Adaptive Denoising in Spectral Analysis by Genetic Programming

Jem J. Rowland and Janet Taylor
Department of Computer Science
University of Wales, Aberystwyth
Aberystwyth,
Wales, SY23 3DB
U.K.
email: jjr@aber.ac.uk; jat@aber.ac.uk

Abstract - This paper relates to supervised interpretation of the infra red analytical spectra of complex biological samples. The aim is to produce a model that can predict the value of a measurand of interest, such as the concentration of a particular chemical constituent in complex biological material. Conventionally, a number of spectra are co-added to reduce measurement noise and this is time consuming. In this paper we demonstrate the ability of evolutionary search to provide adaptive averaging of spectral regions to provide selective tradeoff between spectral resolution and signal-to-noise ratio. The resultant denoised subset of the variables is then input to a proprietary Genetic Programming (GP) package which forms a predictive model that compares well in predictive power with a combination of Partial Least Squares Regression (PLS) and adaptive denoising. This demonstrates the considerable advantage that, given appropriate node functions, the GP could handle the entire process of denoising and forming the final predictive model all in one stage. This reduces or removes the need for co-adding with a consequent reduction in data acquisition time.

I. INTRODUCTION

Infra red (IR) absorbance spectroscopy measures the optical absorbance of a sample at successive wavelengths[1]. Specific regions in the spectrum relate to the vibrational characteristics of specific chemical bonds [1] and the spectrum can therefore reveal the chemical composition of the sample under examination in terms of the constituent components and the relative concentrations of each. For biological specimens that inevitably contain large molecules, interpretation of their IR spectra is difficult by any means other than those based on machine learning. The topic thus provides a fruitful stimulus for the development of appropriate approaches.

The work reported in this paper concerns an investigation of the feasibility of achieving adaptive denoising of infra red spectra via genetic programming (GP) [2] that incorporates appropriate local averaging. This would achieve the appropriate tradeoff between signal-to-noise ratio and resolution for different spectral regions. The inputs are the individual spectra and the output is the predicted concentration of a chemical of interest.

Spectra always contain measurement noise and this interferes with numeric interpretation. Various successful methods have been reported for denoising prior to predictive modelling but part of the aim of the work we report here was to remove the need for preprocessing and to provide a self-contained tool for laboratory use.

A. Supervised learning for spectral interpretation

Perhaps the simplest supervised technique is regression. In its simplest, single dimensional, form this fits a straight line to a set of points as a means of establishing an empirical relationship between a set of observations and a dependant variable, thereby forming a calibration model. In our current context we are concerned with multidimensional data such as the 882 variables that represent a single 'observation' using a Fourier Transform Infra Red (FT-IR) spectrometer[2]. Regression extends to multivariate data in, for example, the form of Multiple Linear Regression (MLR) (e.g. [3]) which, however, does not behave well when there is colinearity between dimensions, as is a common feature with optical spectra. Partial Least Squares Regression (PLS) [4] overcomes this problem and is a regression method based on latent variables whose selection is based on eliminating covariance between 'X' variables (the measured variables) while exploiting covariance between the 'X' and 'Y' variable (the target variable). PLS has performed well in forming calibration models based on optical spectra [5]. Backpropagation neural networks (e.g. [6]) are also successful (e.g. [7]), being able in principle to fit any nonlinear multivariate function. Such supervised methods can of course be used

[1]Spectroscopists prefer to work in terms of wavenumbers, calculated by taking the reciprocal of the wavelength in centimetres.

[2]E.g. A Bruker IFS28 Fourier Transform Infra Red spectrometer. Bruker, Banner Lane, Coventry, UK

0-7803-7282-4/02/$10.00 ©2002 IEEE

both for quantification problems and supervised clustering.

There is considerable evidence in the literature (e.g. [8], [9], [10] that using evolutionary and other techniques to select subsets of the 'X' variables can provide much improved calibration models. For example, Broadhurst [11] formed PLS models on the basis of selection via genetic algorithm of the most important peaks in mass spectra. One of the aspects of our work reported in this paper is the use of evolutionary search to select an effective denoised subset of the variables in infra red spectra.

Williams and co-workers, e.g. [12], [13], explored their method of 'genetic regression'. The original motivation was to eliminate the spectral baseline in calibration of component concentration in visible spectra but they extended their work to other aspects of spectroscopy.

Successes have been achieved in the use of GP for spectral interpretation (e.g. [14], [15]). Johnson et. al. [16] provides an excellent example of an explanatory model that provides insight into the system under study, thereby revealing unexpected new knowledge.

