ELSEVIER

Chemometrics and Intelligent Laboratory Systems 34 (1996) 69–83

Chemometrics and intelligent laboratory systems

Plant seed classification using pyrolysis mass spectrometry with unsupervised learning: The application of auto-associative and Kohonen artificial neural networks

Royston Goodacre *, Joseph Pygall, Douglas B. Kell [1]

Institute of Biological Sciences, University of Wales, Aberystwyth, Dyfed SY23 3DA, Wales, UK

Received 3 November 1995; accepted 29 March 1996

Abstract

Pyrolysis mass spectrometry (PyMS) was used to gain high dimensional (150 m/z values) biochemical fingerprints from *Begonia semperflorens* Summer Rainbow, *Campanula carpatica* White Gem, *Lobelia erinus* White Fountain, and *Lobelia erinus* White Lady plant seeds. Rather than homogenizing the seeds and analysing the extracts, the sample preparation of the seeds in this study was novel and merely involved crimping the metal foil sample carrier around the seeds. Compared to extractive procedures the technique exploited in this study will give a fair representation of the seed, is rapid and thus amenable to the analysis of a high volume of samples. To observe the relationship between these seeds, based on their spectral fingerprints, it was necessary to reduce the dimensionality of these data by unsupervised feature extraction methods. The neural computational pattern recognition techniques of self organising feature maps (SOFMs) and auto-associative neural networks were therefore employed and the clusters observed compared with the groups obtained from the more conventional statistical approaches of principal components analysis (PCA) and canonical variates analysis (CVA). When PCA was used to analyze the raw pyrolysis mass spectra replicate samples were not recovered in discrete clusters; CVA, which minimises the within-group variance and maximises the between-group variance, therefore had to be employed. Although *B. semperflorens* and *C. carpatica* seeds were recovered separately and away from the *L. erinus* plant seeds, the two types of *L. erinus* seeds could still not be discriminated between using this approach. CVA uses a priori information on which spectra are replicates; we therefore encoded this information by employing a novel preprocessing regime where the triplicate mass spectra from each of the seeds were averaged in pairs to produce three new spectra; these were then used by each of the unsupervised methods. PCA still failed to separate the two *L. erinus*; however, auto-associative neural networks could be used successfully to discriminate them. It is likely that this was due to their ability to perform non-linear mappings and hence approximate non-linear PCA. SOFMs could also be used to separate all four seeds unequivocally. To obtain quantitative information regarding the similarity of these seeds from their pyrolysis mass spectra, SOFMs were trained with different numbers of nodes in the Kohonen output layers. The results observed from this procedure are often difficult to report in tables or visualise using topological contour maps; to simplify the graphical representation of the similarity between the seeds we therefore performed the novel construction of a dendrogram from the various SOFMs analyses. This study demonstrates the potential of PyMS for discriminating plant seeds at the genus, species and sub-species level. Moreover the clusters observed were a true reflection

* Corresponding author. Tel.: +44-1970-621947; fax: +44-1970-622354; e-mail: rrg@aber.ac.uk, http://gepasi.dbs.aber.ac.uk / roy / chemom.htm.

[1] E-mail: dbk@aber.ac.uk, http: //gepasi.dbs.aber.ac.uk/home.htm.

PII S0169-7439(96)00021-4

of the known taxonomy of these plants. This approach will be invaluable to the plant taxonomist in representing biological relationships among plant taxa or in describing genomic relationships without the need for cultivation of the propagule.

Keywords: Neural networks; Auto-associative neural networks; Feature extraction; Pyrolysis mass spectrometry; Seed typing; Self organising feature maps

1. Introduction

Morphological diversity is not an obvious property of many plant seeds. To examine the relationships between plants or to effect their identification, without recourse to lengthy cultivation periods it is necessary to type the propagule directly. If the typing procedure were rapid and automated it would be possible to use such an approach for the screening of large populations of plant seeds. Such a method could be exploited by the plant taxonomist for representing biological relationships among plant taxa or in describing genomic relationships without the need for the cultivation of the seeds. If cultivation were not feasible for whatever reason then such a procedure could also be employed for the classification of plants from their non-viable seeds. We therefore sought to assess whether such a method might have the appropriate discrimination.

Pyrolysis mass spectrometry (PyMS) presents itself as a tool capable of satisfying the above criteria. It is a rapid (< 2 min per sample) automated instrument-based physico-chemical technique which can establish the biochemical composition of complex non-volatile material such as plant seeds; it requires minimal sample preparation, can analyze samples directly and permits the acquisition of spectroscopic data from 300 or more samples per working day. Pyrolysis is the thermal degradation of complex material in an inert atmosphere or a vacuum. It causes molecules to cleave at their weakest points to produce smaller, volatile fragments called pyrolysate [1,2]. A mass spectrometer can then be used to separate the components of the pyrolysate on the basis of their mass-to-charge ratio (m/z). Almost all biological materials will produce pyrolytic degradation products such as methane, ammonia, water, methanol and H_2S, whose $m/z < 50$, and fragments with $m/z > 200$ are rarely analytically important for complex non-volatile material such as bacteria [3] unless very special conditions are employed [4]; the analytically useful multivariate data are then constituted by a set of 150 normalised intensities versus m/z in the range 51 to 200. These data may be plotted to produce a pyrolysis mass spectrum [2], which can then be used as a 'chemical profile' or fingerprint of the complex material analyzed.

Since PyMS measures the biochemical composition of samples analyzed it has been exploited by microbial systematics as a chemotaxonomic tool for the fine discrimination between bacteria and fungi at the genus, species and subspecies level [5–7]. PyMS is a very high-resolution technique because it is effectively studying the properties of a system in 150 dimensions (here the m/z values from 51–200) simultaneously; therefore within the plant sciences PyMS has also been used to characterise small changes in the lignin content of genetically engineered tobacco stems [8], to study the developmental stages of maize somatic embryos [9], and for uncovering the biochemical differences between grasses with different levels of resistance to attack by *Labops hesperius* [10].

Multivariate data, such as those generated by PyMS, consist of the results of observations of many different characters or variables (m/z or masses) for a number of individuals or objects (the seeds) [11]. Each mass may be regarded as constituting a different dimension, such that if there are n variables (where $n = 150$ masses) each object may be said to reside at a unique position in an abstract entity referred to as n-dimensional hyperspace. This hyperspace is necessarily difficult to visualise, and an underlying theme of multivariate analysis is thus *simplification* [12,13] or dimensionality reduction, which usually means that we want to summarise a large body of data by means of *relatively* few parameters, preferably the two or three which lend themselves to graphical display, with minimal loss of information.

Conventionally, at least within microbiology and biotechnology, because PyMS has been used as a taxonomic aid [5–7,14], the reduction of the multi-

variate data generated by the PyMS system is normally carried out using principal components analysis. This type of analysis falls into the category of 'unsupervised learning', in which the relevant multivariate algorithms seek 'clusters' in the data [15]. This allows the investigator to group objects together on the basis of their perceived closeness in the *n*-dimensional hyperspace referred to above. Such methods, then, although in some sense quantitative, are better seen as qualitative since their chief purpose is merely to *distinguish* objects or populations.

