Rapid analysis of microbial systems using vibrational spectroscopy and supervised learning methods: application to the discrimination between methicillin-resistant and methicillin-susceptible Staphy

Goodacre, Royston, Rooney, Paul, Kell, Douglas

Royston Goodacre, Paul J. Rooney, Douglas B. Kell, "Rapid analysis of microbial systems using vibrational spectroscopy and supervised learning methods: application to the discrimination between methicillin-resistant and methicillin-susceptible Staphy," Proc. SPIE 3257, Infrared Spectroscopy: New Tool in Medicine, (24 April 1998); doi: 10.1117/12.306087

Event: BiOS '98 International Biomedical Optics Symposium, 1998, San Jose, CA, United States

Rapid analysis of microbial systems using vibrational spectroscopy and supervised learning methods: application to the discrimination between methicillin-resistant and methicillin-susceptible *Staphylococcus aureus*.

Royston Goodacre[1*], Paul J. Rooney[2] and Douglas B. Kell[1]
[1]Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion, SY23 3DA, Wales, UK. & [2]Bronglais General Hospital, Aberystwyth, Ceredigion, SY23 1ER, Wales, UK.

ABSTRACT

FT-IR spectra were obtained from 15 methicillin-resistant and 22 methicillin-susceptible *Staphylococcus aureus* strains using our DRASTIC (Diffuse Reflectance Absorbance Spectroscopy Taking In Chemometrics) approach[1]. Cluster analysis showed that the major source of variation between the IR spectra was not due to their resistance or susceptibility to methicillin; indeed early studies using pyrolysis mass spectrometry[2] had shown that this unsupervised analysis gave information on the phage group of the bacteria. By contrast, artificial neural networks, based on supervised learning, could be trained to recognize those aspects of the IR spectra which differentiated methicillin-resistant from methicillin-susceptible strains. These results give the first demonstration that the combination of FT-IR with neural networks can provide a very rapid and accurate antibiotic susceptibility testing technique.

Keywords: artificial neural networks, chemometrics, drug resistance, FT-IR, *Staphylococcus aureus*

1. INTRODUCTION

For routine purposes the ideal method for microbial characterisation would have minimum sample preparation, would analyse samples directly (i.e. be reagentless), would be rapid, automated, and (at least relatively) inexpensive. With recent developments in analytical instrumentation, these requirements are being fulfilled by physico-chemical spectroscopic methods, often referred to as 'whole-organism fingerprinting' [3]. The most common such methods are pyrolysis mass spectrometry (PyMS)[4], Fourier transform infrared spectroscopy (FT-IR)[5-7] and UV resonance Raman spectroscopy[8].

FT-IR and dispersive Raman microscopy are physico-chemical methods which measure predominantly the vibrations of bonds within functional groups, either through the absorbance of electromagnetic radiation (FT-IR; Figure 1) or from the inelastic scattering of light (Raman shift)[9-13]. Therefore they give quantitative information about the total biochemical composition of a sample. However, the interpretation of these multidimensional spectra, or what is known as hyperspectral data[14-16], has conventionally been by the application of "unsupervised" pattern recognition methods such as principal components (PCA), discriminant function (DFA) and hierarchical cluster (HCA) analyses (see Figure 2 for a flowchart of the usual taxonomic procedure). With "unsupervised learning" methods of this sort the relevant multivariate algorithms seek "clusters" in the data, thereby allowing the investigator to group objects

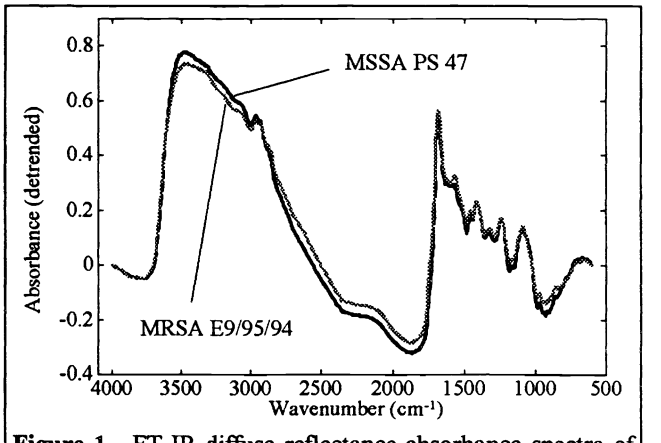
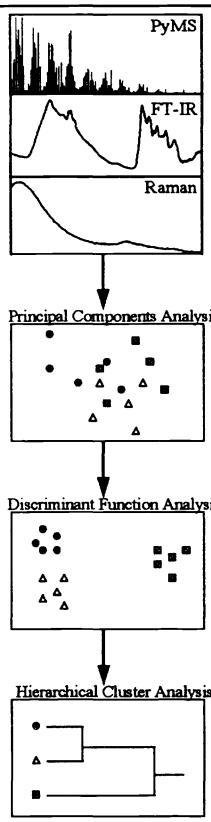


Figure 1 FT-IR diffuse reflectance-absorbance spectra of methicillin susceptible (MSSA) and methicillin resistant (MRSA) *Staphylococcus aureus*.