Kell *et al* [17] used a variant of GP that is embodied in a proprietary tool that was also used in part of the work we describe in this paper. They used liquid chromatography data, that has many similarities to optical spectroscopic data. They were able with 95% accuracy to determine the presence or otherwise of a specific genetic modification in plant tissue.

Predictive models are normally judged and compared as to their effectiveness by means of the root mean square error between the known values (the 'Y' data) that correspond to the data objects and the predicted values produced via the model. We refer to this generally as the RMSEP, or root mean square error of prediction [18].

II. INFRA RED SPECTRA AND NOISE

Spectra normally contain measurement noise. Formation of a model on a relatively small number of individual variables is therefore inappropriate. Preprocessing to remove noise etc. is described in, for example, [19], [20], [21]. Other approaches include filtering of various forms and taking many repeated spectra from each sample and 'co-adding' them to improve the signal to noise ratio. This of course is time consuming, with many tens or hundreds of repeats being needed to achieve 'clean' spectra in some cases. In a typical instrument each spectrum consists of up to 882 absorbance or reflectance measurements taken at successive wavelength (or wavenumber) settings. The spectral resolution is finer than the width of the many peaks that, superimposed, make up the overall spectrum. Thus there is a high degree of covariance between adjacent variables. A simple method of denoising is to replace each variable with the average value of the variables in a fixed width window that is moved across the entire spectrum. This method is, of course, normally unable to adapt to the differences in signal-to-noise ratio in different parts of the spectrum.

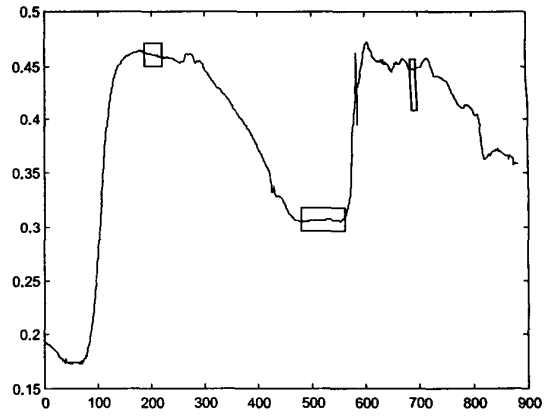

Fig. 1. *Spectral windows. In this example, that is for illustration only, four windows are shown, each of a different width. The window centre positions and widths are evolved by the GA and the spectral variables within them are averaged to remove noise. The extent of noise removal therefore depends on the window width.*

III. BASIS OF THE METHOD

Our method, described below, currently employs a two-stage approach. First, we use a genetic algorithm (GA) to form a simple predictive model from the spectra. This selects the ten most 'important' spectral variables. It also selects a value for the width of a 'window' around each of those variables, within which all variables are averaged and the result used as a new variable (see figure 1). The effect is to provide differing degrees of denoising in different parts of the spectrum. The selected subset of denoised variables is then fed into a GP, which quickly forms a better predictive model, indicating that a GP with suitable functions could provide a 'one-step' tool. The results are compared with PLS on the raw noisy data and on a version of the noisy data that has been (partially) denoised by a moving average filter.

IV. THE DATA SET

The data set is derived from a number of samples taken for the purpose of monitoring an industrial-scale fermentation and is described more fully in [22]. The fermentation is the production of gibberellic acid (GA3). This is a natural regulator of many aspects of plant growth and is of economic importance in agriculture and horticulture. Sixty fermentation samples, collected over a three month period from industrial fermentations, were analysed first by liquid chromatography to determine the gibberelic acid concentrations and thus to provide the 'Y' data for the supervised learning. The remainder were analysed spec-

troscopically via the Bruker IFS28 infra red spectrometer with a diffuse-reflectance attachment. $5\mu l$ of each sample were placed in the wells of a 400 well aluminium plate and oven dried prior to analysis. Mid infra red spectra were collected over the wavenumber range $4000cm^{-1}$ to $600cm^{-1}$ acquired at a rate of 20 s^{-1}, at a spectral resolution of 3.85 cm^{-1}. To improve signal-to-noise ratio, 256 spectra were co-added and averaged. Each sample was analysed in triplicate, resulting in a 180 by 882 data matrix. This was subsequently partitioned into training, and test sets each containing 60 objects and a third set that was not used in the comparitive study reported in this paper but would be used in validating a model that was to be used for prediction on new data sets.

