °C for 3 s, with a temperature rise time of 0.5 s. This pyrolysis temperature was chosen because it has been shown[72,73] to give a balance between fragmentation from polysaccharides (carbohydrates) and protein fractions. The pyrolysate then entered a gold-plated expansion chamber heated to 150 °C, whence it diffused down a molecular beam tube to the ionization chamber of the mass spectrometer. To minimize secondary fragmentation of the pyrolysate, the ionization method used was low-voltage electron impact ionization (25 eV). Nonionized molecules were deposited on a cold trap, cooled by liquid nitrogen. The ionized fragments were focused by the electrostatic lens of a set of source electrodes, accelerated, and directed into a quadrupole mass filter. The ions were

[67] Martin, K. A. *Appl. Spectrosc. Rev.* **1992**, *27*, 325–383.

[68] Liang, Y.-Z.; Kvalheim, O. M.; Manne, R. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 235–250.

[69] Malinowski, E. R. *Factor Analysis in Chemistry*; John Wiley & Sons: New York, 1991.

[70] Brereton, R. G. *Multivariate Pattern Recognition in Chemometrics*; Elsevier: Amsterdam, 1992.

[71] Aries, R. E.; Gutteridge C. S; Ottley T. W. *J. Anal. Appl. Pyrol.* **1986**, *9*, 81–98.

[72] Windig, W.; Kistemaker, P. G.; Haverkamp, J., Meuzelaar, H. L. C. *J. Anal. Appl. Pyrol.* **1980**, *2*, 7–18.

[73] Goodacre, R. Ph.D. Thesis, University of Bristol, 1992.

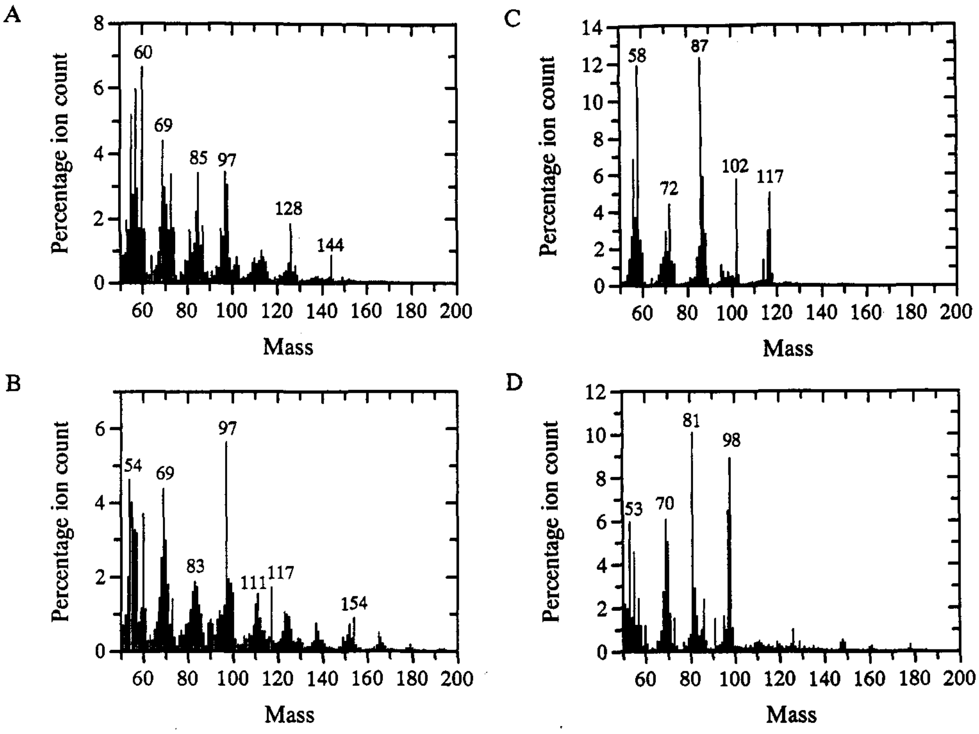

Figure 1. Normalized pyrolysis mass spectra of 20 μg of glycogen (A), 20 μg of lysozyme (B),20 μg of RNA (C), and 20 μg of DNA (D). Experiments were performed exactly as described in the Experimental Section.

separated by the quadrupole, on the basis of their mass-to-charge ratio, and detected and amplified with an electron multiplier. The mass spectrometer scans the ionized pyrolysate 160 times at 0.2-s intervals following pyrolysis. Data were collected over the m/z range 51–200, in one-tenth of a mass unit intervals. These were then integrated to give unit mass. Given that the charge of the fragment was unity the mass-to-charge ratio can be accepted as a measure of the mass of pyrolysate fragments. The IBM-compatible PC used to control the PYMS-200X was also programmed (using software provided by the manufacturers) to record spectral information on ion count for the individual masses scanned and the total ion count for each sample analyzed.

Prior to any analysis, the mass spectrometer was calibrated using the chemical standard perfluorokerosene (Aldrich), such that m/z 181 was one-tenth of m/z 69.

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

For analysis of the pyrolysis mass spectra by ANNs, PLS, or PCR, on the binary mixtures, the training data (ANNs) or the *X* variables (PLS and PCR) were the four normalized replicate pyrolysis mass spectra derived from the mixtures containing 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 μg of determinand, with the output (ANN) or *Y* variables (PLS and PCR) being the actual (true) amount of determinand in the mixtures. In the analysis of the mixtures of the three bacteria the data used to make the models were the normalized triplicate pyrolysis mass spectra derived from mixtures containing x% *S. aureus* (A), y% *B. subtilis* (B), and z% *E. coli* (C), where $x + y + z = 100$ and where the amounts were zero or wholly divisible by 10.

Artificial Neural Networks. All ANN analyses were carried out using a user-friendly, neural network simulation program, NeuralDesk (version 1.2, Neural Computer Sciences, Lulworth Business Centre, Nutwood Way, Totton, Southampton, Hants, SO1 0JR, U.K.), which runs under Microsoft Windows 3.1 on an IBM-compatible PC. To ensure maximum speed, an accelerator board for the PC (NeuSprint) based on the AT&T DSP32C chip, which effects a speed enhancement of some 100-fold, permitting the analysis (and updating) of some 400 000 weights/s, was used. Data were also processed prior to analysis using the Microsoft Excel 4.0 spreadsheet.

The algorithm used was standard back-propagation (BP)[33,74] running on the accelerator board. This algorithm employs processing nodes (neurons or units), connected using abstract interconnections (connections or synapses). Connections each have an associated real value, termed the weight, that scale 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 determinand); together these are called a training pair (or training pattern). An ANN is trained over a number of training pairs; this group is collectively called the training set. 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, we used a learning rate of 0.1 and a momentum of 0.9.

[74] Werbos, P. J. Masters Thesis, Harvard University, 1974.

The structure of the ANN used in this study to analyse pyrolysis mass spectra therefore consisted of three layers containing 159 nodes made up of the 150 input nodes (normalized pyrolysis mass spectra), 1 output node (amount of determinand), and one "hidden" layer containing 8 nodes (i.e., a 150–8-1 architecture). Each of the 150 input nodes was connected to the 8 nodes of the hidden layer which in turn were connected to the output node. In addition, the hidden layer and output node were connected to the bias, making a total of 1217 connections, whose weights will be altered during training. Before training commenced, the values applied to the input and output nodes were normalized between 0 and +1, and the connection weights were set to small random values.[38] In some cases, the output layer was scaled to exploit less than the full range of the normalized scale between 0 and 1.[29] Each epoch represented 1217 connection weight updatings and a recalculation of the root-mean-squared (rms) error between the true and desired outputs over the entire training set. A plot of the rms error vs the number of epochs represents the "learning curve" and was used to estimate the extent of training. Finally after training, all pyrolysis mass spectra of the binary mixtures (0–100 μg) and all mass spectra of the tertiary mixtures (66 knowns and 16 unknowns) were used as the "unknown" inputs (test data); the network then output its estimate (best fit) in terms of the amounts of determinands in the mixtures.

Principal Component Regression and Partial Least-Squares Analyses. All PCR and PLS analyses were carried out using the program Unscrambler II Version 4.0 (CAMO A/S, Olav Tryggvasonsgt. 24, N-7011 Trondheim, Norway; and see ref 8) which runs under Microsoft MS-DOS on an IBM-compatible PC. Data were also processed prior to analysis using the Microsoft Excel 4.0 spreadsheet which runs under Microsoft Windows 3.1 on an IBM-compatible PC.

The first stage is the preparation of the data. This is achieved by presenting the "training set" as two data matrices to the program: **X**, which contains the normalized quadruplicate pyrolysis mass spectra, and **Y**, which represents the amount or percentage content of the determinand(s). For the binary mixtures the **Y** matrix contains one *Y* variable of 0–100 μg (in steps of 10 μg) of lysozyme, RNA or DNA mixed in 20 μg of glycogen (i.e., 44 objects representing 11 quadruplicate concentrations). The **Y** matrix for the analysis of pyrolysis mass spectra from tertiary mixtures contains three variables describing the percentage of *S. aureus*, *B. subtilis*, and *E. coli* (i.e., 198 objects representing 66 triplicates). Unscrambler II also allows the addition of "start noise" (i.e., noise to the **X** data); this option was not used. Finally, the **X** data were scaled in proportion to the reciprocal of their standard deviations.

