BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 7 2005, pages 860–868
doi:10.1093/bioinformatics/bti102

Genome analysis

Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data

Roger M. Jarvis and Royston Goodacre*

Department of Chemistry, UMIST, PO Box 88, Sackville St, Manchester M60 1QD, UK

Received on July 13, 2004; revised on October 7, 2004; accepted on October 18, 2004
Advance Access publication October 28, 2004

ABSTRACT

Motivation: The major difficulties relating to mathematical modelling of spectroscopic data are inconsistencies in spectral reproducibility and the black box nature of the modelling techniques. For the analysis of biological samples the first problem is due to biological, experimental and machine variability which can lead to sample size differences and unavoidable baseline shifts. Consequently, there is often a requirement for mathematical correction(s) to be made to the raw data if the best possible model is to be formed. The second problem prevents interpretation of the results since the variables that most contribute to the analysis are not easily revealed; as a result, the opportunity to obtain new knowledge from such data is lost.

Methods: We used genetic algorithms (GAs) to select spectral pre-processing steps for Fourier transform infrared (FT-IR) spectroscopic data. We demonstrate a novel approach for the selection of important discriminatory variables by GA from FT-IR spectra for multi-class identification by discriminant function analysis (DFA).

Results: The GA selects sensible pre-processing steps from a total of $\sim 10^{10}$ possible mathematical transformations. Application of these algorithms results in a 16% reduction in the model error when compared against the raw data model. GA-DFA recovers six variables from the full set of 882 spectral variables against which a satisfactory DFA model can be formed; thus inferences can be made as to the biochemical differences that are reflected by these spectral bands.

Availability: Supplementary information, datasets and scripts are available from the corresponding author.

Contact: roy.goodacre@manchester.ac.uk

INTRODUCTION

In the biological sciences, high dimensional multivariate datasets are frequently generated by investigations which include transcriptome (Burge, 2001), proteome (Blackstock and Weir, 1999) and metabolic studies (Goodacre *et al*., 2004; Johnson *et al*., 2003, 2004), the characterization and discrimination of microorganisms (Goodacre *et al*., 1994; Maquelin *et al*., 2000; Naumann *et al*., 1991) and bioprocess monitoring (Arnold *et al*., 2000; McGovern *et al*., 2002). Such studies tend to make use of high-throughput multiplexing techniques and physico-spectroscopic methods, the latter of which are finding increasing popularity in the so-called 'omics' studies (Fiehn, 2002). Given the flood of data that is generated from such systems, appropriate analyses need to be performed so that useful knowledge can be induced (Kell and Oliver, 2004). Therefore, multivariate mathematical modelling techniques are often applied to such data in order to expose trends that would be otherwise undetectable.

In this study we are concerned with multivariate spectroscopic data from the analysis of (micro-)biological systems and the application of those data to calibration modelling and discrimination. Supervised linear modelling techniques such as partial least squares (PLS) (Martens and Naes, 1989), a popular calibration model, and discriminant function analysis (DFA) (Manly, 1994), a powerful clustering algorithm, are ideally suited to the analysis of spectroscopic data. In the case of PLS, many studies have shown it to have superb quantitative prediction abilities (McGovern *et al*., 2002; Vaidyanathan *et al*., 2001), whilst DFA has been successfully used for differentiating between a range of microbial libraries (Jarvis and Goodacre, 2004a,b; Lopez-Diez and Goodacre, 2004; Timmins *et al*., 1998).

Essential to such studies are the methods by which the multivariate datasets are obtained. Typically, for the applications described above, spectroscopic methods have been adopted that provide complete biochemical fingerprints from whole cells, known as 'whole-organism fingerprints' or 'metabolic fingerprinting'. Perhaps the most popular methods that have been used so far include the vibrational spectroscopic techniques of Fourier transform infrared (FT-IR) and Raman spectroscopy and a variety of the hyphenated mass spectrometries (Maquelin *et al*., 2000; Naumann *et al*., 1991; Vaidyanathan *et al*., 2002a,b). The latter are proving to be increasingly popular in metabolomics studies, for both metabolic profiling and also the more targeted metabolite analysis (Fiehn *et al*., 2000; Weckwerth, 2003). The benefit of the vibrational techniques is that they are rapid, robust and reproducible in nature, and instruments can be easily adapted for high-throughput analysis, allowing many hundreds or even thousands of samples to be analysed daily.

Biological samples by their nature are inherently variable (Fiehn *et al*., 2000), and so when these are analysed by some physicochemical method it is generally necessary *a priori* to apply data pre-processing to compensate for these differences. In addition, any variability in the spectral acquisition process, e.g. sample size differences and unavoidable baseline shifts in spectra, also needs to be considered. Given some familiarity with the data, the personal experience of the analyst can provide the key to deciding upon the most sensible pre-processing steps. However, there are many different algorithms available to choose from which may require the selection of a range of input variables, making the process of informed guesses and trial and error inherently flawed. For example, given a selection of 50 different algorithms for a set of up to six functions performed serially on the raw data, allowing repeats, the search space would be $50^6 (\sim 10^{10})$ combinations. It would be possible to assess a model

*To whom correspondence should be addressed.

against every combination of pre-processing steps, but this is computationally expensive, depending upon the size of the dataset and the efficiency of the applied transformation algorithm. This generally makes it far too time consuming to do an exhaustive search and therefore optimize a model maximally, so an approach is required that allows the search space to be explored efficiently. One approach is to use a heuristic search algorithm such as a genetic algorithm (GA) (Bäck *et al*., 1997; Goldberg, 1989; Holland, 1992; Mitchell, 1995) to find a range of suboptimal solutions. In the first part of this study we used GA to predict sensible spectral pre-processing steps necessary to reduce the root mean squared error of prediction (RMSEP) for a quantitative PLS model; this model is calibrated to predict the level of the secondary metabolite gibberellic acid produced in an industrial bioprocess (McGovern *et al*., 2002).

PLS and DFA are very powerful quantification and classification modelling techniques; however, whilst loadings plots are generated by these algorithms, for full spectral analyses these are usually complex and so can be considered as 'black box' in nature. Therefore, insight into the main contributing variables or features within the spectra is not readily available. Recovering this information from biological datasets is important if one is to begin to generate new knowledge from the torrents of data that 'post-genomic' science is now able to generate.

