Explanatory analysis of the metabolome using genetic programming of simple, interpretable rules

Helen E. Johnson, Richard J. Gilbert, Michael K. Winson,
Royston Goodacre, Aileen R. Smith, Jem J. Rowland[†], Michael A. Hall and Douglas B. Kell

Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion SY23 3DD, UK
[†]Department of Computer Science, University of Wales, Aberystwyth, Ceredigion SY23 3DB, UK

rcg@aber.ac.uk, hej93@aber.ac.uk, mkw@aber.ac.uk, rrg@aber.ac.uk, jjr@aber.ac.uk,
ars@aber.ac.uk, mzh@aber.ac.uk dbk@aber.ac.uk

Corresponding Author: Richard Gilbert. Tel: +44 (0)1970 622353. Fax: +44 (0)1970 622354. http://gepasi.dbs.aber.ac.uk/rcg

Abstract

Genetic programming, in conjunction with advanced analytical instruments, is a novel tool for the investigation of complex biological systems at the whole-tissue level.

In this study, samples from tomato fruit grown hydroponically under both high- and low-salt conditions were analysed using Fourier-transform infrared spectroscopy (FTIR), with the aim of identifying spectral and biochemical features linked to salinity in the growth environment.

FTIR spectra of whole tissue extracts are not amenable to direct visual analysis, so numerical modelling methods were used to generate models capable of classifying the samples based on their spectral characteristics. Genetic programming (GP) provided models with a better prediction accuracy to the conventional data modelling methods used, whilst being much easier to interpret in terms of the variables used.

Examination of the GP-derived models showed that there were a small number of spectral regions that were consistently being used. In particular, the spectral region containing absorbances potentially due to a cyanide/nitrile functional group was identified as discriminatory. The explanatory power of the GP models enabled a chemical interpretation of the biochemical differences to be proposed. The combination of FTIR and GP is therefore a powerful and novel analytical tool that, in this study, improves our understanding of the biochemistry of salt tolerance in tomato plants.

Introduction

The metabolome is a generic term for the total biochemical composition of a cell or tissue sample at any given time (Oliver *et al*., 1998). Recent advances in DNA sequencing have lead to an explosion in the number of known gene sequences, but the majority of these new genes have never been characterised experimentally, and many have completely unknown functions within the cell (Bork *et al*., 1998; Cole *et al*., 1998; Hinton, 1997). By investigating the changes in the metabolome of biological systems under different conditions, it is hoped that previously undescribed metabolic processes or pathways may be uncovered, leading to functional assignments for many of the newly-discovered genes within the genomic databases (Oliver *et al*., 1998). This area of biology, termed functional genomics (Bouchez and Hofte, 1998), will be a major focus of study over the next decade.

In order to study the metabolome of biological samples, new analytical techniques need to be developed. A typical metabolome study sets out with no prejudices as to which metabolic changes are most significant for any specific bioprocess, and must deconvolute these from potentially hundreds of measurands in a background of thousands of other cellular components. To address this, analytical instrumentation is being developed which is capable of measuring biochemical signatures from whole-tissue or whole-organism samples. This typically results in data sets comprising measurements of many hundreds or thousands of variables. To complicate this task further, the identities of the particular biochemicals to be monitored are frequently unknown at the outset. We here show that

the power of GP to select variables from high dimensional data and to form interpretable predictive models gives it a unique advantage in the analytical interpretation of metabolomic data.

Over the past two decades, tomato as a crop has increased in popularity (Hilhorst *et al*., 1998; Hilhorst and Toorop, 1997). Consequently, much research has been aimed at improving the economic viability of tomato production and post-harvest stability. Environmental stress, such as high salt concentration, is one of the main parameters limiting crop production. The tomato cultivar Edkawy has reduced salt-sensitivity as it grows in the El-Bosaily area of North Egypt, where the soils are saline sands. Edkawy has already been studied in terms of salt tolerance and previous literature provides evidence that this tomato variety may have salt tolerant attributes (Mahmoud *et al*., 1986a; Mahmoud *et al*., 1986b). In this study, Edkawy plants were cultivated using a hydroponic drip irrigation system, allowing precise control of the nutrient conditions within the root zone, including the salinity level. The aim of the study was to identify biochemical constituents (*biomarkers*) within the fruit tissue which are discriminatory for salt-grown tomato plants, and hence to contribute to the understanding of the fundamental biological mechanisms potentially underlying salt tolerance in tomato plants. This in turn may lead to rational improvements in tomato fruit production in conditions of high salinity.

Fourier-transform infrared spectroscopy (FTIR) (Goodacre *et al*., 2000; Goodacre *et al*., 1996b; Griffiths and de Haseth, 1986; Naumann *et al*., 1996; Schrader, 1995; Winson *et al*., 1997) is a physico-chemical analytical technique, which uses the vibrational characteristics of chemical bonds within molecules to obtain a 'fingerprint' spectrum with features defined by the functional chemical groups within the sample. This form of analytical technique is therefore able to give quantitative information about the total biochemical composition of a sample. A thin layer of the biological sample to be analysed is illuminated in the infrared to obtain an *interferogram* (produced by splitting an infrared beam of light, extending the path length of one half by reflecting it off a movable mirror, and recombining the beams optically). Chemical groups within the sample absorb specific frequencies of light within the interferogram due to exchange of energy with their electronic configurations, the precise frequencies absorbed being related to the energies specific to the vibrational modes of each chemical group. The information encoded in the reflected/absorbed light is then recovered by performing a Fourier-transform on the detected signal. The FTIR spectrum so obtained over the range and with the resolution used here comprises 882 variables, each of which indicates the level of absorbance at a particular frequency of infrared light.

A readily accessible interpretation of such extremely high-dimensional spectra, also known as *hyperspectral data* (Aboulseman *et al*., 1994; Goodacre *et al*., 1998; Winson *et al*., 1997), is often very difficult to obtain. Conventional analysis of data of this form falls into two types. The first type, *unsupervised* learning methods, includes principal components analysis (PCA), discriminant function analysis (DFA) and hierarchical cluster analysis (HCA), and seeks to form separable clusters in the data by performing mathematical transforms derived from the variables within the dataset without reference to known classes. The second type, *supervised* learning methods, includes partial least squares (PLS) (Goodacre *et al*., 1994; Martens and Naes, 1989), multivariate rule induction (MRI), inductive logic programming (ILP) (Bratko and Muggleton, 1995) and artificial neural networks (ANNs) (Bishop, 1995; Ripley, 1996), and seeks to refine a model based on the accuracy of its predictions for a set of examples with a known class structure. Although widely used, none of these methods provide models that are readily interpretable in a chemical sense.

