Monitoring of Complex Industrial Bioprocesses for Metabolite Concentrations Using Modern Spectroscopies and Machine Learning: Application to Gibberellic Acid Production

Aoife C. McGovern,[1] David Broadhurst,[1] Janet Taylor,[1,2] Naheed Kaderbhai,[1] Michael K. Winson,[1] David A. Small,[3*] Jem J. Rowland,[2] Douglas B. Kell,[1] Royston Goodacre[1]

[1]Institute of Biological Sciences, Cledwyn Building, University of Wales, Aberystwyth, Ceredigion SY23 3DD, Wales, UK; telephone: +44-(0)-1970-621-947; fax: +44-(0)-1970-622-354; e-mail: rrg@aber.ac.uk
[2]Department of Computer Science, University of Wales, Aberystwyth, Ceredigion SY23 3DD, Wales, UK
[3]Zeneca Bio Products, Billingham, Clevland, UK

Received 10 September 1999; Accepted 3 December 2001

DOI: 10.1002/bit.10226

Abstract: Two rapid vibrational spectroscopic approaches (diffuse reflectance–absorbance Fourier transform infrared [FT-IR] and dispersive Raman spectroscopy), and one mass spectrometric method based on *in vacuo* Curie-point pyrolysis (PyMS), were investigated in this study. A diverse range of *unprocessed,* industrial fed-batch fermentation broths containing the fungus *Gibberella fujikuroi* producing the natural product gibberellic acid, were analyzed directly without *a priori* chromatographic separation. Partial least squares regression (PLSR) and artificial neural networks (ANNs) were applied to all of the information-rich spectra obtained by each of the methods to obtain quantitative information on the gibberellic acid titer. These estimates were of good precision, and the typical root-mean-square error for predictions of concentrations in an independent test set was <10% over a very wide titer range from 0 to 4925 ppm. However, although PLSR and ANNs are very powerful techniques they are often described as ''black box'' methods because the information they use to construct the calibration model is largely inaccessible. Therefore, a variety of novel evolutionary computation-based methods, including genetic algorithms and genetic programming, were used to produce models that allowed the determination of those input variables that contributed most to the models formed, and to observe that these models were predominantly based on the concentration of gibberellic acid itself. This is the first time that these three modern analytical spectroscopies, in combination with advanced chemometric data analysis, have been compared for their ability to analyze a real commercial bioprocess. The results demonstrate unequivocally that all methods provide very rapid and accurate estimates of the progress of industrial fermentations, and indicate that, of the three methods studied, Raman spectroscopy is the ideal bioprocess monitoring method because it can be adapted for on-line analysis. *Biotechnol Bioeng* **78**: 527–538, 2002.
Keywords: evolutionary computing; Fourier transform infrared spectroscopy; dispersive Raman spectroscopy; pyrolysis mass spectrometry

INTRODUCTION

Many process industries are beginning to replace traditional chemical processes with bioprocesses because of their chemical specificity and desirable reaction kinetics. The ability to control a bioprocess is paramount for product yield optimization, and therefore it is imperative that the concentration of the fermentation product (the determinand) is assessed accurately (Kell and Sonnleitner, 1995; Pons, 1991). The development of such monitoring methods (Scheper and Lammers, 1994) is driven by economic and ecological needs, and by the requirements for better process documentation. Whereas many spectroscopic studies have concentrated on measurements of biomass (Harris and Kell, 1985; Validyanathan et al., 1999) and nutrient supply (Brimmer and Hall, 1993), comparatively few have attempted to obtain quantitative information on the product, unless they are non-complex *chemical* processes (Adar et al., 1997; Roberts et al., 1991).

Ideal methods for the rapid, precise, accurate analysis of the biochemical composition of fermentor broths, and the characterization of the organisms that they contain, would permit the simultaneous estimation of multiple

Correspondence to: R. Goodacre
**Current affiliation*: Stiefel Laboratories (UK), Ltd., Maidenhead, Berkshire, UK
Contract grant sponsors: Zeneca Pharmaceuticals; Science in Finance, Ltd.; The Wellcome Trust; UK BBSRC and EPSRC
Contract grant number: 042615/Z/94/Z

determinands; would have minimum sample preparation; would analyze samples directly (i.e., would not require reagents); and would be rapid, automated, accurate, and (at least relatively) inexpensive. With recent developments in analytical instrumentation, these requirements are being fulfilled by spectroscopic methods, and the most common are pyrolysis mass spectrometry (PyMS) (Goodacre et al., 1994b; 1995; McGovem et al., 1999), Fourier transform infrared spectroscopy (FT-IR) (Mattu et al., 1997; McGovern et al., 1999; Timmins, 1998; Winson et al., 1997) and dispersive Raman microscopy (Goodacre et al., 1998; Shaw et al., 1999a). PyMS, FT-IR, and Raman spectroscopies are physicochemical methods that measure predominantly the bond strengths of molecules (PyMS) and the vibrations of bonds within functional groups (FT-IR and Raman) (Ferraro and Nakamoto, 1994; Griffths and de Haseth, 1986; Meuzelaar et al., 1982; Schrader, 1995). They therefore give quantitative information about the total biochemical composition of a sample. However, the extraction of this information typically involves the use of advanced chemometric techniques.

Chemometrics is the application of statistical and other mathematical techniques to analytical chemical data (Lavine, 1998; Massart et al., 1988). These methods, in which we take an input of high dimensionality and allow the extraction of information relevant to the biological question of interest, can be subdivided into two general classes. The first involves those methods that cluster the data with no prior knowledge of the samples analyzed, the so-called unsupervised analyzes. In the second class, known as supervised analyzes or multivariate calibration, one seeks to relate the multivariate spectral inputs to the membership of a predetermined class structure. In the present case, and others of widespread interest, the target classes involve the concentrations of target determinands—that is, generating a *quantitative analysis*. These types of methods therefore exploit multidimensional curve fitting or regression analysis, most commonly (for linear systems in which the number of variables is in excess of the number of samples) using variants of the partial least squares regression (PLSR) algorithm (Martens and Næs, 1989). A related approach (Zupan and Gasteiger, 1993), which has been used to model and control bioprocesses, is the use of (artificial) neural networks (ANNs) (Montague and Morris, 1994).

Gibberellins are important biotechnological products used in agriculture and horticulture for the regulation of plant growth (Brückner and Blechschmidt, 1991). Gibberellic acid 3 is produced industrially by Zeneca Life Science Molecules in fed-batch fermentations of the fungus *Gibberella fujikuroi*. The current protocol for monitoring gibberellin levels involves removal of samples during the course of the fermentation and analysis of extracts by high-performance liquid chromatography (HPLC) analysis. On- or at-line monitoring in real-time of gibberellin levels should allow more accurate control of this bioprocess.

The aim of the present study was to assess the use of FT-IR, Raman, and PyMS, in combination with chemometrics, for at-line monitoring of the gibberellin titer. However, it is known (Martens and Næs, 1989; Zupan and Gasteiger, 1993) that, although PLSR and ANNs are excellent methods for quantitative analysis, they do not lend themselves to easy interpretation; that is, it is not obvious how the mathematical models exploit information specifically in terms of the values of the different inputs (i.e., absorbances or shifts in electromagnetic radiation for FT-IR and Raman analyzes, or intensities of specific mass ions in PyMS spectra). For this it is necessary to develop systems that produce readily comprehensible mathematical models. Toward this end, a number of methods involving evolutionary computation (Bäck et al., 1997; Goldberg, 1989; Holland, 1992; Koza, 1992; Mitchell, 1992), including genetic algorithms (GAs) (Broadhurst et al., 1997) and genetic programming (GP) (Gilbert et al., 1997; Taylor et al., 1998a), were employed to decrease the number of input variables from these otherwise high-dimensional spectra used in forming the models.