In the field of bioinformatics there have been a number of reports showing the capability of GA to effect data reduction in order to improve the performance of predictive models. For example, for classification problems using gene expression data (Li *et al*., 2001; Ooi and Tan, 2003), improved classification accuracy was obtained following GA variable reduction. In a similar study, Chuzhanova *et al*. (1998) used GA with the Gamma (near-neighbour) test for feature selection of genetic sequence data, which again leads to improved classification results. GA optimization has also been applied to other bioinformatics-related problems such as sequence alignment (Notredame *et al*., 1998) and phylogenetic tree construction (Lewis, 1998).

In the post-genomic era where one of the main aims is to elucidate gene function, spectroscopic techniques are being applied in metabolic studies to obtain holistic fingerprints of biological samples, which can then be used for classification or identification. Furthermore, it is necessary to highlight the spectral variables that facilitate the discrimination. Therefore, the same general approach to variable reduction used in bioinformatics studies can be taken here; however, the challenge is to find a much smaller subset of features which may improve discrimination but importantly are also biochemically more informative. GAs have been used in such a way against a variety of different data types (Broadhurst *et al*., 1997; Ellis *et al*., 2002; Goicoechea and Olivieri, 2003; Johnson *et al*., 2003, 2004; Kinoshita *et al*., 1998; Konstam, 1993, 1994; Tapp *et al*., 2003). These approaches have yielded impressive results and have shown GAs to be powerful variable selection methods. However, to date the analysis has been performed in a binary manner, that is to say a 2 class (is it, isn't it) problem. This means that in multi-class problems it is necessary to perform a computationally inefficient pair-wise analysis with disproportionate population sizes. In addition, the variable selection needs to be conducted against a binary algorithm, which is obviously different from that against which the full multi-class discriminatory model has been formulated. For multi-class problems it would be beneficial to use the same modelling algorithm as a measure of fitness in a variable selection GA that is used in the actual modelling process itself. This approach would both reduce computation times and simplify the interpretation of results.

Hence, the aim of the second exercise in this study was to demonstrate the application of a GA to the selection of important discriminatory variables, using a multi-class DFA-based fitness function to differentiate between five types of bacteria commonly implicated in urinary tract infection (UTI). We form models on subsets of the whole dataset by GA optimization, finding combinations of variables that result in accurate DFA models.

SYSTEM AND METHODS

Fourier transform infrared spectroscopy

In both exercises the data were collected using the vibrational spectroscopic technique of FT-IR spectroscopy. Briefly, the FT-IR analyses were performed using a Bruker IFS28 infrared spectrometer equipped with a diffuse-reflectance TLC attachment (Bruker, Ltd., Coventry, UK) and a liquid nitrogen cooled MCT (mercury–cadmium–telluride) detector, as described previously (Goodacre *et al*., 1996, 1998). Mid-infrared spectra over the range 4000–600 cm^{-1} (256 co-adds, spectral resolution 4 cm^{-1}) were collected using an automated high-throughput accessory.

Dataset for spectral pre-processing exercise

The *Gibberella fujikuroi* bioprocess was studied which produces the compound gibberellic acid 3 in a complex undefined medium. Samples were provided by Zeneca Life Science Molecules. As already detailed in McGovern *et al.* (2002), whole broth samples were taken aseptically from the fermentation vessels at different stages of fermentation, and the fresh material was analysed by high-performance liquid chromatography (HPLC) to determine the gibberellin levels. Aliquots of the broth samples were stored at −20°C until a sufficient number and range had been collected for analysis by FT-IR. The full dataset consists of a total of 360 FT-IR spectra, representing gibberellin titre levels between 0 and 4925 parts per million (p.p.m.).

Dataset for spectral feature selection exercise

As previously reported (Goodacre *et al*., 1998), FT-IR spectral fingerprints were recorded from 59 clinical bacterial isolates of UTIs obtained from Bronglais Hospital, Aberystwyth. Typing by the biochemical test API120E and the genetic method of amplified fragment length polymorphism (AFLP) (Kassama *et al*., 2002) showed the isolates to belong to *Escherichia coli* (17, coded C), *Proteus mirabilis* (10, coded P), *Klebsiella* spp. (10, coded K), *Pseudomonas aeruginosa* (10, coded A) and *Enterococcus* spp. (12, coded E).

All strains were cultivated axenically and aerobically on LabM Malthus blood agar base (37 mg ml^{-1}) for 16 h at 37°C. After sub-culturing three times to ensure pure cultures, biomass was carefully collected using sterile plastic loops and suspended in 1 ml aliquots of sterile physiological saline (0.9% NaCl).

Overall methodology

The basic concept in both exercises, typical of many GA applications, is the optimization of a problem where a range of optimal solutions falls within a large search space. The GA is used to find an array of predictor sets which represent potential solutions. The problem is defined within a fitness function (described below) against which each individual is evaluated in order to provide a measure of its accuracy.

ALGORITHMS

Genetic algorithm

The GA is an evolutionary computing technique that can be used to solve problems efficiently for which there are many possible solutions (Holland, 1992). In GAs, the initial step is to generate a random

population (array), consisting of a predefined number of individuals (rows) and variables (columns) Each individual represents a subset of the original variables within the larger superset of data under analysis. The next step in the GA is analogous to the process of Darwinian evolution whereby, through the processes of crossover, mutation and survival of the fittest, individuals are selected for the next generation until a particular stopping criterion has been reached. The GA uses an algorithm known as a fitness function to assess the robustness of the model proposed by each individual. This usually takes the form of a minimization function; therefore the fittest individuals are those with the lowest fitness value.

Typically, the evolutionary stage of a simple GA proceeds as follows: (1) extract a proportion of the fittest individuals from the current (parent) population, (2) recombine the selected offspring (crossover), (3) mutate the mated population, (4) assess the newly evolved offspring for fitness, (5) reinsert a proportion of the offspring into the population, replacing the worst parents, and (6) repeat the process until a stopping criterion is reached. This is perhaps best summarized with pseudocode, as shown below.

```
begin
   create initial population
   evaluate initial population
   gen = 0
   max_gen = N
   while (gen < max_gen) do
         gen+ = 1
         select sub-population from initial population
         recombine 'genes' of selected sub-population
         mutate recombined offspring
         evaluate offspring
         reinsert best offspring replacing worst parents
   end while
```

The stopping criterion can be defined in many different ways. For example, it could be simply defined as a maximum number of generations, a maximum target outcome value for the fitness, or as a set number of generations for which the fitness value for the fittest individual has remained constant.