Genetic programming (Banzhaf *et al*., 1999; Koza, 1992) is an evolutionary technique which uses the concepts of Darwinian selection to generate and optimise a desired computational function or mathematical expression. GP is a supervised learning method, and consequently requires a set of training examples to form predictive models that can then be applied to the classification of a set of previously unseen test samples. We have previously shown that GP performs at least as well as conventional predictive modelling methods for analysing hyperspectral data (Gilbert *et al*., 1998; Gilbert *et al*., 1997; Goodacre *et al*., 2000; Jones *et al*., 1998b; Taylor *et al*., 1998a; Taylor *et al*., 1998b; Woodward *et al*., 1999).

Here we report the use of PCA, PLS, ANN and GP analyses of FTIR hyperspectral data from tomato fruit samples grown under saline and non-saline conditions. We show that GP, unlike the other methods, is able to provide readily-interpretable models. The interpretability of these models enables the identification of potential biological explanations of the mechanisms underlying the classification.

Experimental Methods

Plant Cultivation

The plants were grown in a hydroponic open-drip irrigation system, using perlite as an inert substrate. The use of a hydroponic system is ideal for studies into plant physiology as it allows complete control over the nutrients applied to the plants. The system was arranged to facilitate saline and control treatments. The capacity of this system was 120 plants with 60 replicates per treatment. All plants were irrigated with complete liquid fertiliser and supplementary sodium chloride (4,000 ppm) was applied to the saline treated plants.

Fruit Tissue Preparation

Twenty fully ripe (at stage 10 on the OECD tomato ripening chart, (OECD Scheme for the application of international standards for fruit and vegetables, www.oecd.org/agr/code/cont-e.htm). Edkawy fruits were harvested. Fruit were selected for uniformity to maximise homogeneity between samples. Ten fruit were taken from salt-grown plants and ten from control plants. The seeds and skin were removed, the outer pericarp was crushed using a press and kept on ice. The extract was homogenised using a Polytron blender at speed 5 for 1 minute. After homogenisation, 1ml aliquots of the sample were placed in Eppendorf tubes, and snap-frozen in liquid N_2. These were stored at -70°C until needed.

FTIR Spectroscopy

Ten replicate 5µl samples of each of the 20 fruit tissue samples were applied to wells drilled on a sandblasted aluminium plate. Samples were arranged to minimise the effects of artifactual trends in the data (*e.g.* edge effects). Prior to analysis, the samples were oven-dried at 50 °C for 30 min. The plate was loaded onto the motorised stage of a reflectance thin-layer chromatography (TLC) accessory attached to a Bruker IFS28 FTIR spectrometer (Bruker Ltd.) equipped with a mercury-cadmium-telluride (MCT) detector cooled using liquid N_2.

The diffuse reflectance absorbance FTIR spectra were collected over a wavenumber range from 4,000 cm^{-1} to 600 cm^{-1} under the control of an IBM-compatible personal computer using OPUS 2.1 software running under the IBM OS/2 Warp operating system. Spectra were acquired at a rate of 20 s^{-1} and at a resolution of approximately 3.85 cm^{-1}. To improve the signal-to-noise ratio, 256 spectra were recorded and averaged for each sample. The complete data set therefore comprised 200 averaged spectra, each containing 882 input variables.

Computational Analysis Methods

Principal Components Analysis

PCA reduces the dimensionality of *n*-dimensional data enabling clustering within the data set to be observed (Causton, 1987; Goodacre *et al*., 1998; Jolliffe, 1986). The method extracts a set of uncorrelated variables as linear combinations of the original variables. The new variables are called *components* and are arranged in order of decreasing variance so those with the highest variance are termed *principal components*. The principal components explain a large proportion of the variance within the data set, maximising between cluster variance and minimising within cluster variance. PCA is an unsupervised technique requiring no prior knowledge of class structure within the data set. In this study the entire data set of 200 spectra was used to derive the PCA model.

Partial Least Squares Modelling

Partial least squares (PLS) modelling (Martens and Naes, 1989) is a widely-used supervised learning technique which reduces the dimensionality of multivariate data by using *a priori* knowledge of which spectra were derived from plants grown under saline or control conditions to produce mathematical models comprising linear combinations of variables. This is referred to as discriminant PLS (Alsberg *et al*., 1998). For this study, we used a PLS modelling system written in house by Dr. Alun Jones (Jones *et al*., 1998a). The initial PLS models were formed on a training data set of 50 spectra. In this study a total of 30 PLS models were derived. The generalising ability of these models was then tested using a previously unseen validation data set (50 spectra). The model with the best generalising ability i.e. the smallest error was then selected and tested on a set of 100 spectra also previously unseen.

Artificial Neural Networks

ANNs (Bishop, 1995; Wasserman, 1989) are supervised methods. The standard fully-interconnected feedforward backpropagation net consists of input and output nodes, which employ a squashing function to avoid numerical over flows, linked together by a hidden layer. The models are formed on a training data set, then validated and tested on previously unseen data sets. In training the neural network the numerical inputs are transformed into 'desired' outputs. The transformation of the inputs depends on the connection weight and bias of the nodes. The neural network is trained by presenting the networks with 'known' inputs and outputs (Goodacre *et al*., 1996a). The ANN is said to have generalised when the correct outputs are determined from a previously unseen data set (the validation set). The model can then be tested on a third data set. The ANN model was trained on a data set of 100 samples then validated and tested with two other data sets each containing data from 50 sample spectra. The ANN was unable to form a satisfactory model from the 50 training spectra used for the PLS model, so a larger training set of 100 spectra was used. ANN analysis was carried out using NeuFrame version 3.0.0.0 (Neural Computer Sciences, Luworth Business Centre, Totton, Southampton, UK).