MATERIALS AND METHODS

Bioprocess

Samples were provided by Zeneca Life Science Molecules. Gibberellic acid 3 (GA3) was produced by *Gibberella fujikuroi* in a complex undefined medium. Whole broth samples were taken asceptically from the fermentation vessels and methanol:water extracts were analyzed by HPLC (the typical error in GA3 measurement was 2% to 5%). The remaining unprocessed sample was stored at −20°C prior to analysis at UWA by the three spectrometric methods. Samples were collected over a period of 3 months and assessed using four different HPLC rigs, as detailed in Table I.

Diffuse Reflectance–Absorbance FT-IR

FT-IR analysis was performed using a Bruker IFS28 infrared spectrometer equipped with a diffuse-reflectance TLC attachment (Bruker, Ltd., Coventry, UK) and a liquid N_2-cooled MCT (mercury–cadmium–telluride) detector, as previously described (Goodacre et al., l996; Timmins et al., 1998; Winson et al., 1997). Bioprocess samples (5 μL; three replicates) were placed in the wells of a 10 cm × 10 cm aluminium plate, containing 400 wells. After the samples were oven-dried at 60°C for 20 min, the plate was mounted on a motorized stage and mid-infrared (IR) spectra were collected over the range 4000 to 600 cm^{-1} (see Fig. 1A for typical spectra) with 256 co-adds, and with a spectral resolution of 4 cm^{-1}. Note that, although *near*-IR spectra can be obtained in

Table I. Details of fermentation and GA3 titer[a].

Fermentation	GA3 titer (ppm)	Fermentation	GA3 titer (ppm)	Fermentation	GA3 titer (ppm)
a S^1	2290^{v}	*i* S^1	0^{tr}	*m* S^3	345^{test}
a S^1	4000^{tr}	*i* S^1	20^{test}	*m* S^3	1195^{v}
a S^1	4925^{tr}	*i* S^3	130^{tr}	*m* S^3	1905^{v}
a S^1	1705^{v}	*i* S^3	485^{v}	*m* S^3	2345^{test}
b S^2	755^{v}	*i* S^3	920^{v}	*m* S^3	2960^{tr}
b S^2	1520^{v}	*i* S^1	685^{tr}	*m* S^3	2480^{v}
b S^2	2175^{test}	*i* S^3	2475^{tr}	*n*	1650^{v}
b S^2	3895^{test}	*i* S^1	3050^{tr}	*n*	3620^{test}
b S^2	4395^{test}	*j* S^4	1980^{tr}	*o*	940^{tr}
b S^2	4860^{v}	*j* S^4	3365	*o*	2560^{v}
c S^1	1945^{test}	*k* S^3	1395^{tr}	*o*	3200^{tr}
c S^1	2765^{test}	*k* S^3	1875^{tr}	*p*	1020^{test}
c S^1	3515^{v}	*k* S^3	2515^{test}	*p*	1200^{tr}
c S^1	3780^{v}	*k* S^3	2835^{v}	*p*	1620^{test}
c S^1	3920^{tr}	*k* S^3	3080^{v}	*p*	1640^{tr}
d S^1	3660^{v}	*l* S^1	750^{test}	*p*	3110^{test}
e S^1	2365^{tr}	*l* S^1	2930^{test}	*p*	3440^{tr}
f S^1	4345^{tr}	*l* S^1	3950^{v}	*p*	3170^{test}
g S^1	3770^{test}	*m* S^3	20^{v}	*p*	3730^{test}
h S^1	4230^{v}	*m* S^3	65^{test}	*p*	3680^{tr}

Superscripts: tr, train; v, validate; test, test. Fermentations *n–p* at one production cite. S^{1-4} are labels for the four different sites of HPLC analysis. All fermentation samples were analyzed in triplicate by each of the three spectroscopic methods.

[a]GA3 titer was calculated by HPLC. The sample source is indicated by the bioprocess, labeled a–p.

the presence of H_2O, water interference is concentration-dependent, and thus its influence on the spectra is difficult to remove. Moreover, near-IR absorbance spectra are very broad and lack any obvious detail, whereas interrogation of samples in the mid-IR range allows a wealth of molecular structure information to be collected (Griffiths and de Haseth, 1986).

For chemometric processing, spectral data were converted to ASCII format, using OPUS software that controls the FT-IR instrument and spectral files were imported into MATLAB (The Mathworks, Inc., Natick, MA). Spectra were either: (1) analyzed in raw format; or (2) to minimize problems arising from unavoidable baseline shifts, the spectra were first scaled so that the smallest absorbance was set to 0 and the highest to +1 for each spectrum, and then the first Savitzky–Golay derivative (Savitzky and Golay, 1964) was calculated (see Table II).

Raman Spectroscopy

Spectra were collected using a Renishaw System 100 dispersive Raman spectrometer (Renishaw plc, Gloucestershire, UK) as described previously (Williams et al., l994a, 1994b), with a near-IR 785-nm diode laser with the power at the sampling point typically at 79 mW. The instrument grating was calibrated using neon lines (Tseng et al., 1993) and was routinely checked with a silicon wafer centered at 520 nm. Four milliliters of each sample was pipetted into a 4-mL Supelco vial (Supelco Park, Bellfonte, PA). The vial was placed into a prefixed sample holder such that the laser was focused into the center of the vial (12 mm from the collection lens). A spectrum from each sample was collected for 10 s using the continuous-extended scan according to the instrument software provided by the manufacturer (so that the actual collection time was 60 s). Samples were analyzed in triplicate on 2 separate days.

The GRAMS WIRE software package (Renishaw and Galactic Industries Corp. Salem, NH) running under WINDOWS 95 was employed for instrument control and data capture. Stokes spectra were collected over 100 to 3000 cm^{-1} wavenumber shifts with 1735 data points; therefore, the spectral resolution was ~1.67 cm^{-1}. The data were displayed as intensity of Raman photon counts against the Stokes–Raman shift in wavenumbers (see Fig. 1B for typical spectra). The spectral data were extracted into text files and imported into Matlab. Note that, although the fluorescence is relatively low when biological materials are excited at 785 nm, the system cannot discriminate whether individual photons arise by fluorescence or are scattered via the Raman effect. Although fluorescence is not seen in anti-Stokes–Raman shifts, the strength of this scatter is typically much lower than that compared with Stokes shifts.

Pyrolysis Mass Spectrometry

Clean iron–nickel foils (Horizon Instruments, Heathfield, UK) were inserted, using clean forceps, into clean pyrolysis tubes (Horizon Instruments), so that 6 mm was protruding from the mouth of the tube. Five-microlitre aliquots of crude bioprocess samples were evenly applied onto the foils. Prior to pyrolysis the samples were oven-dried at 60°C for 20 min, then the foils were then pushed into the tube using a stainless-steel depth