Recently there has been an interest in the use of neural computation methods which can also perform unsupervised learning on multivariate data; Wilkins et al. [16] have applied Kohonen [17] maps to multi-dimensional flow cytometric data for the identification of seven species of fresh water phytoplankton, and we have also exploited these self-organising feature maps (SOFMs) successfully to carry out unsupervised learning, and hence the classification of canine *Propionibacterium acnes* isolates [18] and *P. acnes* isolated from man [19]. Another neural network-based method for feature extraction called auto-associative neural networks [20–23] has been used to reduce the dimensionality of the infrared spectra of polysaccharides [24] and to detect plasmid instability using on-line measurements from an industrial fermentation producing a recombinant protein expressed by *Escherichia coli* [25].

The aim of this study was to use PyMS to examine *Begonia semperflorens*, *Campanula carpatica* and two varieties of *Lobelia erinus* plant seeds. Previous workers [26] have employed PyMS for the differentiation of seeds of three Triticeae species. This study, however, used an extremely tedious sample preparation method involving grinding the plant seeds, resuspending the powder in ethanol, sonicating for 30 min, and centrifugation; the supernatant was finally pipetted onto the sample carrier. Procedures of this nature are not only time consuming, and so not amenable to the analysis of a high volume of samples, but inevitably do not give a fair representation of the seed since some of its constituents will have been lost during the process. The sample preparation of the seeds in this study therefore merely involved crimping the metal foil sample carrier around the seeds.

Once data were collected neural computation methods exploiting unsupervised learning, viz. Kohonen's self organising feature maps and auto-associative neural networks, were employed to cluster the spectral data. The results obtained were compared with the conventional approaches of principal components analysis and canonical variates analysis. The clusters seen by all methods were similar, although the neural network-based methods gave less subjective groups. Finally, the classifications observed were in agreement with the known taxonomy of these seeds. Whilst we recognise that the number of objects analyzed in the present study was relatively small, they served more than adequately to illustrate the principles of our approach.

2. Experimental

2.1. Seed types

Details on the family, genus, species and variety of the four seed types that were used in this study are shown in Table 1. All seeds were purchased from Mr. Fothergill's Seeds, Kentford, Newmarket, Suffolk CB8 7QB, UK.

2.2. Pyrolysis mass spectrometry

Approximately 20 seeds were placed onto clean iron–nickel foils (Horizon Instruments, Ghyll Indus-

Table 1
Details of the plant seeds used in this study

Family	Genus	Species	Variety	Identifier in multivariate analyses
Begoniaceae	*Begonia*	*semperflorens*	Summer Rainbow	B
Campanulaceae	*Campanula*	*carpatica*	White Gem	D
Campanulaceae	*Lobelia*	*erinus*	White Fountain	A
Campanulaceae	*Lobelia*	*erinus*	White Lady	C

trial Estate, Heathfield, E. Sussex TN21 8AW, UK), the foils were then crimped around the seeds using clean forceps. These samples were then inserted into clean pyrolysis tubes (Horizon Instruments) and 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. Each of the four seed types were run in triplicate, thereby yielding 12 spectral fingerprints.

The pyrolysis mass spectrometer used in this study was the Horizon Instruments PYMS-200X. 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 [27,28] 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 ionisation chamber of the mass spectrometer. To minimize secondary fragmentation of the pyrolysate the ionisation method used was low voltage electron impact ionisation (25 eV). These conditions were employed because it has been found that the stated expansion chamber temperature gives the most reproducible spectra [29], whilst the spectra from samples ionised at 25 eV are much more robust to small changes in ionisation voltage than are those [2] obtained at lower ionisation voltages, whilst much higher ionisation voltages lead to excessive fragmentation. Non-ionised molecules were deposited on a cold trap, cooled by liquid nitrogen. The ionised fragments were focused by the electrostatic lens of a set of source electrodes, accelerated and directed into a quadrupole mass filter. The ions were separated by the quadrupole, on the basis of their mass-to-charge ratio, detected and amplified with an electron multiplier [30]. The mass spectrometer scans the ionised pyrolysate 160 times at 0.2 s intervals following pyrolysis. Data were collected over the m/z range 51 to 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

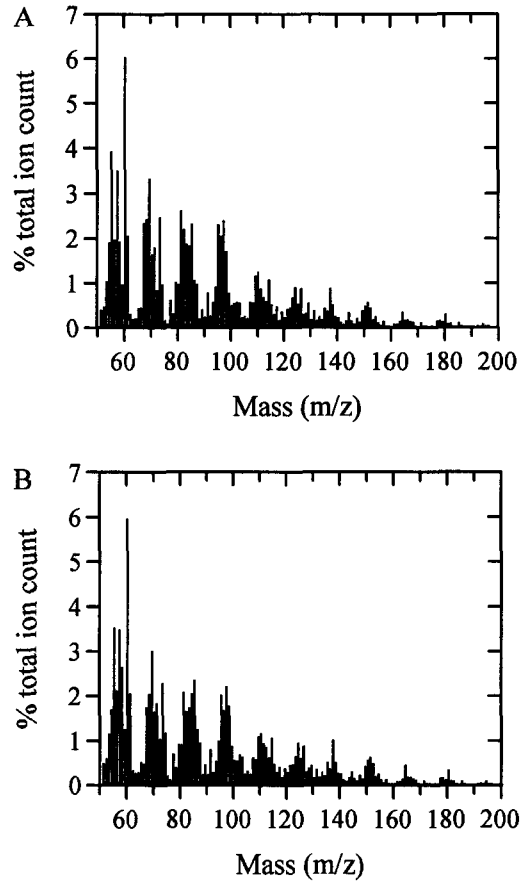

Fig. 1. Representative pyrolysis mass spectra of (A) *Lobelia erinus* White Fountain seeds and of (B) *Lobelia erinus* White Lady seeds.

(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 the abundance of m/z 181 was one tenth of that of m/z 69.

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

2.3. Principal components analysis

Principal components analysis (PCA) is a multivariate statistical technique which can be used to identify *correlations* amongst a set of variables (in this case 150 m/z intensities) and to transform the original set of variables to a new set of *uncorrelated* variables called principal components (PCs). For the present purpose, PCA can be thought of as finding a set of orthogonal axes in 150-dimensional space; these new axes (or PCs) are *linear* combinations of the original variables and are derived in decreasing order of importance; therefore the first PC accounts for the maximum variation among the samples, and subsequent PCs are chosen to account for progressively decreasing variance [11–13,15,31–33].

The objective of PCA is to see if the first few PCs account for most of the variation in the original data. If they do reduce the number of dimensions required to display the observed relationships, then the PCs can be plotted and 'clusters' may be found in the data. PCA is a variable-directed technique and therefore does not use any a priori knowledge of the groupings within samples (objects) in the data set, that is to say it is unsupervised; thus plots of PCs are thought to display the natural relationships between the samples.

To effect PCA the normalised data were processed with the GENSTAT package [34] run under Microsoft DOS 6.2 on an IBM-compatible PC; this has been previously described by MacFie and Gutteridge [35] and Gutteridge et al. [31].