Correspondence to: Dr Royston Goodacre, Institute of Biological Sciences, University of Wales, Aberystwyth, SY23 3DA, Wales, U.K. Telephone: +44 (0)1970 621947, Telefax: +44 (0)1970 622354, E-mail: rrg@aber.ac.uk http://gepasi.dbs.aber.ac.uk/roy/chemom.htm

SPIE Vol. 3257 • 0277-786X/98/$10.00

together on the basis of their perceived closeness[17]; this process is often subjective because it relies upon the interpretation of complicated scatter plots and dendrograms. More recently, various related but much more powerful methods, most often referred to within the framework of chemometrics, have been applied to the "supervised" analysis of these hyperspectral data[1,18-24]; arguably the most significant of these is the application of intelligent systems based on artificial neural networks (ANNs) which effect supervised learning[25,26].



Spectra are high dimensional:

- 150 masses from PyMS
- 882 wavenumbers from FT-IR
- 2283 wavenumbers from Raman

PCA transforms the original set of variables to a new set of uncorrelated variables called PCs. PCA is a data reduction process and the first few PCs will typically account for >95% variance.

DFA has *a priori* information based on spectral replicates and uses this to minimise within group variance and maximise between group variance.

A similarity matrix can be constructed from the DFA space. HCA can then use this to produce a dendrogram, using average linkage clustering.

Figure 2 Flowchart of unsupervised learning analysis used to cluster hyperspectral data

ANNs are a well-known means of uncovering complex, non-linear relationships in multivariate data, whilst still being able to map the linearities. 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[25-29].

For a given analytical system there are some patterns (e.g. FT-IR or Raman spectra) which have desired responses or values which are known (e.g. the identity of a micro-organism or the concentration of target determinands). These two types of data form pairs which are called inputs and targets. The goal of supervised learning is to find a model or mapping that will correctly associate the inputs with the targets.

The relevant principle of "supervised" learning in ANNs is thus 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 (for a diagrammatic representation see Figure 3). 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 non-linear) weighting function referred to as its activation or squashing function. The great power of neural networks stems from the fact that it is possible to "train" them. One can acquire sets of multivariate data (i.e., hyperspectral data) from standard materials of known identities and train ANNs using these identities as the desired outputs. 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[27,30], until the output nodes of the network match the desired outputs to a stated degree of accuracy. The trained ANNs may then be exposed to unknown inputs (i.e. spectra) when they will immediately provide the globally optimal best fit to the outputs.

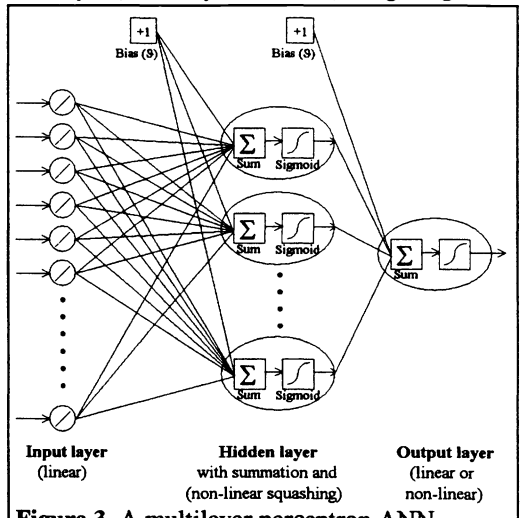


Figure 3 A multilayer perceptron ANN

There has been a dramatic increase in the incidence of nosocomial infections caused by strains of *Staphylococcus aureus* which are resistant to multiple antibiotics, usually due to transfer (acquisition) of resistance genes[31]. Methicillin-resistant *S. aureus* (MRSA) were first isolated in 1961 following the introduction of this ß-lactam for the

treatment of staphylococcal infections[32], and intrinsic resistance in MRSA strains was later found to be due to the presence of a novel additional penicillin binding protein (PBP) PBP 2' [33,34]. PBP 2' is encoded by the *mecA* gene, which is part of the chromosomal DNA *mec* sequence, a 30- to 40- kb piece of DNA whose origin is as yet unknown[35].

In a previous study[2] we used PyMS to discriminate between methicillin-susceptible and -resistant *S. aureus* strains. The aim of this study was to use FT-IR to examine the same collection of MSSA and MRSA strains. Thirty seven strains were examined; these covered a wide range of epidemiologically distinct organisms, and had been classified as either MRSA or MSSA using conventional means. Cluster analysis and artificial neural networks were used to determine whether the infrared spectra could be used to discriminate these strains on the basis of their antibiotic susceptibility.

2. MATERIALS AND METHODS

2.1 Organisms and cultivation

Twenty two methicillin-susceptible (MSSA) and 15 methicillin-resistant *S. aureus* (MRSA) were used in this study; these cultures were chosen to represent a diverse range of MSSA and MRSA strains; moreover, they possessed a wide range of resistances to other antibiotics (data not shown). The National Collection of Type Cultures set of 22 propagating strains for phage typing were used as examples of MSSA[36]. Three recent clinical isolates of MRSA (H34, E9/95/94, and E18/103/94) were supplied by Bronglais General Hospital along with a national standard MRSA (NCTC 10042), whilst the remainder were supplied by Dr Judith Richardson (Laboratory of Hospital Infection, Central Public Health Laboratory, 61 Colindale Avenue, London, NW9 5HT, U.K.). Details of the strain names, phage group, and resistance or susceptibility to methicillin are given in Tables 1 and 2.

Strains were cultured on Mueller Hinton agar (Oxoid-Unipath Ltd., Basingstoke, UK) plus 2% NaCl, which favours the expression of PBP 2' [33,34], and incubated aerobically for 16 h. The bacteria were carefully removed from the agar surface with a plastic loop and suspended in physiological saline (0.9% NaCl) to approximately 20 mg/mL. The samples were then ready for analysis by FT-IR.