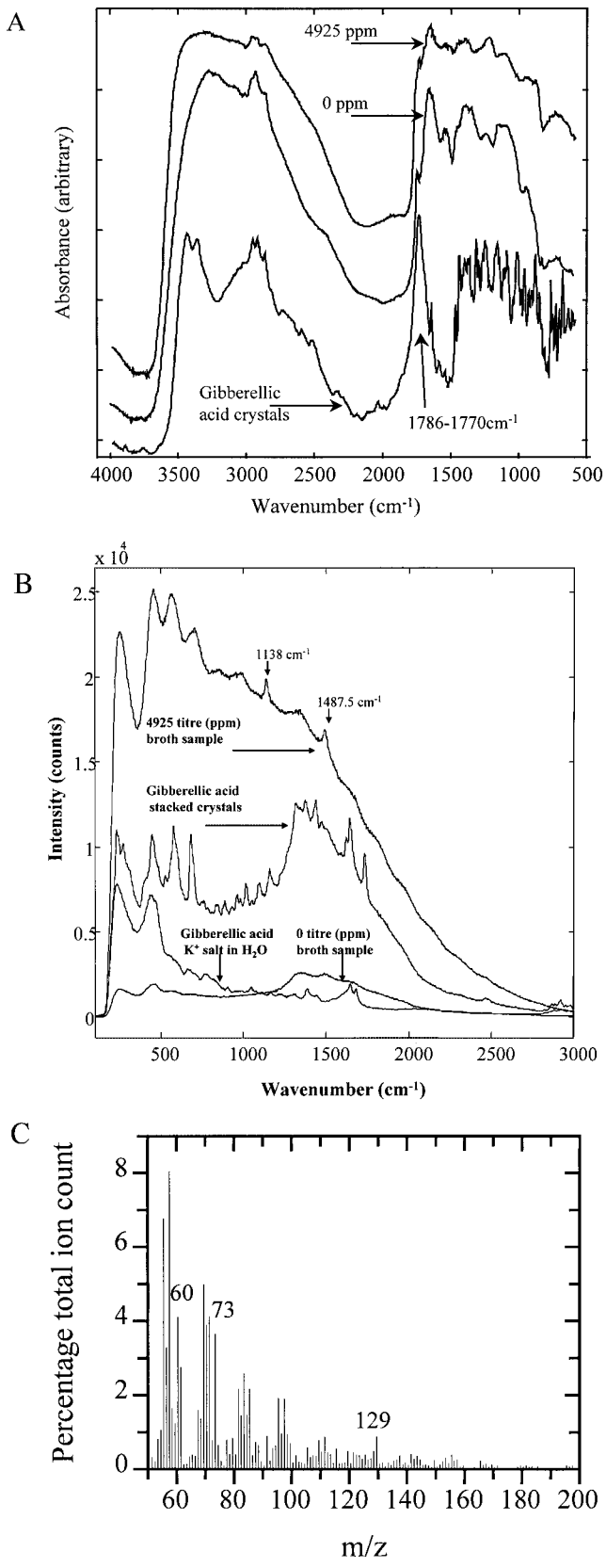

Figure 1. (A) Diffuse reflectance–absorbance FT-IR spectra and (B) dispersive Raman spectra of low- and high-titer GA3 bioprocess samples with pure industrial GA3 crystal product (as its Na salt via the addition of NaOH) and GA3 potassium salt (Sigma). The pyrolysis mass spectrum of 200 μg GA3 is shown in (C).

gauge so as to lie 10 mm from the mouth of the tube. Finally, Viton O-rings (Horizon Instruments) were placed ≈1 m*M* from the mouth of each tube. Samples were run in triplicate.

The pyrolysis mass spectrometer used for this study was a Horizon Instruments PYMS-200X device. The sample tube carrying the foil was heated prior to pyrolysis, at 100°C for 5 s. Curie-point pyrolysis was at 530°C for 3 s, with a temperature rise time of 0.5 s. Data were collected over the mass-to-charge (m/z) range of 51 to 200 and normalized as a percentage of total ion count (see Fig. 1C for a typical spectrum). Full operational details may be found elsewhere (Goodacre et al., 1994a; 1997; Goodacre and Kell, 1996).

Cluster Analysis

The initial stage involved the reduction of the dimensionality of the spectral data by principal components analysis (PCA; Jolliffe, 1986; Causton, 1987). PCA is a well-known technique for reducing the dimensionality of multivariate data while preserving most of the variance, and MATLAB was employed to perform PCA according to the NIPALS algorithm (Wold, 1991). Discriminant function analysis (DFA) was then used to cluster the spectra by discriminating between groups on the basis of the retained PC scores and the *a priori* knowledge of which spectra were replicates (MacFie et al., 1978; Windig et al., 1983), and thus this process did not bias the analysis. DFA was programmed according to Manly's principles (Manly, 1994).

Supervised Analyzes

When the desired responses (targets) associated with each of the inputs (spectra) are known, the system may then be supervised. The goal of supervised learning is to find a mathematical model that will correctly associate the inputs with the targets; this is usually achieved by minimizing the error between the target and the model's response (output). Briefly, a "training" set of spectra with known GA3 titers is used to form the model; the "validation" set, also of spectra with known GA3 titres, is used in conjunction with the training set to establish the optimum model; and an independent "test" set, consisting of spectra not seen by the model creation program at any time, is used to test the effectiveness of the calibrated system.

Creation of Training, Validation, and Test Data Sets for Supervised Learning

It is important that the training data encompass the full range under study (Bishop, 1995; Kell and Sonnleitner, 1995), because, although supervised methods are excellent at being able to interpolate, they are likely to give

Table II. Comparison of FT-IR, Raman, and pyrolysis mass spectrometry in combination with various multivariate calibration methods for the quantification of GA3 in bioprocess samples.

Calibration method	Epochs/factors	Number of runs	% RMS error of data sets			Slope of linear fit line		
			Train	Validation	Test	Train	Validation	Test
Fourier transform infrared spectroscopy[a]								
ANN 882-10-1	4000	10	1.1	9.3	9.0	0.99	0.95	0.87
PLSR1	6	1	1.4	5.1	10.9	0.96	0.95	0.94
PC-ANN 4-2-1	4	10	5.9	5.7	9.2	0.92	0.97	0.90
GAIC (raw data)		60	11.2	8.2	11.8	0.89	0.95	0.93
GA-MLR	4	150	8.0	8.8	8.7	0.90	0.95	0.94
GP		50	2.9	4.4	7.5	0.97	1.00	0.95
Raman spectroscopy[b]								
ANN 1735-12-1	2000	10	13.5	10.8	14.5	0.72	0.83	0.75
PLSR 1	9	1	1.9	8.0	11.9	0.99	0.93	0.82
PC-ANNs 5-2-1		10	8.3	8.6	9.6	0.87	0.85	0.80
GAIC		60	12.1	9.4	13.6	0.75	0.78	0.68
GP		150	7.5	6.8	10.4	0.92	0.89	0.85
Pyrolysis mass spectrometry[c]								
ANNs 150-8-1	7800	10	5.0	9.0	12.2	0.94	0.91	0.82
PLSR 1	8	1	5.7	9.7	13.3	0.93	0.95	0.94
PC-ANNs 15-2-1	15	10	6.8	8.9	14.1	0.89	0.91	0.79
GAIC		60	15.8	9.4	22.3	0.76	0.88	0.5
GA-MLR	4	150	8.7	14.2	10.3	0.89	0.93	0.98
GP		50	7.4	6.2	10.6	0.92	0.91	0.93

[a]Spectra were first scaled so that the smallest absorbance was set to 0 and the highest to +1 for each spectrum, and then the first Savitzky–Golay derivative (Savitzky and Golay, 1964) was calculated. For GAIC, the raw data were analyzed with no preprocessing.
[b]Spectra were normalized so that the smallest photon count was set to 0 and the line at 255 cm^{-1} was scaled to + 1 for each spectrum.
[c]Spectra were normalized so that the total ion count for each spectrum = 1.

poor estimates outside their ''realm of knowledge''; that is, they cannot extrapolate sufficiently well. To achieve this the spectral data from FT-IR, Raman, and PyMS were partitioned using the in-house program Multiplex (Jones et al., 1998). The Multiplex algorithm systematically placed samples into the training, cross-validation, and test sets so that the problem domain (in terms of GA3 titer) was adequately represented. Because the partitioning was based solely on the GA3 titer (rather than sample spectra), the training, cross-validation, and test sets consisted of the same fermentation samples for each of the spectroscopic methods investigated.