2.4. Canonical variates analysis

Canonical variates analysis (CVA) is also a multivariate statistical technique, here carried out using the GENSTAT package. Before CVA is employed PCA is used to reduce the dimensionality of the data and only those principal components (PCs) whose eigenvalues accounted for more than 0.1% of the total variance are used. After the first few PCs, the axes generated will usually be due to random 'noise' in the data; these PCs can be ignored without reducing the amount of useful information representing the data, since each PC is now independent of (uncorrelated with) any other PC.

CVA then separated the objects (samples) into groups on the basis of the retained PCs and the a priori knowledge of the appropriate number of groupings [36,37]; this is achieved by minimising the within-group variance and maximising the between-group variance.

The principle of CVA is similar to PCA, but because the objective of CVA is to maximise the ratio of the between-group to within-group variance, a plot of the first two canonical variates (CVs) displays the best 2D representation of the *group* separation.

2.5. Non-linear principal components analysis

The method of analysing these pyrolysis mass spectral data by non-linear principal components analysis (NLPCA) was by using auto-associative artificial neural networks [20]. All ANN analyses were carried out under Microsoft Windows NT on an IBM-compatible PC. Data were normalised prior to analysis using the Microsoft Excel 4.0 spreadsheet. The back propagation neural network simulation program employed was WinNN version 0.93. (Dr. Yaron Danon, 14 Beman Lane, Troy New York 12180, USA. The program is available via ftp: ftp://sunsite.doc.ic.ac.uk/packages/windows3/programr/, the most recent file name to down load is winnn97.zip.)

The structure of the ANN used in this study to analyze pyrolysis mass spectra consisted of 5 layers containing processing nodes (neurons or units) made up of the 150 input nodes (normalised pyrolysis mass spectra), 150 output nodes (normalised pyrolysis mass spectra), and three 'hidden' layers containing 8, 2 and 8 nodes respectively; this may be represented as a 150–8–2–8–150 architecture (Fig. 2). This ANN can be referred to as a fully interconnected feedforward multilayer perceptron where each of the layers of nodes was connected to the next (hidden) layer using abstract interconnections (connections or synapses). Connections each have an associated real value, termed the weight, that scale signals passing through them. Nodes in the hidden layers and output layer sum the signals feeding to them and output this sum to each driven connection scaled by a 'squashing' function (f) with a sigmoidal shape:

$$f = 1/(1 + e^{-x}), \quad (1)$$

where $x = \Sigma$inputs.

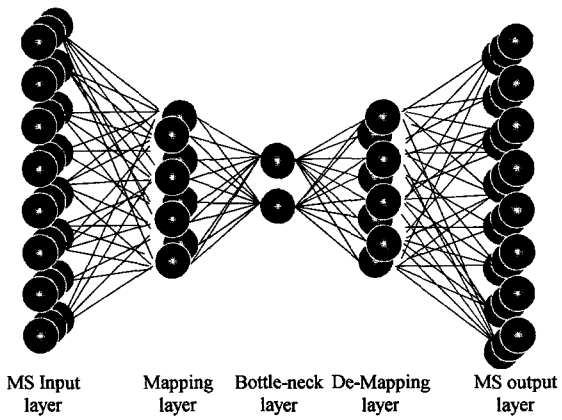

Fig. 2. Architecture of an auto-associative neural network consisting of 5 layers. In the architecture shown, adjacent layers of the network are fully interconnected. The input and output layer are presented with identical PyMS data (in this figure there are 24 nodes in these layers; in the present work the number of nodes was actually 150 inputs/masses). A key feature of the auto-associative network is the data compression in the middle (third) bottle-neck layer of 2 nodes. The second and fourth layers each consisted of 8 nodes and these map and de-map the mass spectra allowing feature extraction in the bottle neck layer; this is equivalent to non-linear principal components analysis.

These signals are then passed to the next layer which sums them and in turn squashed by the sigmoidal activation function (for a diagrammatic representation see Fig. 3); the product of the final layer of nodes was then fed to the 'outside world'. In previous studies [38–42] we have used back-propagation ANNs for the quantification and identification of biological systems from their PyMS spectra. The topology of these ANNs was 150–8–x (where x is the number of determinants). Rather than 8 nodes being chosen randomly it was found that by altering the number of nodes, 8 was normally found to give optimum generalisation. Therefore, in this study 8 nodes were chosen for the mapping and demapping layers of our 150–8–2–8–150 ANNs. It is possible that if more samples were to be analyzed by these auto-associative ANNs, then more nodes may be needed to have enough degrees of freedom. Another study using auto-associative ANNs for the classification of animal cell-lines was also successful when 8 nodes were used [43].

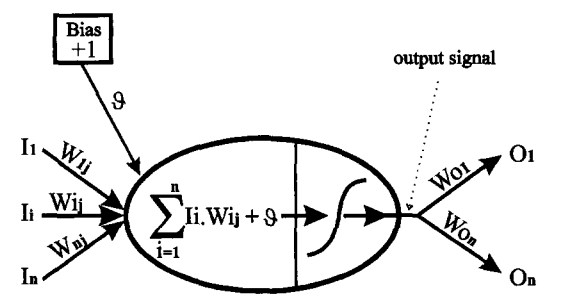

Fig. 3. The information processing by a node in one of the hidden layers or output layer. An individual node sums its input (the Σ function) from nodes in the previous layer, including the bias (ϑ), transforms them via a 'sigmoidal' squashing function, and outputs them to the next node to which it is linked via a connection weight.

Before training commenced the values applied to the input and output nodes were normalised across the whole mass range such that the lowest ion count was set to 0 and the highest to 1. Finally, the connection weights were set to small random values (typically between -0.0001 and $+0.0001$).

The algorithm used to train the neural network was the standard back-propagation (BP) [44–46]. For the training of the ANN each input (i.e., normalised pyrolysis mass spectrum) is paired with a desired output (i.e., the same pyrolysis mass spectrum); 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 $\alpha = 0.05$ and a momentum of 0.9. The reason a relatively small α was employed was because when the learning rate was larger, typically $0.1 \leq \alpha \leq 0.2$, these auto-associative ANNs failed to learn. When the connection weights between the layers were examined it was found that the back-propagation algorithm had set them to either very large positive or very large negative values. This saturation on the weights meant that the ANNs had got stuck and could not learn. The same phenomenon was found when the initial random weights were set $> \pm 0.0001$; this is why these extremely small starting weights were employed.

Each epoch represents the connection weight updatings and a recalculation of the average root mean squared (rms) error between the true and desired outputs (mass spectra) over the entire training set.

average rms error

$$= \sum_{i=0}^{n} \left[\left(\sum_{j=0}^{m} (\text{observed}_i - \text{expected}_i)^2 \right) / 150 \right] / n, \quad (2)$$

where n = number of objects in training set, and m = number of determinants in output layer (for these ANNs $m = 150$).

During training a plot of the error versus the number of epochs represents the 'learning curve', and may be used to estimate the extent of training. Training may be said to have finished when the network has found the lowest error. Provided the network has not become stuck in a local minimum, this point is referred to as the global minimum on the error surface. In the present experiment we trained the auto-associative neural network until the rms error was 0.005; this took approximately 3×10^4 epochs.