Table 1 Identity of the *S. aureus* used in the training set as judged by ANNs

S. aureus strain	Lytic Ø type	Type	ANNs estimates
PS 47	III	S	0.00
PS 81	Misc	S	0.00
PS 84	III	S	0.00
PS 55	II	S	0.01
PS 3C	II	S	0.02
PS 52	I	S	0.01
PS 29	I	S	0.01
PS 83A	III	S	0.00
PS 85X	III	S	0.00
PS 77X	III	S	0.01
PS 42E	III	S	0.01
CRF 634 PS	III	R	1.00
ST 85 1774	NT	R	0.99
CRF 627 PS	III	R	0.98
ST 84 6144	NT	R	0.98
CRF 619 PS	III	R	1.00
H34	NT	R	1.00
E18/103/94	III	R	0.99
NCTC 10442	III	R	0.99

2.2 Diffuse reflectance-absorbance Fourier transform infrared (FT-IR) spectroscopy

Typically, the sample preparation for FT-IR absorbance measurements involves grinding the dried sample to a fine powder and mixing with KBr, although Naumann *et al.*[7] have replaced this slow and rather tedious method with the application of liquid samples (which are then dried) to one of 16 ZnSe windows on a rotating disc. However, we consider that a much more elegant approach, which is automated and can allow many more samples (>>100) to be analysed in one data collection run, is to use reflectance methods. Diffuse reflectance-absorbance can be achieved by applying the sample onto a sand-blasted metal plate which can then be loaded onto a motorised stage of a reflectance TLC accessory[37]. It is noteworthy that such an approach also allows spectra to be obtained as a function of spatial location. Moreover, it has been shown that reflectance methods can be both more sensitive and discriminatory than absorbances[38,39].

Ten microlitres of the above bacterial samples were evenly applied onto a sand-blasted aluminium plate. Prior to analysis the samples were oven-dried at 50°C for 30 min. Samples were run in triplicate. The FT-IR instrument used was the Bruker IFS28 FT-IR spectrometer (Bruker Spectrospin Ltd., Banner Lane, Coventry, UK) equipped with an MCT (mercury-cadmium-telluride) detector cooled with liquid N_2. The aluminium plate was then loaded onto the motorised stage of a reflectance TLC accessory[1,23,40].

Table 2 Identity of the *S. aureus* used in the test set as judged by ANNs

S. aureus strain	Lytic Ø type	Type	ANNs estimates
PS 54	III	S	0.00
PS 71	II	S	0.00
PS 95	Misc	S	0.00
PS 6	III	S	0.00
PS 52A/79	I	S	0.00
PS 3A	II	S	0.00
PS 80	I	S	0.00
PS 96	V	S	0.01
PS 94	V	S	0.00
PS 75	III	S	0.99
PS 53	III	S	0.05
CRF 631 PS	I	R	0.99
ST 84 6255	III	R	1.00
CRF 621 PS	NT	R	1.00
ST 84 6983	-	R	1.00
CRF 633 PS	III	R	1.00
ST 85 3566	NT	R	0.94
E9/95/94	NT	R	1.00

The averages of the 10 ANNs are shown.

The IBM-compatible personal computer used to control the IFS28, was also programmed (using OPUS version 2.1 software running under IBM OS/2 Warp provided by the manufacturers) to collect spectra over the wavenumber range 4000 cm^{-1} to 600 cm^{-1}. Spectra were acquired at a rate of 20 s^{-1}. The spectral resolution used was 4 cm^{-1}. To improve the signal-to-noise ratio, 256 spectra were co-added and averaged. Each sample was thus represented by a spectrum containing 882 points and spectra were displayed in terms of absorbance as calculated from the reflectance-absorbance spectra using the Opus software. Typical FT-IR spectra are shown in Figure 1.

2.3 Cluster analysis

ASCII data were exported from the OPUS software used to control the FT-IR instrument and imported into Matlab version 4.2c. 1 (The MathWorks, Inc., 24 Prime Par Way, Natick, MA, USA), which runs under Microsoft Windows NT on an IBM-compatible personal computer. To minimize problems arising from baseline shifts the following procedure was implemented: (i) the spectra were first normalised so that the smallest absorbance was set to 0 and the highest to +1 for each spectrum, (ii) next these normalised spectra were detrended by subtracting a linearly increasing baseline from 4000 cm^{-1} to 600 cm^{-1}, (iii) finally the smoothed first derivatives of these normalised and detrended spectra were calculated using the Savitzky-Golay algorithm[41] with 5-point smoothing.

To reduce the dimensionality of the FT-IR data Matlab was also employed to perform PCA (according to the NIPALS algorithm[42]); of the original 882 spectral points 96.8 % of the total variance was retained in the first 30 principal components (PCs). Figure 4a is a plot of explained variance versus PCs extracted from these FT-IR data and highlights the power of this technique and shows that 30 PCs adequately describe the majority of the varaince from the original data. Next these 30 PCs were used as inputs to the DFA algorithm with the a priori knowledge of which spectra were replicates. DFA was programmed according to Manly's principles[43].