The next stage is the generation of the calibration model; this first requires the user to specify the appropriate algorithm. The Unscrambler II program has one PCR algorithm and two PLS algorithms; PLS1 which handles only one *Y* variable at a time, and PLS2 which will model several *Y* variables simultaneously.[8] For the analysis of the pyrolysis mass spectra from binary mixtures, predicting only one *Y* variable (amount 0–100 μg of lysozyme, RNA or DNA mixed in 20 μg of glycogen), the PCR and PLS1 algorithms were used. In addition to these two algorithms, we also employed the PLS2 algorithm for the prediction of the percentages of bacteria in the tertiary mixture.

The method of validation used was full cross-validation, via the leave-one-out method. This technique sequentially omits one sample from the calibration; the PCR or PLS model is then redetermined on the basis of this reduced sample set. The amount (micrograms) or percentage of the omitted sample is then predicted with the use of the model. This method is required to determine the optimal size of the calibration model, so as to obtain good estimates of the precision of the multivariate calibration method (i.e., neither to under- nor over-fit predictions of unseen data).[8,75–77] Unscrambler also has reasonably sophisticated outlier detection methods; although these were employed, we did not find it necessary to delete any of the objects from the calibration models formed.

Cross-validation can indicate the optimal number of principal components (PCs) or PLS factors to use in predictions after the model is calibrated. To establish the accuracy of the suggestions produced by Unscrambler, we therefore calculated the rms error between the true and desired concentrations over the entire calibration model, both for the known and unknown mass spectra, and plotted these rms errors vs the number of PCs or PLS factors used in predictions. We also generated plots of the rms error of the test set vs the error in the training set to assist in determining the calibration model that best generalized. Using this approach, after calibration, to choose the optimal number of PCs or PLS factors to use in the prediction, all pyrolysis mass spectra of the binary mixtures (0–100 μg) and all mass spectra of the tertiary mixtures (66 knowns and 16 unknowns) were used as the "unknown" inputs (test data); the model then gave its prediction in terms of the amount or percentage of determinand(s) in the three binary and one tertiary mixture.

RESULTS AND DISCUSSION

Binary Mixture Analyses. Examples of the pyrolysis mass spectra obtained from the analysis of binary mixtures are shown in Figure 1. These materials were chosen to be representative of the various classes of compounds found in biological samples; glycogen is representative of carbohydrates, lysozyme of proteins, and RNA and DNA of nucleic acids. Although the mass spectra of glycogen and lysozyme are fairly complex there are some distinguishing peaks: these were notably m/z 60, 69, 85, 97, 128, and 144 for glycogen (previously seen to be characteristic by Goodacre et al.[29]), and for lysozyme m/z 54, 69, 83, 97, 111, 117, and 154. The relatively less complex pyrolysis mass spectra of RNA (Figure 1C) and DNA (Figure 1D) show a characteristic cyclic shape with the centre of the peaks at m/z 58, 72, 87, 102, and 117 for RNA and m/z 53, 70, 81, and 98 for DNA. It is noteworthy that the pyrolysis mass spectra of RNA shows five major peaks and for DNA only four; it is possible that this was because the fifth mean peak center for DNA was less than m/z 51 and hence was not detected because only the mass range 51–200 was used. The difference (in m/z) between the centers of the major peaks between the mass spectra of DNA and RNA were 19, 17, 21, and 19 m/z (average 19 m/z), this is relatively stable and it seems that the pyrolysis mass spectrum

[75] Haaland, D. M.; Thomas, E. V. *Anal. Chem.* **1988**, *60*, 1193–1202.
[76] Brown, P. *J. Chemom.* **1992**, *6*, 151–161.
[77] Seasholtz, M. B.; Kowalski, B. *Anal. Chim. Acta* **1993**, *277*, 165–177.

of DNA was *in essence* the same as that of RNA but shifted by 19 amu. This may be due to the different nucleic acid salts supplied (RNA as the (diethylamino)ethanol salt and DNA as the sodium salt), which may have led to chemical ionization. However, for present purposes the detailed interpretation of the individual mass spectra is not significant, and will be elaborated elsewhere. It is worthwhile remarking that these pyrolysis mass spectra (Figure 1), especially of RNA (Figure 1C) and lysozyme (Figure 1B), are significantly different from those displayed in the well-known compendium and atlas of Meuzelaar and colleagues.[3] This is because the mass spectrometric conditions used were significantly different; in particular, we used an ionizing voltage of 25 eV, while the spectra in the atlas were obtained with an ionizing voltage of 14 eV; the pyrolysis conditions also differ. In addition, different salts of the RNA were almost certainly used; in this study we used the diethylaminoethanol salt while Meuzelaar et al.[3] used RNA "from bakers yeast" (salt unspecified).

The three binary mixtures were analysed by PyMS. Data from each of the binary mixtures were split into two sets. The training set contained the normalized quadruplicate ion intensities from the pyrolysis mass spectra from 0, 10, 20, ..., 90, and 100 μg of either lysozyme, RNA, or DNA in 20 μg of glycogen, while the test set contained both the training set and the 10 "unknown" pyrolysis mass spectra (5, 15, 25, 35, 45, 55, 65, 75, 85, and 95 μg of determinand in 20 μg of glycogen). The next stage was to use ANNs, PCR, and PLS, as outlined above, to predict the amount of determinand in all three binary mixtures.

We therefore trained ANNs, using the standard back-propagation algorithm, with the normalized PyMS data from the training sets as the inputs and the amount of determinand (0–100 μg) mixed in 20 μg of glycogen as the output, the latter being scaled between −50 and 150. The effectiveness of training was expressed in terms of the rms error between the actual and desired network outputs; using lysozyme mixed in glycogen as an example this "learning curve" is shown in Figure 2A (open circles). The "learning curve" of the test data (closed circles) is also shown in Figure 2A; it can be seen that whereas the learning curve of the training set continues to decrease during training, the test set's learning curve initially decreases for approximately 10^4 epochs and then increases. This indicates that the ANN was being overtrained, and it is important not to overtrain ANNs since (by definition) the network will not generalize well.[28] This overtraining appears even more marked when the rms error of the test set is plotted against the rms error of the training set (Figure 2B); the minimum rms error in the test set was reached (1.36%) when the rms error of the training set was 0.5% and optimal training had occurred. The ANN was then interrogated with the training and test sets and a plot of the network's estimate versus the true amount of lysozyme in 20 μg of glycogen (Figure 3) gave a linear fit which was indistinguishable from the expected proportional fit (i.e., $y = x$). It was therefore evident that the network's estimate of the quantity of lysozyme in the mixtures was very similar to the true quantity, both for spectra that were used as the training set and, most importantly, for the "unknown" pyrolysis mass spectra.

In other studies ANNs were set up using the standard back-propagation algorithm with the same architecture as that used above except that the output node was scaled from 0 to 100. The network was still able to converge, but took longer (7.6 × 10^4 epochs) to reach an rms error of 0.5% in the training set. However, the network did not generalize as well; after training to the same rms error in the training set (0.5%) the rms error in the test set was 5.27% compared with

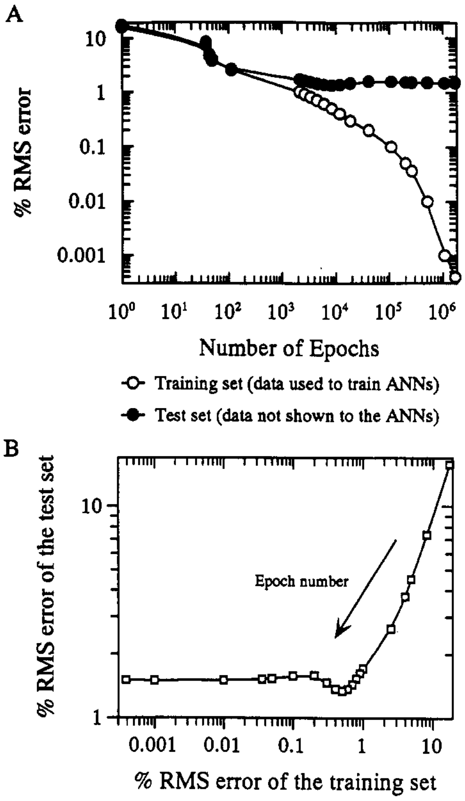

Figure 2. Typical learning curves for the ANN, using the standard back propagation algorithm and with one hidden layer consisting of eight nodes, trained to estimate the amount of lysozyme (micrograms) in 20 μg of glycogen (A). The open circles represent the percentage rms error of the data used to train the neural network (the training set) and the closed circles the data from the test set. A plot of the percentage rms error of the test set versus the percentage rms error of the training set (B) shows that optimal training occurred at 0.5% rms error; the number of epochs (and hence extent training) increases from right to left.