Partial least squares

PLS (Martens and Naes, 1989) is a supervised linear modelling technique that can be used to formulate empirical models from multivariate datasets. The general representation of the least squares model is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{B} + \boldsymbol{C}, \tag{1}$$

where $\mathbf{Y}$ is a $n \times m$ matrix of dependent variables (e.g. gibberellic acid titres), $\mathbf{X}$ is a $n \times p$ matrix of independent variables (e.g. FT-IR spectra), $\boldsymbol{B}$ is a $p \times m$ matrix of regression parameters and $\boldsymbol{C}$ are the differences between the measured and predicted data, or the residuals.

To formulate the model, $\boldsymbol{B}$ needs to be calculated; this involves computing the inverse of X. For a non-square $\mathbf{X}$ this calculation is not trivial, and therefore methods have been developed by which the regression parameters can be accurately estimated. In this study the dependent variables are represented by an $n \times 1$ column vector. For models of this type where $\mathbf{Y}$ consists of a single constituent, the dependent variables are used directly as the estimate of $\boldsymbol{B}$.

Discriminant function analysis

A powerful supervised clustering method used frequently in microbial classification and discrimination exercises is DFA (Manly, 1994), also known as canonical variate analysis (CVA). The method generates a number of linear discriminant functions for separating groups by finding the eigenvalues and eigenvectors of the expression

$$\mathbf{W}^{-1}\mathbf{B}, \tag{2}$$

where $\mathbf{W}$ is the within-sample matrix of sums of squares and cross products and $\mathbf{B}$ is the difference between the total-sample matrix sums of squares and cross products and $\mathbf{W}$. The group structure is specified to the algorithm *a priori*. This implementation of DFA using Manly's principles (Manly, 1994) finds the linear combination of discriminant functions that maximize the Fisher (F) ratio (ratio of between-group to within-group variance) for the dataset.

Model validation

To test the reproducibility of the models obtained from analysis by both PLS and DFA, the method of validation by projection was employed. When a model is formulated there is a possibility that it will over-fit the data; this means that a relationship is found between the data and the target class structure or dependent variables that does not hold for subsequent predictions, i.e. the model has learnt the training data perfectly and is not able to generalize. Therefore, the datasets were randomly split into three groups, a model training set, a cross validation set and an independent test set. The models were built and optimized on the training and cross validation sets and then independently tested against the third set of 'hold-out' data.

IMPLEMENTATION AND DISCUSSION

Optimization of spectral data pre-processing

The purpose of this exercise was to minimize the RMSEP (Allen, 1971) for a quantitative PLS regression model predicting gibberellic acid titre concentrations in *G.fujikuroi* bioprocesses (McGovern *et al*., 2002). The RMSEP is given as

$$\text{RMSEP} = \sqrt{\sum_{i=1}^{n}(y_{\text{act}} - y_{\text{pred}})^2/n}, \tag{3}$$

where y_{act} are the actual dependent variables, y_{pred} are the predicted dependent variables and n is the number of objects.

A GA was used to recover subsets of pre-processing functions from a total of 50 alternatives (Table 1). These are by no means a comprehensive list of spectral pre-processing algorithms available to the analyst but represent the broad range of tools which in general can be split into scaling, filtering, baseline correction and derivatization categories. A maximum of 6 pre-processing functions (50^6 or $\sim 10^{10}$ possible combinations) could be selected by the algorithm, and an incremental penalty was applied to any individuals selecting more than three functions. Briefly, experimentally determined penalties of 4, 9 and 16 points were added to the objective function scores of those individuals selecting more than three, four and five pre-processing functions respectively. The consequence of determining these values empirically is that the same penalty system may not be relevant for different sets of data; therefore in future an improvement needs to be made to the way this problem is approached. The reason for applying penalties to the objective scores of certain individuals is based

Table 1. The 50 pre-processing functions made available for evaluation by the genetic algorithm for the PLS optimization study

Category	Description	Variables	Coded
Baseline correction[a]	Normalize first bin to zero	—	B_1
Baseline correction[a]	Subtract average of the first and last bin	—	B_2
Baseline correction[a]	Detrend by subtracting a linearly increasing baseline	—	B_3
Derivatization[a]	Derivatizes a straight-line fit approximated across a sliding window of size M	$M = 3$	D_1
		$M = 4$	D_2
		$M = 5$	D_3
		$M = 6$	D_4
		$M = 7$	D_5
		$M = 8$	D_6
		$M = 9$	D_7
		$M = 10$	D_8
		$M = 11$	D_9
		$M = 12$	D_{10}
		$M = 13$	D_{11}
		$M = 14$	D_{12}
		$M = 15$	D_{13}
Derivatization[a]	Savitzky–Golay derivatization or order K and frame size F	$K = 3; F = 3$	D_{14}
		$K = 5; F = 5$	D_{15}
		$K = 7; F = 7$	D_{16}
		$K = 9; F = 9$	D_{17}
		$K = 9; F = 11$	D_{18}
		$K = 9; F = 13$	D_{19}
		$K = 9; F = 15$	D_{20}
Filtering[a]	One-dimensional mean filter of order N	$N = 2$	F_1
		$N = 3$	F_2
		$N = 4$	F_3
		$N = 5$	F_4
		$N = 6$	F_5
		$N = 7$	F_6
Filtering[b]	Smoothes using a Savitzky–Golay (polynomial) smoothing filter of order K and frame size F	$K = 3; F = 5$	F_7
		$K = 3; F = 7$	F_8
		$K = 3; F = 9$	F_9
		$K = 3; F = 11$	F_{10}
		$K = 3; F = 13$	F_{11}
		$K = 3; F = 15$	F_{12}
		$K = 4; F = 7$	F_{13}
		$K = 4; F = 9$	F_{14}
		$K = 4; F = 11$	F_{15}
		$K = 4; F = 13$	F_{16}
		$K = 4; F = 15$	F_{17}
		$K = 5; F = 7$	F_{18}
		$K = 5; F = 9$	F_{19}
Scaling[a]	Normalize maximum and minimum between 0 and 1	—	S_1
Scaling[a]	Autoscale—mean centre sample and scale to unit variance of the variable	—	S_2
Scaling[a]	Mean centre	—	S_3
Scaling[a]	Normalize to total (sum equals unity)	—	S_4
Scaling[a]	Vector normalization	—	S_5
Scaling[a]	Autoscale—mean centre variables and scale to unit variance of the samples	—	S_6

[a]MATLAB script developed in-house.
[b]MATLAB Signal Processing Toolbox.

on the assumption that pre-processing regimes using fewer functions are more likely to reflect sensible answers. From preliminary work it was clear that applying GA-derived combinations of large numbers of functions to the data could dramatically reduce the model RMSEP; but these solutions were always difficult if not impossible to scientifically justify. Therefore, by penalizing such individuals one can be assured that only those with exceptional performance survive and that the population does not become dominated by excessive pre-processing events, a strategy adopted by the genetic programming field to avoid 'bloat', a phenomenon in which the GP function trees gets so huge that it lacks explanatory power (Langdon and Poli, 2002; Podgorelec and Kokol, 2000).