After training each of the pyrolysis mass spectra were applied in turn to the input layer and the activation on the two nodes in the 'bottle-neck' layer calculated. The compression of the 150 inputs through only two nodes in the middle layer allows NLPCA to be performed; a biplot of the activations of the first node in the 'bottle-neck' layer against the second node's activations therefore allow 'clusters' to be found in the data. For a more detailed account of this data compression through the 'bottle-neck' layer please refer to Kramer [20].

2.6. Kohonen artificial neural networks

All self-organizing feature maps (SOFMs) analyses were run under Microsoft Windows NT on an IBM-compatible PC using software written by Dr. Mark Neal (Institute of Biological Sciences, University of Wales, Aberystwyth, Dyfed SY23 3DA, Wales, UK) in Microsoft Visual C^{++} according to the general principles outlined by Kohonen [17].

KANNs provide a way of classifying data through self-organising networks of artificial neurons. The SOFMs used in this work consisted of a two-dimensional network of neurons arranged on a square or rectangle grid (Fig. 4). Each neuron was connected to its eight nearest neighbours on the grid. The neurons store a set of weights (a weight vector) each of which corresponds to one of the inputs in the data. Thus, for PyMS data consisting of 150 ion counts each node stores 150 weights in its weight vector. Upon presentation of a mass spectrum (represented as a vector consisting of the 150 ion counts) to the network each neuron calculates its 'activation level'. A node's activation level is defined as

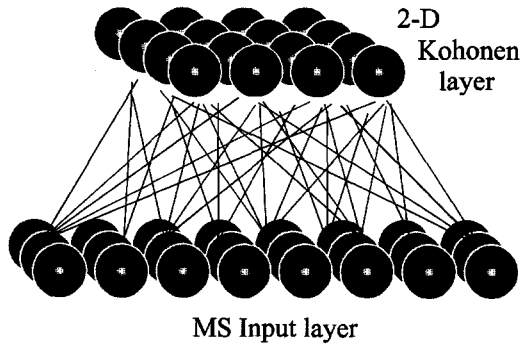

Fig. 4. A simplified Kohonen artificial neural network. Nodes in the two-dimensional Kohonen layer are interconnected with each other, such that an activation node tends to activate surrounding nodes also. The PyMS data are applied to the input layer (represented here by 24 nodes; in the present work the number of nodes was actually 150 inputs/masses) which activates a node or group of neighbouring nodes in the Kohonen layer (represented here as having 4×4 nodes; the number of nodes was varied to allow quantitative information to be extracted).

$$\sqrt{\sum_{i=0}^{n} (\text{weight}_i - \text{input}_i)^2}. \quad (3)$$

This is simply the Euclidean distance between the points represented by the weight vector and the input vector in n-dimensional space. Thus a node whose weight vector closely matches the input vector will have a small activation level, and a node whose weight vector is very different from the input vector will have a large activation level. The node in the network with the smallest activation level is deemed to be the 'winner' for the current input vector.

During the training process the network is presented with each input pattern in turn, and all the nodes calculate their activation levels as described above. The nodes included in the set which are allowed to adjust their weights are said to belong to the

'neighbourhood' of the winner. The winning node and some of the nodes around it are then allowed to adjust their weight vectors to match the current input vector more closely by an amount depending upon the distance from the most active node, the current size of the neighbourhood and the current value of α. This is the usual triangular shape [47]; thus if the neighbourhood size is 2 then the winning node can update its weights by $1 \times \alpha$, and the surrounding 8 nodes by $0.5 \times \alpha$; likewise if the neighbourhood size is 3 then the winning node can update its weights by $1 \times \alpha$, and the surrounding 8 nodes by $0.66' \times \alpha$, and the 16 outer nodes by $0.33' \times \alpha$. The size of the winner's neighbourhood is varied throughout the training process. Initially all of the nodes in the network are included in the neighbourhood of the winner, but as training proceeds the size of the neighbourhood is decreased linearly after each presentation of the complete 'training set' (all the mass spectra being analyzed), until it includes only the winner itself. The amount by which the nodes in the neighbourhood are allowed to adjust their weights is also reduced linearly through the training period.

The factor which governs the size of the weight alterations is known as the learning rate and is represented by α. The adjustments to each item in the weight vector (where δw is the change in the weight) are made in accordance with the following:

$$\delta w_i = -\alpha(w_i - i_i). \quad (4)$$

This is carried out for $i = 1$ to $i = n$ where in this case $n = 150$. The initial value for α is 1 and the final value is 0.

The effect of the 'learning rule' (weight update algorithm) is to distribute the neurons evenly throughout the region of n-dimensional space populated by the training set [17,48,49]. This effect is displayed in Fig. 5 which shows the distribution of a square network over an evenly populated two-dimensional square input space. The neuron with the weight vector closest to a given input pattern will win for that pattern and for any other input patterns that it is closest to. Input patterns which allow the same node to win are then deemed to be in the same group, and when a map of their relationship is drawn a line encloses them. By training with networks of increasing size a map with several levels of groups or 'contours' can be drawn. These contours, however, may sometimes cross — this appears to be due to failure of the SOFM to converge to an even distribution of neurons over the input space [50].

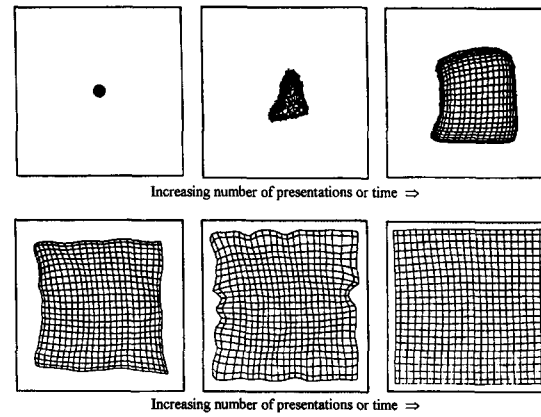

Fig. 5. Provided the input data (mass spectra) are evenly populated then after training the input space will be evenly covered with nodes in the Kohonen layer; thus as training proceeds the input space is mapped. This is represented here as a projection into two-dimensional space of a 24×24 square SOFM distributed across an evenly distributed input space (i.e., the samples evenly populate the input data). The weights to 24×24 output nodes from two input nodes (i and j) change as the number of presentations increases and a feature map is formed. The abscissa represents the value of the weight from input i and the ordinate represents the value of the weight from input j. Line intersections specify the two weights from each node. Lines connect weights for nodes that are nearest neighbours.

Construction of these maps allows close examination of the relationships between the items in the training set, which in this case consisted of the 12 normalised pyrolysis mass spectra derived from the seeds. Networks on grids of 1×1, 1×2, 2×2, 3×3, 4×4, 5×5, 6×6 and 7×7 nodes were trained for 750 epochs and used to group the samples. The SOFMs were allowed to 'wrap around' so that they formed toroidal structures; this was in order to avoid the edge effects which would otherwise tend to corrupt very small networks of this type. Although using toroidal KANNs means that the maximum topological distance for a $N \times N$ SOFM is decreased from N to $N/2$. This does not limit the discriminatory ability because a succession of larger KANNs was employed to assign quantitative differences and hence clustering. Indeed, it is true to say that it is very difficult to assign quantitative meaning to a single KANN.