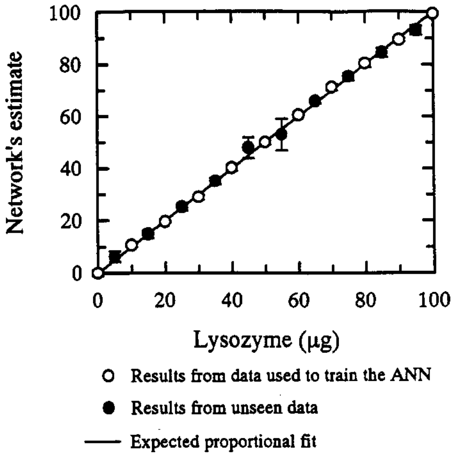

Figure 3. Estimates of trained 150-8-1 neural networks vs the true amount of lysozyme (0–100 μg in 20 μg of glycogen). Networks were trained using the standard back propagation algorithm, to 0.5% rms error (the point at which Figure 2 had indicated that optimal training took place). Data points are the averages of the quadruplicate pyrolysis mass spectra. Open circles represent spectra that were used to train the network and closed circles indicate "unknown" spectra which were not in the training set. Error bars show standard deviation. The expected proportional fit is shown.

Table 2. Comparison of Artificial Neural Network Calibration with Principal Component Regression and Partial Least Squares in the Analysis of Pyrolysis Mass Spectra from Binary Mixtures

Lysozyme (0–100 μg) Mixed in 20 μg of Glycogen

artificial neural networks

architecture	no. of epochs[a]	test set error
150–8–1[c]	75000	5.27
150–8–1[d]	9870	1.36
150–1[d]	238000	1.76

partial least squares 1

no. of PLS factors	% rms error of	
	training set	test set
3[b]	4.97	5.89
5	3.19	4.20
7	1.86	3.71
10	0.96	4.15

principal component regression

no. of PCs	% rms error of	
	training set	test set
3[b]	6.89	8.07
5	5.24	5.68
7	4.92	5.23
10	3.77	5.80
15	3.22	4.96
20	2.92	4.53
25	2.46	4.24
30	2.13	3.81
35	1.44	4.19
40	1.14	4.43
43(max)	0.00	4.71

RNA (0–100 μg) Mixed in 20 μg of Glycogen

artificial neural networks

architecture	no. of epochs[a]	test set error
150–8–1[c]	461150	5.50
150–8–1[d]	103000	1.16
150–1[d]	>10^{6}[e]	1.76

partial least squares 1

no. of PLS factors	% rms error of	
	training set	test set
3[b]	3.59	4.62
5	2.46	4.03
7	1.45	4.09
10	0.72	4.38

principal component regression

no. of PCs	% rms error of	
	training set	test set
3[b]	5.83	6.63
5	3.35	4.91
7	3.28	4.80
10	3.23	4.60
15	2.77	4.37
20	2.34	3.57
25	1.57	4.04
30	1.35	4.19
35	0.75	4.51
40	0.41	4.92
43(max)	0.00	4.55

DNA (0–100 μg) Mixed in 20 μg of Glycogen

artificial neural networks

architecture	no. of epochs[a]	test set error
150–8–1[c]	136200	5.53
150–8–1[d]	36000	2.49
150–1[d]	163000	2.77

partial least squares 1

no. of PLS factors	% rms error of	
	training set	test set
3[b]	4.25	8.08
5	2.18	5.78
7	1.27	5.68
10	0.72	5.74

principal component regression

no. of PCs	% rms error of	
	training set	test set
3[b]	6.84	8.03
5	4.43	7.96
7	3.63	6.82
10	2.68	5.78
15	2.07	6.30
20	1.57	6.10
25	1.42	5.67
30	1.29	5.96
35	1.06	5.43
40	0.47	5.73
43(max)	0.00	5.93

[a] When the error of the training set was 0.5%. [b] Optimal number of factors, predicted by Unscrambler II. [c] Output node was scaled from 0 to 100. [d] Output node was scaled from –50 to 150. [e] Trained to an rms error of 0.538%.

1.36% obtained previously. This result was not perhaps surprising since it has been shown previously[29] that increasing the scaling range on the output layer for quantification of 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.

Further studies were set up using direct linear feedthrough; this is where in addition to the 150–8-1 architecture the input and output layers are also connected directly, which is analogous to including the linear terms in a polynomial expansion.[78,79] The output node was scaled from –50 to 150. The network was able to converge but took twice as many epochs (typically 2.0×10^{4}) as did the 150–8-1 ANN to reach an rms error of 0.5% in the training set. The network did generalize well and the error of the test set was usually 1.2%, not significantly better than the 150–8-1 ANNs (1.36% error of the test set). This result was not suprising since the output scaling which was used on the 150–8-1 ANNs was not between 0 and 1 but a more restricted range (giving a buffer of ±50%) and so already has a strong linearizing function. Further, a number of workers[59–63] have shown mathematically that a neural network consisting of only one hidden layer, albeit with an arbitrarily large number of nodes, can learn any, arbitrary (and hence nonlinear) mapping of a continuous function to an arbitrary degree of accuracy.

Finally, in addition to these ANNs others were employed using the standard back-propagation algorithm with *no* hidden layer and in which the output node was scaled from –50 to

[78] Werbos, P. J. *Proc. IEEE* **1990**, *78*, 1550–1560.
[79] Wilson, E.; Rock, S. M. *Proc. World Congress on Neural Networks* **1993**, *3*, 157–162.

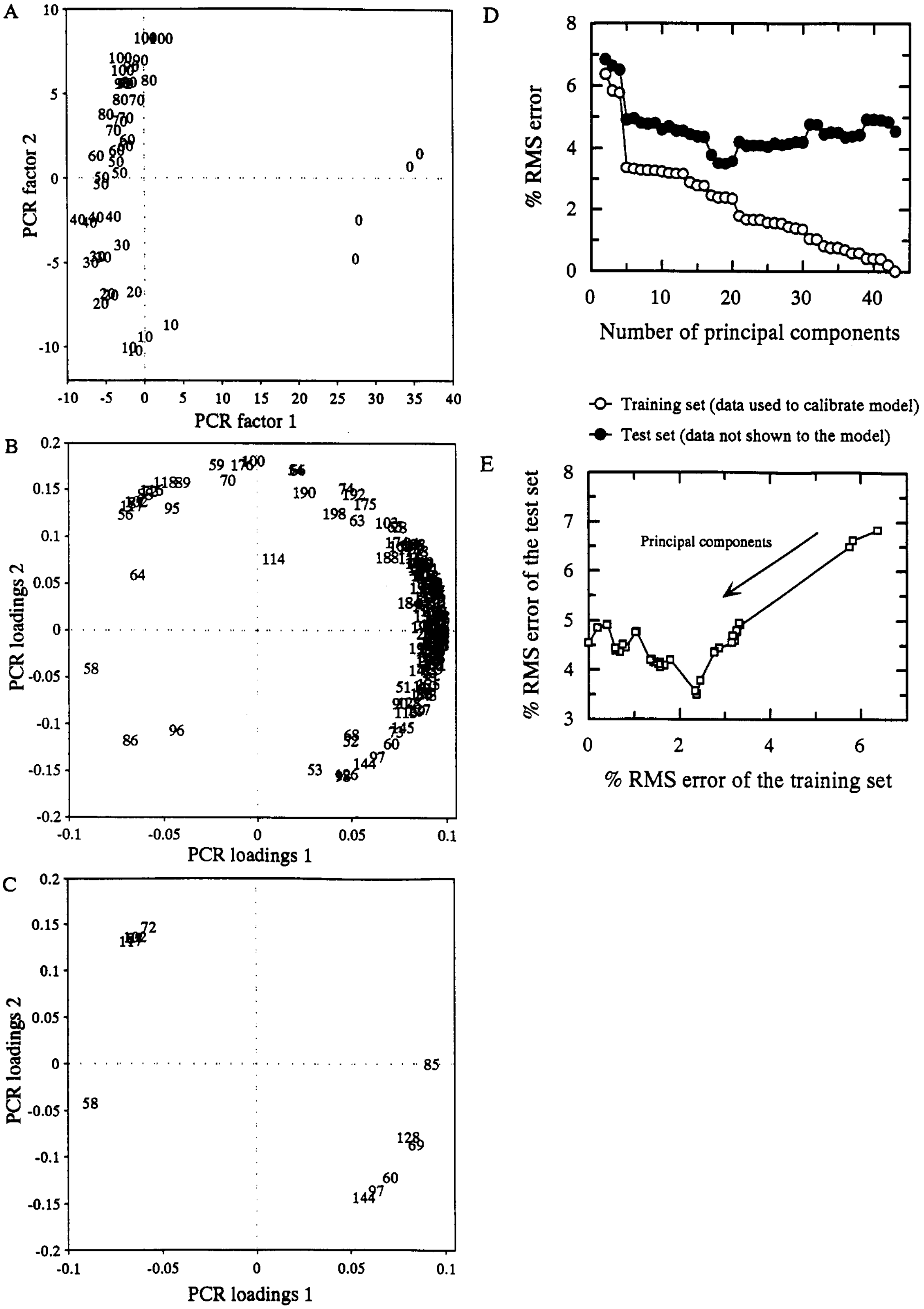

Figure 4. PCR scores plot based on PyMS data for RNA of the stated number of micrograms mixed with 20 μg of glycogen (A). The first two PCR factors account for 72 and 20% of the total variation, respectively. (B) PCR loadings plot showing all 150 masses; a reduced number of masses are also shown (C); these masses are representative of RNA (*m/z* 58, 72, 87, 102, and 117) and glycogen (*m/z* 60, 69, 85, 97, 128, and 144). Effect of the number of principal components on the PCR calibration models formed to estimate the amount of RNA (micrograms) in 20 μg of glycogen (D). The open circles represent the percentage rms error of the data used to create the model (the training set) and the closed circles the data from the test set. (E) A plot of the percentage rms error of the test set versus the percentage rms error of the training set; the optimal calibration model was formed using 20 principal components, and the number of latent variables used in predictions increases from right to left.