Common Supervised Analysis Methods

PLSR was used following the pseudocode given by (Martens and Næs (l989); the inputs and outputs were scaled to a standard deviation of 1 and mean centered (Martens and Næs, 1989). Two types of ANNs were trained by gradient descent using the standard back-propagation (BP) algorithm (Rumelhart et al., 1986), and these differed by the representation of their input patterns. ANNs were trained with either: (1) full spectral inputs; or (2) the scores of the first *n* principal components as inputs. To determine the optimum number of PCs that would represent a spectrum, a number of PC-ANNs were trained with between 1 and 15 PCs. For PyMS, the structure of ANN used consisted of three layers containing 150 input nodes, 1 output node (amount of GA3), and one ''hidden'' layer containing 8 nodes (a 150–8–1 topology), whereas, for PC-ANNs, the architecture was 15–2–1. For FT-IR 883–10–1 ANNs and 4–2–1 PC-ANNs were employed, and for Raman 1735–12–1 ANNs and PC-ANNs with a 5–2–1 topology were used. Prior to training, each input and output variable was scaled between 0.2 and 0.8.

During calibration of these models, the RMS (root-mean-square) error, between the true and desired concentrations for the validation data, was calculated; the lowest RMS error for this was used to find the optimal calibration that would give the best general predictive model. PLSR, ANNs, and PC-ANNs were carried out using an in-house package developed by Dr. Alun Jones (Jones et al., 1998), which runs under Microsoft WINDOWS NT on an IBM-compatible PC.

Evolutionary Computation

When dealing with mathematical models that are built with little, if any, *a priori* information about the system under analysis it is often difficult to decide how many variables to measure to build an adequate model. The experimental methods described in this study provide measurement of a large number of variables automatically. It is very easy for the modeler to create a model using all available variables. However, often many of the measured variables contribute little to (or even de-

grade) the predictive abilities of the final model produced (Broadhurst et al., 1997; Shaw et al., 1997). It has been shown that it is advantageous to select the "best" variables prior to the modeling process (Kell and Sonnleitner, 1995; Miller, 1990; Seasholtz and Kowalski, 1993) and, as a rule, adequate local solutions can be found in a relatively short time.

Until recently, the most popular optimization strategies were univariate in their approach (forward selection, backward selection, and stepwise multiple regression [Wonnacott and Wonnacott, 1981]); that is, each available variable is studied independently and ranked appropriately, and no consideration is given to variable interaction. To study the importance of uncorrelated groups of variables as well as individual variables, a more global optimization method needs to be used, where variables are selected or rejected simultaneously. These strategies are known as multivariate optimization methods.

Three multivariate variable selection methods based on evolutionary computation are presented in what follows.

Genetic Algorithm–multiple Linear Regression (GA-MLR) Methodology

The GA-MLR variable selection methodology as described by Broadhurst et al., (1997) uses a genetic algorithm (GA) (Goldberg, 1989; Holland, 1992) to determine the optimal subset of variables with a predetermined (from PLS calibration on the same data) maximum RMS error in an MLR (Wonnacott and Wonnacott, 1981) model.

In the GA a population of n subsets is created (the chromosome population size, $n = 400$), each containing a random combination of variables. Each subset is considered as a string of m 1's and 0's, where m is the total number of variables to choose (where $m = 882$ for FT-IR, 1735 for Raman, and 150 for PyMS). The state of each variable is represented by a "1" (selected to be in the model) or "0" (not selected). In genetic terms, each variable is called a gene and a set of variables is called a chromosome. For example, in a variable selection problem starting with 8 variables, one possible chromosome would be 00110101. This can be translated such that variables 3,4, 6, and 8 are to be used in the modeling process and variables 1, 2, 5, and 7 are to be omitted.

The five steps of (1) encoding into chromosomes, (2) initial population selection, (3) evaluation of the cost function, (4) reproduction, and (5) testing for the stopping criterion are the basic building blocks for all GAs. However, there are various ways of carrying out each step. In the current methodology, two-point crossover was used. The selection of parent chromosomes for the next generation was carried out using a rank-based scheme (Whitley, 1994), where the top 20% of each generation was included in the next generation to aid algorithm efficiency. The probabilities of crossover and mutation were set to 0.7 and 0.01, respectively, and the evolution in silico took place for 400 generations.

Genetic Algorithm Identification of Calibration (GAIC) Method

A second GA decoding method (Taylor et al., 1998b) adapted from Williams and Paradkar (1997) was employed wherein the chromosome was composed of an array of integers rather than binary digits. A chromosome comprised 10 genes, where a single gene was composed of 4 consecutive integers that encode an expression term. This gene comprises the average measurement value of a continuous region of spectral variables (integers "b" and "c"); integer "a" was a weighting value applied to this average, and this value was then linked to the next gene by integer "d", which encoded a simple arithmetic operator (+, −, *, or /) to form an expression.

For FT-IR and Raman, the GA selects the weighting values (integer "a"), the position (integer "b") and width (integer "c") of each of the spectral regions, and the operators (integer "d") linking the terms together by the use of standard one point, two-parent crossover and various mutation strategies implemented as follows. Standard, single-point random replacement mutation was employed for the weighting value (integer "a") and the arithmetic operator modification (integer "d"). A positional mutation (applied to integer "b") was used to exploit the continuous nature of the data. The point selected was moved one variable place either to the left or the right of the current position. This creates a "sliding" region to scan the spectra. To optimize the denoising effect of signal averaging, a resize mutation was used (applied to integer "c") to increase or decrease by one variable the size of the continuous region to be averaged. This resizing mutation strategy was switched off when analyzing PyMS data because the spectrum obtained from this instrument is composed of noncontinuous variables. Therefore, each region was given a constant region size of 1, and the mutation function was disabled.

Genetic Programming (GP)

A GP is an application of the GA approach to derive mathematical equations, logical rule, or program functions automatically (Banzhaf et al., 1998; Koza, 1992; 1994). Rather than representing the solution to the problem as a string of parameters, as in a conventional GA, a GP uses a tree structure. The leaves of the tree, or *terminals*, represent input variables or numerical constants. Their values are passed to *nodes*, at the junctions

of branches in the tree, which perform some numerical or program operation before passing on the result further toward the root of the tree. Mutations are performed by selecting a parent and modifying the value or variable returned by a terminal, or changing the operation performed by a node. Crossovers are performed by selecting two parents and grafting subtrees at randomly selected nodes within their trees. The new individuals so generated replace less-fit members of the population.

For the GP implementations used here only the node operator functions "add," "substract," "multiply," and "protected divide" (where $n/0 = 1$) were employed. All GP analyzes with a population size of 5000 were carried out using an in-house program (Gilbert et al., 1997), following a procedure similar to that of Singleton (1994), running under Microsoft WINDOWS NT on an IBM-compatible PC.

RESULTS AND DISCUSSION

Typical FT-IR, Raman, and PyMS spectra from *Gibberella fujikuroi* fermentations accumulating gibberellic acid are shown in Figure 1. The FT-IR spectra (Fig. 1A) from the industrial fermentation broths show broad complex spectral features and, although quantitative differences were observed, it was difficult to relate a single peak to that observed from pure GA3. From the infrared spectra of known (bio)chemicals (Schrader, 1989; Stuart, 1997), the strongest vibrations observed in GA3 can be assigned to hydroxyl groups (3450 to 3038 cm^{-1}), carbon hydrogen bonds (2940 cm^{-1}), and carboxylate group (1786 to 1770 cm^{-1}) vibrations. Moreover, many of the sharp peaks seen between 1400 and 600 cm^{-1} may also be assigned to the bending and stretching of aromatic rings, alcohol and carbon hydrogen bonds, and ether bonds, all of which appear in GA3. Two distinct Raman peaks that appear only in the higher GA3 titer samples (Fig. 1B), at 1138 cm^{-1} and 1487.5 cm^{-1}, may be attributed to GA3. Finally, whereas the complex pyrolysis mass spectrum of gibberellic acid (Fig. 1C) shows several dominant peaks, none was found to scale with the GA3 titer (data not shown).