2.4 Creation of training and test data sets for artificial neural network analyses.

It is well known that if the number of weights in a neural network is significantly higher than the number of exemplars in the training set then over-fitting can more easily occur[26,44]. In this case if the full spectra were used as input data then the 882-12-1 ANN would contain 10609 weights (10584 between input and hidden layers + 12 weights between hidden and output layer + 13 between the bias and the nodes in the hidden and output layers). Therefore, to obey the parsimony principle as described by Seasholtz & Kowalski[44] and to account for any further baseline effects the next stage was to reduce the number of inputs to the ANNs the averages of the first 15 discriminant function (DF) scores constructed from the first 30 PCs and were used as the input data. The fifteen DFs were used because when too few DFs are used not enough information is present, and when more DFs are employed the later DFs contribute only noise to the model, thus increasing the probability of chance correlations between input and ouptut data. This is shown diagrammatically in Figure 4b.

In addition, it is important that the training data encompass the full range under study[26,45,46] since although supervised methods are excellent at being able to interpolate they are likely to give poor estimates outside their 'realm of knowledge', i.e. they can not extrapolate sufficiently well. Since the 37 strains of S. aureus encompassed a diverse range of epidemiologically distinct strains it was imperative that the training set of MSSA and MRSA represented multidimensional space sufficiently well to allow interpolation.

Duplex is a method for choosing an optimal split between training and test data sets[47], and an extension to this methodology called "Multiplex" has been developed in-house (Jones A., Kell D.B. & Rowland J., in preparation). Briefly

this method starts by placing the two most separated samples into the training set. It then places the next two most separated remaining samples into the test set. This is performed iteratively until all samples have been split. This ensures that the training set range covers the test set range, and that both sets are representative. The first fifteen DF scores from the IR spectra were sorted using "Multiplex" so that the training and test data were split in the ratio 1:1. Data may be split on both the X-matrix (DF scores) and the Y-matrix (bacterial type); so as not to bias the partitioning process data were split on the X-matrix only. The reason that the "Multiplex" method is superior to standard techniques such as n-fold cross validation (CV) is because, although both methods will lead to comparable solutions, n-fold CV is more computationally intense and especially when ANNs are used this process will take appreciably longer.

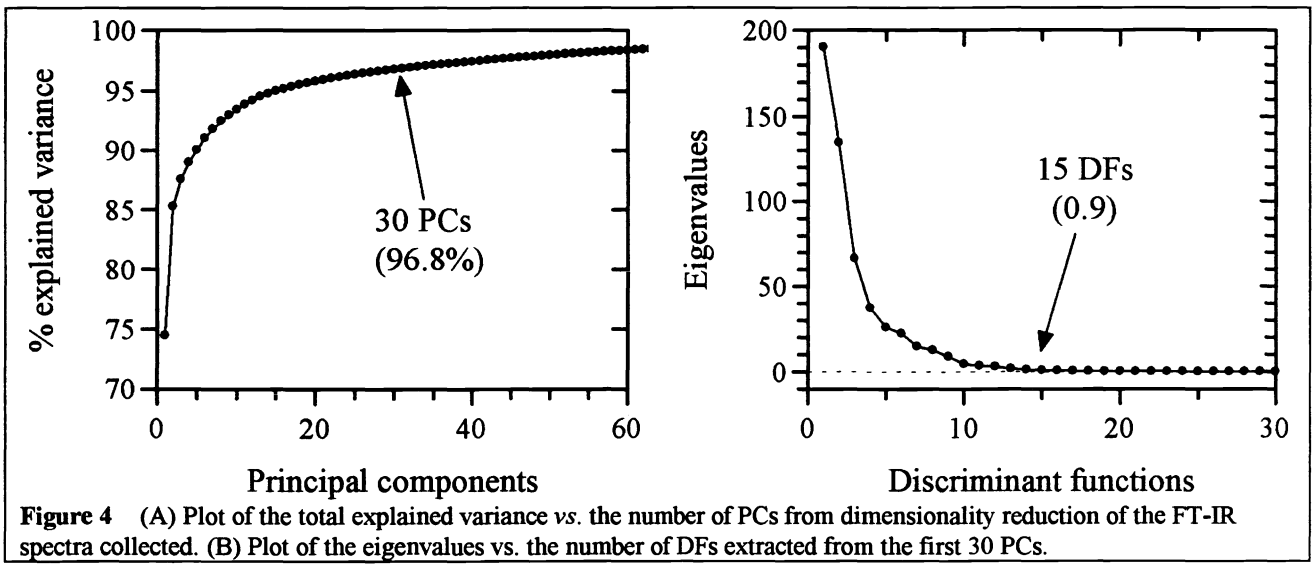


Figure 4 (A) Plot of the total explained variance *vs.* the number of PCs from dimensionality reduction of the FT-IR spectra collected. (B) Plot of the eigenvalues vs. the number of DFs extracted from the first 30 PCs.

2.5 Artificial neural networks (ANNs)

All ANN analyses were carried out with a user-friendly, neural network simulation program, NeuFrame version 1,1,0,0 (Neural Computer Sciences, Lulworth Business Centre, Nutwood Way, Totton, Southampton, Hants), which runs under Microsoft Windows NT on an IBM-compatible PC. In-depth descriptions of the *modus operandi* of this type of multilayer perceptron (MLP) analysis are given elsewhere[19,48-50].