For the purposes of the present investigation into adaptive denoising, the co-added spectra were artificially contaminated with noise. This was done by adding to each spectral variable a random signed value of maximum amplitude equal to 20% of the average range of the spectra in the data set. For examples of 'clean' and 'noisy' spectra, see figure 2.

V. GA FOR SELECTION OF DENOISED VARIABLES

The work reported here uses a genetic algorithm to select variables, constants and arithmetic functions to form a simple model that predicts the value of the measurand (in this case the concentration of gibberelic acid). The GA is a modification of the supervised predictive system based on linear chromosomes previously described by Taylor *et. al.* [23]. The slightly modified version of that GA that we have used in this work incorporates mutation operators that modify the location and width of windows, as illustrated above, but the 'sliding' mutation used in [23] is not used here.

The overall effect is to select, on the basis of predictive ability, a set of spectral variables and corresponding windows across which the values are averaged in order to achieve denoising. The resulting window widths provide different degrees of denoising depending on the required level of relevant detail in the various spectral regions, traded off against the noise level.

Each window mid-point could take a random value between 5 and 877 (the maximum number of variables is 882, but the mid point value may not be set beyond half the width of the maximum window size). The window width could take random values between 1 and 10; the maximum of 10 was chosen on the basis of initial trials in which the limit was higher, but in which the largest window width selected by the GA was 9. Restricting the maximum in this way reduces the 'end effects' at the extremities of the spectrum.

The fitness function was constructed so as to minimise the raw RMSEP of the predicted values and thus maximise the prediction accuracy.

A. Termination criterion

The GA training was terminated after an arbitrary number of generations that always proceeded past the point at which the learning curves of the training and validation sets diverged (see figure 3). The divergence indicates the onset of overtraining, where the model begins to incorporate the noise in the training set and ceases to generalise. The RMSEP value of the test set at divergence was taken as the value predicted by the model.

B. Evolution conditions

The GA was run on a Sun 9500 Enterprise server under Solaris. Initially, many configurations of the GA parameters, particularly mutation and crossover probabilities, were tested. Sensitivity to different crossover and mutation rates was not unduly large. However, based on these initial trials, a population size of 1000 was used, with mutation probability of 0.7 and crossover of 0.3. The seemingly high mutation rate reflects the complexity of the chromosome and the different types of mutation incorporated (see [23]), so that the effective mutation rate is rather lower than the figure suggests.

Ten runs were then undertaken, with a chromosome length that selected ten spectral variables with a window of GA-selected width surrounding each of them. Out of these ten runs the three that performed best were identified, as judged by the test set RMSEP at which the learning curves for training set and test set diverged (see figure 3). 'Best' variables and corresponding window widths from each of these three runs were then used to form a new subset of the original data. Because of ambiguity in determining the divergence point (for an example see the left hand plot in figure 4) in two of these three runs, windows specified by three chromosomes were taken from two of the runs, and by one chromosome from the third. Each new variable was the average of the spectral variables in the windows selected by the GA.

This resulted in a subset with 61 denoised variables; ten of these windows were identical and so the final subset contained 51 variables for each of the 60 samples in the training set and the 60 in the test set. The GP, and PLS for comparison, were then applied to this data subset.

VI. GENETIC PROGRAMMING

This used a proprietary package[3] marketed for applications in the bio-technology and pharmaceuticals industries. Currently it does not have inherent denoising capability, beyond the obvious ability of a GP to combine variables in various ways, and part of the aim of the current work was to evaluate the potential benefit of incorporating specialised node functions for denoising. The GP was run on a 1GHz Pentium under Windows 2000.