For all three spectral types there was very little qualitative difference between the spectra, although, as highlighted earlier, some complex quantitative differences between them were observed. Such spectra, essentially uninterpretable by the naked eye, readily illustrate the need to employ multivariate statistical techniques for the analysis of FT-IR, Raman, and PyMS data. The next stage was to employ unsupervised learning to cluster the fermentation samples. Using FT-IR as an example, the DFA plot shown in Figure 2 shows that the only sample accumulating no product of interest is clearly separated in the second discriminant function (DF2) from all other samples; moreover, DF1

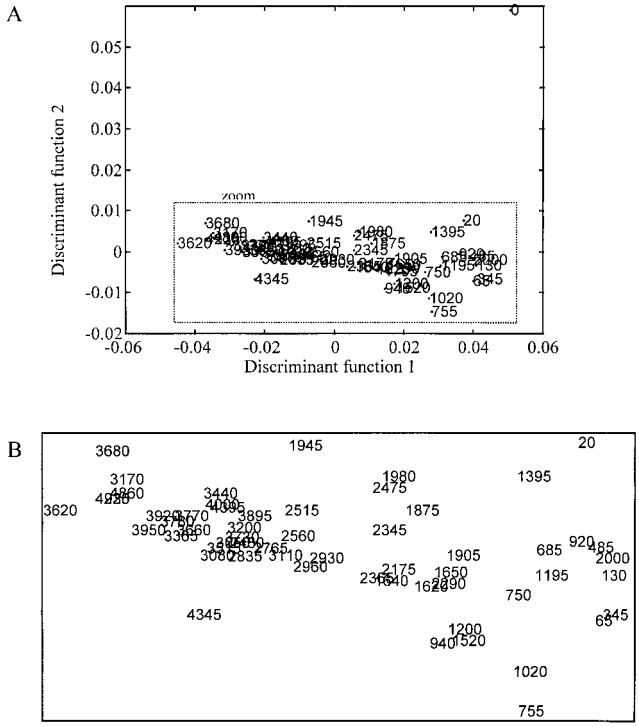

Figure 2. (A) Discriminant function analysis of the FT-IR spectra of the 60 bioprocess samples. The labels shown are the GA3 titers. (B) Enlarged version of the DFA separation on samples >0 ppm (GA3).

shows a clear linear trend from right to left for the low to medium to high samples. It is significant that at least some quantitative information was observed in DF1 because it was extracted by the DFA algorithm to contain the most overall variance. This suggests that supervised learning should be able to quantify these fermentation broths in terms of their GAS titer. Similar results were observed for DFA plots on Raman and PyMS data (data not shown).

The next stage was therefore to use linear regression and neural computation. As detailed earlier, the 60 fermentation samples, in replicate, were evenly split using the multiplex algorithm into training, validation, and independent test data sets. Table I gives details of how the samples were segregated, and also shown is the origin of the broths from 16 different fermentations indicating the wide diversity of samples that covered the range of 0 to 4925 ppm. PLSR, ANNS, and PC-ANNs were carried out as described earlier, and the RMS errors for each of the three data sets are shown in Table II. It was clear that all the models were satisfactory because the percent RMS for the independent test was between 9% and 14%. The slopes of the best linear fit lines for each of the three data sets are also shown in Table II, and in most instances these are close to the true ($y = x$) slope of 1.

Although a PLSR and the ANN methods were all able to quantify accurately the GA3 titer, is preferable to understand the modeling process in terms of which

variables, either masses or wavenumbers, are of importance. For ANNs, the information used can nominally be found in the weights (the connections between the input, hidden, and output layers); however, this information is very abstract and almost impossible to extract realistically, especially when these ANNs are *interconnected*, and for the full spectral ANNs trained with the PyMS, FT-IR, and Raman data sets contained 1217, 8841, and 20,845 weights, respectively. Whereas using PC scores as input means that the number of weights is much lower, PCs are also abstract and very difficult to interpret. For PLSR the interpretation is potentially simpler because the PLSR model is a summation of the dot products of linear weighting vectors (latent variable loadings) and the original spectral data. However, when these latent variable loadings were plotted against the original spectral data (not shown), for PyMS and FT-IR they were as complex as the original spectra and no single absorbance or m/z intensity was seen to be especially important. By contrast, although the latent variable loading plots from the Raman spectra were complex, they did show that shifts at 1138 and 1487.5 cm^{-1} were important for the formation of PLSR models. Therefore, a need to exploit supervised learning based on methods that produce rules or equations that can readily be interpreted. We have implemented a number of methods based on evolutionary computation.

As detailed previously, using the same three data sets, GAIC, GA-MLR, and GPs were evolved successfully to quantify the level of GA3 in the fermentation broths (typical rules and equations generated by these methods are detailed in Table III). Table II also contains RMS errors and slopes of best-fit lines for the training, validation, and test sets and these compare very favorably with the more classical chemometric techniques used. To highlight the success of these methods the estimates from 10 GPs versus the true GA3 titer (as judged by HPLC) are plotted in Figure 3. It can be seen that the estimates are indeed very close to the true titer and, most significantly, for the independent test set. Figure 3 contains the estimates of 10 separately evolved GPs and one can clearly see that the estimates are very similar and from the residuals plots have a precision of between 20 and 100 ppm. The next stage was therefore to inspect these rules, and the expressions from the two GA-based methods, to ascertain if any single or combination of

Table III. Examples of rules and functions produced from the evolutionary computation methods.

FT-IR data

GP rule:

$$A_{1778}\frac{\left(\frac{(9.44-A_{3638})A_{1778}}{A_{3602}(A_{3491}-.266-A_{829}}-1.45\frac{{A_{1778}}^2}{A_{3606}(A_{3692}-A_{3865})}\right)}{A_{1762}-A_{595}}+43.8\frac{\left((A_{1778}+7.45){A_{1778}}^2(A_{1763}-A_{2430})\left(\frac{A_{1778}}{A_{3869}}+32.1\right)\right)}{A_{829}A_{3919}}$$

GAIC function:

$$7.4(A_{1790})+9.99(A_{2716})+7.9(A_{1366})+9.99(A_{3834})/1.4(A_{2573})-9.99(A_{3572})+9.99(A_{3368})-9.99(A_{3931})/4.7(A_{2307})-9.99(A_{1319})$$

GA-MLR function:

$$y=0.956A_{3591}-1.65A_{1819}+3.30A_{1782}-0.887A_{1335}$$

where A = absorbance per wavenumber (cm^{-1})

Raman data

GP rule:

$$\frac{2S_{1140}-S_{2723}-2S_{1349}+S_{1487}-S_{1614}+S_{1489}}{-1.60\frac{S_{1031}}{S_{2666}}}$$

GAIC function:

$$3.7(S_{1883})-4.9(S_{1379})*9.02(S_{1029})/3.9(S_{224})*8.5(S_{1369})+8.4(S_{693})/3.9(S_{1342})/0.6(S_{2631})-1.1(S_{1634})+1.5(S_{1487})-0.4(S_{1710})$$

GA-MLR function: not computable (see text)

where S = photons per scatter (cm^{-1})

PyMS data

GP rule:

$$\frac{M_{92}}{M_{194}}+\frac{M_{103}}{M_{184}}-9.93-\frac{\frac{2M_{161}}{M_{195}}-12.90}{M_{75}}-\left(\frac{M_{161}}{M_{195}}-76.1+\frac{M_{114}}{M_{181}}-\frac{0.388}{M154}\frac{M_{114}}{M_{192}}+7.49M_{161}\right)\left(\frac{7.49}{M_{75}}-(10.72M_{85})(M_{55}-5.84)\right)$$