The comprehensive Genetic Algorithm Toolbox for MATLAB (Chipperfield and Fleming, 1995; Chipperfield *et al*., 1994a,b) (http://www.shef.ac.uk/~gaipp/ga-toolbox/) was configured as a simple GA (SGA). Populations of 30 individuals with six variables were encoded as real values, with evolution driven by single point crossover and mutation; the selection function used at each generation was stochastic universal sampling (SUS). The stopping criterion terminated the GA after the fittest individual remained unchanged for 20 consecutive generations. A total of 50 independent GA runs were performed, resulting in an average of 67 (range 40–100) generations per run. Due to the large size of the dataset (360×882), the cumulative processing time was ~5 days in total, processing ~100 000 individuals at up to six calculations per individual. Therefore, an exhaustive search of every possible subset of pre-processing function combinations would require $\sim 0.5 \times 10^6$ days computation time on a 1.6 GHz PC.

In Table 2 the best results from each of the independent runs are provided with the selected pre-processing functions encoded as detailed in Table 1. For comparison, details are shown for the model constructed against the raw data and for exhaustive searches using a single function. Included in the latter are the result for the best model and the results for the models generated after pre-processing by S_2 and S_5, which are the functions most frequently used by the GA in this example. The values for RMSEP of the training, cross validation and test datasets ($RMSEP_{train}$, $RMSEP_{cval}$ and $RMSEP_{test}$ respectively) are listed, the number of PLS factors was optimized using the cross validation data and it was the $RMSEP_{cval}$ that was used in the fitness function. Finally, the $RMSEP_{test}$ was calculated following projection into each model of the independent test data. The most frequent pre-processing combination selected by the GA is the two-step process of autoscaling (S_2) followed by vector normalization (S_5), and this produces an $RMSEP_{test}$ that is ~16% better than that obtained using the raw data. As can be seen from Table 2, this result is also better than that obtained by using either S_2 or S_5 independently.

In Figure 1 the raw and pre-processed datasets and their associated PLS models are plotted. A close inspection of the PLS model in Figure 1B shows that there is much greater convergence on the ideal linear line compared to the model using the raw data (Fig. 1A). However, given the high dimensionality of the spectra and the complex nature of PLS loadings plots we had seen previously that it was not possible to pick out a single region in the loading plot that was contributing most to the analysis (McGovern *et al*., 2002). Therefore, as a final step the established technique of variable selection by GA-PLS (Broadhurst *et al*., 1997) was performed on the processed *G. fujikuroi* FT-IR spectra, in order to determine the most important variables for PLS calibration modelling of this bioprocess.

Table 2. The highest ranking results, giving the order of FT-IR spectra pre-processing, for each of the 50 independent GA runs evolved to quantify gibberellic acid from the *Gibberella fujikuroi* fermentation broth

Pre-processing steps						Number of steps	RMSEC (%)[a]	$RMSEP_{eval}$ (%)[b]	$RMSEP_{test}$ (%)[c]	Number of factors
Raw data										
—	—	—	—	—	—	0	3.75	4.87	5.85	10
Result for one-step exhaustive search										
B_3	—	—	—	—	—	1	3.96	4.64	5.57	9
S_2	—	—	—	—	—	1	4.25	4.57	5.92	8
S_5	—	—	—	—	—	1	3.74	4.75	5.92	8
GA optimization order										
S_1	B_2	S_2	S_1	S_6	—	5	3.46	4.30	4.85	9
S_2	S_4	B_3	S_3	S_6	—	5	3.52	4.29	4.86	8
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_4	S_5	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	S_5	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	S_2	S_5	—	—	2	3.21	4.31	4.92	9
S_2	S_5	S_2	S_5	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_5	—	—	—	—	2	3.21	4.31	4.92	9
S_2	S_3	S_5	—	—	—	3	3.54	4.34	5.02	8
S_2	S_5	S_3	—	—	—	3	3.54	4.34	5.02	8
S_1	S_5	S_2	—	—	—	3	3.57	4.36	5.15	8
S_5	S_2	—	—	—	—	3	3.57	4.36	5.15	8
S_2	S_1	B_3	S_5	—	—	4	3.59	4.29	5.25	8
S_2	S_2	B_1	—	—	—	3	4.18	4.30	5.27	9
S_2	S_5	D_{13}	S_1	S_6	—	5	3.94	4.37	5.39	7
S_2	B_1	S_5	—	—	—	3	3.66	4.35	5.43	8
S_2	S_5	S_2	—	—	—	3	3.63	4.33	5.47	8
B_3	S_3	—	—	—	—	2	4.07	4.56	5.51	8
S_2	S_5	B_2	—	—	—	3	3.63	4.34	5.54	8
D_{13}	B_3	S_5	—	—	—	3	4.13	4.54	5.76	5
D_{12}	S_5	—	—	—	—	2	4.09	4.57	5.79	5
S_2	B_3	—	—	—	—	2	4.37	4.38	5.86	8
S_2	B_3	—	—	—	—	2	4.37	4.38	5.86	8
D_{12}	S_1	—	—	—	—	2	4.56	4.57	6.03	6
S_1	S_3	S_2	D_8	—	—	4	4.50	4.46	6.03	5
S_5	D_{11}	—	—	—	—	2	3.95	4.56	6.20	5

Table 2. *Continued*

Pre-processing steps						Number of steps	RMSEC (%)[a]	$RMSEP_{eval}$ (%)[b]	$RMSEP_{test}$ (%)[c]	Number of factors
S_4	S_2	—	—	—	—	2	4.15	4.52	6.28	8
S_4	S_2	—	—	—	—	2	4.15	4.52	6.28	8
S_2	D_7	—	—	—	—	2	5.41	4.67	6.56	5

[a]Root mean squared error of calibration ($RMSEP_{train}$)
[b]Root mean squared error pf prediction for the cross validations
[c]Root mean squared error of prediction for the independent test set.
Each pre-processing option was applied to the data individually, and the transformed array was assessed against the fitness function used in the GA. The function giving the best performance against the fitness function is listed (B_3) along with the results for the S_2 and S_5 which are most frequently selected by the GA in the main study.

