Genetic Programming in the Interpretation of Fourier Transform Infrared Spectra: Quantification of Metabolites of Pharmaceutical Importance

†Janet Taylor
Jem J Rowland
Dept. Computer Science
University of Wales,
Aberystwyth SY23 3DB
United Kingdom
†jjt95@aber.ac.uk
jjr@aber.ac.uk

Royston Goodacre
Richard J Gilbert
Biological Sciences
University of Wales,
Aberystwyth SY23 3DD
United Kingdom
rrg@aber.ac.uk
rcg@aber.ac.uk

Michael K Winson
Douglas B Kell
Biological Sciences
University of Wales,
Aberystwyth SY23 3DD
United Kingdom
mkw@aber.ac.uk
dbk@aber.ac.uk

ABSTRACT

This paper describes the use of genetic programming (GP) to evolve a concise and explicit relationship between a complex and noisy set of infrared spectra taken from biological samples (*Escherichia coli*) and the concentrations of a specific antibiotic (ampicillin) in the samples. The work relates to an investigation of the use of diffuse reflectance absorbance infrared spectroscopy as a novel method of rapid screening for metabolite overproduction in the context of process improvement in the pharmaceutical industry. The results show that models generated by genetic programming are comparable, in terms of the accuracy of prediction, with those produced by a 'standard' multivariate calibration method, Partial Least Squares (PLS). However, the GP models also show tolerance to spectral noise and facilitate a consequent reduction in the need for data pre-processing. Furthermore, the output expressions produced readily allow the identification of significant spectral variables, or wavenumbers, and thus directly aid interpretation of the original spectra in molecular terms.

1. Introduction

Genetic programming, the automated generation of computer programs *via* evolution (Koza, 1992; Bäck *et al.*, 1997), has been applied in various domains, such as robot planning (Handley, 1993), engineering design (McKay *et al.*, 1996) and pattern recognition (Koza, 1994).

One area of particular interest is the optimisation and empirical discovery of relationships in data, such as the optimisation of parameters for digital signal processing (Sharman *et al.*, 1995) and chemical process control (McKay *et al.*, 1996). The ability to take multiple input values into a program makes GP an ideal candidate for multivariate analysis, where many measured (x) variables bear a relationship to one property (y variable). Another benefit of GP is that the output of a run is a procedural program whose instructions can elucidate the variables used, and their manipulation. These features of the GP paradigm, being a supervised learning method, made it an ideal candidate to apply to datasets where different properties of samples are related to biological and biochemical features of interest. PLS and ANNs (Goodacre *et al.*, 1996; Gemperline *et al.*, 1991; Timmins *et al.*, 1997) are current 'standard' methods that have been applied to such data for generating predictive models for qualitative and quantitative analysis.

There is also an interest in identifying the important variables selected in model formation. Various algorithms for variable selection (Eshuis *et al.*, 1977; George and McCulloch, 1993; Kubinyi, 1994a; b) have been applied with varying success. However, many documented methods rely on assumptions concerning the data, such as linearity, normal distribution, or absence of colinearity between variables. Consequently these methods are unreliable for some datasets (Goodacre *et al.*, 1996).

2. Data Acquisition

Three datasets were acquired for analysis. For dataset A the bacterial samples were prepared as in (Winson *et al.*, 1997) with antibiotic added in the concentration range 0 to 20 mM. Each sample was represented by a spectrum containing 882 data points or variables.

For datasets B and C, the bacterium used was a similar strain as above, grown under identical conditions. The

antibiotic was added over a different concentration range (0 to 5000 μg/ml) as detailed in (Goodacre *et al.*, 1995).

3. Data Preprocessing

The spectra for all datasets were exported from the FT-IR spectrometer and converted to ASCII format.