3. Results and discussion

After the collection of pyrolysis mass spectra the first stage was to perform the unsupervised learning method (linear) principal components analysis (PCA), as described above, using the GENSTAT package to establish the *natural* relationships between the spectra of the twelve samples. The resulting PCA plot is shown in Fig. 6; the first two principal components (PCs) are displayed and they account for 53.5% and 17.9% (71.4% total) of the total variation respectively. In this figure it can be seen that the replicate spectra do not group together well; in particular, replicates from *L. erinus* White Fountain (A) and *C. carpatica* (D) are disperse and overlap other samples. PCA alone could therefore not be used to analyze these data and additional methods are needed.

Classically when PyMS has been used to discriminate between bacteria, PCA is first used as a linear dimensionality reduction step, which is thought to remove any collinearity or noisy variables (masses). After PCA those PCs whose eigenvalues accounted for more than 0.1% of the total variance are used by the canonical variates analysis (CVA) procedure together with knowledge of which samples are replicates. CVA separates the samples into groups by minimising the within-group variance and maximising the between-group variance. This procedure was performed on our data set using the first six PCs; although this dimensionality reduction was from 150 to only six, 97.6% of the total (linear) variance was preserved. Next new canonical variates (CVs) were calculated and plotted (Fig. 7). In this ordination plot the replicates now cluster together allowing the true differences between the seeds to be observed. As one might expect from the known taxonomy of these plants three groups are observed; *B. semperflorens* (B) and *C. carpatica* (D) are recovered separately from the closely related *L. erinus* plants (A and C). On closer inspection of this plot *L. erinus* White Fountain (A) and *L. erinus* White Lady (C) might appear to be discriminated. However, CVA results can be interpreted statistically to discriminate populations based on the 95% tolerance region constructed around each population mean by the χ-squared distribution on two degrees of freedom [18,51]. This area can be represented by drawing a circle of radius 2.448 canonical variates (CV) units, which means that group means must be separated by more than 4.9 CV units. That A and C were separated by only 2.05 CVs indicates that the separation of these groups is not statistically valid.

The next stage was to attempt to use unsupervised learning methods based on artificial neural networks

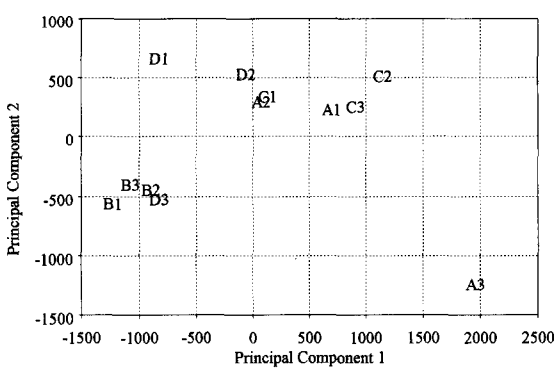

Fig. 6. Principal components biplot based on PyMS data analyzed by GENSTAT showing the relationship between the four seed types. The first two principal components are displayed and they account for 53.5% and 17.9% (71.4% total) of the total variation respectively. A = *L. erinus* White Fountain, B = *B. semperflorens* Summer Rainbow, C = *L. erinus* White Lady, and D = *C. carpatica* White Gem; the 1, 2 and 3 represent the replicates.

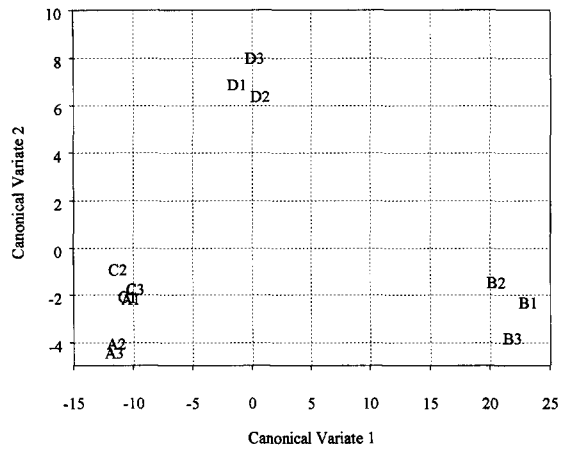

Fig. 7. Canonical variates analysis biplot based on the first 6 principal components (those whose eigenvalues accounted for more than 0.1% of the total variance, the total variance accounted for in these 6 PCs was 97.6%) based on PyMS data analyzed by GENSTAT showing the relationship between the four seed types. A = *L. erinus* White Fountain, B = *B. semperflorens* Summer Rainbow, C = *L. erinus* White Lady, and D = *C. carpatica* White Gem; the 1, 2 and 3 represent the replicates.

to cluster the pyrolysis mass spectra of the seeds. Kohonen neural networks (KANNs) and non-linear principal components analysis (NLPCA) (the latter calculated using auto-associative neural networks) were applied to the 12 spectra. The results showed (data not shown) that the three replicate spectra from each of the plant seeds failed to group together; moreover the clusters obtained were similar to the results from PCA (Fig. 6). KANNs, NLPCA, and PCA are based on unsupervised methods, where the system is shown a set of inputs (i.e., mass spectra) and the relevant multivariate algorithms seek 'clusters' in the data [15], thereby allowing one to group objects together on the basis of their perceived closeness. Thus the chief purpose of such methods is merely to *distinguish* objects or populations in an un-biased way. Since PCA also failed to give the correct clustering, rather than a failure of the unsupervised clustering by neural networks these results may have arisen because of spectral variation; this was attributed either to spectral drift, which is unlikely given the rather short analysis time, or more likely noise associated with the lower intensity masses.

To account for the spectral variation seen above so as to observe the differences between the seeds (and not the differences between the replicate samples), we can, like CVA, use the a priori information on which spectra are replicates. Therefore the triplicate spectra from each of the seeds were averaged in pairs to produce three new spectra. That is to say, for replicate spectra 1, 2 and 3 new spectra are generated from the average (1, 2), average (1, 3) and average (2, 3).

After this preprocessing step the new spectra were first analyzed by PCA using the GENSTAT package. The first three PCs are displayed as a pseudo-3D plot in Fig. 8 and they account for 64.8%, 17.6% and 7.9% (90.3% total) of the total variation, respectively. In contrast with PCA on the raw spectra (Fig. 6), but in agreement with CVA (Fig. 7), three groups are seen; the *B. semperflorens* (B) and *C. carpatica* (D) are recovered separately from the two *L. erinus* (A and C). These two *L. erinus* species cluster closely and can only really be separated when a priori information is at hand, that is to say there is a human subjective bias. This PCA plot reflects the taxonomy of these plants and demonstrates that the preprocessing has allowed the relationships between the seeds to be

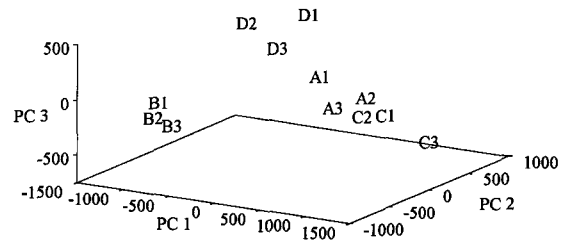

Fig. 8. Pseudo-three-dimensional principal components plots based on PyMS data analyzed by GENSTAT showing the relationship between the four seed types after the mass spectra had been averaged, as detailed in the text. The first three principal components are displayed and they account for 64.8%, 17.6% and 7.9% (90.3% total) of the total variation respectively. A = *L. erinus* White Fountain, B = *B. semperflorens* Summer Rainbow, C = *L. erinus* White Lady, and D = *C. carpatica* White Gem; the 1, 2 and 3 represent the replicates.

elucidated. It is noteworthy that before preprocessing the first two PCs accounted for 71.4% of the total variance and 82.4% after processing; it is likely that this is because the noise associated with the low intensity masses is removed partially (or wholly) by the averaging procedure.