[3]gmax-bio, www.abergc.com

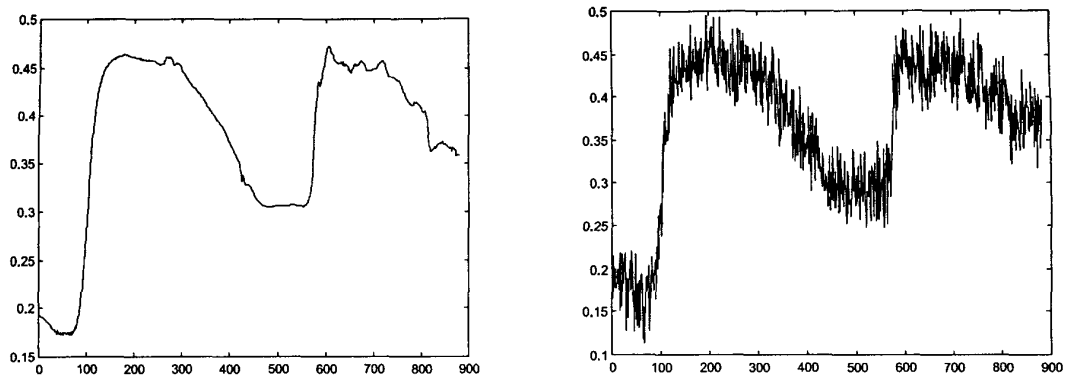

Fig. 2. *On the left, a virtually noise-free Fourier Transform Infra Red spectrum taken from the set of spectra upon which this paper is based; on the right, the same spectrum to which noise was subsequently added artificially. The axes are absorbance vs. wavenumber.*

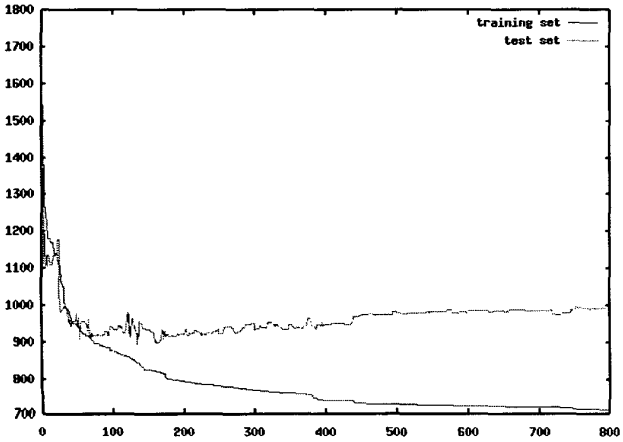

Fig. 3. *An example of training and test set learning curves from the GA modelling process. The vertical axis is the RMS Error and the horizontal axis is the number of generations. In this example the point at which the two curves diverge, and overtraining commences, is relatively easy to determine.*

Given the expected linearity of the relationship between the concentration of the chemical of interest and the height of the relevant spectral features, the GP was restricted to using only the four simple arithmetic functions; it was also permitted to use random constants as additional terminals. The population size was 1000, crossover 80% and mutation 10%. The package's default proprietary fitness function was used. Convergence was rapid, always being achieved in well under a minute. For each run, the termination point was judged on the basis of divergence of the model performance on the training and independent test sets. Typically, the models used only three or four variables from the denoised subset.

	PLS	Best GP	GP Mean (10 runs)	GP Std dev
Noisy data	940	N/A	N/A	N/A
Noisy data - denoised subset	744	739	757	12.9
Noisy data, filtered by moving average filter (width 10)	1410	N/A	N/A	N/A

TABLE I

Test Set RMSEP values that illustrate the results of the technique presented in this paper in comparison with PLS and conventional filtering. The table entries marked as N/A are not central to the point at issue.

VII. OVERALL RESULTS AND COMPARISONS

The RMSEP values obtained from the GP were compared with those obtained using a PLS model at the number of factors (3) at which the learning curves for train and test sets diverged. They were also compared with the PLS model produced on a version of the noisy data set that was filtered by a simple 10 element moving average filter. The results are shown in Table I.

The RMSEP values obtained by the GP and PLS, when applied to the subset of the original spectral variables that had been denoised by the GA, compared very well and this indicates that the approach based entirely on evolutionary search is able to compete with the more well-established PLS. Both results illustrate the benefit of adaptive denoising.

It is particularly interesting to note that the combination of the simple moving average filter and PLS gave a very poor prediction error (1400), which demonstrates