Genetic Programming

The GP implementation used in this study was capable of performing non-linear multivariate regressions with automatic variable selection. It was written in C, and was run on IBM-compatible PCs under Windows NT 4.0, and on DEC Alpha-based PCs under Linux 5.1.

The GP used the arithmetic operator functions *add*, *subtract*, *multiply*, and *protected divide* and a Boolean '*if greater than or equal to'* function. The *if* function returned a value of 1.0 if the first argument was greater than or equal to the second argument, 0.0 otherwise. To avoid possible numeric overflows, a protected *divide* function was used which returned a numerical value of 10^{15} for divisions with a denominator $\leq 10^{-15}$. Additional protection from floating-point errors was enforced by clipping the return value of each node into the range $\pm 10^{15}$.

Terminals comprised either floating-point constants (initialised randomly in the range –10.0 to 10.0) or input variables (corresponding to one of the 882 absorbance measurements which comprised each spectrum).

The GP generated initial individuals with random function trees of depth 2 to 6, and assessed their fitness using a scoring function that compared e_i (the model's estimate of the output for example i) with o_i (the experimentally-observed value) by calculating the root-mean-square error of prediction (RMSEP) for n training examples:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (o_i - e_i)^2}{n}}$$

The fittest individuals were those that gave the lowest RMSEPs for the training set examples.

Since the dataset contains two classes (fruit from plants grown under either saline or control conditions), class membership was defined in the training examples by assigning a target output value of 1.0 to members of the saline-grown class, and 0.0 to members of the control class. A correct classification was assigned when the output value of the GP-derived rule was within 0.01 of the target output for any given spectrum.

GP-generated rules, if allowed to evolve unchecked, tend to become longer and more arithmetically complex as the evolution proceeds, a phenomenon known as *bloat* (Angeline, 1998; Langdon, 1998; Langdon and Poli, 1998). This increase in complexity reduces the ready interpretability of the expressions generated and is likely to lead to overfitting (Seaholtz and Kowalski, 1993). To combat this, a penalty of $0.01 \times N$, where N represents the number of nodes in the function tree, was added to the fitness calculation. This ensured that, for a given RMSEP, a shorter tree would be chosen over a longer one. In addition, a maximum tree depth of 10, and a maximum node count of 100 was enforced during the evolution.

The size constraints on the GP rules meant that even the longest rules could use only a small subset of the available input variables comprising the dataset. The GP was therefore compelled to perform an automatic variable selection, resulting in predictive models with significantly lower dimensionality (*i.e.* using far fewer variables) than the dataset as a whole. The automatic variable-selection ability of the GP approach is one of the main benefits of using this as a predictive modelling method (Gilbert *et al.*, 1998). Since the GP-derived models are much more readily interpretable than say those produced by ANNs, analysis of the selected variables can lead to a rationalisation of the mechanism underlying the model.

The GP used five demes (sub-populations) each of 7,500 individuals. Each run was allowed to continue for 5,000 generations, but the final model was typically produced after about 2,200 generations. The GP was also allowed to terminate when the RMSEP fell below 0.01, which did not occur in practise. Every 10 generations, the best 5% of the individuals in each of four satellite demes replaced the worst 5% in a central deme. The best 5% from this deme then replaced the worst 5% in the satellite demes, resulting in every deme containing the best 5% of the population as a whole. This divergent evolution and migration strategy has been shown to be more effective at solving high-dimensional problems than a conventional single-population (Whitlock and Barton, 1997).

During each generation, 1,500 new individuals were created by single-point mutation, and 3,000 by single-point crossover. Parental selection was proportional to fitness, and new individuals were retained in the deme if their fitness was higher than that of the current-worst individual, maintaining a population size of 7,500. In this study 100 spectra (50 from saline-grown and 50 from control fruit samples) were used by the GP as a training set to derive the models, and the remaining 100 spectra were used to test their predictive ability. No information from the test set was used to guide the evolution of the GP.

Correlation Analysis

An analysis was performed to investigate the correlation between the GP-selected input variables and the known class structure of the data. *Product moment correlation* (PMC) is a method which uses linear transformations to quantify which variables (x) are most strongly related to the output data (y) being modelled.

The PMC (R) value ranges from -1 to +1, indicating a perfect negative to a perfect positive correlation, with a value of 0 indicating that the variable is uncorrelated with the class structure. R takes the sign of C_{xy}. For n examples, R can be calculated as follows:

$$R = \frac{C_{xy}}{\sqrt{C_{xx} \cdot C_{yy}}}$$

where

$$C_{xy} = \left(\sum_{i=1}^{n} x_i \cdot y_i \right) - n \cdot \left(\bar{x} \cdot \bar{y} \right)$$

$$C_{xx} = \left(\sum_{i=1}^{n} x_i^{\,2} \right) - n \cdot \left(\bar{x} \right)^2$$

$$C_{yy} = \left(\sum_{i=1}^{n} y_i^{\,2} \right) - n \cdot \left(\bar{y} \right)^2$$

The PMC between each of the 882 input variables in the training set and the known class structure was calculated to investigate whether the GP was selecting only well-correlated variables.

Quantum Mechanics and Infrared Spectral Analysis

The semi-empirical quantum mechanics program PM1, part of the HyperChem 5.1 molecular modelling package (HyperCube, Inc.) was used to calculate infrared vibrational spectra and molecular vibrational modes for potential metabolites identified during the analysis of the data. The infrared spectral analysis package IR Mentor Pro 2.0

(Bio-Rad Laboratories) was used to suggest candidate chemical groups responsible for the particular spectral features selected as discriminatory by the GP models.