One problem with PCA and CVA is that these clustering algorithms (a) rely on linear (orthogonal) transformations of the raw multivariate data and so cannot provide the truly best analytical discriminations for systems that contain non-linearities, and (b) are subjective because they rely on the interpretation of complicated scatter plots. The next stage was therefore to investigate new neural computational methods based on auto-associative neural networks which calculate non linear PCA, and Kohonen's self-organising feature maps which have the potential to group pyrolysis mass spectra both automatically and relatively objectively [18,19].

The architecture of the auto-associative neural networks employed was 150–8–2–8–150 (illustrated in Fig. 2), the training pairs consisted of the *same* normalised pyrolysis mass spectrum, and the 12 spectra were then applied in turn to the 150 input and 150 output nodes. These multi-layered perceptrons were trained as described above until the rms error was 0.005; this took approximately 3×10^4 epochs.

After training to this point each of the pyrolysis mass spectra were applied to the input layer of the auto-associative ANN and the activation on the two nodes in the 'bottle-neck' layer calculated. Fig. 9 is a

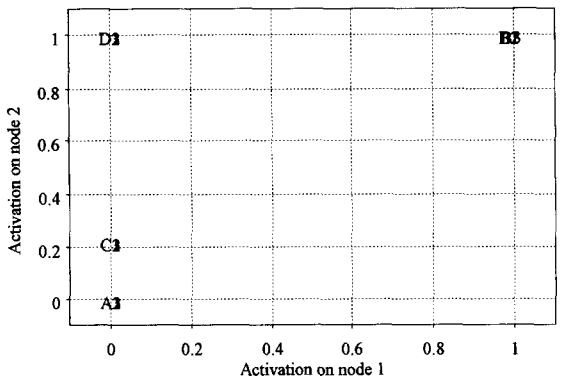

Fig. 9. Non-linear principal components plot based on PyMS data analyzed by the 150–8–2–8–150 auto-associative neural network showing the relationship between the four seed types. The activations of the two nodes in the bottle-neck layer are shown. A = *L. erinus* White Fountain, B = *B. semperflorens* Summer Rainbow, C = *L. erinus* White Lady, and D = *C. carpatica* White Gem; the 1, 2 and 3 (which are virtually superimposed) represent the replicates. ANNs were trained using the standard-back propagation algorithm, to a RMS error of 0.005 which typically took 3×10^4 epochs.

plot of the activations of the first node against the activations of the second node for each of the 12 spectra. It can be seen that the three replicates for each seed now superimpose and that all *four* seeds are easily discriminated. This NLPCA plot also highlights that the two *L. erinus* seeds A and C are closely related and different from the *B. semperflorens* (B) and the *C. carpatica* (D) seeds. This result demonstrates that PyMS and auto-associative neural networks can be used to classify these plant seeds.

The next stage was to use KANNs to cluster the seeds based on their mass spectra. Provided that the training conditions for SOFMS are kept constant, then KANNs can be used to create automatic grouping of data, thereby removing any human (which may be potentially biased) interpretation. Although a single KANNs provides no truly quantitative information about the similarity of samples within groups, they do provide qualitative information about the groups present. By using Kohonen layers of increasing sizes, finer discriminations may be sought and therefore some quantitative information can be gained [18,19]. As described above, networks with Kohonen layers of 1×1, 1×2, 2×2, 3×3, 4×4, 5×5, 6×6 and 7×7 nodes were used to group the samples. These SOFMS were trained for 750 epochs and details of the clusters formed at the eight different discrimination levels are given in Table 2.

In the past the clusters formed using SOFMs have been displayed either tabulated as detailed in Table 2, or as a rather complex 'topological contour map' shown in Fig. 10. When analysing many spectra these contour maps are often difficult to interpret and it is therefore necessary to display the results in a more simplified graphical representation. It is evident from Table 2 and Fig. 10 that quantitative information on

Table 2
Groups produced by self-organizing feature maps trained on pyrolysis mass spectral data of the plant seeds

Plant seed type		Groups formed at the following Kohonen layer sizes [a]							
		1 × 1	1 × 2	2 × 2	3 × 3	4 × 4	5 × 5	6 × 6	7 × 7
Lobelia WF	A1	1	1	1	1	1	1	1	1
	A2	1	1	1	1	1	2	2	2
	A3	1	1	1	1	1	3	3	3
Begonia	B1	1	2	2	2	2	4	4	4
	B2	1	2	2	2	2	4	5	5
	B3	1	2	2	2	2	4	6	6
Lobelia WL	C1	1	1	1	3	3	5	7	7
	C2	1	1	1	3	3	6	7	8
	C3	1	1	1	3	4	7	8	9
Campanula	D1	1	2	3	4	5	8	9	10
	D2	1	2	3	5	6	9	10	11
	D3	1	2	3	5	7	10	11	12

[a] The identifiers indicate which spectra were described by the same node in the Kohonen layer, hence which pyrolysis mass spectra group together.

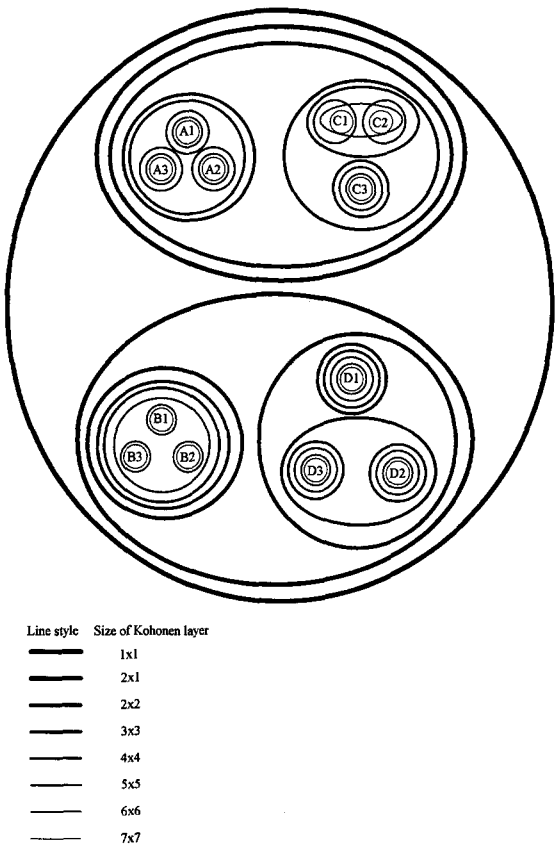

Fig. 10. Topological contour map of groups from SOFMs trained with PyMS data of the plant seeds. A = *L. erinus* White Fountain, B = *B. semperflorens* Summer Rainbow, C = *L. erinus* White Lady, and D = *C. carpatica* White Gem; the 1, 2 and 3 represent the replicates. The map is correct only in topology.

seed relationships can be elucidated, therefore it should be possible to depict these details in a dendrogram format.