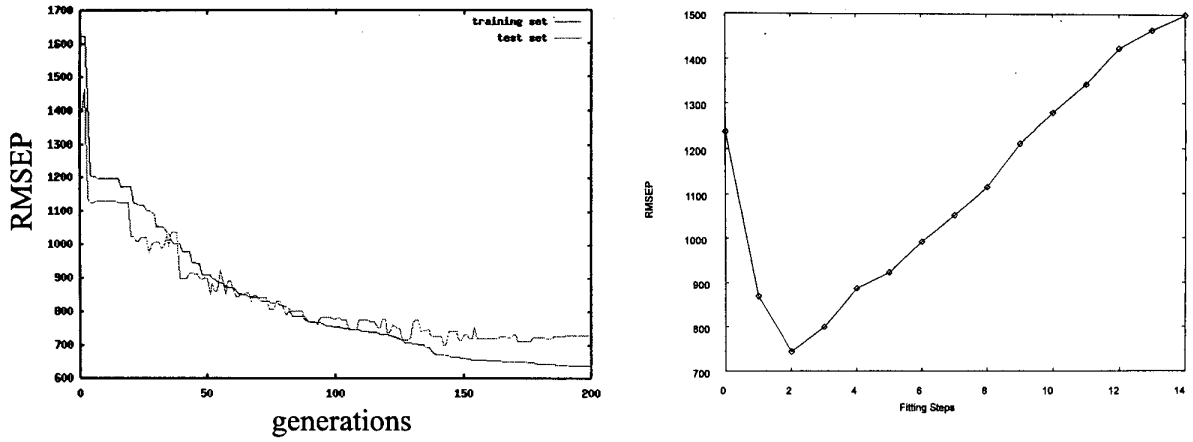

Fig. 4. *On the left, training and test set learning curves from the GA modelling process, illustrating the difficulty of identifying the onset of overtraining, compared with the clear-cut minimum in the test set curve of RMSEP vs number of factors produced by PLS (on the right).*

very clearly the benefits of the GA-based selection and denoising, in comparison with this naive and non-adaptive approach, in reducing the prediction error by an amount approaching 50% of that value.

The results demonstrate that a GP with an appropriate 'window averaging' node function and, given its inherent variable selection capability, would be capable of performing the entire process in one step. This is advantageous from the point of view of extending the capability of a tool for laboratory use that is relatively easy to use.

Each window midpoint can be expressed as a wavenumber using the formula

$$n = 4000 - (3.85 * (x + 1)) \quad (1)$$

where n is the wavenumber, x is the variable number, 3.85 is the spectral resolution and 4000 is the starting wavenumber of the IR scan. The spectral regions selected can thus be related back to the original IR spectra. A broad region centred on variable number 580 translates to a region of the spectrum centred at 1767 cm^{-1} and windows in this region were selected repeatedly by both the GA and the GP. It is known that vibrations in this region of the IR spectrum can be attributed to carboxylate groups, and GA3 contains two such groups.

The GP results on the denoised and variable-selected subset are marginally inferior to those obtained with PLS on the same subset. However, this could at least in part be attributable to the difficulty of determining the appropriate termination point in chemometric applications of evolutionary computing; the point at which the training and test set learning curves diverge is often difficult to judge correctly, whereas in PLS there is normally a much more easily identifiable point. (See figure 4). Although the final RMSEP achieved is still significantly higher than the value of 425 achieved by PLS (for example) on the 'clean' data that results from 256 co-adds (see the left hand plot in figure 2), the results obtained indicate that far fewer co-adds would be required to achieve prediction accuracy comparable to that on the 'clean' data.

VIII. CONCLUSIONS

We have shown that evolutionary search can provide effective adaptive denoising of infra red spectra and that the resultant denoised subset of variables when input to a proprietary GP can provide predictive models that compare with results from PLS applied to the same subset. However, since the GP inherently performs variable selection, the addition of node functions analogous to those implemented in the separate GA described above could provide a one-step tool for forming predictive models on the basis of rather fewer co-adds than is normally the case. More significantly, it would provide good predictions from new spectra with a higher noise content than would normally be the case; this would reduce significantly the data acquisition time in time-critical spectroscopic analyses such as those used in high throughput screening.

References

[1] P.R. Griffiths and J.A. de Haseth. *Fourier transform infrared spectrometry.* John Wiley, New York, 1986.

[2] J.R. Koza. *Genetic programming: on the programming of computers by means of natural selection.* MIT Press, Cambridge, Mass, 1992.

[3] B. F. J. Manly. *Multivariate Statistical Methods : A Primer.* Chapman and Hall, London, 1994.

[4] H. Martens and T. Naes. *Multivariate calibration.* John Wiley, Chichester, 1989.

[5] M. K. Winson, R. Goodacre, É. Timmins, A. Jones, B. K. Alsberg, A. M. Woodward, J. J. Rowland, and D. B. Kell. Diffuse Reflectance Absorbance Spectroscopy Taking In Chemometrics (DRASTIC). a hyperspectral FT-IR-based approach to rapid screening for metabolite overproduction. *Anal. Chim. Acta*, 348:273 – 282, 1997.

[6] C.M. Bishop. *Neural Networks in Pattern Recognition.* Oxford University Press, Oxford, U.K., 1995.