Results and Discussion

The hyperspectral data were analysed using four different chemometric techniques (Figure 1). Figure 1A shows a two-dimensional PCA plot using principal components 1 and 2 since these account for the greatest proportion of the variance within the data set, 85.75 % and 12.3% respectively. PCA was unable to separate the control and the saline treated fruit samples into 2 distinct clusters although a slight trend can be observed in the plot. The PLS model (Figure 1B), using 11 PLS factors (these resulted in the best generalising ability) correctly classified 88% of the samples comprising the validation and test data sets. A correct prediction for the discriminant PLS was taken to be <0.5 for control sample and ≥0.5 for saline-treated samples. The second supervised technique to be applied was ANNs (Figure 1C). The whole data set was divided into 3 sub-sets, training, validation and test. The optimum model was derived after 10,500 epochs resulting in prediction accuracy of 100% for the training set and 84% for the validation and test sets. A correct prediction of class was determined using the same criteria as with PLS. The GP derived expressions were capable of correctly classifying the examples in both the training and test data sets with an average accuracy of 88.9%. Figure 1D shows the results obtained by the best GP model (rule 28) with a predictive accuracy of 95% for the test set data. Due to the use of the *if* function which enforces a strict separation it is difficult to determine the degree of misclassification for the incorrectly classified samples.

PLS, ANN and GP derived models all classified the control and saline-treated tomato samples with similar prediction accuracy. However, the models derived by PLS and ANN, in common with the other widely used statistical modelling methods, are not readily interpretable, whereas GP produces mathematical rules that enable the easy identification of wavenumbers selected to perform the classification.

The GP was run 30 times. The wavenumbers used by each of the GP rules and their prediction accuracies are shown in Table 1. The wavenumbers indicated in bold fall within the critical region from 2270-1960 wavenumbers identified in Figure 2. At least one wavenumber (typically 3-4 wavenumbers) from this critical region is used in every rule highlighting its importance. These 30 independent GP rules used a total of 112 out of the 882 input variables. The entire data set contained wavenumbers with PMC values ranging from 0.000332 to 0.4401. The GP-selected waveumbers had PMC values ranging from 0.000642 to 0.4296, indicating that the GP selected wavenumbers with both high and low correlations with the known class structure. The wavenumber used most frequently (found in five models) had a PMC value of 0.2876, so is reasonably well-correlated with the class structure. Although the most-correlated wavenumber in the data set was not used, the GP selected an adjacent wavenumber on two occasions. As has been observed before, the GP uses the low-correlated wavenumbers as internal constants in the data set, enabling it to perform normalisation and baseline shift correction (Gilbert *et al*., 1998).

The GP-derived models typically used 5 variables, with the smallest rule using 4 and the largest using 13. No single variable could be used to classify the spectra with an accuracy approaching the 90% value of the GP-derived rules. Despite the reasonably high PMC values for most of the variables within the data set PLS was unable to separate the two classes completely based on the spectral data (Figure 1B). This indicates that the data set does not contain enough information to allow a very high degree of separation of the two classes to be made without using non-linear combinations of variables. The *if* operator was used in every GP rule, a clearly essential function for a classification problem which is not readily available to standard feed-forward, back-propagation neural networks or the conventional statistical modelling methods.

The GP models were all different. The prediction accuracy for classification of the test data set of 100 spectra ranged from 83% to 95% correct. Runs 9, 13, 18 and 27 produced remarkably similar models, with the same logical structure and using very similar wavenumbers. The predictive accuracy is shown for the test set:

Run 9: (91% correct).

IF (A_{2164}-A_{2245}) ≥ (A_{2060}-A_{2098}) THEN [Saline] ELSE [Control]

Run 13: (90% correct).

IF (A_{2171}-A_{2245}) ≥ (A_{1963}-A_{2106}) THEN [Saline] ELSE [Control]

Run 18: (92% correct).

IF (A_{2168}-A_{2257}) ≥ (A_{2029}-A_{2114}) THEN [Saline] ELSE [Control]

Run 27: (92% correct).

IF (A_{2179}-A_{2230}) ≥ (A_{2025}-A_{2118}) THEN [Saline] ELSE [Control]

In the above rules, A_n represents the measured absorbance at wavenumber ***n***. These rules may be indicative of the nature of the globally-optimal rule derivable from this data set. The best performing rule, with a 95% predictive accuracy for the test set data, used a similar logical construct and selected similar variables. However it included an additional term associated with a spectral feature at 2480 wavenumbers which enabled a better class prediction for some of the saline-grown samples which were incorrectly classified by the simpler rules:

Run 28: (95% correct)

$$\mathbf{IF}\left[\left(\frac{A_{3476}}{A_{3499}}\right)\times \mathbf{IF}\left(A_{2480}\geq A_{883}\right)\right]\geq\left[\left(A_{2268}-A_{2133}\right)\times\left(A_{1558}+A_{3638}\right)+A_{2017}-A_{2110}\right]$$

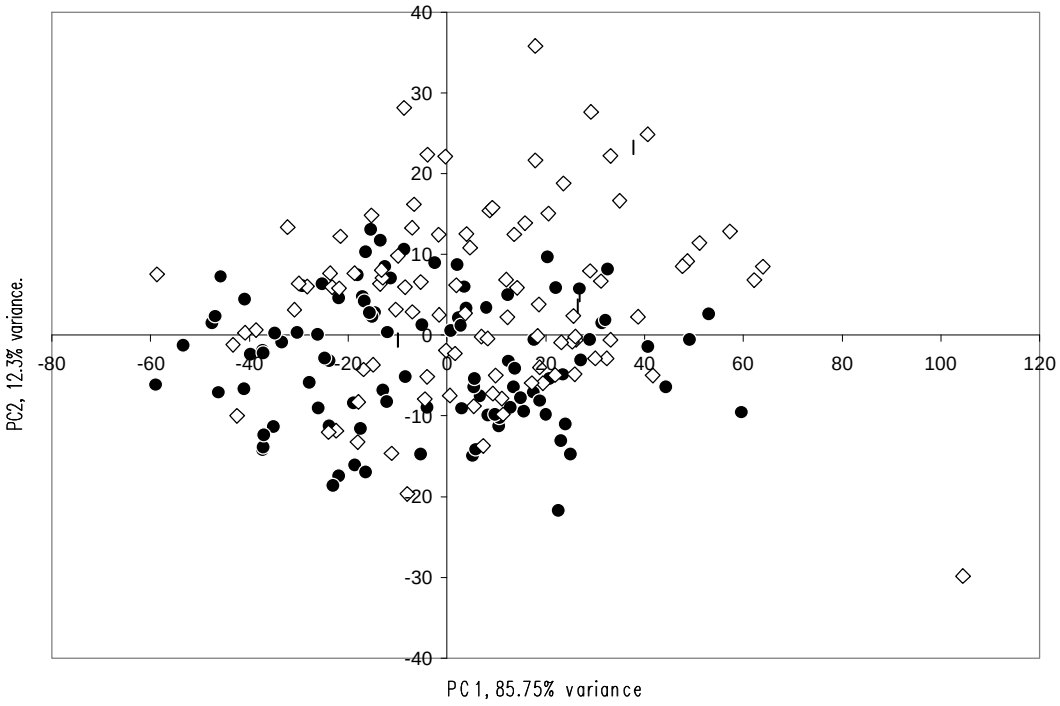
THEN [Saline] ELSE [Control]