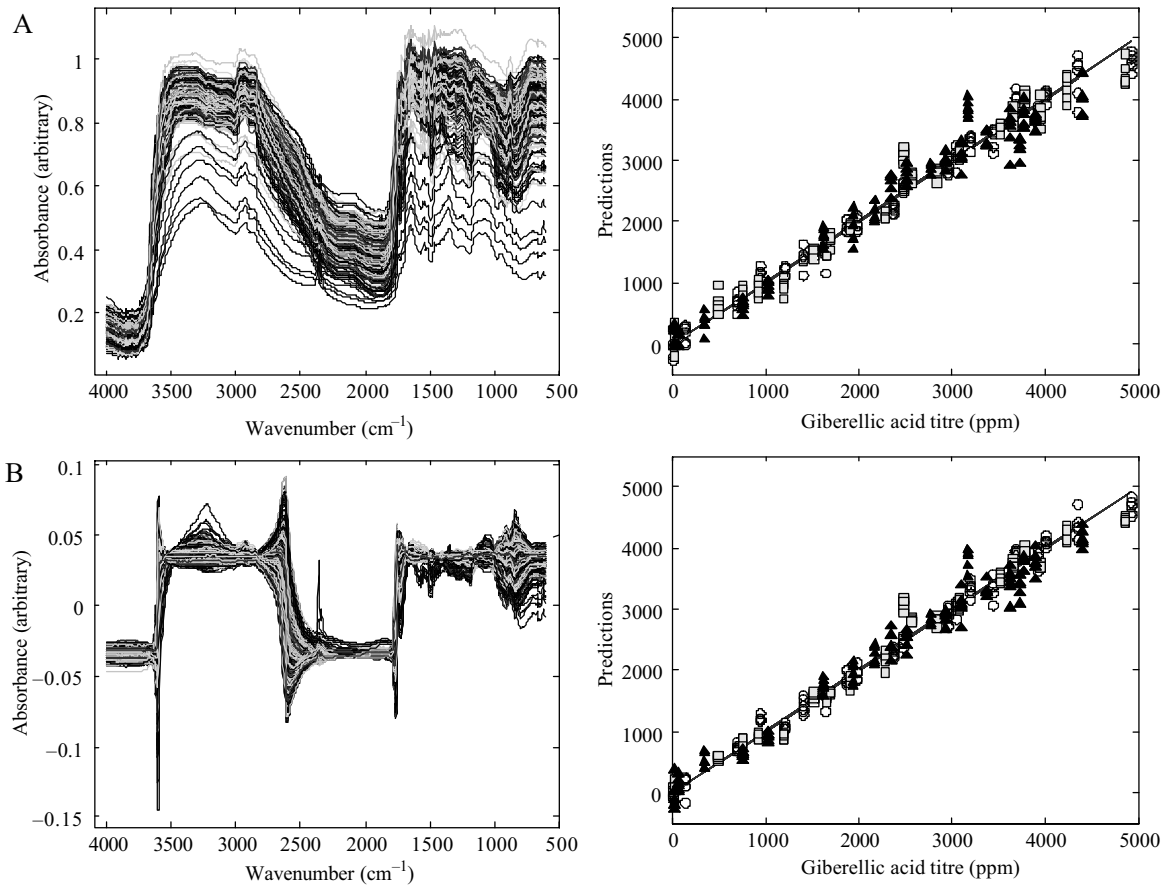

Fig. 1. The fermentation spectra and their associated PLS models, with the training objects shown as open circles, cross validation objects as shaded squares and test objects as black triangles, for the following: (**A**), the raw FT-IR data ($RMSEP_{test}$ 5.97%); (**B**), the most frequent model generated by the GA using functions S_2 and S_5 in that order ($RMSEP_{test}$ 4.92%). In both A and B the solid line shows the expected linear fit.

The GA was configured to select a subset of five variables, a total population size of 250 individuals was used and 50 independent GA-PLS runs were performed on the pre-processed FT-IR spectra. This analysis revealed the spectral region from 1778 to 1770 cm^{-1} as being particularly important for modelling this bioprocess (data not shown), which corresponds with GA-MLR results on the same data (McGovern *et al*., 2002). The identified region is rich in carbonyl vibrations, a functional group that is present in gibberellic acid 3, and provides confirmation that FT-IR analysis of this bioprocess is measuring the formation of the product rather than the disappearance of any substrate(s).

Spectral feature selection

The second exercise sought to demonstrate the application of a DFA-based fitness function for spectral feature selection by GA, which we will now refer to as GA-DFA. When the full FT-IR dataset from UTI was analysed by principal component analysis (PCA) followed by DFA (PC-DFA) as detailed in Goodacre *et al*. (1998), each of

the five different organisms could be differentiated. However, the spectral bands that most contribute to this discrimination could not be revealed by the DFA algorithm.

Since one cannot feed collinear variables or too many variables into DFA, it is usually necessary with hyperspectral data to perform some sort of data reduction step prior to modelling. The reason for this is that $\mathbf{W}^{-1}$ term in Equation (2) above can only exist when the input matrix is non-singular, i.e. its determinant is other than zero, which implies that it is of full rank (Dixon, 1975; MacFie *et al*., 1978); this is generally the case if

$$(N_s - N_g - 1) > N_v \tag{4}$$

where N_s is the number of samples, N_g is the number of groups and N_v is the number of inputs.

An aspect of performing data reduction is that a small subset of variables can be recovered, which still give a predictive model, and these variables will be indicative of which spectral bands are discriminatory and hence of biological or biochemical importance. This GA-DFA implementation uses the eigenvalues computed by DFA as an approximation for the F-ratio, with the GA fitness function configured to minimize the inverse sum of the eigenvalues in order to formulate increasingly better DFA models. The GA was configured as a multi-population genetic algorithm (MGA) in order to more efficiently search a solution space 2 orders of magnitude larger than that in the first study. Populations of 20 individuals across five subpopulations used real-value encoding, with evolutionary and termination criteria set as for the SGA. Duplicate variables were not allowed in any one individual. In addition, a migration parameter was defined in order to establish the fittest individuals in a single subpopulation every 20 generations.

FT-IR spectra often contain many local collinearities; that is to say, a peak is sampled many times and this may have arisen from the same chemical species. Thus, in order to remove these local collinearities, the full set of 882 FT-IR spectral variables were reduced to 98 variables by selecting only the maxima, minima and stationary points from the spectra, as detailed in Johnson *et al*. (2003, 2004), and performed using a MATLAB script written in-house.