GAIC function:

$$9.2(M_{66})*3.8(M_{177})/4.04(M_{156})*4(M_{92})*205(M_{143})-9.5(M_{74})/3.4(M_{121})*7.5(M_{79})+9.9(M_{80})$$

GA-MLR function:

$$y=-0.277M_{75}-0.494M_{114}+1.04M_{153}-0.53M_{166}$$

where M = mass intensity (m/z ratio)

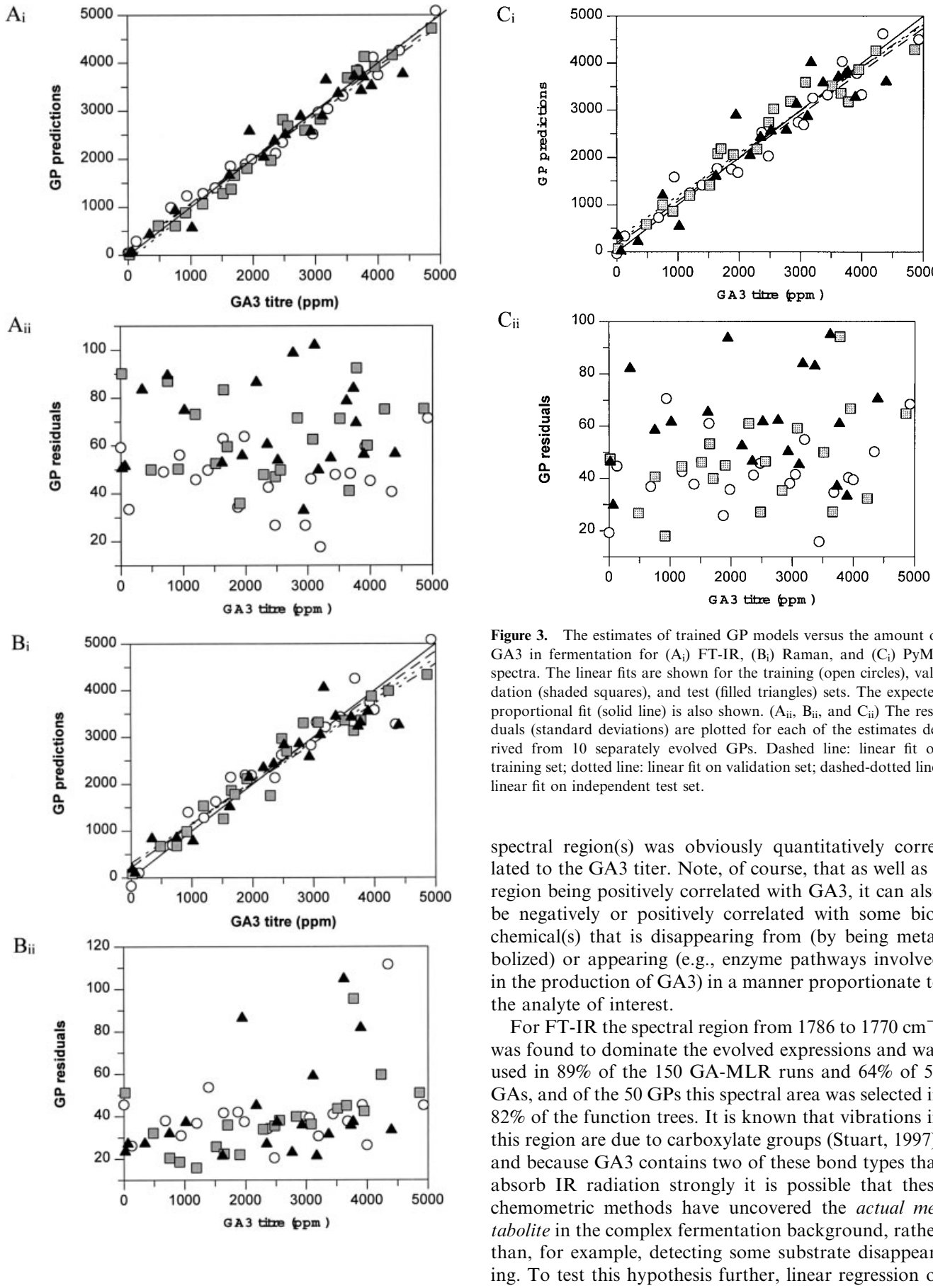

Figure 3. The estimates of trained GP models versus the amount of GA3 in fermentation for (A_i) FT-IR, (B_i) Raman, and (C_i) PyMS spectra. The linear fits are shown for the training (open circles), validation (shaded squares), and test (filled triangles) sets. The expected proportional fit (solid line) is also shown. (A_{ii}, B_{ii}, and C_{ii}) The residuals (standard deviations) are plotted for each of the estimates derived from 10 separately evolved GPs. Dashed line: linear fit on training set; dotted line: linear fit on validation set; dashed-dotted line: linear fit on independent test set.

spectral region(s) was obviously quantitatively correlated to the GA3 titer. Note, of course, that as well as a region being positively correlated with GA3, it can also be negatively or positively correlated with some biochemical(s) that is disappearing from (by being metabolized) or appearing (e.g., enzyme pathways involved in the production of GA3) in a manner proportionate to the analyte of interest.

For FT-IR the spectral region from 1786 to 1770 cm^{-1} was found to dominate the evolved expressions and was used in 89% of the 150 GA-MLR runs and 64% of 50 GAs, and of the 50 GPs this spectral area was selected in 82% of the function trees. It is known that vibrations in this region are due to carboxylate groups (Stuart, 1997), and because GA3 contains two of these bond types that absorb IR radiation strongly it is possible that these chemometric methods have uncovered the *actual metabolite* in the complex fermentation background, rather than, for example, detecting some substrate disappearing. To test this hypothesis further, linear regression of

each wavelength was carried out onto the GA3 titer using product moment correlation (PMC) as detailed by Gilber et al. (1997). Although their individual PMC values were low it is noteworthy that the spectral region from 1786 to 1770 cm^{-1} was the *only area positively correlated* with GA3. This region was investigated for its usefulness in forming linear regression predictive models; as expected from the PMC, calculations using independent (single) wavenumbers gave very poor predictions (data not shown). However, when multiple wavenumbers in this region were used for MLR, this combination gave very satisfactory results, and a definite linear trend between predicted and expected GA3 titer was observed.

Many masses in the PyMS were selected by the evolutionary methods, and the consensus from this was that m/z 75 and 153 were chosen most frequently with a percentage frequency of 32 and 43, respectively. Some of the masses selected (e.g., m/z 55, 58, 67, 71, 85, 86, and 114), including m/z 75, were negatively correlated with GA3 titer, and whereas it would be valid to model on something (a substrate) that is disappearing in a manner proportionate to the analyte of interest, this approach may be more hazardous because the disappearance of a substrate does not guarantee its appearance in a product. Moreover, PyMS has significant disadvantages in that: (1) the *in vacuo* thermal degradation step means that essentially all information on the structure or identity of the molecules producing the pyrolysate is lost; and (2) molecular reactions in the melt or pyrolysate–pyrolysate interactions in the gas phase can yield new molecular species (Goodacre et al., 1994b). Therefore, it is not sensible to use this destructive technique to attempt to elucidate precise structural information when the target analyte is a very complex, high-molecular-weight molecule.

