Genetic Programming-Based Variable Selection for High-Dimensional Data

Richard J. Gilbert[†]
Royston Goodacre
Beverly Shann
Douglas B. Kell

Institute of Biological Sciences
University of Wales
Aberystwyth
Ceredigion SY23 3DD
UNITED KINGDOM

Janet Taylor
Jem J. Rowland

Department of Computer Science
University of Wales
Aberystwyth
Ceredigion SY23 3DB
UNITED KINGDOM

[†]rcg@aber.ac.uk, +44 (0)1970 622353
http://gepasi.dbs.aber.ac.uk/rcg

ABSTRACT

A major advantage of the genetic programming [GP] approach to data modeling is the automatic ability of the GP to select input variables that contribute beneficially to the model and to disregard those that do not. GPs are thus able to reduce substantially the dimensionality of the model, with consequent interpretation benefits.

Experimental analytical techniques frequently generate data with very high dimensionality, typically measuring many tens or even hundreds of variables per sample. It is often not apparent which of the measured variables can best be used to derive a predictive model describing the data. The identification of these variables often provides a better understanding of the physical, chemical or biological mechanism underlying the experimental observations.

The ability of a GP to perform variable selection is assessed with regard to a binary classification of the sporulation state of bacterial strains. The analytical technique used, Curie-point pyrolysis mass spectrometry, generates data for 150 variables per sample. The GP-derived predictive rules for these data contain a substantially smaller subset of these variables, typically just 6-9. Inspection of these rules leads to the somewhat counter-intuitive conclusion that the best predictive models use both highly characteristic and highly *non*-characteristic variables.

1. Introduction

Cells from the genus *Bacillus*, which are rod-shaped, Gram-positive bacteria, respond to slowed growth or starvation by initiating the process of *sporulation*. In doing this, the bacteria undergo cellular differentiation to form resting bodies, called spores, which are morphologically and biochemically distinct from the normal (*vegetative*) bacterial cell type. Spores are highly resistant to adverse conditions such as dehydration, extremes of temperature and low nutrient availablility, and so sporulation provides the bacteria with a survival mechanism. After a period of time the spores may germinate, each producing a single vegetative cell which is then able to grow and divide in the normal way. Members of the genus *Bacillus* are widely distributed in soil, water and air and therefore, because of the resiliance of their spores, an understanding of the mechanisms underlying sporulation is of considerable importance for the preparation of sterile products, particularly in the medical apparatus and food processing industries.

A collection of 36 *Bacillus* strains (spanning 7 species) were used in this study. The strains were grown under both vegetative and spore-inducing conditions, after which their biomasses (harvested cellular material) were analysed by Curie-point pyrolysis mass spectrometry [PyMS][1-3].

PyMS is a high-resolution experimental technique which, in combination with supervised learning techniques such as artificial neural networks [ANNs] and partial least squares [PLS], has been shown to enable the derivation of accurate and precise models for both the qualitative and quantitative analysis of complex biological samples[4-12]. Only recently have the first applications of GP-based analytical modeling to data of this type been described, by ourselves[13-15].

To analyse a sample using PyMS, a small sample is dried onto an iron-nickel foil. The foil is heated rapidly by magnetic induction *in vacuo* until its magnetic properties undergo a phase change which prevents any further temperature rise. The exact temperature at which this occurs (the Curie-point temperature) depends on the alloy composition of the foil. The foils are held at the Curie-point temperature for a few seconds, during which volatile molecules are released from the sample and non-volatile molecules are thermally degraded into volatile breakdown products (*pyrolysis*). The volatilised molecules (*pyrolysate*) are charged electrostatically and passed into a quadrupole mass spectrometer where their relative abundances are measured, typically over a mass:charge (*m/z*) range from 51 to 200.