The spectra from the samples for set A were averaged, resulting in a 40 by 882 data matrix. Two subsets of these data were taken, the first containing 25 variables from the characteristic ampicillin peak and the second consisting of 200 variables containing the infrared fingerprint region. The original dataset and both subsets were split into training and test sets, each of 20 data objects, using in-house software based on the Duplex algorithm (Snee, 1977).

For datasets B and C, 21 samples were analysed in triplicate resulting in a 63 by 882 data matrix. These data were again split into a training set containing 11 triplicate spectra (33 data objects) and a test set containing 10 triplicate samples (30 data objects) by the same Duplex (Snee, 1977) based software. A copy of dataset B was made prior to its separation. The replicate copy was subjected to baseline correction via a genetic algorithm (GA) package, written in C and run on a Pentium Pro 200 IBM compatible PC running NT 4.0. The variance of the variables in all spectra were calculated and the GA optimised 8 parameters, such that areas with high variance were maximised and areas of low variance were minimised by adjusting vertical shift and scale parameters.

a: **b:**

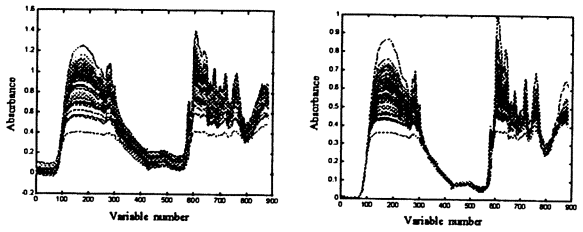


Figure 1: FT-IR spectra of Ampicillin in *E.coli*.
a. Raw averaged spectra; b: Processed spectra.

Curve fitting to correct for baseline trends was also carried out by the optimisation of a range of values of x in the equation $ax^3 + bx^2 + cx + d$. The fitness function for the GA was the minimisation of 10% of the spectral regions with the lowest variance, and maximisation of 50% of the regions with the highest variance. The result of this correction may be seen in Figure 1.

4. Genetic Programming

The genetic programming implementation used in this study is similar to that described by Gilbert *et al.*, (1997). The breeding strategy implemented the following genetic operators: crossover of two individuals at a proportion of 80% of the total population, reproduction of one individual at a proportion of 10% of total population, and random node mutation at a proportion of 10% of total population. The package was written in ANSI C following a procedure similar to Singleton (1994) and run on a 486DX66 PC under NT 4.0. The demes were implemented as follows: five populations were evolved in parallel, one 'main' and four 'satellite' populations. Every 10 generations the best individuals from the main population were copied into the satellite populations, and replaced with the worst individual from each of the satellites, to provide a migration effect which aided population diversity, thus preventing premature convergence to a local minimum. Each population comprised 500 individuals, each of which had a depth of 17 and 100 nodes as development limits. The fitness function for all populations was the raw root mean squared error of prediction (RMSEP) of the individual calculated via:

$$RMSEP = \sqrt{\frac{\sum (y - \hat{y})^2}{n}}$$ where y = measured output,

$\hat{y}$ = predicted output and n = number of examples.

Termination of the GP run was at 250 generation cycles. The most fit individual from all populations was deemed the solution at this point. Two function sets were considered for use with the data, an arithmetic set consisting of {add, subtract, multiply, protected divide} and a functional set consisting of the arithmetic set and additionally of {sin, cos, tan, log}.

5. PLS

Partial least squares regression is a supervised latent-variable multivariate regression technique (Hoskuldsson, 1988; Martens and Næs 1989). PLS was implemented using in house calibration software (Jones *et al.*, 1998).

6. Results and Discussion

Initial experiments were conducted to determine the potential benefits of reducing each dataset so as to contain only the 25 variables that form the spectral peak around 1767cm^{-1} that is a characteristic of ampicillin, or the 200 variables that hold the majority of ampicillin related information. The GP described previously (Gilbert *et al.*, 1997) was used to analyse dataset A and its two subsets. This experiment was also used to determine the required components of the GP function set. Two function sets were used, one with simple arithmetic operators, and the other containing these plus log, tan, cos, sin. The results show the test RMS error for the smaller dataset was significantly higher (see table 1), suggesting insufficient information to model the relationship accurately. There was no significant difference between using the 200 variable subset and the full dataset, and as the time taken to generate a model was comparable (data not shown) the full dataset was used in subsequent experiments. As can be seen in table 1, the data are adequately modelled using the arithmetic operators.