One of the advantages that Raman has over infrared and PyMS is that measurements can be made on wet samples, because, in contrast to FT-IR, the aqueous media do not strongly absorb the interrogating beam at 785 nm (Ferraro and Nakamoto, 1994). However, it is necessary for the far red laser beam to penetrate the broth and thus, like any other optical technique, some density-dependent phenomena will be observed. Most notable (Fig. 1B) is that the total Raman signal increases with increasing titer and biomass, and therefore the PMC calculations for all variables were similar and positively correlated at >0.7. The GA-MLR method could not be applied to the Raman data because the ratio of the number of objects to the variables was such that the algorithm became unstable, and therefore optimization of simple MLR was not possible in this instance. Although the GAIC method did not select any single spectral region with any great frequency ($>50\%$), and Raman shifts were highlighted across the whole spectra, the most popular shift chosen was 1138 cm^{-1} and this was selected 14% of the time. By contrast, the GP-approach used the Raman shifts at 1138 and 1487 cm^{-1} with a frequency of 56% and 40%, respectively. These variables were clearly visible as two strong peaks in the Raman spectrum of the fermentation broth containing 4925 ppm GA3, and were absent from the zero titer broth (Fig. 1B). It is likely that these vibrations are from the GA3 itself as the spectrum of the industrial product (Fig. 1B) showed vibrations in these areas. Moreover, the Stokes shift at 1487 cm^{-1} can be attributed to ring and CH stretching in aromatic rings, and the vibration at 1138 cm^{-1} can be from ether bonds (C—O—C) (Schrader, 1989) and/or CH stretching in aromatic rings. All these molecular bonds are found in gibberellic acid.

CONCLUSIONS

The three rapid spectroscopic approaches of Curie-point PyMS, diffuse reflectance–absorbance FT-IR, and dispersive Raman spectroscopy were used to analyze a diverse range of unprocessed fed-batch fermentations broths containing the fungus *Gibberella fujikuroi* producing the natural product gibberellic acid. To obtain quantitative information in terms of the gibberellic acid titer, the modern chemometric-based techniques of PLSR and ANNs were employed. However, although these are very powerful techniques, the precise information they use to construct the calibration is effectively inaccessible. Therefore, a variety of evolutionary computational-based methods, including genetic algorithms and genetic programming, were used to decipher the spectra. The results from the FT-IR and Raman studies show that the models formed were based on spectral features derived from the GA3 molecule itself, and which could be used to quantify the product in these industrial fermentations.

The typical accuracy by all methods when analyzed by GP was between 7.5% and 10.6% (Table II), and considering that these spectroscopic approaches need to be trained with "primary reference data" (gold standards) based on HPLC measurements, which themselves have a typical error in GA3 measurement of 2% to 5%, these error measurements are suitably low. Moreover, because HPLC takes 30 min plus additional extraction and sample preparation time, the speed of the measurements from PyMS (2 min), FT-IR (10 s), and Raman spectroscopy (1 min) makes them particularly attractive to industry (Shaw et al., 1999b; Validyanathan et al., 1999).

Raman spectroscopy has the advantage over PyMS and FT-IR in that it is capable of being used noninvasively. By contrast, PyMS is highly destructive because the sample is thermally degraded *in vacuo* and, although FT-IR is not destructive *per se*, mid-IR does require that the sample is dry. These and the other features of the three spectroscopic methods are detailed in Table IV. FT-IR has the advantage of speed and, particularly with our diffuse reflectance–absorbance approach, easily

Table IV. Features of the spectroscopic approaches investigated.

	Curie-point PyMS	Diffuse reflectance–absorbance FT-IR	Dispersive Raman
Destructive	Yes	No	No
Sample dried	Yes	Yes	No
On- or at-line	At-line	At-line	At- or on-line
Reproducibility	Poor	Good	Good
Sample size for at-line	5 μL	5–10 μL	Any volume
Typical speed for collection of spectra	2 min	10 s	~1 min
Automatable	Yes	Yes	Yes
Complex data capture	No	No	No
Typical dimensionality	150	882	1735
Data analysis	Easy	Easy	Easy
Relative expense of equipment	Moderate	Low	High
Consumable costs	High	Moderate	Low
Other significant problems	Unable to analyze volatiles	None, other than that H_2O vibrations swamp mid-IR	Stokes–Raman scattered photons can not be distinguished from fluorescence

allows acquisition of 400 samples per hour on a single 10×10 cm aluminum plate. On the other hand, Raman has a slightly slower sample throughput than FT-IR, because the sample does not need to be dried and spectral acquisition can be made directly on the fermentation broth. Indeed, due to its confocal nature it is possible to focus the laser into a liquid sample through a window in the fermentation vessel, rather than introducing a probe. Thus, Raman spectroscopy appears to be the ideal bioprocess-monitoring method as it is rapid, on-line, noninvasive, and gives interpretable answers of good precision.

References

Adar F, Geiger R, Noonan J. 1997. Raman spectroscopy for process/quality control. Appl Spectrosc Rev 32:45–101.

Bäck T, Fogel DB, Michalewicz Z. 1997. Handbook of evolutionary computation. Oxford, UK: IOP Oxford University Press.

Banzhaf W, Nordin P, Keller RE, Francone FD. 1998. Genetic programming: An introduction. San Francisco, CA: Morgan Kaufmann.

Bishop CM. 1995. Neural networks for pattern recognition. Oxford, UK: Clarendon Press.

Brimmer PJ, Hall JW. 1993. Determination of nutrient levels in a bioprocess using near-infrared spectroscopy. Can J App Spectrosc 38:155–162.

Broadhurst D, Goodacre R, Jones A, Rowland JJ, Kell DB. 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 348:71–76.

Brückner B, Blechschmidt D. 1991. The gibberellin fermentation. Crit Rev Biotechnol 11:163–192.

Causton DR. 1987. A biologist's advanced mathematics. London: Allen and Unwin.

Ferraro JR, Nakamoto K. 1994. Introductory Raman spectroscopy. London: Academic Press.

Gilbert RJ, Goodacre R, Woodward AM, Kell DB. 1997. Genetic programming; a novel method for the quantitative analysis of pyrolysis mass spectral data. Anal Chem 69:4381–4389.

Goldberg DE. 1989. Genetic algorithms in search, optimization and machine learning. Reading, MA: Addison-Wesley.

Goodacre R, Hammond D, Kell DB. 1997. Quantitative analysis of the adulteration of orange juice with sucrose using pyrolysis mass spectrometry and chemometrics. J Anal Appl Pyrol 40/41:35–158.

Goodacre R, Kell DB. 1996. Pyrolysis mass spectrometry and its applications in biotechnology. Curr Opin Biotechnol 7:20–28.

Goodacre R, Neal MJ, Kell DB. 1994a. Rapid and quantitative analysis of the pyrolysis mass spectra of complex binary and tertiary mixtures using multivariate calibration and artificial neural networks. Anal Chem 66:1070–1085.

Goodacre R, Timmins ÉM, Burton R, Kaderbhai N, Woodward A, Kell DB, Rooney PJ. 1998. Rapid identification of urinary tract infection bacteria using hyperspectral, whole organism fingerprinting and artificial neural networks. Microbiology 144:1157–1170.

Goodacre R, Timmins ÉM, Rooney PJ, Rowland JJ, Kell DB. 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, Trew S, Wrigley-Jones C, Neal MJ, Maddock J, Ottley TW, Porter N, Kell DB. 1994b. Rapid screening for metabolite overproduction in fermentor broths using pyrolysis mass spectrometry with multivariate calibration and artificial neural networks. Biotechnol Bioeng 44:1205–1216.

Goodacre R, Trew S, Wrigley-Jones C, Saunders G, Neal MJ, Porter N, Kell DB. 1995. Rapid and quantitative analysis of metabolites in fermentor broths using pyrolysis mass spectrometry with supervised learning: Application to the screening of *Penicillium chrysogenum* fermentations for the overproduction of penicillins. Anal Chim Acta 313:25–43.