An analysis of which input variables (in terms of absorbances at particular wavenumbers) were selected showed that there were a few regions of the spectra that were consistently being used to form the different models (Figure 2). In particular, the spectral region covering 2270 to 1960 cm^{-1} was used by most of the rules, and all of the best-performing models were based on a few small but distinct features within this critical region. The absolute differences between saline and non-saline grown samples in this region are very small (*ca.* 0.005 absorbance units), and so would not have been selected in a direct visual analysis such as by using a difference spectrum.

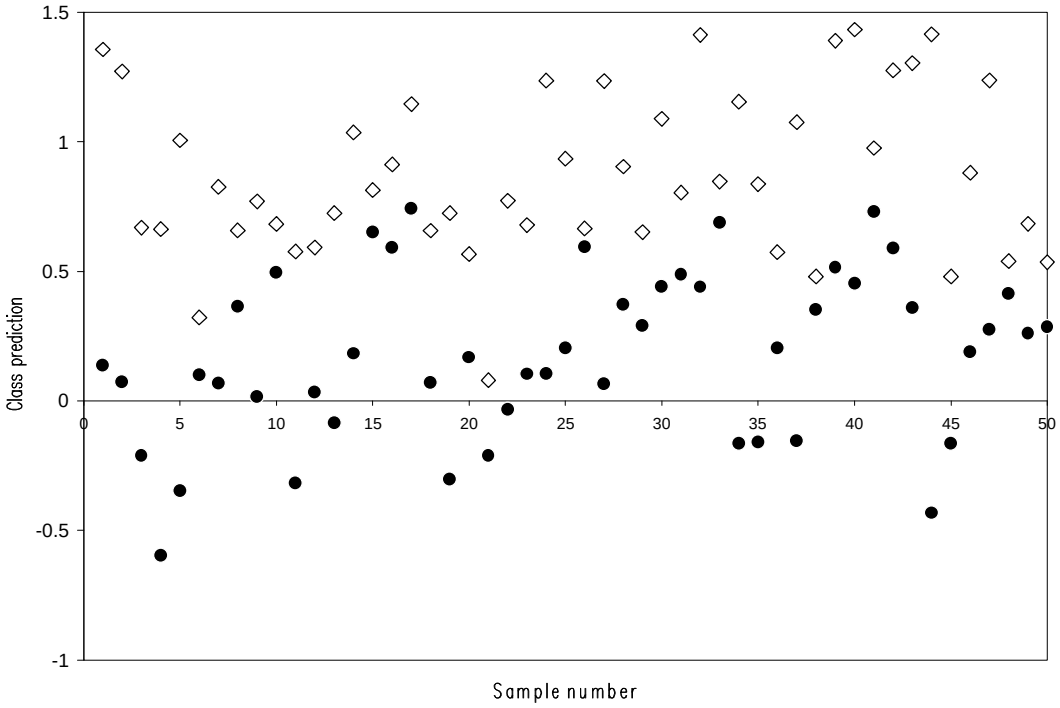
Quantum mechanical calculations and spectral libraries showed that the only biochemically-reasonable functional groups that absorb strongly in this critical part of the IR spectrum are acetylenes ($R-C\equiv C-R'$) and cyanides or nitriles ($R-C\equiv N$), with the absorption due to a periodic stretching motion of the triple-bond. Acetylenes have a second characteristic vibration at approximately 3300 wavenumbers, a region unused by any of the GP-derived rules. If an acetylene group were responsible for the characteristic spectral features, it would be expected that the GP models would also use this region. Therefore, the most likely candidate chemical moiety being identified by the predictive models as characteristic for tomatoes grown under saline conditions is a cyanide or nitrile group.

Nitrile and cyanide groups occur in many compounds found in plants. Cyanide groups often result from the detoxification of hydrogen cyanide (HCN) which is toxic to biological systems if allowed to accumulate. HCN is a by-product of ethylene production, the biosynthesis of which is known to increase in plants in response to stress and during the ripening of climacteric fruit (Hulme, 1970), such as tomato. It has previously been reported that tomato plants grown under saline conditions show enhanced ethylene production (Mizrahi, 1982), with a consequent increase in hydrogen cyanide production. The hydrogen cyanide is detoxified *via* a series of chemical reactions, increasing the concentration of cyanide-containing compounds within the salt-stressed tomato fruit. It can be hypothesised that the GP selected variables correspond to small spectral differences due to a change in the concentration of cyanide-containing compounds. Although no definite conclusions can yet be drawn as to the specific chemical identity of these compound(s) the GP has enabled the identification of potential biomarkers for saline treated fruit. This work highlights the use of GP as an exploratory tool able to identify critical regions within hyperspectral data, unlike the other numerical methods. Using the GP models in conjunction with knowledge of the biological system, a preliminary identification of important biochemical differences between the samples can be made, providing direction for future biological investigations.