The raw RMS error for the GP solutions using the functional operators sometimes show a larger error, suggesting that the increase in complexity can inhibit prediction ability.

Table 1: Raw RMS errors (mM) for 3 runs each of the 25 and 200 variable subsets and the full spectrum GP analysis. In each case the most fit individual is shown.

no. Variables		Arithmetic Set	
	1	2	3
25	9.4	9.4	9.6
200	0.5	0.7	1.6
882	0.9	0.5	0.5
		Function Set	
	1	2	3
25	10.1	10	10
200	0.7	0.6	0.8
882	2.6	1.2	2.1

The following experiments therefore used the full spectrum as input to the GP, with the arithmetic function set. Each output expression was converted from Polish notation to standard form via an in-house Perl program (run on a 500MHz DEC Alpha Workstation under Digital Unix 4.0), then simplified using Maple (Waterloo Maple Inc., 450 Phillip Street, Waterloo, Ontario Canada N2L 5J2) run on a DEC Alpha 3000/700 workstation.

Each dataset was also analysed using PLS regression noting the optimum number of factors to form a predictive model. The use of neural networks to analyse these data has been documented in (Winson *et al.*, 1997); results are comparable with PLS regression methods and therefore only the PLS results are presented here.

The results in table 2 show the test set RMS errors of prediction for the GP and for PLS, one of the current 'standard' methods of spectroscopic analysis (Baroni and Clementi, 1992; Fuller *et al.* 1988).

Table 2: Comparison of PLS and Genetic Programming (best individual at 250 generation cycles). Raw RMS errors (A - mM, B and C -μg/ml) shown.

Method		Dataset	
	A	B	C
PLS	0.79	149.8	135.6
GP	0.52	148.6	117.9

As seen, using a GP shows an improvement in the accuracy of the models of all datasets compared with PLS. Test predictions for each method are shown in Figure 3. Detailed examination of the simplified output expressions from each GP model clearly identifies the variables used in model formation (a variable is designated as P*x*, where *x* is the bin number in the dataset). The bin number is directly related to a wavenumber in the original spectrum via 4000 - (3.85 * *x*) cm^{-1} where *x* is the bin number. 4000 cm^{-1} is one extremity of the spectral range, and there are 3.85 cm^{-1} per bin. A variable listed in bold typeface is located in the characteristic peak due to the β-lactam of ampicillin (centred at 1767 cm^{-1}):

Dataset A:

$$2.9\frac{\left(5.8-5.8P117-3.8+21.8\frac{P579}{P598}\right)P579}{P598}+100916.1\left(\frac{P579^6\,P844P310P16}{P598P19}\right)$$

Variables selected: 16, 19, 117, 310, **579**, **598**, 844

Wavenumbers(cm^{-1}): 3938.4, 3926.9, 3549.6, 2806.5, **1770.9**, **1697.7**, 750.6.

Dataset B:

$$\left(\frac{(P617+15.7P577)P70}{P564P418}+4.1\right)\left(-15.7\frac{P577P654}{P524}+9.2\right)$$
$$\left(\frac{-5.2+P815}{P524+P355}+P367-0.6\left(\frac{7.8-(4.2/577)-P831}{P577}\right)-4.3-\left(\frac{19.3-(9.3/P577)+\frac{(P201-57.9P577)P70}{P654P418}}{2P831-P685}\right)\right.$$
$$\left.*P103\frac{P524}{P654}\right)$$