150. Networks with no hidden layer are unable to separate objects in different classes unless such classes are linearly separable.[80] These networks took over 2×10^5 epochs to reach 0.5% rms error in the training set but had converged well and

the rms error in the test set was typically 1.8%. It was interesting to observe that these networks also generalized well, indicating that the differences in pyrolysis mass spectra due to the addition of lysozyme to glycogen could be fitted in a linear model in 150-dimensional space.

Neural networks of the above types were also used to estimate the amount of RNA or DNA mixed in 20 μg of glycogen, with the results shown in Table 2. Both groups of the mass spectra of the binary mixtures were found to be quantified accurately by ANNs; in addition, the effects of different network architectures were similar to those found in the lysozyme/glycogen case described above.

PCR was also used as outlined above to create calibration models, using the training sets, to predict the amount of determinand (0–100 μg) mixed in 20 μg of glycogen. The example used is the mixture of RNA in glycogen. After the model was created the first two PCR factors for the training sets were plotted (Figure 4A). The first two PCR factors accounted for the majority of the variance in the samples, viz. 72% and 20%, respectively. It is evident (Figure 4A) that the first PCR factor largely served to account for (or describe) the difference between the spectra of pure glycogen and samples containing RNA, while the second PCR factor largely accounted for the difference in the amount of RNA in the glycogen backgrounds. To ascertain which masses contributed positively to the formation of the PCR model their loadings may be viewed (Figures 4B,C). Figure 4B is the PCR (mass) loadings plot of all 150 masses. This shows that all masses were influential on the model formed because no masses plot at the origin. Also, all the masses have approximately the same influence on the model because they are equidistant from the origin; presumably this is simply due to the normalization procedure used (where the masses were scaled as a percentage of the total), such that if a given mass is relatively greater in a particular spectrum then all of the other masses are necessarily relatively lower. When only the masses that predominate in the mass spectra of glycogen (m/z 60, 69, 85, 97, 128, and 144) and RNA (m/z 58, 72, 87, 102, and 117) are plotted (Figure 4C) they form two distinct groups which map opposite each other; this implies, as one should expect, that these two sets of masses are negatively correlated. Finally, when the mass loadings plot (Figure 4C) is overlaid on the factor plot (Figure 4A) the masses that predominate the spectra of glycogen and RNA lie in the same area as the samples from pure glycogen and 100 μg of RNA in 20 μg of glycogen respectively; this shows that these masses are indeed influential in the discrimination between the two species.

To assess the accuracy of the predictions after the model was calibrated, a varying number of latent variables was used in the predictions, and the rms error between the true amount of determinand and the predicted amount for both the training and test set was calculated (Table 2) and plotted against the number of PCs used to form the calibration model (Figure 4D). The open circles represent the percentage rms error of the data used to create the model (the training set) and the closed circles the data from the test set. It can be seen that the error in the training set continues to decrease with increasing number of latent variables but that the lowest value of rms error for the test set, indicating optimal calibration, was formed using 20 PCs and was 3.57%; the error in the training set at this point was 2.34%. This can also be clearly seen in Figure 4E which is a plot of the percentage rms error of the test set versus the percentage rms error of the training set. This was perhaps surprising because Unscrambler II stated that the optimal model should be formed using only three PCs; at this point the percentage rms error in the training and test sets were 5.83 and 6.63, respectively (Table 2), and one would have presumed that using more than three factors would cause overfitting (i.e., inaccurate predictions on the test data). The model's predictions were then made using 20 PCs and a plot of the PCR's predictions versus the true amount of RNA mixed in glycogen is shown in Figure 5. This graph also gave a proportional fit (i.e., $y = x$). It was therefore evident that the PCR calibration model's predictions of the quantity of RNA in the mixtures was very similar to the true quantity, both for spectra that were used as the training set and for the "unknown" pyrolysis mass spectra.

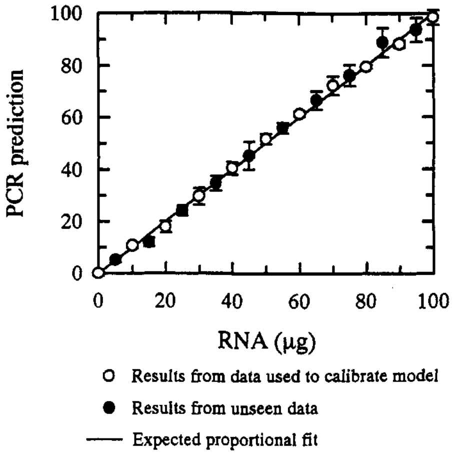

Figure 5. PCR predictions plotted against the true amount of RNA (0–100 μg in 20 μg of glycogen), using 20 principal components in the calibration model (the point at which Figure 4 had indicated that the optimal model was formed). Data points are the averages of the quadruplicate pyrolysis mass spectra. Open circles represent spectra that were used to calibrate the PCR model and closed circles indicate "unknown" spectra which were not in the training set. Error bars show standard deviation. The expected proportional fit is shown.

PCR models were also formed to estimate the amount of lysozyme or DNA mixed in 20 μg of glycogen and the results are shown in Table 2. Both groups of the mass spectra of these binary mixtures were found to be predicted accurately using PCR.

Finally, the pyrolysis mass spectra of the three binary mixtures were analyzed using PLS as outlined above. The example illustrated is from the spectra of 0–100 μg of DNA mixed in 20 μg of glycogen. After the model was created the first two PLS factors for the training sets were plotted (Figure 6A). Again the majority of the variation was preserved, in that the first two PLS factors accounted for 56 and 31% of the variance, respectively. It was again evident that the major difference was between the spectra of pure glycogen and samples containing DNA; a combination of the first and second PLS factors accounted for the difference in the amount of DNA in the glycogen backgrounds. Figure 6B is the PLS (mass) loadings plot of all 150 masses; as with the PCR model of RNA in glycogen, this plot shows that all masses were again approximately equally influential in forming the PLS model. When only the masses that predominate in the mass

(80) Minsky, M. L.; Papert, S. A. *Perceptrons*; MIT Press: Cambridge, MA, 1969.