For training the ANNs, each of the inputs was the averages of the first fifteen DF scores split above using the multiplex program (details of the training and test sets are given in Tables 1 and 2) and was paired with each of the desired outputs. These were binary-encoded such that the MSSA strains were coded as 0 and MRSA coded as 1 at the output node. These training pairs collectively made up the training set. The structure of the ANN used in this study consisted of 3 layers; 15 input nodes, 1 output node, and one "hidden" layer containing 4 nodes (i.e., a 15-4-1 architecture). Before training commenced the values applied to the output nodes were normalised between 0 and 1. The scaling regime used for the input layer was to scale each node such that the lowest DF was set to 0 and the highest to 1. For present purposes these ANNs were trained to a RMSEF (RMS error of formation of the model) of 0.01, which typically took ca. 3.10^2 epochs. Initially MLPs were trained until the RMSEF was 0.005 (0.5%), and their ability to generalise was assessed on the test set. It was found that MLPs trained until the RMSEF was 0.01 (1%) were still able to generalise well, and since these MLPs obviously took less time to train and were less likely to overfit the input data (i.e., fitting to noise or the fitting of a model to outliers[46,50]), all MLPs were trained until the RMS was 0.01 (1%). Training was conducted 10 times (a) to observe whether this process was reproducible and (b) to use the "committee" approach for prediction[26], where the outputs from the ten 15-4-1 ANNs were averaged.

3. RESULTS AND DISCUSSION

Typical normalised FT-IR spectra for methicillin-susceptible *S. aureus* strain PS 47 and methicillin-resistant *S. aureus* strain E9/95/94 are shown in Figure 1. These and all the FT-IR spectra of the other *S. aureus* show broad and complex contours and although there was very little qualitative difference between these spectra, small complex quantitative differences between the spectra were observed. Such spectra, uninterpretable by the naked eye, readily illustrate the need to employ multivariate statistical techniques for the analysis of FT-IR data.

After collection of the infrared spectra, each of the 37 strains, each represented by three replicate spectra, were coded to give 37 individual groups, and analysed by DFA; the resulting ordination plot is shown in Figure 5. The coding in this plot is simply for whether the strain is MSSA (indicated by a 'S') or MRSA ('R'), and shows that DFA cannot be used to cluster these bacteria according to whether they were resistant or susceptible to methicillin because two distinct groups were not formed. It is of course possible that this differentiation may happen if the lower discriminant functions were viewed; however, these graphs were plotted (data not shown) and separation based on drug resistance or susceptibility was not evident.

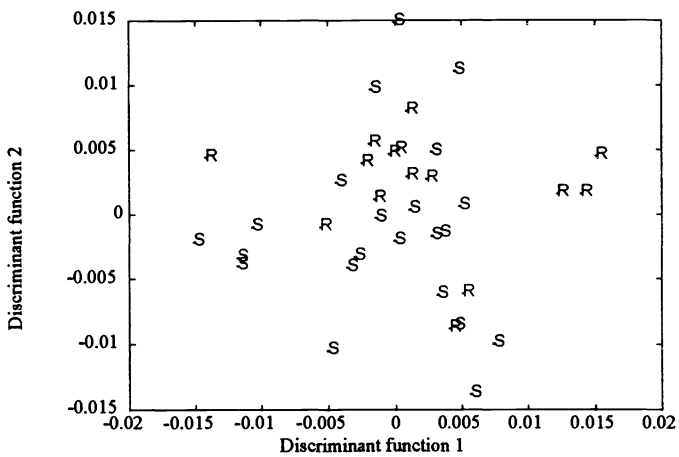


Figure 5 Discriminant function plot based on FT-IR data showing the relationship between the 37 *S. aureus* strains. The first 30 PCs were used as the inputs for the DFA algorithm and accounted for 96.8% of the variance.
'S' refers to MSSA and 'R' to MRSA.

The next stage was therefore to examine the ability of artificial neural networks (ANNs), a supervised method which should uncover non-linear relationships between the two classes of bacteria and which has been demonstrated to be greatly superior to clustering techniques in classification problems of this type[2,51,52].

As detailed above, the 37 strains encompassed a diverse range of epidemiologically distinct strains and it is thus imperative that the training set for the ANNs will represent the FT-IR multidimensional space sufficiently well to allow interpolation, and that the range of test set was enclosed by the training set. Therefore the program "Multiplex" was used to split the data equally into training and test sets. Details of the two data sets can be found in Tables 1 and 2.

ANNs were trained with the first fifteen discriminant function scores from the infrared spectra (as processed above) from the training set; the 11 MSSA were coded 0 at the output node, and the 8 MRSA were coded 1. The 15-4-1 ANNs were trained using the standard back-propagation algorithm, and the effectiveness of training was expressed in terms of the RMS error between the actual and the desired outputs; training was stopped after the RMS error had reached 0.01 (or 1%). Training was effected ten times, using randomised, small initial values for the starting weights; the ten learning curves were seen to superimpose (data not shown) and it was clear that despite the randomised starting connection weights, training was executed (i.e. the error surface in weight space was negotiated) in a rather reproducible manner. Moreover, these ANNs typically took 320 epochs to train to an RMS error of 0.01 within a spread of only ±10 epochs.

When training had ceased (i.e. as determined by the attainment of an RMS error of 0.01 averaged over the training set) the ten neural networks were interrogated with the FT-IR spectra from both data sets. Not surprisingly, the network's estimate of the resistance or susceptibility to methicillin of the training set was the same as those known in all ten trainings (Table 1). The results of the ANN's analyses of the unknown test set is also shown in Table 2. This table is the average of the ANN's predictions for each of the replicates of the 37 strains; small standard deviations were calculated (the largest was only 0.011) indicating that training was indeed reproducible. Rather than using a simple 'crisp' identification criterion where if the output is > 0.5 then the strain is a MRSA and if the output is < 0.5 then the strain is a MSSA, a correct identification was taken to be that to belong to MRSA the output must be ≥ 0.9 and for MSSA to be ≤ 0.1. This procedure

allows a more rigid classification to be used since if any output is close to 0.5 the ANN would be taken to indicate that it is 'undecided' about the identification, and hence unable to discriminate that bacterial sample sufficiently well.