Variables selected: 70, 103, 355, 367, 418, 524, **577**, 617, 654, 685, 815, 831

Wavenumbers(cm^{-1}): 3730.5, 3603.5, 2633.25, 2587.1, 2390.7, 1982.6, **1778.6**, 1624.6, 1482.1, 1362.8, 862.3, 800.7

Dataset C:

$$483\frac{(1.7/P8-5.3)(P815-4.8)P58}{P817P158}+16.7P282$$
$$\left(\frac{(P579-P104)P163}{P579^2}+75.8\frac{P583^2}{P260^2P457}-3.96\frac{P123P741P158^2}{(P579-4.8)P58}-0.57\right.$$
$$\left.\frac{(P579-P104)\left(-8.7\frac{P583}{P260P457}-8.7P614\right)}{P86}+P583/P260\right)$$

Variables selected: 8, 58, 86, 104, 123, 158, 163, 187, 260, 282, 457, **579**, **583**, 614, 741, 815

Wavenumbers(cm^{-1}): 3969.2, 3776.7, 3668.9, 3599.6, 3526.5, 3391.7, 3372.5, 3280.1, 2999.0, 2914.3, 2240.6, **1770.9**, **1755.5**, 1636.1, 1147.2, 862.3

Although these expressions are quite complex, it can be seen that the variables are taken from differing regions of the spectra, confirming that there is information in more regions than just the characteristic peak. From an initial input dataset of 882 variables, GP has reduced the dimensionality of each dataset to a number of variables comparable with the optimal number of PLS factors (7 - 16)

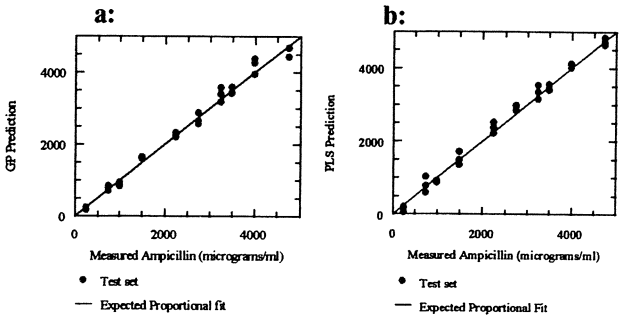


Figure 3: Example test set prediction plots for GP and PLS models for dataset C.

7. Conclusions

Genetic Programming gave results comparable with those from PLS regression in the analysis of biological spectra. By identifying significant variables it also gives insight into the relationship between a spectrum and the analyte of interest. This study has led to further work aimed at constraining the output expression so as to define more clearly the predictive relationship.

Acknowledgements

We are indebted to Dr A. Edmonds of Science in Finance Ltd, BBSRC, EPSRC, HEFCW, GlaxoWellcome and the Wellcome Trust (grant no. 042615/Z/94/Z).

Bibliography

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

Baroni, M., et al., Predictive Ability of Regression Models. Part II: Selection of the Best Predictive PLS Model, *Journal of Chemometrics* **6**, 347 - 356 (1992).

Eshuis, W., Kistemaker, P. G., Meuzelaar, H. L. C., in *Analytical Pyrolysis* C. E. R. Jones, C. A. Cramers, Eds. (Elsevier, Amsterdam, 1977) pp. 151-156.

Fuller, M. P., Ritter, G. L., Draper, C. S., Partial Least Squares Quantitative Analysis of Infrared Spectroscopic Data. Part 1: Algorithm Implementation, *Applied Spectroscopy* **42**, 217 - 227 (1988).

Gemperline, P. J. *et al.*, Non linear multivariate calibration using principal components regression and artificial neural networks, *Analytical Chemistry* **63**, 2313-2323 (1991).

George, E. I., McCulloch, R. E., Variable Selection Via Gibbs Sampling, *Journal of the American Statistical Association* **88**, 881 - 889 (1993).