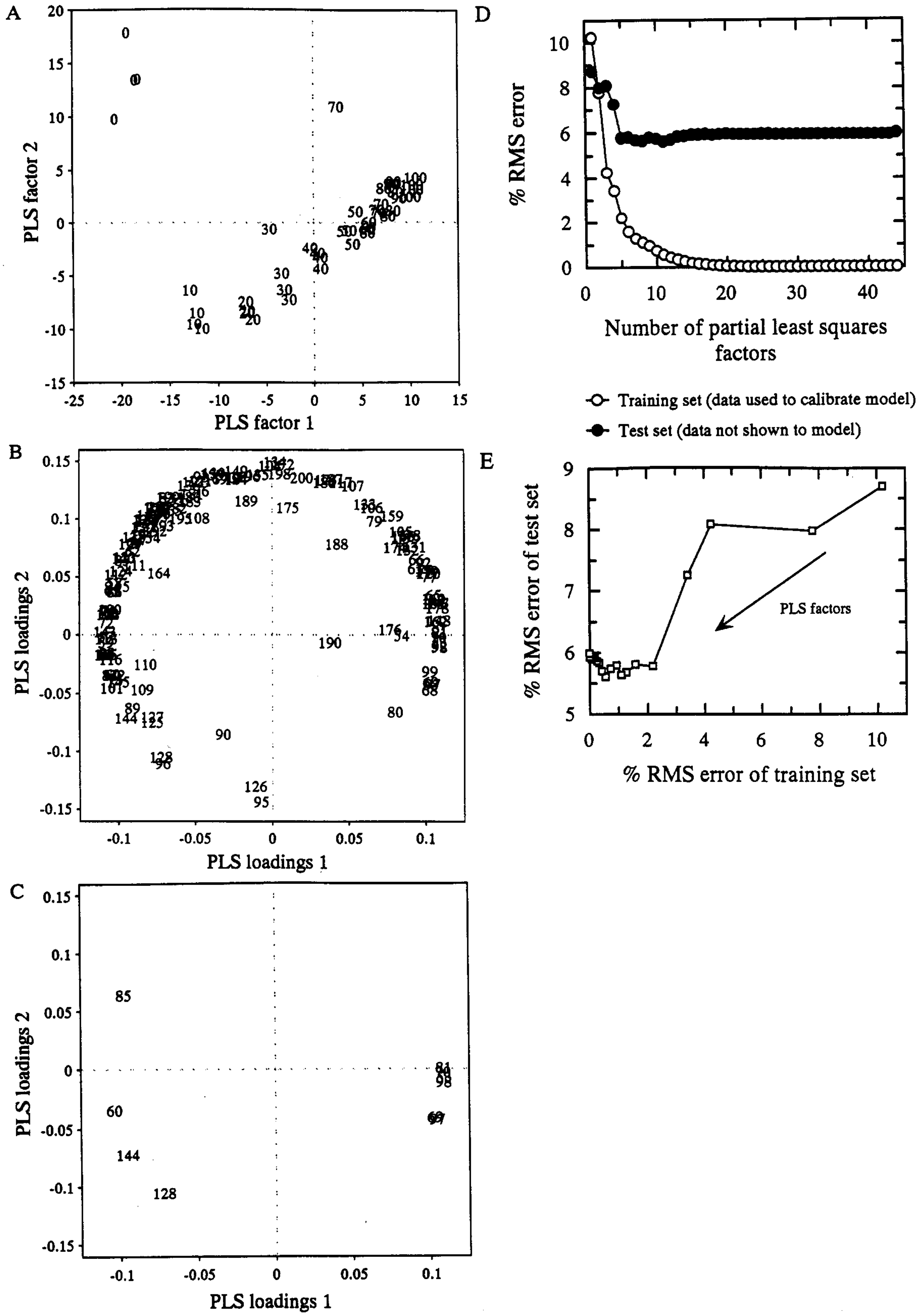

Figure 6. PLS scores plot based on PyMS data for DNA of the stated number of micrograms mixed with 20 μg of glycogen (A). The first two PLS factors account for 56 and 31% of the total variation, respectively. (B) PLS loadings plot showing all 150 masses; a reduced number of masses are also shown (C); these masses are representative of DNA (*m/z* 53, 70, 81, and 98) and glycogen (*m/z* 60, 69, 85, 97, 128, and 144). Effect of the number of PLS factors on the accuracy of the PLS calibration models used to estimate the amount of DNA (μg) in 20 μg of glycogen (D). The open circles represent the percentage rms error of the data used to create the model (the training set) and the closed circles the data from the test set. (E) A plot of the percentage rms error of the test set versus the percentage rms error of the training set; the optimal calibration model was formed using seven PLS factors, and the number of latent variables used in predictions increases from right to left.

spectra of glycogen (*m/z* 60, 69, 85, 97, 128, and 144) and DNA (*m/z* 53, 70, 81, and 98) are plotted (Figure 6C) they form two distinct groups (although *m/z* 69 and 97 from glycogen group with the masses from the spectrum of DNA)

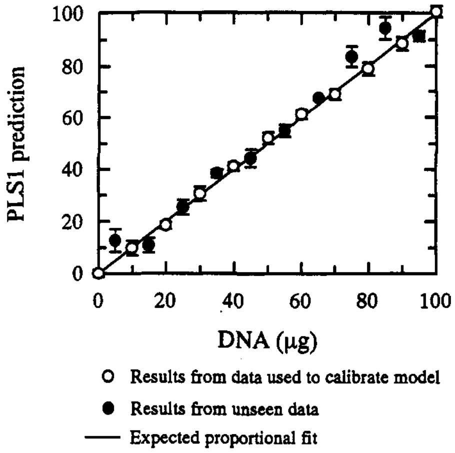

Figure 7. PLS1 predictions plotted against the true amount of DNA (0–100 μg in 20 μg of glycogen), using five PLS factors in the calibration model (the point at which Figure 6 had indicated that optimal calibration took place). Data points are the averages of the quadruplicate pyrolysis mass spectra. Open circles represent spectra that were used to calibrate the PLS1 model and closed circles indicate "unknown" spectra which were not in the training set. Error bars show standard deviation. The expected proportional fit is shown.

which map opposite each other, implying that these two sets of masses are mainly negatively correlated. Finally, when the mass loadings plot (Figure 6C) is overlaid on the factor plot (Figure 6A), it may be seen that the masses that predominate the spectra of glycogen and DNA lie in the same area as the samples from pure glycogen and 100 μg of DNA in 20 μg of glycogen, respectively.

To assess the accuracy of the predictions after the model was calibrated, different numbers of latent variables were used in the predictions, and the rms error between the true amount of determinand and the predicted amount for both the training and test set was calculated (Table 2) and plotted against the number of PLS factors used to form the calibration model (Figure 6D). The open circles represent the percentage rms error of the data used to create the model (the training set) and the closed circles the data from the test set. It can be seen that the error in the training set very rapidly decreases to below 1% with the first 10 latent variables and then gradually continues to decrease to 0%. The error in the test set, however, decreases rapidly to 5.68% with the first seven PLS factors and then does not decrease any further. The optimal calibration model therefore appears to be formed using only seven PLS factors; this is also clearly seen when the percentage rms error of the test set is plotted against the percentage rms error of the training set (Figure 6E). (Unscrambler in fact stated that optimal calibration occurred using only three factors; from the values of rms error it could be seen that this was not the case (Table 2).) Once again it is presumably to be considered that when more than three latent variables were used in predictions, this would cause overfitting (i.e., give inaccurate predictions on the test data) and may be due to influences of "noise" in the training data. A plot of the PLS predictions using seven latent variables versus the true amount of DNA mixed in glycogen is shown in Figure 7. It can be seen that the pyrolysis mass spectra of the training set (open circles) were accurately predicted and that the estimates for the "unknown" spectra were also predicted well with the exception of 5, 75, and 85 μg of DNA mixed in 20 μg of glycogen which were estimated to be 12, 83, and 94 μg of DNA respectively. When ANNs and PCR were calibrated to predict the amount of DNA, the 75- and 80-μg mixes were still inaccurately predicted as 83 and 94 μg of DNA, respectively; however, the 5-μg mix was accurately estimated. Rather than these inaccuracies being a failure of the multivariate methods to quantitate the pyrolysis mass spectra of DNA mixed in glycogen, it is more likely to be due to erroneous experimental measurements since when DNA was mixed in water, it formed a very viscous colloid, and it was difficult to prepare concentrations above 12 μg μL^{-1}. In any event, Figure 7 shows a proportional fit (i.e., $y = x$), and it was evident that the PLS calibration model's predictions of the quantity of DNA in the mixtures was in most cases very similar to the true quantity.

PLS models were also formed to estimate the amount of lysozyme or RNA mixed in 20 μg of glycogen and the results are shown in Table 2. Both groups of the mass spectra of these binary mixtures were found to be predicted accurately using PLS.

Table 2 gives the percentage rms error on the predictions produced by ANNs, PCR, and PLS on both the training and test sets for all three sets of binary mixtures. The number of hidden nodes, scaling on the output layer, and the number of latent variables used are also given. It can be seen in all instances the number of PCs used to give optimal calibration models in PCR was higher than the number of PLS factors needed but that the percentage rms error on prediction was approximately the same, typically between 1 and 2% for the training set and 3.5 and 6% for the "unknown" mass spectra in the test set. In addition, optimal calibration models were produced using more latent variables than Unscrambler stated should give optimal predictions; this usually implies that there are *nonlinear* relationships within the pyrolysis mass spectral data.[8] When ANNs were trained to analyze the pyrolysis mass spectra of the binary mixtures the results were more accurate, and it is likely that this method was indeed mapping *nonlinear* relationships in the data. The best estimates were obtained when ANNs were trained using the standard back-propagation algorithm, with one hidden layer of eight nodes (150–8-1), and the output node scaled between –50 and 150. ANNs of the same architecture but containing no hidden layers (150–1) were also trained until an rms error between the desired and true output in the training set of 0.5% was reached. ANNs of this type should map *linearly* and therefore possibly to the same degree of accuracy as PCR and PLS. Although training was significantly slower (in terms of the number of epochs) and not quite as accurate as for the 150–8-1 ANNs, the estimates obtained were still more accurate than the best predictions produced by the PCR and PLS calibration models. In conclusion, these results indicate that ANNs, which can be used to uncover *nonlinear* as well as *linear* properties in data, produced superior results as compared to the *linear* mapping techniques of PCR and PLS, in terms of the ability to quantitate the pyrolysis mass spectra of either 0–100 μg of lysozyme, RNA or DNA mixed in 20 μg of glycogen.

In other studies ANNs and PLS were set up using data that were scaled differently. Previously data from each pyrolysis mass spectrum were scaled across the *whole* mass range such that the lowest ion count was set to 0 and the highest to 1. In this set of studies the mass spectra were

Table 3. Comparison of Artificial Neural Network Calibration to the Analysis of Pyrolysis Mass Spectra from Binary Mixtures[a]

	scaling used in Table 2[b]	each input node was scaled and mass values lower than those below were set to zero[c]				
		0	32	65	130	650
	Protein (0–100 μg) Mixed in 20 μg of Glycogen					
no. of epochs	9870	7809	5986	17833	7008	1252
% time taken[d]	100	79.12	60.65	180.68	71.00	12.68
% rms error of training set	0.50	0.25	0.25	0.10	0.25	1.00
% rms error of test set	1.36	1.11	1.33	1.49	1.83	4.62
	RNA (0–100 μg) Mixed in 20 μg of Glycogen					
no. of epochs	103000	11624	7166	952	7689	22349
% time taken[d]	100	11.29	6.94	0.92	7.47	21.70
% rms error of training set	0.50	0.25	0.50	1.00	0.75	0.75
% rms error of test set	1.16	1.36	1.90	1.88	1.69	1.58
	DNA (0–100 μg) Mixed in 20 μg of Glycogen					
no. of epochs	36000	18539	2113	1080	219	2684
% time taken[d]	100	51.50	5.87	3.00	0.61	7.46
% rms error of training set	0.50	0.10	0.50	0.75	1.00	1.00
% rms error of test set	2.49	2.57	2.77	3.16	3.56	3.01