To decide on the minimum number of variables that could be extracted from the dataset whilst maintaining the maximum level of discrimination, initially 10 independent GA-DFA runs were performed for the selection of subsets of 3, 4, 5, 6, 7, 8, 9 and 10 variables. From these results the average inner (within-group) variance and the average distance between group centres were plotted (Fig. 2). This plot indicates that when selecting small numbers of variables the training groups are well separated but that the models are not predictive for the test set data. In contrast, when more variables are selected the training and test groups are more disperse and therefore the mean distance between them is lower. This is particularly the case with the selection of 9 or 10 inputs to DFA where the within-group variance is very high.

It is clear from Figure 2 that the optimal number of variables to be selected is 6, and consequently further experiments were performed to select this number of inputs to DFA. Thus, for the main study, 100 independent GA-DFA runs were performed. The plot in Figure 3 shows the frequency with which variables were selected based on the best result from each independent run, superimposed on the mean spectrum of the whole dataset. These variables are defined in Table 3 along with their general FT-IR wavenumber assignment,

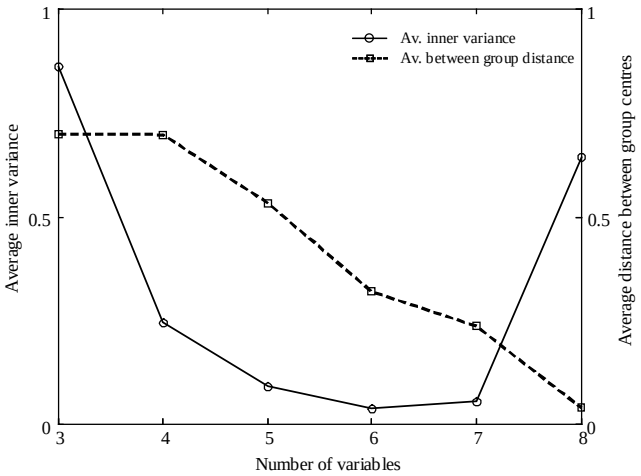

Fig. 2. Results of the GA-DFA models programmed to select subsets of 3, 4, 5, 6, 7, 8, 9 and 10 variables for differentiating between bacteria from their FT-IR spectra. Each subset is performed 10 times independently. From these results the average inner variance and average between group distances are plotted in order to determine the optimal number of variables to be selected in the main study. The results from subsets of 3–8 variables only are shown here since the much greater magnitude of results from 9 and 10 variables skews the scaling of the plot.

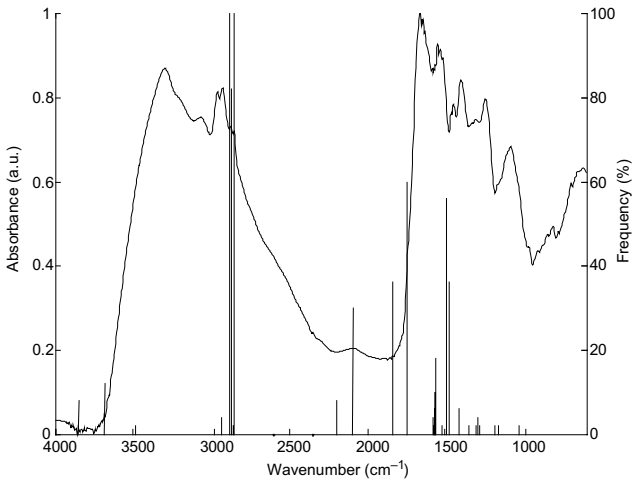

Fig. 3. A plot showing the FT-IR variables selected by GA-DFA for differentiating between urinary tract isolates. The frequency plot resulting from 100 independent GA runs reveals important discriminatory variables in the fatty acid region of the spectra (Table 3). The GA was programmed to select six variables, and total computation time for the 100 runs was 60 min.

which shows that bands within the fatty acid and amide regions are predominantly selected by GA-DFA.

Using only the variables in Table 3 that were identified by GA-DFA analysis, the fully validated DFA model in Figure 4 was generated. The full FT-IR spectra consist of 882 variables. In Figure 4 we show that six key variables can generate a good discriminatory model and therefore point to interesting (bio)-chemical differences between complex samples such as bacteria. Inspection of the DFA loading plot (data not shown) indicated that no single input was

Table 3. Frequencies of occurrence and general wavenumber assignments of the six most frequently used variables in GA-DFA analysis

Frequency (%)	Wavenumber (cm^{-1})	General assignment
100	2885	CH_3 symmetric stretch in fatty acids
100	2854	CH_2 or CH_3 symmetric stretch in fatty acids
82	2874	CH_3 symmetric stretch of methyl in fatty acids
60	1747	C═O stretch from amide I region
56	1497	Imide─CO─NH─CO─ deformation in nucleic acids
36	1477	Ring stretch or CH_3 Symmetric stretch in fatty acids

Assignments from Degen (1997) and Naumann (2001).

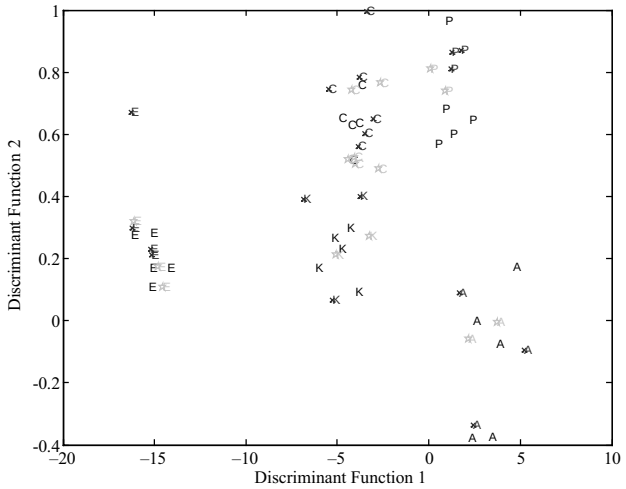

Fig. 4. DFA bi-plot where DFA was performed directly on the FT-IR UTI data for the six most frequently occurring variables (2885, 2874, 2854, 1747, 1497 and 1477 cm^{-1}) as selected by the GAs. The resultant ordination plot (mean replicate centres are shown for clarity) clearly shows that five clusters are recovered. The isolates are encoded as follows; A, *Pseudomonas aeruginosa*; C, *Escherichia coli*; E, *Enterococcus* spp.; K, *Klebsiella* spp.; P, *Proteus mirabilis*. The cross validation data are denoted by crosses and the independent test data by pentagram markers.

important for the discrimination, implying that the combination of these six variables was necessary. The discrimination between the Gram-positive *Enterococcus* spp. and the other isolates which are all Gram-negative is accounted for by variances across all six variables identified by the GA. Certainly, removing each variable in turn from the DFA model does not adversely affect the distance between this organism and the other four bacteria. This is not surprising given the vast differences in cell wall structure that can be observed between Gram-positive and Gram-negative microbes, which are consequently heavily reflected in their FT-IR spectra.