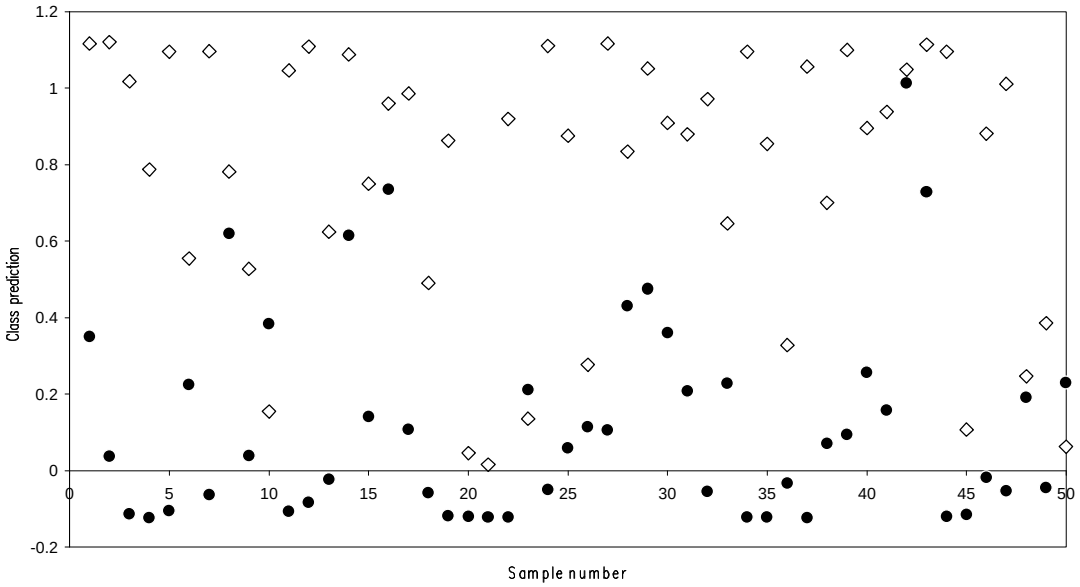
A


40
30
20
10
0
-10
-20
-30
-40
-80
-60
-40
-20
0
20
40
60
80
100
120
PC2, 12.3% variance.
PC 1, 85.75% variance

B


1.5
1
0.5
0
-0.5
-1
0
5
10
15
20
25
30
35
40
45
50
Class prediction
Sample number

C



D

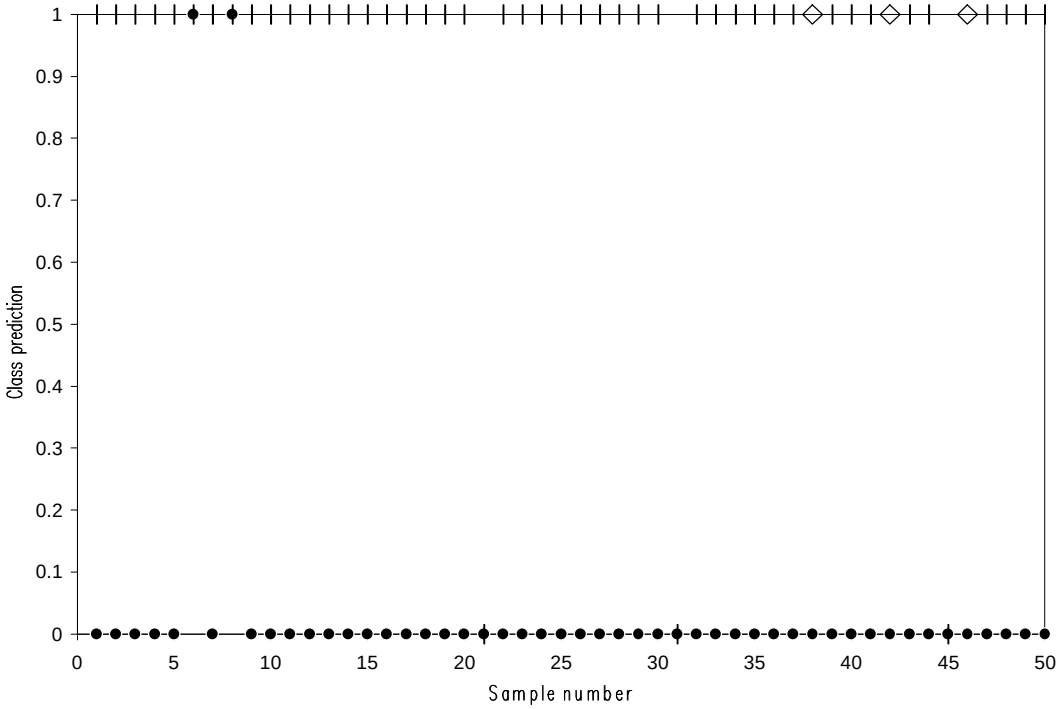


Figure 1

The prediction outputs from the four chemometric modelling methods are shown. Principal component analysis (PCA), presented in plot A, is the only unsupervised method used and was unable to separate the control and saline-treated samples satisfactorily. Partial least squares (PLS) regression, artificial neural network (ANN) and genetic programming (GP) predictions are shown in plots B to D respectively. With the exception of PCA, which used all 200 samples to form its model, the data was split into training and test/validation sets. The results shown are for the 100 test/validation samples. It is apparent that the GP's use of the discontinuous *if* function dramatically improves the class separation.

Rule	Wavenumbers Used	Test Set % Correct
1	**2245, 2172, 2037, 2029**	86
2	2824, **2257, 2122, 2087, 1963**, 1678	90
3	**2068, 1967**, 841, 806	89
4	3653, **2233, 2164, 1967**	90
5	**2249, 2149, 2122, 2017**, 1022, 999	90
6	**2245, 2179, 2152, 2010**, 1227	93
7	3734, **2257, 2145, 2087, 2041**	91
8	**2241, 2152, 1975**	89
9	**2245, 2164, 2098, 2060**	91
10	3695, 3217, 3055, 2754, 2446, 2411,2338, **2071**, 1782, 849, 633	94
11	3113, **2249, 2214, 2152, 2098, 2021,** 1199	92
12	3935, 3487, 2986, 2704, **2118, 2002,** 1936, 1813, 1339, 1265, 1180, 1007, 829	87
13	**2245, 2172, 2106, 1963**	90
14	2936, 2774, **2245, 2187, 2083, 2029**	83
15	3668, 3236, 2685, **2230, 2129**, 1173, 876, 640	89
16	2750, **2249, 2125,** 1501	90
17	2287, **2264, 2152, 2114, 2098, 2010**	92
18	**2257, 2168, 2114, 2029**	92
19	**2218, 2187, 2098, 1967**	92
20	**2268, 2226, 2172, 2114, 2025**	91
21	2272, **2087, 2029, 2025,** 1304	91
22	2299, **2083, 2048,** 1890	93
23	2291, **2160, 2137, 2133,** 1863	91
24	**2233, 2141, 2110, 2025,** 1836, 1659, 1512	88
25	**2257, 2149, 2006**	91
26	**2241, 2133, 2118, 2021**	93
27	**2230, 2179, 2118, 2025**	92
28	3678, 3499, 3476, 2480, **2268, 2133, 2110, 2017,** 1558, 883	95
29	2916, 2708, **2095, 2002, 1987**	88
30	**2264, 2145, 2118, 2025**	92

Table 1

The wavenumbers used by the rules derived from 30 GP runs are presented. All the rules use at least one wavenumber which falls in the critical region identified in Figure 2, ranging from 2270 to 1960 wavenumbers, these are indicated in bold. The prediction accuracies are shown for the test set data indicating that irrespective of rule length classification is greater then 83%.