The construction of the dendrogram (Fig. 11) begins when only a single node is used in the Kohonen layer and all 12 spectra necessarily group together; this is drawn as a single line in the 1×1 size of Kohonen layer zone. The information in Table 2 tells us that when two nodes are employed spectra from A and C group together in a single group and B and D form another discrete cluster (this is also displayed in Fig. 10); therefore in the dendrogram the single line in the 1×1 zone can be split into two in the 2×1 size of Kohonen layer region. The first line contains all the spectra from A and all the spectra from C, likewise the second line contains all the spectra from B and D. Next the number of nodes is increased to four, the two *L. erinus* seeds (A and C) still cluster together but the *B. semperflorens* (B) and *C. carpatica* (D) seeds are now separated (Table 2, Fig. 10); in the dendrogram this information is depicted as the line from the A and C spectra staying together and the line from the B and D spectra splitting in the 2×2 zone. Only when the number of nodes in the Kohonen layer is increased to nine (3×3 zone in Fig. 11) can the two *L. erinus* seeds be discriminated. By progressively increasing the number of nodes in the output layer of the SOFM more detailed discriminations are found, these are shown in the dendrogram (Fig. 11) and the contour map, (Fig. 10), as well as in Table 2. Finally, when the number of nodes in the output layer was 49 all the spectra were recovered separately.

It was interesting to observe the rather peculiar behaviour of C1 and C2; in the dendrogram (Fig. 10) these spectra were recovered separately with KANNs using 16 and 25 nodes, however when 36 nodes are used they clustered together and then ungrouped when 49 nodes are used. One could speculate as to possible reasons that this behaviour occurs: (1) different weight randomisations, this is unlikely because the behaviour observed was reproducible; al-

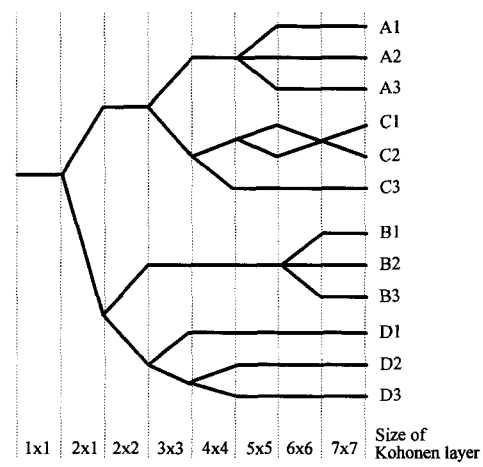

Fig. 11. Dendrogram produced using self organizing feature maps trained with PyMS data showing the relationship between the four seed types. Networks on grids of 1×1, 1×2, 2×2, 3×3, 4×4, 5×5, 6×6 and 7×7 nodes were trained for 750 epochs. Details of how the dendrogram was constructed are given in the text.

ternatively it could be due to (2) different input order of objects, (3) an improper variation of the learning rate α, and finally (4) an insufficient training period. A combination of the above four reasons could cause the odd behaviour of C1 and C2, by contrast this could just be a data dependent phenomenon.

The construction of this dendrogram is novel in that it is the first demonstration of the visual simplification of the groups from employing a range of SOFMs with different Kohonen layers for the analysis of pyrolysis mass spectra. Furthermore, results of feature extraction depicted in dendrograms are easier to interpret than either tabulated results or topological contour maps.

This results clearly show that PyMS and neural networks, carrying out unsupervised learning, can be employed to discriminate between plant seeds at the genus and species level and that the classification observed is congruent with the known plant taxonomy. Furthermore, it was encouraging that our very simple, albeit crude, method of sample preparation allowed the seeds to be classified successfully, without recourse to the laborious protocol used by other workers [26], which only analyses part of the plant seed.

4. Concluding remarks

Artificial neural network pattern recognition techniques based on unsupervised learning were compared with the statistical approaches of principal components analysis and the supervised method of canonical variates analysis (also referred to as discriminant analysis) for the analysis of the pyrolysis mass spectra of *Lobelia erinus* White Fountain, *Lobelia erinus* White Lady, *Begonia semperflorens* Summer Rainbow, and *Campanula carpatica* White Gem plant seeds.

When PCA was used on the raw pyrolysis mass spectra replicate samples were not recovered in discrete clusters; CVA, which minimises the within-group variance and maximises the between-group variance, had to be employed. Although *B. semperflorens* and *C. carpatica* seeds were recovered separately and away from the *L. erinus* plant seeds, the two sub-species of *L. erinus* seeds could still not be discriminated between using CVA.

CVA uses a priori information on which spectra are replicates; we therefore encoded this information by preprocessing the mass spectra prior to analysis by each of the unsupervised methods; the triplicate spectra from each of the seeds were averaged in pairs to produce three new spectra. PCA still failed to separate the two *L. erinus*; however, auto-associative neural networks could be used successfully to discriminate them. It is likely that this was due to their ability to perform non-linear mappings and hence approximate non-linear PCA [20].

Self-organizing feature maps could also be used to separate all four seeds unequivocally; by increasing the size of the Kohonen output layer, and the novel construction of a dendrogram, quantitative information regarding the similarity of this group of seeds was elucidated.

This study demonstrates the potential of PyMS for discriminating plant seeds down to the sub-species level. Moreover the clusters observed were a true reflection of the known taxonomy of these plants, we feel therefore that this approach will be valuable to the plant taxonomist in representing biological relationships among plant taxa or in describing genomic relationships without the need for cultivation of the seed. Indeed, if cultivation were not possible then this approach could still be used to type the non-viable or dormant plant.

The major advantages that PyMS offers over more conventional methods are its speed, sensitivity and the ability to analyze many hundreds of samples per day. This study also analysed the whole seed without recourse to lengthy sample preparation procedures, therefore, one could employ this technique to screen larger populations of plant seeds. We conclude that the combination of PyMS and ANNs can provide an objective, rapid and accurate discriminatory technique for plant seed typing.

Acknowledgements

We thank Dr. Mark Neal for very useful discussions. We are also grateful to the reviewers for their useful comments on this paper. R.G. is funded as a research fellow by the Wellcome Trust grant number

042615/Z/94/Z. D.B.K. thanks the Chemicals and Pharmaceuticals Directorate of the UK BBSRC for financial support under the terms of the LINK scheme in Biochemical Engineering, in collaboration with Horizon Instruments, Neural Computer Sciences and Zeneca Bioproducts plc.