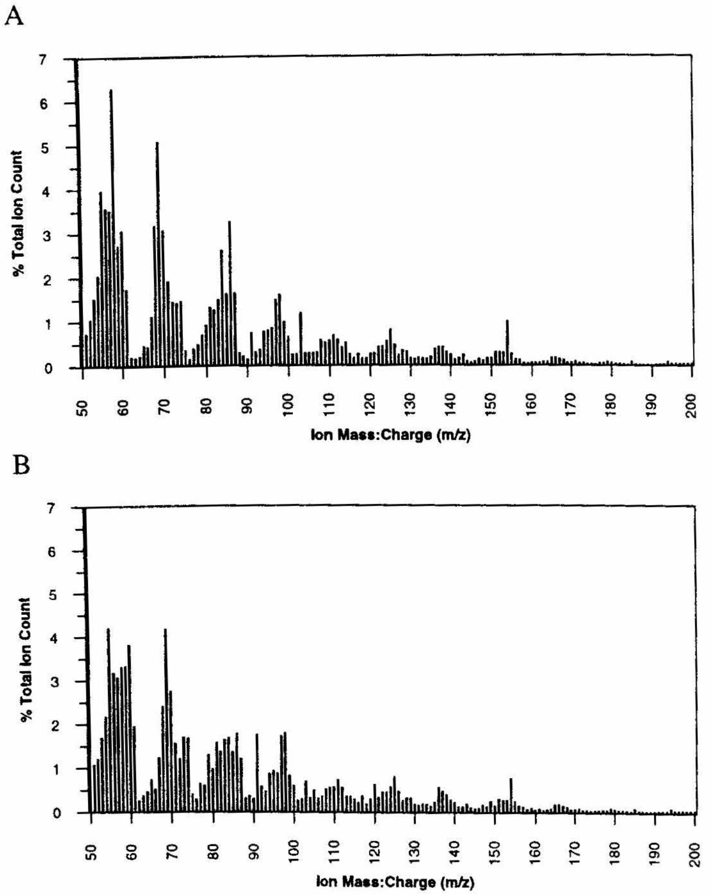

Figure 1 Typical PyMS spectra for a single *Bacillus* strain grown under either vegetative (A) or sporulating (B) conditions.

The PyMS spectra were too complex to interpret visually (Figure 1), and so a GP system was used to derive rules for the classification of the sporulation state of the samples. By constraining the complexity of these rules using a function tree node-count penalty in the GP fitness function, it was possible to reduce substantially the dimensionality (*i.e.* number of variables) of the data used by the models, thereby enabling their interpretation in a chemically meaningful way.

2. Materials and Methods

2.1. Cultivation of Bacterial Samples

The 36 *Bacillus* strains used in this study have previously been determined, using a combination of conventional biochemical tests and nucleic acid sequencing technologies, to belong to the species *B. amyloliquefaciens* (five strains), *B. cereus* (five strains), *B. licheniformis* (five strains), *B. megaterium* (five strains), *B. subtilis* (seven strains, including two *B. niger* and one *B. globigii*), *B. sphaericus* (five strains) and *B. laterosporus* (four strains). The collection therefore contained examples of seven distinct species of *Bacillus*.

Vegetative cells were obtained by incubating the 36 bacterial strains on Lab M blood agar base plates (without blood) at 37°C for 10 hours.

Spores were prepared by incubating the strains on Lab M blood agar base plates, with the addition of 5mg.l^{-1} $MnSO_4$, at 30°C for 7 days.

After incubation, the bacterial cells were harvested and stored in suspension in physiological saline (0.9% NaCl) at -20°C until required for analysis.

2.2. Pyrolysis Mass Spectrometry

5 μl aliquots (samples) of the bacterial cell suspensions were evenly applied to iron-nickel foils (50:50 Fe:Ni) to give a thin, uniform surface coating. Prior to pyrolysis, the samples were oven-dried at 50°C for 30 min. The sample tube carrying the foil was heated prior to pyrolysis at 100°C for 5 sec. Curie-point pyrolysis was at 530°C for 3 sec, with a temperature rise-time of 0.5 sec. The PyMS spectra were collected over an *m/z* range from 51 to 200. For full operational procedures see Goodacre *et al.* (1993, 1994, 1995).

The 36 strains were analysed in triplicate, having been grown under both sporulated and vegetative conditions, to provide a total data set comprising 216 spectra, each containing 150 data points.