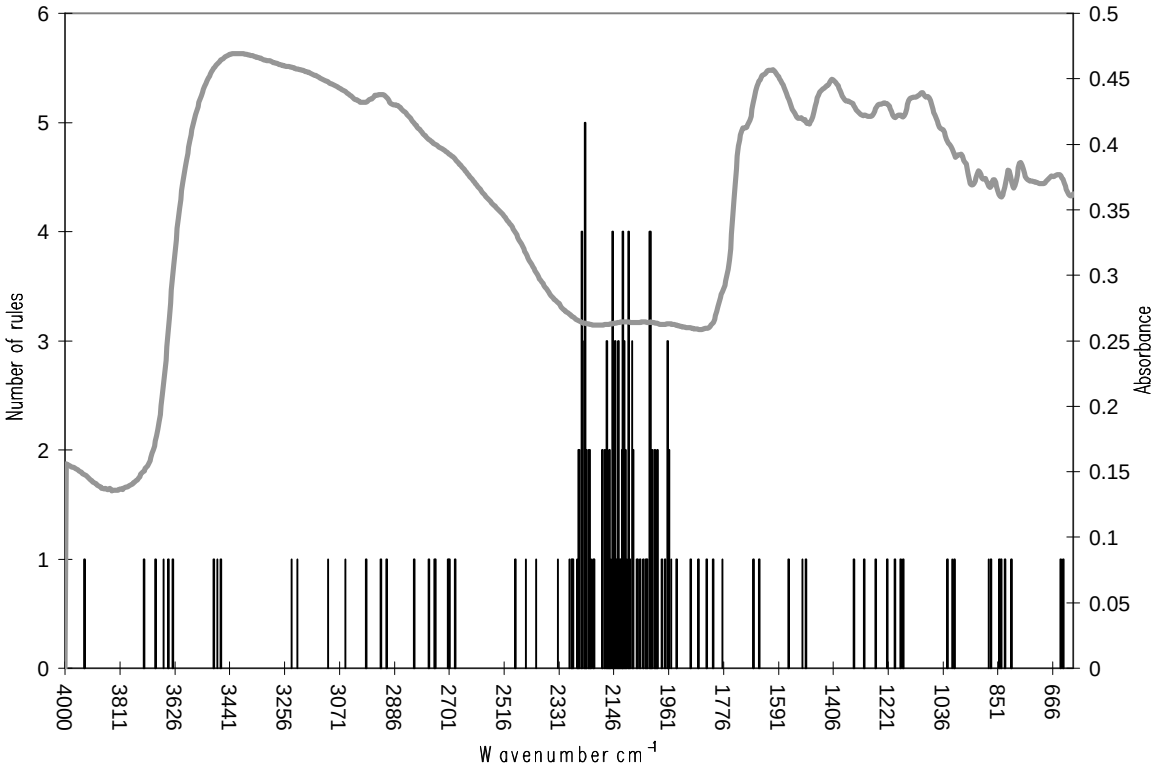


Figure 2

The wavenumbers selected by the 30 GP rules are shown in reference to a spectrum averaged from the whole data set. The vertical lines represent the number of GP-derived rules that use particular wavenumbers to form a predictive model. The region from 2270 to 1960 wavenumbers is clearly important for producing good predictive models.

Conclusions

This study has shown that GP, in combination with FTIR, is a powerful new tool for the analysis of whole-tissue biological samples at the metabolome level. The combined technique is sensitive enough to detect changes in the levels of a single metabolite against the background of the entire cellular components, and can provide chemical information which can lead to the identification of the biochemicals which may be involved in metabolic processes under investigation.

The main benefit of the GP approach is that, unlike more conventional numerical analyses, it provides readily interpretable models which enable mechanistic explanations of the underlying biological systems. Additionally, GPs are able to analyse effectively extremely high-dimensional data that are generally not amenable to simpler analytical algorithms. The GP method therefore has the promise of becoming an extremely sensitive and discriminatory analytical tool that may be of crucial importance in the emerging field of functional genomics, and so help to advance the understanding of metabolic processes as yet unexplored.

Acknowledgements

RJG, JJR and DBK thank the UK EPSRC for financial support. MKW, JJR and DBK thank the UK BBSRC for financial support. HEJ, ARS and MAH acknowledge the financial support of the European Union INCO-DC programme. We thank Pat Causton and Dave Summers for their help in plant cultivation. We also thank Dr. Gary Salter for his assistance and encouragement.

References

Aboulseman, G.P., Gifford, E. and Hunt, B.R. (1994) *Optical Engineering*, **33**, 2562-2571.

Alsberg, B.K., Kell, D.B. and Goodacre, R. (1998) *Analytical chemistry*, **70**, 4126-4133.

Angeline, P.J. (1998) In R. J. Koza, W. Banzhaf, K. Chellapilla, K. Deb, M. Dorigo, D. B. Fogel, M. H. Garzon, D. E. Goldberg, H. Iba and R. L. Riolo (eds.), *Subtree crossover causes bloat.* Madison, Wisconsin, USA: Morgan Kaufmann, pp. 745-752.

Banzhaf, W., Nordin, P., Keller, R. and Francone, F. (1999) *Genetic programming - An introduction.* Academic Press, San Franscisco.

Bishop, C.M. (1995) *Neural networks for pattern recognition.* Clarendon Press, Oxford.