Eliminating the enterococci from the analysis and sequentially remodelling on just five variables (i.e. generating new models using each possible combination of five of the six GA-DFA selected variables) allows us to identify which variables are of importance in discrimination between each of the groups. *P. mirabilis E. coli* and *Klebsiella* spp. all belong to the family Enterobacteriaceae, whilst *P. aeruginosa* belongs to a different family altogether and is dissimilar phenotypically; therefore discrimination between these two groups remains reasonably consistent upon removal of any one variable. However, discrimination between *E. coli* and *Klebsiella* is lost upon removal of either of variables 2885 or 2874 cm^{-1}, Whereas discrimination between *Proteus*, *E. coli* and *Klebsiella* breaks down when variable 2854 cm^{-1} is removed from the analysis. This shows that these variables reflect biological or biochemical differences between those organisms. The implication of obtaining such knowledge from FT-IR spectra is that further targeted spectroscopic analysis by gas chromatography mass spectrometry (GC-MS) or liquid chromatography mass spectrometry (LC-MS) could be more rapidly performed in order to exactly determine the discriminatory biochemistry which enables hypothesis induction. There is also an opportunity, given this information, to look into developing inexpensive and rapid instrumentation or bioassays, specifically for the discrimination of these strains.

CONCLUSIONS

Most, if not all, calibration or classification studies of biological samples conducted with vibrational spectroscopic techniques use some form of data pre-processing prior to input into an appropriate mathematical model to compensate for experimental and machine variability. To date, any such pre-processing has been applied using algorithms favoured by the analyst often using the moistened finger raised vertically approach and/or already proven on a particular type of data. However, the large numbers of algorithms available, many of which are subtle variations on the same theme, make routine exhaustive searches for the best pre-processing solution impractical. We clearly demonstrate that GA optimization of spectral pre-processing enables the rapid search for optimal or near-optimal solutions that can then allow for the most robust and accurate models to be generated.

The outputs from supervised multivariate modelling methods do not readily reveal the most important variables used to formulate a model. In post-genomic studies more emphasis is being placed on obtaining useful information from data, as opposed to just being satisfied with a well-constructed model. For this purpose, in order to reveal important variables in large hyperspectral datasets GAs have already been adopted. However, these studies either have been interested in simple binary classification problems, or have used GA coupled with binary classifiers such as LDA (Fisher DA) or PLS1 to perform variable selection on multi-class problems by having multiple binary classifiers (that is to say, one for each class). We believe that a much more robust approach can be made when all sample types are modelled simultaneously, as this will remove any bias due to pair-wise analysis and disproportionate population sizes. Using this strategy we show that GA-DFA can be used to select small numbers of variables from the full dataset to formulate robust models using a multi-class modelling algorithm. This gives an indication of those variables that are most important for discrimination and therefore of (bio-)chemical interest.

ACKNOWLEDGEMENTS

We are very grateful to Dr. Helen E. Johnson and Dr. David Broadhurst for useful discussions and encouragement. R.M.J. is indebted to Renishaw plc. and the UK EPSRC for his CASE PhD

studentship. R.G. also thanks the UK BBSRC Engineering and Biological Systems Committee and EPSRC for financial support.

REFERENCES

Allen,D.M. (1971) Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13**, 469–475.

Arnold,S.A., Crowley,J., Vaidyanathan,S., Matheson,L., Mohan,P., Hall,J.W., Harvey,L.M. and McNeil,B. (2000) At-line monitoring of a submerged filamentous bacterial cultivation using near-infrared spectroscopy. *Enzyme Microb. Technol.*, **27**, 691–697.

Bäck,T., Fogel,D.B. and Michalewicz,Z. (1997) *Handbook of Evolutionary Computation*. IOPPublishing/Oxford University Press, Oxford.

Blackstock,W.P. and Weir,M.P. (1999) Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.*, **17**, 121–127.

Broadhurst,D., Goodacre,R., Jones,A., Rowland,J.J. and Kell,D.B. (1997) Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Anal. Chim. Acta*, 71–86.

Burge,C.B. (2001) Chipping away at the transcriptome. *Nat. Genet*, **27**, 232–234.

Chipperfield,A.J. and Fleming,P.J. (1995) The MATLAB Genetic Algorithm Toolbox. *IEE Colloquium Applied Control Techniques Using MATLAB*. pp. 10/11–10/14.

Chipperfield,A.J., Fleming,P.J. and Fonseca,C.M. (1994b) Genetic Algorithm Tools for Control Systems Engineering. *Proceedings of Adaptive Computing in Engineering Design and Control*. Plymouth Engineering Design Centre, pp. 128–133.

Chipperfield,A.J., Fleming,P.J. and Pohlheim,H. (1994a) A Genetic Algorithm Toolbox for MATLAB. *Proceedings of International Conference on Systems Engineering*. Coventry, UK, pp. 200–207.

Chuzhanova,N.A., Jones,A.J. and Margetts,S. (1998) Feature selection for genetic sequence classification. *Bioinformatics*, **14**, 139–143.

Degen,I.A. (1997) *Tables of Characteristic Group Frequencies for the Interpretation of Infrared and RAMAN Spectra*. Acolyte Publications, Harrow, UK.

Dixon,W. (1975) *Biomedical Computer Programs.* University of California Press, Los Angeles.

Ellis,D.I., Broadhurst,D., Kell,D.B., Rowland,J.J. and Goodacre,R. (2002) Rapid and quantitative detection of the microbial spoilage of meat by Fourier transform infrared spectroscopy and machine learning. *Appl. Environ. Microbiol.*, **68**, 2822–2828.

Fiehn,O. (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155–171.

Fiehn,O., Kopka,J., Dörmann,P., Altmann,T., Trethewey,R.N. and Willmitzer,L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, **18**, 1157–1161.

Goicoechea,H.C. and Olivieri,A.C. (2003) A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy. *J. Chemometr.*, **17**, 338–345.

Goldberg,D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.

Goodacre,R., Neal,M.J. and Kell,D.B. (1994) Rapid identification using pyrolysis mass spectrometry and artificial neural networks of *Propionibacterium acnes* isolated from dogs. *J. Appl. Bacteriol.*, **76**, 124–134.