2.3. Genetic Programming

A genetic algorithm (GA) is an optimisation method based on the principles of Darwinian selection [18-20]. A population of individuals, each representing the parameters of the problem to be optimised as a string of numbers or binary digits, undergoes a process analogous to evolution in order to derive an optimal or near-optimal solution. The parameters stored by each individual are used to assign it a *fitness*, a single numerical value indicating how well the solution using that set of parameters performs. New individuals are generated from members of the current population by processes analogous to biological asexual and sexual reproduction.

Asexual reproduction, or *mutation*, is performed by randomly selecting a parent with a probability related to its fitness, then randomly changing one or more of the parameters it encodes. The new individual then replaces a less-fit member of the population, if one exists. Sexual reproduction, or *crossover*, is achieved by randomly selecting two parents at a rate related to their fitnesses, and generating two new individuals by copying parameters from one parent, and switching to the other parent after a randomly-selected point. The two new individuals then replace less fit members of the population as before. The above procedure is repeated, with the overall fitness of the population improving at each generation, until an acceptably-fit individual is produced.

A genetic program (GP) is an application of the GA approach to derive mathematical equations, logical rules or program functions automatically [21,22]. 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 *terminal*s, 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 towards 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 sub-trees at randomly-selected nodes within their trees. The new individuals so generated again replace less-fit members of the population.

The GP system used in this study was implemented in *C* following a procedure similar to Singleton (1994). For the results presented here, the GP used four types of operator node (add, subtract, multiply and protected divide), and two types of terminal (a floating-point ephemeral random constant type and an input variable type representing the data for a specific *m/z* spectral peak). The configuration of the GP was defined using mainly arbitrary parameters. Five sub-populations (*demes*), each comprising 500 individuals, were used. At each generation, and for each deme, 50 new individuals were created by mutation and 50 by crossover, which then replaced less-fit members of the population using a ranked (*i.e.* sorted by decreasing fitness) selection method. The five demes were allowed to evolve independently for 10 generations, after which all 2500 individuals were sorted according to their fitness scores and the best 100 individuals from the total pooled population were used to replace the worst 100 in each of the five sub-populations. This process was repeated at 10-generation intervals throughout the GP run.

In order to conduct a GP-based supervised learning regime on these data it was first necessary to partition the samples into training and test sets. The training set consisted of the triplicate results for two representatives from each of the seven *Bacillus* species grown under both vegetative and sporulating conditions: a total of 84 PyMS spectra, and the test set comprised the remaining 132 spectra.

The fitness function used in the GP returned the RMS (root mean square) error for the output expression as compared with the known class (encoded as 0.0 for vegetative and 1.0 for sporulated samples) for the training set samples. An optimal GP rule would threfore return a numerical value of 0.0 when presented with a vegetative sample, and 1.0 when presented with a sporulated one.

In addition to a node count limit of 64, a penalty of 0.01 × [the number of nodes in the tree] was added to the fitness to reduce the complexity of the GP-derived rules. The GP optimised the rules by minimising the fitness function for the training set examples. The run terminated when the fitness value of the rule reached less than 0.01 or when the rule was able to classify correctly all 132 members of the test set.

3. Results and Discussion

A GP approach was used to model the data from a PyMS analysis for a group of 36 *Bacillus*, spanning seven species. In order to compare several GP-derived predictive models, multiple GP runs were performed using a dataset comprising PyMS spectra for 216 examples, representing triplicate analyses of the 36 *Bacillus* strains grown under both vegetative and sporulating conditions.

The GP-derived rules all differ in their detailed mathematical forms, but many share several features in common. This suggests that there are mathematical relationships between certain input variables and the desired output which different GP runs are consistently finding. For example, most of the rules contain a simple linear relationship between the desired output and m_{105}, m_{64} and, albeit with a lower frequency, with m_{76}. It is probable that frequently-occurring mathematical relationships in the rules indicate an equivalent relationship in the actual physico-chemical or biological processes underlying the experimental observations.