Bork, P., Dandekar, T. and Diaz-Lazcoz, Y. (1998) *Journal of Molecular Biology*, **283**, 707-725.

Bouchez, D. and Hofte, H. (1998) *Plant Physiology*, **118**, 725-732.

Bratko, I. and Muggleton, S.H. (1995) *Comm. ACM*, **38**, 65-70.

Causton, D.R. (1987) *A biologist's advanced mathematics.* Allen and Unwin, London.

Cole, S.T., Brosch, R. and Parkhill, J. (1998) *Nature*, **393**, 537-544.

Gilbert, R.J., Goodacre, R. and Shann, B. (1998) In R. J. Koza, W. Banzhaf, K. Chellapilla, K. Deb, M. Dorigo, D. B. Fogel, M. H. Garzon, D. E. Goldberg, H. Iba and R. L. Riolo (eds.), *Genetic programming-based variable selection for high-dimensional data.* Madison, Wisconsin, USA: Morgan Kaufmann.

Gilbert, R.J., Goodacre, R., Woodward, A.M. and Kell, D.B. (1997) *Analytical Chemistry*, **69**, 4381-4389.

Goodacre, R., Neal, M.J. and Kell, D.B. (1994) *Analytical Chemistry*, **66**, 1070-1085.

Goodacre, R., Neal, M.J. and Kell, D.B. (1996a) *Zentralblatt Fur Bakteriologie-International Journal of Medical Microbiology Virology Parasitology and Infectious Diseases*, **284**, 516-539.

Goodacre, R., Shann, B., Gilbert, R.J., Timmins, E.M., McGovern, A.C., Alsberg, B.K., Kell, D.B. and Logan, N.A. (2000) *Analytical Chemistry*, **72**, 119-127.

Goodacre, R., Timmins, E.M., Burton, R., Kaderbhai, N., Woodward, A.M., Kell, D.B. and Rooney, P.J. (1998) *Microbiology-UK*, **144**, 1157-1170.

Goodacre, R., Timmins, E.M., Rooney, P.J., Rowland, J.J. and Kell, D.B. (1996b) *FEMS Microbiology Letters*, **140**, 233-239.

Griffiths, P.R. and de Haseth, J.A. (1986) *Fourier transorm infrared spectrometry.* John Wiley, New York.

Hilhorst, H.W.M., Groot, S.P.C. and Bino, R.J. (1998) *Acta Botanica Neerlandica*, **47**, 169-183.

Hilhorst, H.W.M. and Toorop, P.E. (1997) *Advances in Agronomy*, **61**, 111-165.

Hinton, J.C.D. (1997) *Molecular Microbiology*, **26**, 417-422.

Hulme, A.C. (1970) *The biochemistry of fruits and their products.* Academic Press, London.

Jolliffe, I.T. (1986) *Principal component analysis.* Spring-Verlag, New York.
Jones, A., Shaw, A.D. and Salter, G.J. (1998a) In R. J. Hamilton (ed.) *The exploitation of chemometric methods in the analysis of spectroscopic data: application to olive oils.* Chapman and Hall, London, pp. 317-376.
Jones, A., Young, D., Taylor, J., Kell, D.B. and Rowland, J.J. (1998b) *Biotechnology and Bioengineering*, **59**, 131-143.
Koza, J.R. (1992) *Genetic programming: On the programming of computers by means of natural selection.* MIT Press, Cambridge, MA.
Langdon, W.B. (1998) *Genetic programming and data structures.* Kluwer Academic Publishers, Holland.
Langdon, W.B. and Poli, R. (1998) In W. Banzhaf, R. Poli, M. Schoenauer and T. C. Fogarty (eds.), *Fitness causes bloat: mutation.* Springer, Paris, France, pp. 37-48.
Mahmoud, M.H., El-Beltagy, A.S., Helal, R.M. and Maksoud, M.A. (1986a) *Acta Horticulturae*, **190**, 559-565.
Mahmoud, M.H., Jones, R.A. and El-Beltagy, A.S. (1986b) *Acta Horticulturae*, **190**, 533-543.
Martens, H. and Naes, T. (1989) *Mutlivariate calibration.* John Wiley, Chichester.
Mizrahi, Y. (1982) *Plant Physiology*, **69**, 966-970.
Naumann, D., Schultz, C.P. and Helm, D. (1996) In H. H. Mantsch and D. Chapman (eds.), *What can infrared spectroscopy tell us about the structure and composition of intact bacterial cells?* John Wiley, New York, pp. 279-310.
Oliver, S.G., Winson, M.K., Kell, D.B. and Baganaz, F. (1998) *Trends in Biotechnology*, **16**, 373-378.
Ripley, B.D. (1996) *Pattern recognition and neural networks.* Cambridge University Press, Cambridge.
Schrader, B. (1995) *Infrared and Raman spectroscopy: methods and applications.* Verlag Chemie, Weinheim.
Seaholtz, M.B. and Kowalski, B. (1993) *Analytica Chimica Acta*, **277**, 165-177.
Taylor, J., Rowland, J.J. and Goodacre, R. (1998a) In J. R. Koza, W. Banzhaf, K. Cellapilla, K. Deb, M. Dorigo, D. B. Fogel, M. H. Garzon, D. E. Goldberg, H. Iba and R. L. Riolo (eds.), *Genetic programming in the interpretation of Fourier transform infrared spectra: quantification of metabolites of pharmaceutical importance.* Morgan Kaufman: San Francisco., pp. 377-380.
Taylor, J., Winson, M.K., Goodacre, R., Rowland, J.J. and Kell, D.B. (1998b) *FEMS Microbiology Letters*, **160**, 237-246.
Wasserman, P.D. (1989) *Neural computing: Theory and practice.* Van Nostrand Reinhold, New York.
Whitlock, M.C. and Barton, N.H. (1997) *Genetics*, **146**, 427-441.
Winson, M.K., Goodacre, R., Timmins, E.M., Jones, A., Alsberg, B.K., Woodward, A.M., Rowland, J.J. and Kell, D.B. (1997) *Analytica Chimica Acta*, **348**, 273-282.
Woodward, A.M., Gilbert, R.J. and Kell, D.B. (1999) *Bioelectrochemistry and Biogenetics*, **48**, 389-396.