[a] The architecture used was 150–8–1, employing the back-propagation algorithm, with the output node scaled from −50 to 150. Two scaling methods were used on the input layer and some masses removed from the analysis to assess whether generalization improved. The values shown are from *optimum* ANN generalization. This was judged as the point at which the lowest % rms error in the test set was obtained. The numbers shown are the averages of three ANN training runs. [b] The input layer was scaled across the *whole* mass range such that the lowest mass was set to 0 and the highest mass to 1. [c] The masses that were removed from the analyses (by setting the values to 0) were chosen to reflect a percentage of the total ion count (2^{16}); 0 was the control where no masses were removed, 32 was equivalent to 0.05%, 65 to 0.1%, 130 to 0.2%, and 650 was equivalent to 1% of the total ion count. The input layer was scaled for *each* input node such that the lowest mass was set to 0.4 and the highest mass to 0.6. [d] Based on the number of epochs needed to train ANNs which were scaled across the whole mass range.

scaled for *each* mass such that the lowest mass was set to 0.4 and the highest mass to 0.6. This alternative scaling had no effect on PLS1 calibrations (data not shown); this was not surprising because before calibration the X data were scaled in proportion to the reciprocal of their standard deviations (and so were scaled for each mass anyway). This method of scaling did however have a *very* marked effect on the neural net analyses; most notable was the decrease in the time required to train the networks to a particular rms error (Table 3). The network was trained and the extent of training monitored by calculating the percent rms error in the test set (data not shown, but for an example see Figure 2); when this error had reached a minimum the network was said to have reached optimal generalization. For the analysis of protein mixed in glycogen, the time taken was less than 80% compared to the scaling method previously used, and in addition to this increase of speed the network also generalized (to the test set) better, viz., to 1.11% rms error in the test compared to 1.36% obtained earlier. For RNA and DNA mixed in glycogen the time taken was 11.3% and 51.5%, respectively, although the error in the test set was slightly greater than the values previously observed.

It is known in some cases that removing "unimportant" X data can significantly improve both the speed of learning and of generalization.[76,77] In addition to employing the above "individual" method of scaling, therefore, masses were removed to ascertain the effects both on network generalization and on the speed of training required to obtain the optimal calibration models for both ANNs and PLS1. The masses were removed from the analyses by setting their values to 0 and were chosen to reflect a percentage of the total ion count. The pyrolysis mass spectra were initially normalized so that the total ion count was 2^{16}; 0 therefore represents the control where no masses were removed, then masses of value 32 or less were removed which is equivalent to 0.05% of the total ion count, 65 or less (0.1%), 130 (0.2%), and masses less than or equal to 650 were set to 0 which represented 1% of the total. The results of such ANN analyses are shown in Table 3, where it can be seen that in general the minimal percent rms error of the test set increases when more masses are removed. It is noteworthy that although this error rises, it is however not as bad as the error level found when PLS1 or PCR were used to analyze the three mixtures (Table 2). Also with the exception of removing mass values of 65 or less for protein in glycogen the time taken to train the ANN is significantly more rapid than ANNs using all the masses. In fact, when masses whose values are lower than 65 and 130 were removed for the analysis of the binary mixtures of RNA and DNA respectively the time taken is *so fast* that it is less than *1%* of the total time to train the ANNs using all the masses and scaled across the entire mass range.

The same masses were removed from calibration models formed with the PLS1 algorithm using Unscrambler. The models were calibrated using 1, 2, 3, ..., 10, and 20 latent variables, and to assess the accuracy of the calibrations the numbers of latent variables used in the three sets of predictions were plotted against the percent rms error between the true amount of determinand and the predicted amount for the test set (Figure 8). For the analysis of the mixture of protein and glycogen it can be seen in Figure 8A that better calibration models were formed when some masses were removed from the analyses. When no masses were removed the lowest percent rms error of the test was 3.71% which was formed with seven factors; however, when mass intensities of 0.05% (≤32) of the total or lower were removed the lowest error found was 3.42% with seven factors, and an error level of 3.71% was found when five factors were used in the predictions. A further reduction in information by removing 65 (≤0.1%) masses were studied and the best model was formed using only three latent variables and was 3.62%. The models then deteriorated when more masses were removed (Figure 8A). These results strongly imply that PLS models are not very robust to noise and that removing *small* amounts of data improves calibration as

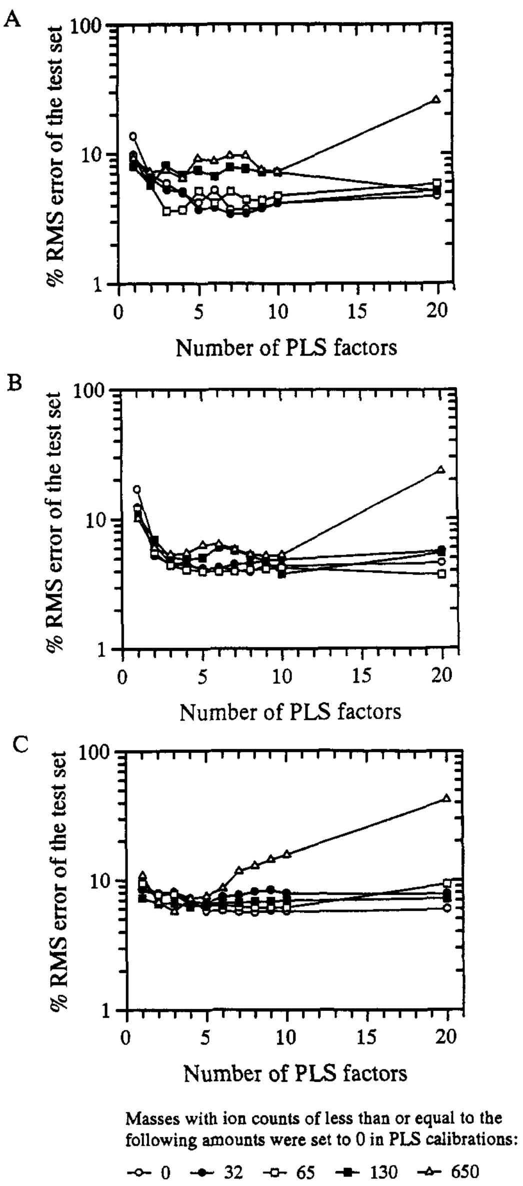

Figure 8. Effect of the number of PLS factors and the effective number of masses used on the percentage rms error of the test sets for the calibration models formed for the quantification of 0–100 μg of protein (A), RNA (B), or DNA (C) in 20 μg of glycogen. Masses were removed from the analyses (by setting the values to 0) if the ion count (after normalization of the total ion count to 65 536, i.e., 2^{16}) was less than or equal to 0 (this was the control where no masses were removed), 32 (equivalent to 0.05% of the total ion count), 65 (0.1%), 130 (0.2%), or 650 (1% of the total).

indicated by the lower error levels found in the test set and, perhaps more importantly, fewer latent variables were used.[77] When the pyrolysis mass spectra of RNA mixed in glycogen were analyzed with PLS (Figure 8B) similar, but less obvious, results were seen and the best model was formed when mass intensities of 0.1% (≤65) of the total ion count or lower were removed: an rms error of 3.99% was found when five factors were used compared with 4.03% with five factors when all of the data were used in calibrations. For the quantification of 0–100 μg of DNA in 20 μg of glycogen the best calibration using all the masses was formed with seven factors in the prediction and was 5.68% rms error of the test set. It was mentioned above that when large number of latent variables are used in the calibration model it implies that the model formed was overfitting the data and may be influenced by noise in the data. It was therefore very significant that when masses of intensity of ≤1% of the total (650) were removed from PLS analyses, the best calibration model was formed using *only* three factors and the rms error in the test set was 5.77%.

In conclusion, it was observed that removing low-intensity masses when calibration models were formed using the PLS algorithm gave better results because lower (or at least very similar) error levels in the pyrolysis mass spectra of the test sets were formed when predictions were made using fewer latent variables. Removing data for ANN analyses did neither improve network generalization nor cause a significant deterioration in calibrations (and was still superior to PLS calibration). However, scaling each input node (as compared to scaling *across* the mass range) did speed up network training significantly. This may well be due to the premature saturation[81] of the nodes associated with very large and very small peaks in the spectra; scaling for each node individually reduces the likelihood of the initial (random) weights leading the nodes to be saturated (fully on or fully off).

Analysis of a Tertiary Mixture. The three-way mixture of *S. aureus* NCTC6571, *B. subtilis* DSM10, and *E. coli* W3110 were analyzed by PyMS. The spectral data were split into two sets. The training set contained the normalized triplicate ion intensities from the pyrolysis mass spectra from tertiary mixtures describing the percentage of *S. aureus*, *B. subtilis*, and *E. coli* (i.e., 198 objects representing 66 triplicates, and see Figure 9A). The test set contained both the training set and the 16 "unknown" pyrolysis mass spectra (Table 1 and Figure 9A). The next stage was to use ANNs, PCR, and PLS, as outlined above, to predict the proportional amount of bacteria in the three-way mixture.