Gilbert, R. J. *et al.*, Genetic Programming: a Novel Method for the Quantitative Analysis of Pyrolysis Mass Spectral Data, *Analytical Chemistry* **69**, 4381 - 4389 (1997).

Goodacre, R., et al., Identification and Discrimination of Oral Assacharolytic Eubacterium spp. by Pyrolysis Mass Spectroscopy and Artificial Neural Networks., *Current Microbiology* **32**, 77-84 (1996).

Goodacre, R. *et al.*, Rapid identification of *Streptococcus* and *Enterococcus* species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks, *FEMS Microbiology Letters* **140**, 233-239 (1996).

Goodacre, R., et al., 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, *Analytica Chimica Acta* **313**, 25-43 (1995).

Handley, S., The Genetic Planner: The automatic generation of plans for a mobile robot via genetic programming, in 1993 International Symposium on Intelligent Control, Illinois (IEEE Press, 1993).

Hoskuldsson, A., PLS Regression Methods, *Journal of Chemometrics* **2**, 211 - 228 (1988).

Jones, A. *et al.*, Quantification of Microbial Productivity via Multi-Angle Light Scattering and Supervised Learning, *Biotechnology and Bioengineering* **in Press** (1998).

Koza, J. R., Automated discovery of detectors and iteration-performing calculations to recognise patterns in protein sequences using genetic programming, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (IEEE Press, 1994).

Koza, J. R., *Genetic Programming: On the Programming of computers by Means of Natural Selection* (MIT Press, Cambridge, MA, 1992).

Kubinyi, H., Variable Selection in QSAR Studies II. A Highly Efficient Combination of Systematic Search and Evolution, *Quantitative Structure Activity Relationships* **13**, 393 - 401 (1994).

Kubinyi, H., Variable Selection in QSAR Studies. I: An Evolutionary Algorithm, *Quantitative Structure Activity Relationships* **13**, 285 - 294 (1994).

Martens, H., Næs, T., *Multivariate calibration* (John Wiley, Chichester, 1989).

McKay, B., Elsey, J., Willis, M. J., Barton, G. W. Evolving input output models of chemical process systems using genetic programming, in IFAC '96, San Fransisco, USA (1996).

McKay, B., Lennox, B., Willis, M., Barton, G. W., Montague, G. Extruder modelling: A comparison of two paradigms, *http://lorien.ncl.ac.uk/sorg/* (1996).

Sharman, K. C., Esparcia, A. I., Li, Y., Evolving digital signal processing algorithms by genetic programming, in First IEE/IEEE International Conference on Genetic Algorithms in Engineering Systems: Innovations and Applications, Sheffield, UK (1995).

Singleton, A., Genetic Programming with C++ *Byte* **19**, 171 (1994).