References

[1] W.J. Irwin, Analytical Pyrolysis: A Comprehensive Guide (Marcel Dekker, New York, 1982).
[2] H.L.C. Meuzelaar, J. Haverkamp and F.D. Hileman, Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials (Elsevier, Amsterdam, 1982).
[3] R.C.W. Berkeley, R. Goodacre, R.J. Helyer and T. Kelley, Lab. Pract. 39 (1990) 81–83.
[4] P.B. Smith and A.P. Snyder, J. Anal. Appl. Pyrolysis 24 (1993) 199–210.
[5] J.T. Magee, Whole-organism fingerprinting, in: Handbook of New Bacterial Systematics, ed. M. Goodfellow and A.G. O'Donnell (Academic Press, London, 1993) pp. 383–427.
[6] A.P. Snyder, P.B.W. Smith, J.P. Dworzanski and H.L.C. Meuzelaar, Pyrolysis-gas chromatography-mass spectrometry – detection of biological warfare agents, ACS Symposium Series, Vol. 541 (1994) pp. 62–84.
[7] R. Goodacre, Microbiol. Eur. 2 (1994) 16–22.
[8] C. Halpin, M.E. Knight, G.A. Foxon, M.M. Campbell, A.M. Boudet, J.J. Boon, B. Chabbert, M.T. Tollier and W. Schuch, Plant J. 6 (1994) 339–350.
[9] A.M.C. Emons, M.M. Mulder and H. Kieft, Acta Bot. Neerl. 42 (1993) 319–339.
[10] W. Windig, H.L.C. Meuzelaar, B.A. Haws, W.F. Campbell and K.H. Asay, J. Anal. Appl. Pyrolysis 5 (1983) 183–198.
[11] H. Martens and T. Næs, Multivariate calibration (John Wiley, Chichester, 1989).
[12] C. Chatfield and A.J. Collins, Introduction to Multivariate Analysis (Chapman and Hall, London, 1980).
[13] I.T. Jolliffe, Principal Component Analysis (Springer-Verlag, New York, 1986).
[14] C.S. Gutteridge, Methods Microbiol. 19 (1987) 227–272.
[15] B.S. Everitt, Cluster Analysis (Edward Arnold, London, 1993).
[16] M.F. Wilkins, L. Boddy and C.W. Morris, Binary Comput. Microbiol. 6 (1994) 64–72.
[17] T. Kohonen, Self-Organization and Associative Memory (Springer-Verlag, Berlin, 1989).
[18] R. Goodacre, M.J. Neal, D.B. Kell, L.W. Greenham, W.C. Noble and R.G. Harvey, J. Appl. Bacteriol. 76 (1994) 124–134.
[19] R. Goodacre, S.A. Howell, W.C. Noble and M.J. Neal, Zentralbl. Bakteriol. (1996), in press.
[20] M.A. Kramer, AIChE J. 37 (1991) 233–243.
[21] M.A. Kramer, Comput. Chem. Eng. 16 (1992) 313–328.
[22] J.A. Leonard and M.A. Kramer, IEEE Expert Intelligent Systems Appl. 8 (1993) 44–53.
[23] D.R. Kuespert and T.J. McAvoy, Chem. Eng. Commun. 130 (1994) 251–264.
[24] S.P. Jacobsson, Anal. Chim. Acta 291 (1994) 19–27.
[25] G. Montague and J. Morris, Trends Biotechnol. 12 (1994) 312–324.
[26] R. Valcarce, G.G. Smith, D.N. Stevenson and K.H. Asay, Chemom. Intell. Lab. Syst. 9 (1990) 95–105.
[27] W. Windig, P.G. Kistemaker, J. Haverkamp and H.L.C. Meuzelaar, J. Anal. Appl. Pyrolysis 2 (1980) 7–18.
[28] R. Goodacre, Ph.D. thesis, University of Bristol (1992).
[29] W. Windig, P.G. Kistemaker, J. Haverkamp and H.L.C. Meuzelaar, J. Anal. Appl. Pyrolysis 1 (1979) 39–52.
[30] J.R. Chapman, Practical Organic Mass Spectrometry (Wiley and Sons, New York, 1993).
[31] C.S. Gutteridge, L. Vallis and H.J.H. MacFie, Numerical methods in the classification of microorganisms by pyrolysis mass spectrometry, in: Computer-assisted Bacterial Systematics, ed. M. Goodfellow, D. Jones and F. Priest (Academic Press, London, 1985) pp. 369–401.
[32] D.R. Causton, A Biologist's Advanced Mathematics (Allen and Unwin, London, 1987).
[33] B. Flury and H. Riedwyl, Multivariate Statistics: A Practical Approach (Chapman and Hall, London, 1988).
[34] J.A. Nelder, Genstat Reference Manual, Scientific and Social Service Program Library, University of Edinburgh (1979).
[35] H.J.H. MacFie and C.S. Gutteridge, J. Anal. Appl. Pyrolysis 4 (1982) 175–204.
[36] H.J.H. MacFie, C.S. Gutteridge and J.R. Norris, J. Gen. Microbiol. 104 (1978) 67–74.
[37] W. Windig, J. Haverkamp and P.G. Kistemaker, Anal. Chem. 55 (1983) 81–88.
[38] R. Goodacre, A.N. Edmonds and D.B. Kell, J. Anal. Appl. Pyrolysis 26 (1993) 93–114.
[39] R. Goodacre and D.B. Kell, Anal. Chim. Acta 279 (1993) 17–26.
[40] R. Goodacre, S. Trew, C. Wrigley-Jones, G. Saunders, M.J. Neal, N. Porter and D.B. Kell, Anal. Chim. Acta 313 (1995) 25–43.
[41] R. Goodacre, M.J. Neal and D.B. Kell, Anal. Chem. 66 (1994) 1070–1085.
[42] R. Goodacre and D.B. Kell, Current Opinion Biotechnol. 7 (1996) 20–28.
[43] R. Goodacre, D.J. Rischert, P.M. Evans and D.B. Kell, Cytotechnology, (1996) in press.
[44] Y. Chauvin and D.E. Rumelhart, ed., Backpropagation: Theory, Architectures, and Applications (Erlbaum, Hove, 1995).
[45] D.E. Rumelhart, J.L. McClelland and The PDP Research Group, Parallel Distributed Processing, Experiments in the Microstructure of Cognition, Vol. I and II (MIT Press, Cambridge, MA, 1986).
[46] P.J. Werbos, The roots of back-propagation: from ordered derivatives to neural networks and political forecasting (John Wiley, Chichester, 1994).

[47] J. Zupan and J. Gasteiger, Neural Networks for Chemists: An Introduction (VCH Verlagsgesellschaft, Weinheim, 1993).
[48] R. Hecht-Nielsen, Neurocomputing (Addison-Wesley, MA, 1990).
[49] J. Hertz, A. Krogh and R.G. Palmer, Introduction to the Theory of Neural Computation (Addison-Wesley, CA, 1991).
[50] E. Erwin, K. Obermayer and K. Schulten, Biol. Cybernetics 67 (1992) 47–55.
[51] W.J. Krzanowski, Principles of Multivariate Analysis: A User's Perspective (Oxford University Press, Oxford, 1988).