[7] A.D. Shaw, M.K. Winson, A.M. Woodward, A. McGovern, H.M. Davey, N. Kaderbhai, D.I. Broadhurst, R.J. Gilbert, J. Taylor, Timmins E.M., B.K. Alsberg, J.J. Rowland, R. Goodacre, and D.B. Kell. Rapid analysis of high-dimensional bioprocesses using multivariate spectroscopies and advanced chemometrics. *Adv. Biochem. Eng.*, 66:83–113, 2000.

[8] C. B. Lucasius, M. L. M. Beckers, and G. Kateman. Genetic algorithms in wavelength selection - a comparative-study. *Analytica Chimica Acta*, 286(2):135–153, 1994.

[9] D. Jouan-Rimbaud, D. L. Massart, R. Leardi, and O. E. De Noord. Genetic algorithms as a tool for wavelength selection in multivariate calibration. *Analytical Chemistry*, 67(23):4295–4301, 1995.

[10] A. S. Bangalore, R. E. Shaffer, and G. W. Small. Genetic algorithm based method for selecting wavelengths and model size for use with partial least squares regression: Application to near-infrared spectroscopy. *Analytical Chemistry*, 68(23):4200 – 4212, 1996.

[11] D. Broadhurst, R. Goodacre, A. Jones, J.J. Rowland, and D.B. Kell. 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*, 348:71–86, 1997.

[12] M. Mosley and R. Williams. Determination of the accuracy and efficiency of genetic regression. *Applied Spectroscopy*, 52(9):1197–1202, 1998.

[13] R.P. Paradkar and R.R. Williams. Correcting fluctuating baselines and spectral overlap with genetic regression. *Applied Spectroscopy*, 51(1):92–100, 1997.

[14] J. Taylor, J.J. Rowland, R. Goodacre, R.J. Gilbert, M.K. Winson, and D.B. Kell. Genetic programming in the interpretation of fourier transform infrared spectra: quantification of metabolites of pharmaceutical importance. In J.R. Koza, W. Banzhaf, K. Chellapilla, K. Deb, M. Dorigo, D.B. Fogel, M.H. Garzon, D.E. Goldberg, H. Iba, and R.L. Riolo, editors, *Genetic Programming 1998: Proceedings of the Third Annual Conference,*, pages 377–380. Morgan Kaufmann, San Francisco, 1998.

[15] P.J. Rauss, J.M. Daida, and S. Chaudhary. Classification of spectral imagery using genetic programming. In Darrell Whitley, David Goldberg, Erick Cantu-Paz, Lee Spector, Ian Parmee, and Hans-Georg Beyer, editors, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000)*, pages 726–733, Las Vegas, Nevada, USA, 10-12 July 2000. Morgan Kaufmann.

[16] H.E. Johnson, R.J. Gilbert, M.K. Winson, R. Goodacre, A.R. Smith, J.J. Rowland, M.A. Hall, and D.B. Kell. Explanatory analysis of the metabolome using genetic programming of simple interpretable rules. *Genetic Programming and Evolvable Machines*, 1(3):243–258, 2000.

[17] Douglas B. Kell, Robert M. Darby, and John Draper. Genomic computing. explanatory analysis of plant expression profiling data using machine learning. *Plant Physiology*, 126:1–9, 2001.

[18] D. M. Allen. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3):469–475, 1971.

[19] R. E. Shaffer, G. W. Small, and M. A. Arnold. Genetic algorithm-based protocol for coupling digital filtering and partial least-squares regression: Application to the near-infrared analysis of glucose in biological matrices. *Analytical Chemistry*, 68(15):2663–2675, 1996.

[20] B. M. Smith and P. J. Gemperline. Wavelength selection and optimization of pattern recognition methods using the genetic algorithm. *Analytica Chimica Acta*, 423(2):167–177, 2000.

[21] B.K. Alsberg, A.M. Woodward, M.K.Winson, J.J. Rowland, and D.B. Kell. Wavelet denoising of infrared spectra. *Analyst*, 122(7):645 – 652, 1997.

[22] A.C. McGovern, D. Broadhurst, J. Taylor, R.J. Gilbert, N. Kaderbhai, M.K. Winson, D.A. Small, J.J. Rowland, D.B. Kell, and R. Goodacre. Monitoring of complex industrial bioprocesses for metabolite concentrations using modern spectroscopies and machine learning: application to gibberellic acid production. *Biotechnology and Bioengineering.*, 2002. In Press.

[23] Janet Taylor, Jem J. Rowland, and Douglas B. Kell. Spectral analysis via supervised genetic search with application-specific mutations. In *IEEE Congress on Evolutionary Computation (CEC)*, volume 1, pages 481–486, Seoul, Korea, 2001. IEEE.