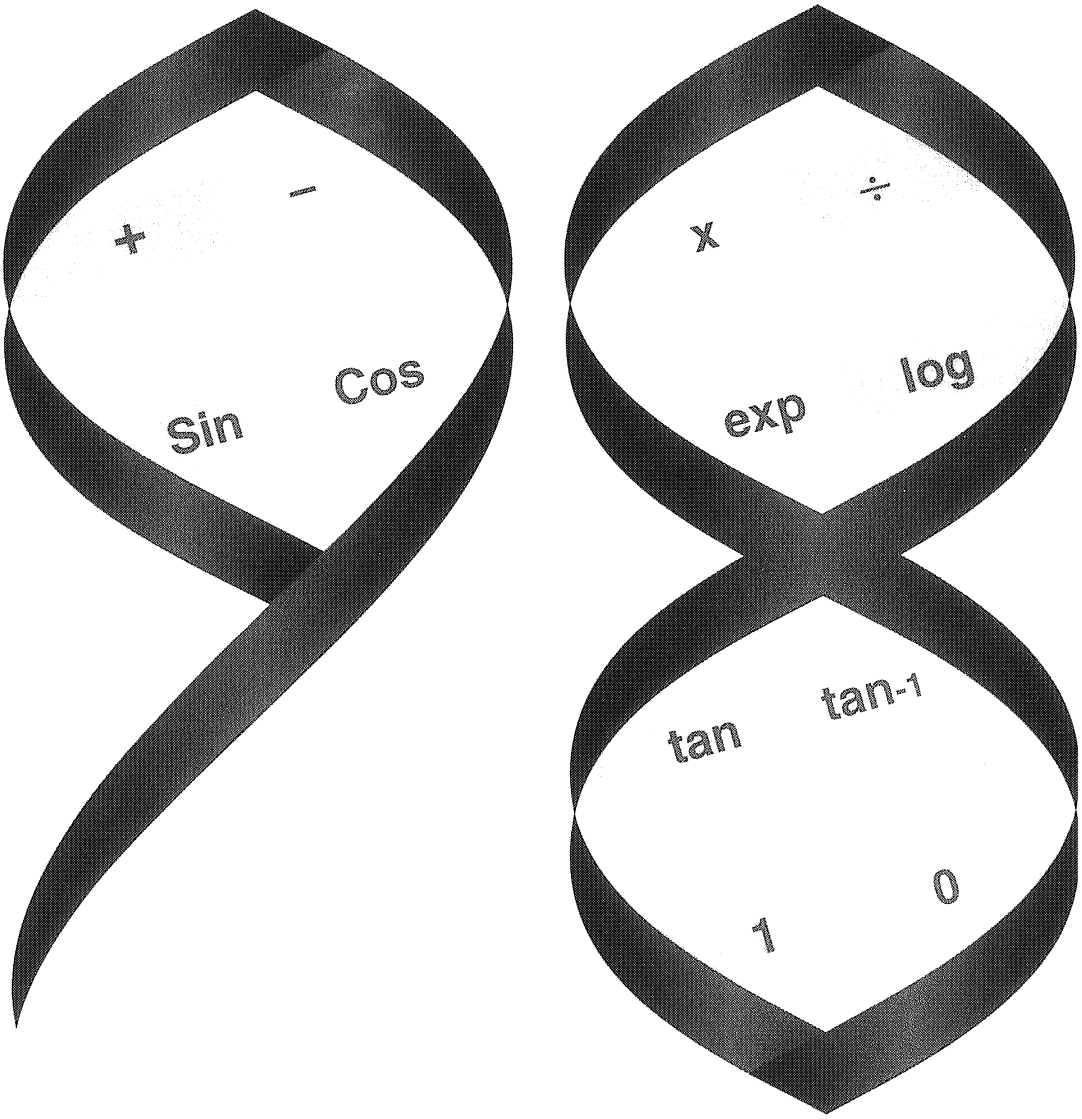
Snee, R. D., Validation of Regression Models: Methods and Examples., *Technometrics* **19**, 415-428 (1977).

Timmins, É. M., Goodacre, R., Rapid quantitative analysis of binary mixtures of *Escherichia coli* strains using pyrolysis mass spectrometry with multivariate calibration and artificial neural networks, *Journal of Applied Microbiology* **83**, 208 - 218 (1997).

Winson, M. K., et al., Diffuse reflectance absorbance spectroscopy taking in chemometrics (DRASTIC) A hyperspectral FT-IR based approach to rapid screening for metabolite overproduction, *Analytica Chema Acta* **348**, 273- 282 (1997).

CONFERENCE PROCEEDINGS

Genetic Programming



edited by
John R. Koza
Wolfgang Banzhaf
Kumar Chellapilla
Kalyanmoy Deb
Marco Dorigo
David B. Fogel
Max H. Garzon
David E. Goldberg
Hitoshi Iba
Rick L. Riolo

Proceedings of the Third Annual Genetic Programming Conference

July 22–25, 1998
University of Wisconsin, Madison, Wisconsin