We therefore initially trained three ANNs using the standard back-propagation algorithm, with PyMS data normalized to 2^{16} from the training sets as the inputs (and scaled across the *whole* mass range such that the lowest mass was set to 0 and the highest mass to 1) and the percentage amount of the three bacteria as the outputs, the latter were scaled between −50 and 150. The training set therefore consisted of 198 objects each with 150 variables (masses) and training was consequently very slow. ANNs were therefore trained for 1 × 10^5 epochs (this was equivalent to approximately 12 h of training time). After training, the test data were applied to the network's input nodes and the predictions are shown in Figure 9B where it can be seen that both the training (closed circles) and test sets (closed crosses) were accurately estimated. This plot may be difficult to interpret so plots of the network's estimates versus the true percentage amounts of bacteria are shown in Figure 10. All three plots gave a proportional fit (i.e., $y = x$), although the estimates of 75% *B. subtilis* in the binary mixtures with *S. aureus* or *E. coli* were rather inaccurate as judged by the large error bars. The rms error in the three networks were for *S.aureus* 2.42%, *B. subtilis* 5.24%, and for *E. coli* 1.51% (Table 4), and it was therefore evident that the network's estimate of the percentage of the three bacteria was very similar to the true quantity, both for spectra that were used as the training set and for the "unknown" pyrolysis mass spectra.

In other studies ANNs were set up to estimate the percentage of all three bacteria simultaneously. The same

[81] Lee, Y.; Oh, S.-H.; Kim M. W. *Neural Networks* **1993**, *6*, 719–728.

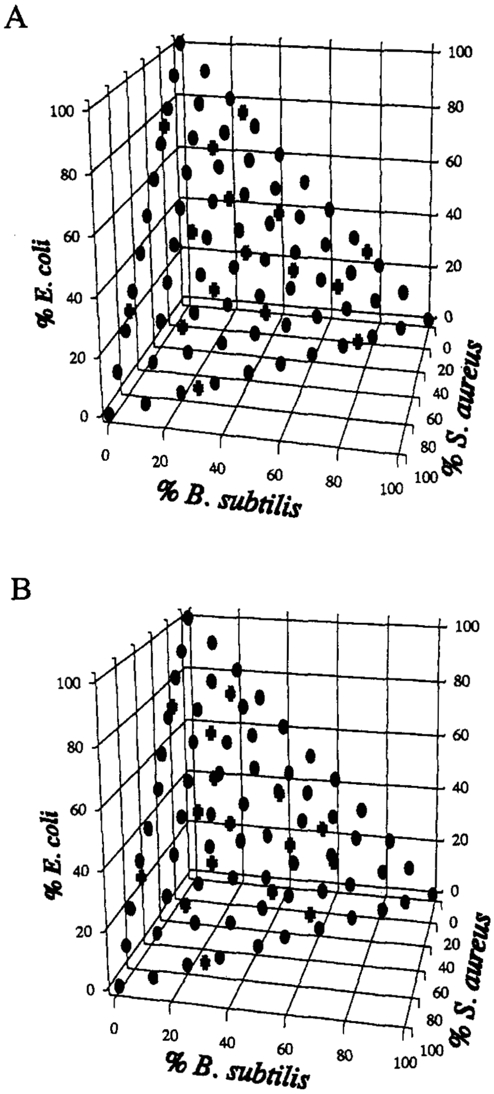

Figure 9. Expected (A) and actual results (B) of the estimates of 150-8-1 neural networks trained to analyse the tertiary mixtures of *S. aureus*, *B. subtilis*, and *E. coli*. Closed circles represent spectra that were used to train the network and crosses indicate "unknown" spectra which were not in the training set.

training and test sets were used, the network employed the standard back-propagation algorithm and the output layer of three nodes was scaled from −50 to 150. This network trained very slowly so was trained for 2.5 × 10^5 epochs (training took about 36 h). The ANN was then interrogated and the percent rms error of the training and test set calculated (Table 4). From Table 4 it can be seen that this network had not generalized as well as the three 150–8-1 ANNs, although these network's estimates were reasonably accurate. Arguably the main problem is that in trying to teach fully interconnected feedforward networks more than one thing at a time conflicting error messages are back-propagated from the output layer during the learning process.[82,83] That is to say, the error that is fed back from one of the three output nodes is fed to *all* eight nodes in the preceding hidden layer, which also contains information pertinent to learning the other two targets; if one target is failing to be learned and thus sends the algorithm off in a different direction in weight space, it will inevitably hinder the learning of the other targets.

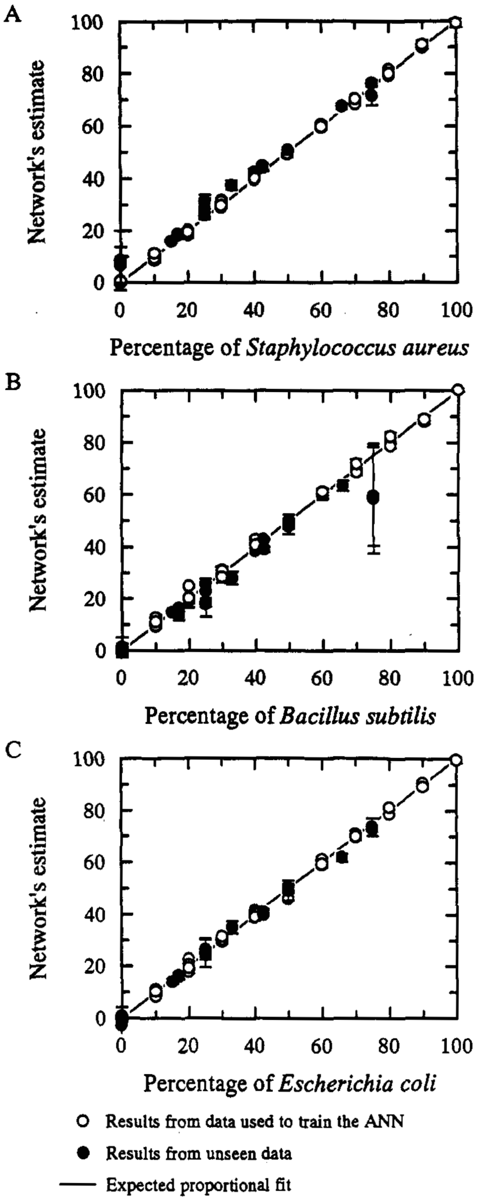

Figure 10. Estimates of trained 150-8-1 neural networks against the true percentage of *S. aureus* (A), *B. subtilis* (B), and *E. coli* (C) in mixures of the three organisms. ANNs were trained using the standard-back propagation algorithm for 1 × 10^5 epochs (which took about 12 h). Data points are the averages of the triplicate pyrolysis mass spectra. Open circles represent spectra that were used to train the network and closed circles indicate "unknown" spectra which were not in the training set. Error bars show standard deviation. The expected proportional fits are shown.

The next stage was to evaluate the accuracy of the analysis of the tertiary mixture using PCR and PLS. Both PCR and PLS were set up as detailed above and models calibrated using the same training data as used for ANNs analyses. In the first instance PCR and PLS were calibrated to predict only the percentage of one of the bacteria (i.e., calibrated using one variable in the Y matrices); in separate calibrations, all three *Y* variables were used for in model callibration using the PCR and PLS2 algorithms. Table 4 gives the percentage rms error on the predictions produced by PCR (calibrated using one or three *Y* variables) and PLS1 and -2 on both the training and test sets for the tertiary mixture. It can be seen that PCR and PLS can be used to gain quantitative information from the tertiary mixture whether models are formed to predict one or three variables in the Y matrix. For PCR optimal

[82] Jordan, M. I. *J. Math. Psychol.* **1992**, *36*, 396–425.
[83] Jordan, M. I.; Rumelhart, D. E. *Cognit. Sci.* **1992**, *16*, 307–354.

Table 4. Comparison of Artificial Neural Network Calibration with Principal Component Regression and Partial Least Squares in the Analysis of Pyrolysis Mass Spectra from Tertiary Mixtures

Percentage of *Staphylococcus aureus*

									principal component regression				
artificial neural networks				partial least squares 1			PLS2			calibrated using 1 Y variable		calibrated using all 3 Y variables	
		% rms error of		no. of	% rms error of		% rms error of			% rms error of		% rms error of	
architecture	no. of epochs	training set	test set	PLS1 or 2 factors	training set	test set	training set	test set	no. of PCs	training set	test set	training set	test set
150–8–1[a]	100 000	0.50	2.42	3[c]	2.19	5.76	2.16	6.14	3	4.57	13.29	4.57	13.29
150–8–3[a]	250 000	2.18	6.99	4[d]	1.81	4.85	2.10	5.44	4[b]	2.32	5.56	2.32	5.56
				5	1.29	5.31	1.43	4.52	5	1.95	4.75	1.95	4.75
				7	0.83	5.75	0.98	5.77	7	1.63	5.27	1.63	5.27
				10	0.59	5.98	0.80	5.86	10	1.02	5.90	1.02	5.90
									15	0.88	5.70	0.88	5.70
									20	0.82	5.81	0.82	5.81
									40	0.69	5.92	0.69	5.92
									79 (max)	0.51	5.76	0.51	5.76