Griffiths PR, de Haseth JA. 1986. Fourier transform infrared spectrometry. New York: John Wiley & Sons.

Harris CM, Kell DB. 1985. The estimation of microbial biomass. Biosensors 1:17–84.

Holland JH. 1992. Adaption in natural and artificial systems. Cambridge, MA: MIT Press.

Jolliffe IT. 1986. Principal component analysis. New York: Springer.

Jones A, Young D, Taylor J, Kell DB. 1998. Quantification of microbial productivity via multi-angle light scattering and supervised learning. Biotechnol and Bioeng 52:131–l43.

Kell DB, Sonnleitner B. 1995. GMP—Good modelling practice: An essential component of good manufacturing practice. Tibtech 13:481–492.

Koza JR. 1992. Genetic programming: On the programming of computers by means of natural selection. Cambridge, MA: MIT Press.
Koza JR. 1994. Genetic Programming II: Automatic discovery of reusable programs. Cambridge, MA: MIT Press.
Lavine BK. 1998. Chemometrics. Anal Chem 70:R209–R228.
MacFie HJH, Gutteridge CS, Norris JR. 1978. Use of canonical variate analysis in differentiation of bacteria by pyrolysis gas–liquid chromatography. J Gen Microbiol 104:67–74.
Manly BFJ. 1994. Disriminant function analysis. In: Multivariate statistical methods. London: Chapman & Hall.
Martens H, Næs T. 1989. Multivariate calibration. New York: John Wiley & Sons.
Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kaufmann L. 1988. Chemometrics: A textbook. Amsterdam: Elsevier.
Mattu MJ, Small GW, Arnold MA. 1997. Application of multivariate calibration techniques to quantitative analysis of bandpass-filtered Fourier transform infrared interferogram data. Appl Spectrosc 51:1369–1376.
McGovern AC, Ernill R, Kara BV, Kell DB, Goodacre R. 1999. Rapid analysis of the expression of heterologous proteins in *Escherichia coli* using pyrolysis mass spectrometry and Fourier transform infrared spectroscopy with chemometrics: Application to α2-interferon production. J Biotechnol (in press).
Meuzelaar HLC, Haverkamp J, Hileman FD. 1982. Pyrolysis mass spectrometry of recent and fossil biomaterials. Amsterdam: Elsevier.
Miller AJ. 1990. Subset selection in regression analysis. London: Chapman & Hall.
Mitchell M. 1995. An introduction to genetic algorithms. Boston: MIT Press.
Montague G, Morris J. 1994. Neural network contributions in biotechnology. Trends Biotechnol 12:312–324.
Pons M-N, editor. 1991. Bioprocess monitoring and control. Munich: Hanser.
Roberts MJ, Garrison AA, Kercel SW, Muly EC. 1991. Raman spectroscopy for on-line, real-time, multipoint industrial chemical analysis. Proc Control Quality 1:281–291.
Rumelhart DE, McClelland JL, Group P. 1986. Parallel distributed processing, experiments in the microstructure of cognition. Cambridge, MA: MIT Press.
Savitzky A, Golay MJE. 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem 36:1627–1633.
Scheper TH, Lammers F. 1994. Fermentation monitoring and process control. Curr Opin Biotechnol 5:187–191.
Schrader B. 1989. Raman infrared atlas of organic compounds. New York: Verlag Chemie.
Schrader B. 1995. Infrared and Raman spectroscopy: Methods and applications. Weinheim: Verlag Chemie.
Seasholtz MB, Kowalski B. 1993. The parsimony principle applied to multivariate calibration. Anal Chim Acta 277:165–177.
Shaw AD, diCamillo A, Vlahov G, Jones A, Bianchi G, Rowland J, Kell DB. 1997. Discrimination of the variety and region of origin of extra virgin olive oils using C-13 NMR and multivariate calibration with variable reduction. Anal Chim Acta 348:557–374.
Shaw AD, Kaderbhai N, Jones A, Woodward AM, Goodacre R, Rowland JJ, Kell DB. 1999a. Non-invasive, on-line monitoring of the biotransformation by yeast of glucose to ethanol using dispersive Raman spectroscopy and chemometrics. Appl Spectrosc 53:1419–1428.
Shaw AD, Winson MK, Woodward AM, McGovern AC, Davey, HM, Kadherbhai N, Broadhurst D, Gilbert RJ, Taylor J, Timmins EM, Alsberg BK, Rowland JJ, Goodacre R, Kell DB. 1999b. In: Advances in biochemical engineering biotechnology. Berlin: Springer. p 83–111.
Singleton A. 1994. Genetic programming with C^{++}. Byte 19:171.
Stuart B. 1997. Biological applications of infrared spectrocopy. Chichester, UK: John Wiley & Sons.
Taylor J, Goodacre R, Wade WG, Rowland JJ, Kell DB. 1998a. The deconvolution of pyrolysis mass spectra using genetic programming: Application to the identification of some *Eubacterium* species. FEMS Microbiol lett 160:237–246.
Taylor J, Rowland JJ, Gilbert RJ, Jones A, Winson MK, Kell DB. Genetic algorithm decoding for the interpretation of infra red spectra in analytical biotechnology. Technical report CSRP-98-10 1998b. Birmingham, UK: University of Birmingham.
Timmins ÉM, Howell SA, Alsberg BK, Noble WC, Goodacre R. 1998. Rapid differentiation of closely related *Candida* species and strains by pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. J Clin Microbiol 36:367–374.
Timmins EM, Quain DE, Goodacre R. 1998. Differentiation of brewing yeast strains by pyrolyis mass spectrometry and fourier transform infrared spectroscopy. Yeast 14:885–893.
Tseng CH, Ford JF, Mann CK, Vickers TJ. 1993. Wavelength calibration of a multichannel spectrometer. Appl Spectrosc 47:1808–1813.
Validyanathan S, Macaloney G, McNeill B. 1999. Fundamental investigations on the near-infrared spectra of microbial biomass as applicable to bioprocess monitoring. Analyst 124:157–162.
Whitley D. 1994. A genetic algorithm tutorial. Stat Comput 4:65–85.
Williams KPJ, Pitt GD, Batchelder DN, Kip BJ. 1994a. Confocal Raman microspectroscopy using a stigmatic spectrograph and CCD detector. Appl Spectrosc 48:232–235.
Williams KPJ, Pitt GD, Smith BJE, Whitley A, Batchelder DN, Hayward IP. 1994b. Use of a rapid-scanning stigmatic Raman-imaging spectrograph in the industrial environment. Raman Spectrosc 25:131–138.
Williams RR, Paradkar RP. 1997. Correcting fluctuating baselines and spectral overlap with genetic regression. Appl Spectrosc 51:92–100.
Windig W, Haverkamp J, Kistemaker PG. 1983. Interpretation of sets of pyrolysis mass spectra by discriminant analysis and graphical rotation. Anal Chem 55:81–88.
Winson MK, Goodacre R, Woodward AM, Timmins ÉM, Jones A, Alsberg BK, Rowland JJ, Kell DB. 1997. Diffuse reflectance absorbance spectroscopy taking in chemometrics (DRASTIC). A hyperspectral FT-IR-based approach to rapid screening for metabolite overproduction. Anal Chimi Acta 348:273–282.
Wold S. 1991. Chemometrics, why, what and where to next? J Pharm Biomed Analysis 9:589–596.
Wonnacott TH, Wonnacott RJ. 1981. Regression: A second course in statistics. New York: John Wiley & Sons.
Zupan J, Gasteiger J. 1993. Neural networks for chemists: An introduction. Weinheim: VCH.