It is evident from Table 2 that the ANNs had assessed correctly the seven methicillin-resistant *S. aureus* strains from the unseen test set, as being MRSA isolates; the networks' output was >0.98 for all isolates. With the exception of the methicillin-susceptible strain PS 75, all the MSSA isolates were also characterised correclty; the networks' output was <0.05 for all isolates. These results show that there were no false negatives and a single false positive (PS 75).

4. CONCLUSIONS

DF cluster analyses of the FT-IR spectra from 37 *S. aureus* failed to separate these bacteria on the basis of their resistance or susceptibility to the antibiotic methicillin. By contrast, however, ANNs could be trained to assess whether an unknown strain was resistant to methicillin, and with one exception was able to assess the methicillin-susceptible *S. aureus* in an unseen test set. Although there was a single false positive, more importantly from the physicians point of view there were no false negatives.

The application of FT-IR to microbiology is undoubtedly useful in the discrimination between bacteria and fungi at the genus, species and subspecies level, as has previously been demonstrated by Naumann and colleagues[5,6,53]. FT-IR has the advantage of speed and particularly with our diffuse reflectance-absorbance approach[1,23,54] easily allows the acquisition of 400 samples per hour on a single 10x10 cm aluminium plate.

ANNs have proved very advantageous in the analysis of FT-IR data. These mathematical techniques based on artificial 'intelligence' have allowed us to identify clinical isolates from hospital isolates of *Enterococcus faecalis*, *E. faecium*, *Streptococcus bovis*, *S. mitis*, *S. pneumoniae*, or *S. pyogenes*[23], to discriminate between common infectious agents associated with urinary tract infection[55], to assess the physiological state of a wide range of *Bacillus* species[56], and to determine the concentration of secondary metabolites in titre improvement programmes[1,57,58].

In conclusion, ANNs can be used to extract very subtle physiological differences between strains of the same species of *S. aureus* from their FT-IR data, and in this case for the rapid and accurate methicillin susceptibility testing.

Neural networks are only one group of techniques in the huge chemometrics tool box. Of the unsupervised cluster analysis methods the most common used are principal components analysis (PCA), Kohonen's self organizing feature maps, autoassociative neural networks (which effect non-linear PCA), and discriminant function analysis (DFA, or canonical variates analysis). The number of supervised learning algorithms is ever increasing and include methods based on linear regression (PCR and PLS[59]), rule induction methods that are based on crisp or fuzzy rules, classification and regression trees (CART), and radial basis functions. Some exciting methods which are emerging are based on evolutionary computing, and these include genetic algorithms and genetic programming which can be used to deconvolute (IR and MS) spectra in terms of which wavenumbers are important either in classification or quantification studies[60-62]. In-depth tutorials and reviews on the above methods can be found on our WWW site via http://gepasi.dbs.aber.ac.uk/home.htm.

ACKNOWLEDGMENTS

We would like to thank Dr Judith Richardson for helpful advice and provision of strains. R.G. is indebted to the Wellcome Trust (grant number 042615/Z/94/Z), and D.B.K. to the Chemicals and Pharmaceuticals Directorate of the UK BBSRC, for financial support.