Goodacre,R., Timmins,É.M., Burton,R., Kaderbhai,N., Woodward,A., Kell,D.B. and Rooney,P.J. (1998) Rapid identification of urinary tract infection bacteria using hyperspectral, whole organism fingerprinting and artificial neural networks. *Microbiology*, **144**, 1157–1170.

Goodacre,R., Timmins,E.M., Rooney,P.J., Rowland,J.J. and Kell,D.B. (1996) 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.

Goodacre,R., Vaidyanathan,S., Dunn,W.B., Harrigan,G.G. and Kell,D.B. (2004) Metabolomics by numbers – acquiring and understanding global metabolite data. *Trends Biotechnol.*, **22**, 245–252.

Holland,J.H. (1992) *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA.

Jarvis,R.M. and Goodacre,R. (2004a) Ultra-violet resonance Raman spectroscopy for the rapid discrimination of urinary tract infection bacteria. *FEMS Microbiol. Lett.*, **232**, 127–132.

Jarvis,R.M. and Goodacre,R. (2004b) Rapid discrimination of bacteria using surface enhanced Raman spectroscopy. *Anal. Chem.*, **76**, 40–47.

Johnson,H.E., Broadhurst,D., Goodacre,R. and Smith,A.R. (2003) Metabolic fingerprinting of salt-stressed tomatoes. *Phytochemistry*, **62**, 919–928.

Johnson,H.E., Broadhurst,D., Kell,D.B., Theodorou,M.K., Merry,R.J. and Griffith,G.W. (2004) High-throughput metabolic fingerprinting of legume silage fermentations via Fourier transform infrared spectroscopy and chemometrics. *Appl. Environ. Microbiol.*, **70**, 1583–1592.

Kassama,Y., Rooney,P.J. and Goodacre,R. (2002) Fluorescent amplified fragment length polymorphism probabilistic database for identification of bacterial isolates from urinary tract infections. *J. Clin. Microbiol.*, **40**, 2795–2800.

Kell,D.B. and Oliver,S.G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, **26**, 99–105.

Kinoshita,E., Ozawa,Y. and Aishima,T. (1998) Differentiation of soy sauce types by HPLC profile pattern recognition – isolation of novel isoflavones. In *Flavonoids in the Living System*. Plenum Press, New York, pp. 117–129.

Konstam,A.H. (1993) Linear discriminant analysis using genetic algorithms. *Proceedings of the 1993 ACM/SIGAPP Symposium on Applied computing: States of the Art and Practice*. ACM Press, Indianapolis, IN, pp. 152–156.

Konstam,A.H. (1994) N-Group classification using genetic algorithms. *Proceedings of the 1994 ACM Symposium on Applied Computing*. ACM Press, Phoenix, AZ, pp. 212–216.

Langdon,W. and Poli,R. (2002) *Foundations of Genetic Programming*. Springer-Verlag, Berlin.

Lewis,P. (1998) A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Mol. Biol. Evol.*, **15**, 277–283.

Li,L., Weinberg,C.R., Darden,T.A. and Pederson,L.G. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131–1142.

Lopez-Diez,E.C. and Goodacre,R. (2004) Characterization of microorganisms using UV resonance Raman spectroscopy and chemometrics. *Anal. Chem.*, **76**, 585–591.

MacFie,H., Gutteridge,C. and Norris,J. (1978) Use of canonical variates in differentiation of bacteria by pyrolysis gas-liquid chromatography. *J. Gen. Microbiol.*, **104**, 67–74.

Manly,B.F.J. (1994) *Multivariate Statistical Methods: A Primer*, 2nd edn. Chapman & Hall/CRC, New York.

Maquelin,K., Choo-Smith,L.P., van Vreeswijk,T., Endtz,H.P., Smith,B., Bennett,R., Bruining,H.A. and Puppels,G.J. (2000) Raman spectroscopic method for identification of clinically relevant microorganisms growing on solid culture medium. *Anal. Chem.*, **72**, 12–19.

Martens,H. and Naes,T. (1989) *Multivariate Calibration*. Wiley, Chichester, UK.

McGovern,A.C., Broadhurst,D., Taylor,J., Kaderbhai,N., Winson,M.K., Small,D.A., Rowland,J.J., Kell,D.B. and Goodacre,R. (2002) Monitoring of complex industrial bioprocesses for metabolite concentrations using modern spectroscopies and machine learning: application to gibberellic acid production. *Biotechnol. Bioeng.*, **78**, 527–538.

Mitchell,M. (1995) *An Introduction to Genetic Algorithms*. MIT Press, Boston, MA.

Naumann,D. (2001) FT-infrared and FT-Raman spectroscopy in biomedical research. *Appl. Spectrosc. Rev.*, **36**, 239–298.

Naumann,D., Helm,D. and Labischinski,H. (1991) Microbiological characterizations by FT-IR spectroscopy. *Nature*, **351**, 81–82.

Notredame,C., Holm,L. and Higgins,D. (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, **14**, 407–422.

Ooi,C.H. and Tan,P. (2003) Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, **19**, 37–44.

Podgorelec,V. and Kokol,P. (2000) Fighting program bloat with the fractal complexity measure. *Lecture Notes in Computer Science, Genetic Programming Proceedings*, **1802**, 326–337.

Tapp,H.S., Defernez,M. and Kemsley,E.K. (2003) FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils. *J. Agric. Food Chem.*, **51**, 6110–6115.

Timmins,E.M., Quain,D.E. and Goodacre,R. (1998) Differentiation of brewing yeast strains by pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. *Yeast*, **14**, 885–893.

Vaidyanathan,S., Kell,D.B. and Goodacre,R. (2002a) Rapid, high-throughput microbial characterization by metabolite and protein profiling of whole cells using soft-ionization mass spectrometry. *Abstr. Pap. Am. Chem. Soc.*, **224**, 011-BIOT.

Vaidyanathan,S., Kell,D.B. and Goodacre,R. (2002b) Flow-injection electrospray ionization mass spectrometry of crude cell extracts for high-throughput bacterial identification. *J. Am. Soc. Mass Spectrom.*, **13**, 118–128.

Vaidyanathan,S., Macaloney,G., Harvey,L.M. and McNeil,B. (2001) Assessment of the structure and predictive ability of models developed for monitoring key analytes in a submerged fungal bioprocess using near-infrared spectroscopy. *Appl. Spectrosc.*, **55**, 444–453.

Weckwerth,W. (2003) Metabolomics in systems biology. *Ann. Rev. Plant Biol.*, **54**, 669–689.