Percentage of *Bacillus subtilis*

									principal component regression				
artificial neural networks				partial least squares 1			PLS2			calibrated using 1 Y variable		calibrated using all 3 Y variables	
		% rms error of		no. of	% rms error of		% rms error of			% rms error of		% rms error of	
architecture	no. of epochs	training set	test set	PLS1 or 2 factors	training set	test set	training set	test set	no. of PCs	training set	test set	training set	test set
150–8–1[a]	100 000	0.67	5.24	3[c]	2.42	4.83	3.15	6.83	3	5.98	14.18	5.98	14.18
15–8–3[a]	250 000	2.25	6.52	4[d]	2.00	3.84	2.09	3.86	4[b]	2.41	3.97	2.41	3.97
				5	1.59	4.67	2.09	3.88	5	2.35	3.78	2.35	3.78
				7	1.28	4.59	1.57	4.54	7	1.78	4.59	1.78	4.59
				10	1.00	4.85	1.16	4.85	10	1.46	4.50	1.46	4.50
									15	1.41	4.67	1.41	4.67
									20	1.35	4.93	1.35	4.93
									40	1.13	4.36	1.13	4.36
									79 (max)	0.93	4.80	0.93	4.80

Percentage of *Escherichia coli*

									principal component regression				
artificial neural networks				partial least squares 1			PLS2			calibrated using 1 Y variable		calibrated using all 3 Y variables	
		% rms error of		no. of	% rms error of		% rms error of			% rms error of		% rms error of	
architecture	no. of epochs	training set	test set	PLS1 or 2 factors	training set	test set	training set	test set	no. of PCs	training set	test set	training set	test set
150–8–1[a]	100 000	0.58	1.51	3[c]	2.65	3.70	3.31	3.24	3	3.26	3.15	3.26	3.15
150–8–3[a]	250 000	2.78	3.73	4[d]	2.35	3.87	2.75	3.73	4[b]	2.87	3.82	2.87	3.82
				5	1.93	2.71	2.26	3.16	5	2.78	3.44	2.78	3.44
				7	1.53	3.21	1.62	3.21	7	2.43	3.06	2.43	3.06
				10	1.16	2.86	1.29	2.79	10	1.69	3.21	1.69	3.21
									15	1.65	2.88	1.65	2.88
									20	1.57	2.84	1.57	2.84
									40	1.35	3.36	1.35	3.36
									79 (max)	1.02	2.78	1.02	2.78

[a] Output node was scaled from −50 to 150. [b] Optimal number of principal components, predicted by Unscrambler II. [c] Optimal number of factors using PLS1, predicted by Unscrambler II. [d] Optimal number of factors using PLS2, predicted by Unscrambler II.

calibration occurs for *S. aureus* and for *B. subtilis* when five latent variables are used, the percent rms error of the test set was 4.75% and 3.78%, respectively. It is not surprising that it does not matter if the latent variables from the **X** matrix were regressed onto one or three variables in the **Y** matrices, this is because there is *no* reduction of the **Y** matrix in PCR analysis and whether three different matrices or one matrix containing all the **Y** data were used, the model regresses onto each of the *Y* variables separately. For *E. coli* when five principal components were used in the prediction models, the rms error of the test set was 2.71%; however, models employing 20 latent variables gave optimal estimation of the 2.34% rms error of the test set. Very similar percent rms error values were also seen when PLS1 and -2 were used to create calibration models (Table 4). In this instance because PLS forms a reduction on the **Y** matrix and *then* regresses the latent variables from the **X** matrix on to the *reduced* **Y** matrix, better predictions occurred when the **Y** matrix contained only one variable. It is therefore evident that although the test set error was higher than the error from ANN analyses, these linear regression methods could also be used to analyse three-way mixtures. In fact PCR and PLS assessed the pyrolysis mass spectra in terms of the amount of *B. subtilis* in a more accurate fashion than did either 150–8–1 or 150–8–3 ANNs.

The training sets used above used 198 pyrolysis mass spectra (66 samples triplicated) because it was found previously[29] that for the quantitative analysis of binary mixtures the concentration range of any determinand was advantageously

Table 5. Comparison of Artificial Neural Network Calibration with Partial Squares in the Analysis of Pyrolysis Mass Spectra from Tertiary Mixtures[a]

Percentage of *Staphylococcus aureus*

artificial neural networks				partial least squares 2		
		rel error of			rel error of	
architecture	no. of epochs[b]	training set	test set	no. of PLS2 factors	training set	test set
150–8–3[c]	2700	0.56	4.63	2	8.74	14.00
150–3[c]	5410	0.65	4.35	3[d]	2.01	7.59
150–8–1[c]	5460	0.50	5.11	4	1.32	6.19
150–1[c]	7470	0.50	4.44	5	0.57	6.57

Percentage of *Bacillus subtilis*

artificial neural networks				partial least squares 2		
		rel error of			rel error of	
architecture	no. of epochs[b]	training set	test set	no. of PLS2 factors	training set	test set
150–8–3[c]	2700	0.41	5.65	2	6.86	11.18
150–3[c]	5410	0.37	5.05	3[d]	3.51	7.73
150–8–1[c]	2080	0.50	5.50	4	1.20	4.56
150–1[c]	3130	0.50	4.93	5	0.89	4.77

Percentage of *Escherichia coli*

artificial neural networks				partial least squares 2		
		rel error of			rel error of	
architecture	no. of epochs[b]	training set	test set	no. of PLS2 factors	training set	test set
150–8–3[c]	2700	0.52	5.95	2	3.36	5.50
150–3[c]	5410	0.43	3.77	3[d]	2.12	4.10
150–8–1[c]	1170	0.50	7.07	4	1.13	4.30
150–1[c]	5050	0.50	4.12	5	1.06	4.37

Average of All Three Predictions

artificial neural networks				partial least squares 2		
		rel error of			rel error of	
architecture	no. of epochs[b]	training set	test set	no. of PLS2 factors	training set	test set
150–8–3[c]	2700	0.50	5.44	2	6.70	10.82
150–3[c]	5410	0.50	4.42	3[d]	2.64	6.69
150–8–1[c]	1170–5460	0.50	5.95	4	1.22	5.09
150–1[c]	3130–7470	0.50	4.51	5	0.87	5.33

[a] All models were calibrated using only the spectra of axenic suspensions of *S. aureus*, *B. subtilis*, and *E. coli*. Relative errors are given in % rms. [b] When the error of the training set was 0.5% rms. [c] Output node was scaled from −50 to 150. [d] Optimal number of PLS2 factors, predicted by Unscrambler II.

split into 10 equal parts (i.e., 0, 10, 20, ..., 100%) in order to get such neural networks to generalize well. To collect the pyrolysis mass spectra of 198 samples takes 5 h (198 × 1 min 50 s) and this is obviously not desirable. The question therefore arises as to whether fewer exemplars would produce accurate calibration models.

PLS2 with three variables in the Y matrix, and ANNs of architectures 150–8–3, 150–3, 150–8–1, and 150–1 were therefore calibrated with only the spectral data from the axenic (pure) bacterial suspensions of *S. aureus*, *B. subtilis*, and *E. coli*. The results, evaluated by calculating the rms error between the model's estimates and the true amount of determinand, for both the training and test sets for the individual bacteria and the average of all three predictions, are shown in Table 5. This table also includes the number of epochs needed to train the ANNs to 0.5% rms error in the training set, and not surprisingly after the great reduction in exemplars used to train the ANNs there was a huge decrease in training time. It can be seen that the percent rms error in the test sets were comparable for both ANNs and PLS2 and that although this error was higher than those observed from using all 66 samples (Table 4) ANNs and PLS2 could be used to give *relatively* accurate but *very rapid* estimates of the percentage amounts of *S. aureus*, *B. subtilis*, and *E. coli* present in the tertiary mixture.

In summary, we have shown that the combination of PyMS with multivariate analyses of ANNs and the linear regression methods of PLS and PCR were able quantitatively to analyze the PyMS of three binary mixtures of the protein lysozyme, RNA or DNA in glycogen, and accurately to predict the amounts of *S. aureus*, *B. subtilis*, and *E. coli* in a tertiary mixture. ANNs gave consistently better predictions than did the linear regression methods, presumably due to their ability to uncover *nonlinear* as well as *linear* relationships in pyrolysis mass spectral data. These nonlinearities may have arisen because during pyrolysis intermolecular reactions may have taken place in the pyrolysate.[26,27] It was also found that scaling the input layer on ANNs had a very marked effect on the time taken for the network to reach optimal generalization and that the preferred method for the quantitative analysis of the mass spectra from binary mixtures is to scale *each input node* individually so that the minimum and maximum values lay between 0.4 and 0.6. Finally, removing low-intensity masses had little effect on the accuracy of ANN predictions (but a substantial effect on the speed of learning) and slightly improved the ability of PLS calibrations. This implies that there was some noise in the mass spectral data to which the ANNs were robust but which the linear regression analyses must have incorporated into their calibration models.

We conclude that the combination of PyMS and ANNs constitutes a powerful and exciting methodology for the rapid analysis of mixtures and would be applicable to assessing the concentrations of appropriate substrates, metabolites, and products in biochemical processes generally.

ACKNOWLEDGMENT

We thank Liz Goodacre for her help with PyMS and Dr. Andy Woodward for useful discussions concerning the interpretation of multivariate linear regression techniques. This work is supported by the Biotechnology Directorate of the UK SERC, under the terms of the LINK scheme in Biochemical Engineering, in collaboration with Horizon Instruments, Neural Computer Sciences and Zeneca plc.

Received for review September 2, 1993. Accepted January 11, 1994.[⊗]

[⊗] Abstract published in *Advance ACS Abstracts*, February 15, 1994.