REFERENCES

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

2. R. Goodacre, P.J. Rooney, and D.B. Kell, "Discrimination between methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* using pyrolysis mass spectrometry and artificial neural networks," *J. Antimicr. Chemother.* **41**, in press, 1998.
3. J.T. Magee, "Whole-organism fingerprinting," in *Handbook of New Bacterial Systematics*, edited by M. Goodfellow and A.G. O'Donnell, pp. 383-427, Academic Press, London, 1993.
4. R. Goodacre and D.B. Kell, "Pyrolysis mass spectrometry and its applications in biotechnology," *Cur. Opin. Biotechnol.* **7**, 20-28, 1996.
5. D. Helm, H. Labischinski, G. Schallehn, and D. Naumann, "Classification and identification of bacteria by Fourier transform infrared spectroscopy," *J. Gen. Microbiol.* **137**, 69-79, 1991.
6. D. Naumann, D. Helm, and H. Labischinski, "Microbiological characterizations by FT-IR spectroscopy," *Nature* **351**, 81-82, 1991.
7. D. Naumann, D. Helm, H. Labischinski, and P. Giesbrecht, "The characterization of microorganisms by Fourier-transform infrared spectroscopy (FT-IR)," in *Modern techniques for rapid microbiological analysis*, edited by W.H. Nelson, pp. 43-96, VCH Publishers, New York, 1991.
8. W. H. Nelson, R. Manoharan, and J. F. Sperry, "UV resonance Raman studies of bacteria," *Appl. Spectrosc. Revs.* **27**, 67-124, 1992.
9. H.L.C. Meuzelaar, J. Haverkamp, and F.D. Hileman, *Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials*, Elsevier, Amsterdam, 1982.
10. P.R. Griffiths and J.A. de Haseth, *Fourier transform infrared spectrometry*, John Wiley, New York, 1986.
11. N.B. Colthup, L.H. Daly, and S.E. Wiberly, *Introduction to infrared and Raman spectroscopy*, Academic Press, New York, 1990.
12. J.R. Ferraro and K. Nakamoto, *Introductory Raman Spectroscopy*, Academic Press, London, 1994.
13. B. Schrader, *Infrared and Raman spectroscopy : methods and applications.*, Verlag Chemie, Weinheim, 1995.
14. A.F.H. Goetz, G. Vane, J. Solomon, and B.N. Rock, "Imaging spectrometry for earth remote sensing," *Science* **228**, 1147-1153, 1985.
15. G. P. Abousleman, E. Gifford, and B. R. Hunt, "Enhancement and compression techniques for hyperspectral data," *Optical Engineering* **33**, 2562-2571, 1994.
16. T. A. Wilson, S. K. Rogers, and L. R. Myers, "Perceptual-based hyperspectral image fusion using multiresolution analysis," *Optical Engineering* **34**, 3154-3164, 1995.
17. B.S. Everitt, *Cluster Analysis*, Edward Arnold, London, 1993.
18. J. Chun, E. Atalan, A. C. Ward, and M. Goodfellow, "Artificial neural network analysis of pyrolysis mass spectrometric data in the identification of *Streptomyces* strains," *FEMS Microbiol. Lett.* **107**, 321-325, 1993.
19. R. Goodacre, M. J. Neal, D. B. Kell, L. W. Greenham, W. C. Noble, and R. G. Harvey, "Rapid identification using pyrolysis mass spectrometry and artificial neural networks of *Propionibacterium acnes* isolated from dogs," *J. Appl. Bacteriol.* **76**, 124-134, 1994.
20. R. Freeman, R. Goodacre, P. R. Sisson, J. G. Magee, A. C. Ward, and N. F. Lightfoot, "Rapid identification of species within the *Mycobacterium tuberculosis* complex by artificial neural network analysis of pyrolysis mass spectra," *J. Med. Microbiol.* **40**, 170-173, 1994.
21. P. R. Sisson, R. Freeman, D. Law, A. C. Ward, and N. F. Lightfoot, "Rapid detection of verocytotoxin production status in *Escherichia coli* by artificial neural network analysis of pyrolysis mass spectra," *J. Anal. Appl. Pyrol.* **32**, 179-185, 1995.
22. R. Goodacre, S.J. Hiom, S.L. Cheeseman, D. Murdoch, A.J. Weightman, and W.G. Wade, "Identification and discrimination of oral asaccharolytic *Eubacterium* spp. using pyrolysis mass spectrometry and artificial neural networks," *Cur. Microbiol.* **32**, 77-84, 1996.
23. R. Goodacre, É.M. Timmins, P.J. Rooney, J.J. Rowland, and D.B. Kell, "Rapid identification of *Streptococcus* and *Enterococcus* species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks," *FEMS Microbiol. Lett.* **140**, 233-239, 1996.
24. B.K. Alsberg, R. Goodacre, J.J. Rowland, and D.B. Kell, "Classification of pyrolysis mass spectra by fuzzy multivariate rule induction - comparison with regression, K-nearest neighbour, neural and decision-tree methods," *Anal. Chim. Acta* **348**, 389-407, 1997.
25. P.D. Wasserman, *Neural Computing: Theory and Practice*, Van Nostrand Reinhold, New York, 1989.
26. C.M. Bishop, *Neural networks for pattern recognition*, Clarendon Press, Oxford, 1995.

27. D.E. Rumelhart, J.L. McClelland, and The PDP Research Group, *Parallel Distributed Processing, Experiments in the Microstructure of Cognition*, MIT Press, Cambridge, Mass., 1986.
28. R. Hecht-Nielsen, *Neurocomputing*, Addison-Wesley, Massachusetts, 1990.
29. J. Zupan and J. Gasteiger, *Neural Networks for Chemists: An Introduction*, VCH Verlagsgeesellschaft, Weinheim, 1993.
30. P.J. Werbos, *The roots of back-propagation: from ordered derivatives to neural networks and political forecasting.*, John Wiley, Chichester, 1994.
31. B.R. Lyon and R. Skurray, "Antimicrobial resistance of *Staphylococcus aureus*: genetic basis," *Microbiol. Revs.* **51**, 88-134, 1987.
32. M.P. Jevons, "'Celbenin' resistant staphylococci," *Br. Med. J.* **1**, 124-125, 1961.
33. P.E. Reynolds and C. Fuller, "Methicillin resistant strains of *Staphylococcus aureus*; prescence of identical additional penicillin binding protein in all strains examined," *FEMS Microbiol. Lett.* **33**, 251-254, 1986.
34. D.M. O'Hara and P.E. Reynolds, "Antibody used to identify penicillin binding protein 2' in methicillin resistant strains of *Staphylococcus aureus* (MRSA)," *FEBS Lett.* **212**, 237-241, 1987.
35. F.H. Kayser, "Methicillin and glycopeptide resistance in staphylococci: a new threat?," *Curr. Opin. Infect. Dis.* **8**, S7-S11, 1995.
36. J.F. Richardson, P. Aparicio, R.R. Marples, and B.D. Cookson, "Ribotyping of *Staphylococcus aureus*: an assessment using well-defined strains," *Epidemiol. Infect.* **112**, 93-101, 1994.
37. G. Glauninger, K.A. Kovar, and V. Hoffmann, "Possibilities and limits of an online coupling of thin-layer chromatography and FTIR spectroscopy," *Fresenius J. Anal. Chem* **338**, 710-716, 1990.
38. M.P. Fuller and P.R. Griffiths, "Infrared microsampling by diffuse reflectance Fourier transform spectrometry," *Appl. Spectrosc.* **34**, 533-539, 1980.
39. M.B. Mitchell, "Fundamentals and applications of diffuse reflectance infrared fourier transform (DRIFT) spectroscopy," *Adv. Chem. Ser.* **236**, 351-375, 1993.
40. S.P. Bouffard, J.E. Katon, A.J. Sommer, and N.D. Danielson, "Development of microchannel thin layer chromatography with infrared microspectroscopic detection," *Anal. Chem.* **66**, 1937-1940, 1994.
41. A. Savitzky and M.J.E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.* **36**, 1627-1633, 1964.
42. H. Wold, "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis*, edited by K. R. Krishnaiah, pp. 391-420, Academic Press, New York, 1966.
43. B.F.J. Manly, *Multivariate Statistical Methods : A Primer*, Chapman & Hall, London, 1994.
44. M. B. Seasholtz and B. Kowalski, "The parsimony principle applied to multivariate calibration," *Anal. Chim. Acta* **277**, 165-177, 1993.
45. R. Goodacre, A. N. Edmonds, and D. B. Kell, "Quantitative analysis of the pyrolysis mass spectra of complex mixtures using artificial neural networks - application to amino acids in glycogen," *J. Anal. Appl. Pyrol.* **26**, 93-114, 1993.
46. D.B. Kell and B. Sonnleitner, "GMP - Good Modelling Practice: an essential component of good manufacturing practice," *Trends Biotechnol.* **13**, 481-492, 1995.
47. R.D. Snee, "Validation of regression models: Methods and examples," *Technometrics* **19**, 415-428, 1977.
48. R. Dybowski and V. Gant, "Artificial neural networks in pathological and medical laboratories," *Lancet* **346**, 1203-1207, 1995.
49. W.G. Baxt and J. Skora, "Prospective validation of artificial neural network trained to identify acute myocardial infarction," *Lancet* **347**, 12-15, 1996.
50. R. Goodacre, M.J. Neal, and D.B. Kell, "Quantitative analysis of multivariate data using artificial neural networks: a tutorial review and applications to the deconvolution of pyrolysis mass spectra," *Zbl. Bakt. - Int. J. Med M.* **284**, 516-539, 1996.
51. R. Goodacre, D. B. Kell, and G. Bianchi, "Neural networks and olive oil," *Nature* **359**, 594-594, 1992.
52. R. Goodacre, D. B. Kell, and G. Bianchi, "Rapid assessment of the adulteration of virgin olive oils by other seed oils using pyrolysis mass spectrometry and artificial neural networks," *J. Sci. Food Agric.* **63**, 297-307, 1993.
53. D. Naumann, V. Fijjala, H. Lavischinski, and P. Giesbrecht, "The rapid differentiation and identification of pathogenic bacteria using fouriter transform infrared spectroscopic and multivariate statistical analysis," *Molecular Structure* **174**, 165-170, 1988.

54. É.M. Timmins, S.A. Howell, B.K. Alsberg, W.C. Noble, and R. Goodacre, "Rapid differentiation of closely related *Candida* species and strains by pyrolysis mass spectrometry and fourier transform infrared spectroscopy," *J. Clin. Microbiol.* , in press, 1998.
55. R. Goodacre, É.M. Timmins, R. Burton, N. Kaderbhai, A. Woodward, D.B. Kell, and P.J. Rooney, "Rapid identification of urinary tract infection bacteria using hyperspectral, whole organism fingerprinting and artificial neural networks," *Microbiology* , submitted, 1998.
56. R. Goodacre, B. Shann, R.J. Gilbert, É.M. Timmins, A.C. McGovern, B.K. Alsberg, N.A. Logan, and D.B. Kell, "The characterisation of *Bacillus* species from PyMS and FT IR data," presented at the Proc. 1997 ERDEC Scientific Conference on Chemical and Biological Defense Research, Aberdeen Proving Ground, 1998.
57. D.B. Kell, M.K. Winson, R. Goodacre, A.M. Woodward, B.K. Alsberg, A. Jones, É.M. Timmins, and J.J. Rowland, "DRASTIC (Diffuse Reflectance Absorbance Spectroscopy Taking In Chemometrics). A novel, rapid, hyperspectral, FT-IR-based approach to screening for biocatalytic activity and metabolite overproduction," in *Screening for Novel Biocatalysts,* 1998.
58. M. K. Winson, M. Todd, B. A. M. Rudd, A. Jones, B. K. Alsberg, A. M. Woodward, R. Goodacre, J. J. Rowland, and D. B. Kell, "A DRASTIC (Diffuse Reflectance Absorbance Spectroscopy Taking in Chemometrics) approach for the rapid analysis of microbial fermentation products: quantification of aristeromycin and neplanocin A in Streptomyces citricolor broths," in *Screening for Novel Biocatalysts,* 1998.
59. H. Martens and T. Næs, *Multivariate Calibration*, John Wiley, Chichester, 1989.
60. D. Broadhurst, R. Goodacre, A. Jones, J.J. Rowland, and D.B. Kell, "Genetic algorithms as a method for variable selection in PLS regression, with application to pyrolysis mass spectra," *Anal. Chim. Acta* **348**, 71-86, 1997.
61. R.J. Gilbert, R. Goodacre, A.M. Woodward, and D.B. Kell, "Genetic programming : a novel method for the quantitative analysis of pyrolysis mass spectral data," *Anal. Chem.* **69**, 4381-4389, 1997.
62. J. Taylor, R. Goodacre, W.G. Wade, J.J. Rowland, and D.B. Kell, "The deconvolution of pyrolysis mass spectra using genetic programming: application to the identification of some *Eubacterium* species," *FEMS Microbiol. Lett.*, in press, 1998.