Variable selection is a procedure which substantially reduces the size of the search space for problems of extremely high dimensionality by selecting a small subset of variables from those available. By doing this, the dimensionality of the problem may be reduced substantially, with consequent benefits for the rapid derivation of readily-interpretable predictive models, but there is the potential penalty of losing any useful information provided by the variables which are being discarded. The task of selecting a small subset of variables which provide a maximally-useful amount of information for the derivation of predictive models whilst minimising the dimensionality of the problem search space is a common one for many practical experimental methods capable of recording data for a large number of variables.

Table 1 The rules and predictive accuracies from 10 example GP runs.

Run	GP Rule	Number correctly classified (out of 132)
1	$output = m_{105} + m_{89} + 2.40m_{76} - 0.43 + \frac{2.46m_{152}m_{76}}{m_{135}m_{58}} - 2m_{156} - m_{142} - m_{198}$	128
2	$output = m_{105}\left(\frac{m_{64}}{m_{118}} - m_{105}{}^{2}m_{117} + 0.33m_{79} - m_{72}m_{134}\right) - 0.60m_{114}$	124
3	$output = \frac{m_{105}m_{64}m_{76}m_{97}}{m_{134}} - m_{178}$	124
4	$output = m_{105} + \frac{m_{79} - m_{80}}{m_{55}} + 2m_{64} - m_{134} - m_{142} - m_{146} - m_{118}$	122
5	$output = \frac{1.53m_{105} - m_{167}}{0.87(m_{84} - m_{151})} - m_{134} + (m_{96} + 0.47m_{79})m_{64} - m_{157}$	132
6	$output = m_{105} + m_{76} - 3m_{168} + m_{125}m_{64}{}^{2} + m_{64} - m_{142} - 0.09$	128
7	$output = 1.66m_{64} + m_{89} - \frac{m_{164}m_{154}}{m_{105}} - \frac{m_{162}}{m_{105}}$	129
8	$output = (m_{89} + m_{76})^{2} + 2m_{64} - m_{135} - m_{118} - m_{134}$	120
9	$output = 2.59m_{105} + m_{51}m_{64} + m_{179} - 2m_{156} - 4.19m_{134}$	130
10	$output = m_{105} - m_{114} + (m_{76} + m_{64}m_{94})m_{94}$	120

Conventional variable selection is performed by calculating some statistical metric relating the input variables to the desired outputs and choosing those inputs ranked as most *characteristic*. A characteristic variable is one which shows a high degree of correlation between its experimentally-measured values and the known target values of the problem to be solved. For the sporulation problem described here, a highly characteristic variable would represent a *m/z* peak which consistently showed a high/low or low/high pattern to its values for bacterial samples grown under either vegetative or sporulating conditions.

The most commonly-used statistical metrics are shown in Equations 1-3[16,17], where σ_i is the standard deviation for peak *i* in all classes, $\sigma_{(i,k)}$ is the standard deviation for peak *i* in class *k* (*k* = 0 for vegetative and 1 for sporulated samples), n_k is the number of members in class *k*, $\bar{x}_i$ is the mean of peak *i* for all classes, and $x_{(i,k)}$ is the mean of peak *i* in class *k*.

The rules from 10 example GP runs are shown in Table 1, along with their predictive accuracies for the 132-sample test set, which comprised 66 vegetative and 66 sporulated samples all previously unseen by the GP. The term m_n refers to the ion count for a *m/z* ratio of n, expressed as a

$$F_i = \frac{\sigma_{(i,k)}{}^{2}}{\sigma_i{}^{2}}$$

Equation 1 **The Fisher Test (*F*-test)**

$$t_i = \frac{\bar{x}_{(i,0)} - \bar{x}_{(i,1)}}{\sqrt{\left(\frac{\sigma_{(i,0)}{}^{2}}{n_0}\right) + \left(\frac{\sigma_{(i,1)}{}^{2}}{n_1}\right)}}$$

Equation 2 **The Student Test (*t*-test)**

$$c_i = \frac{(\bar{x}_{(i,0)} - \bar{x}_i)^2 + (\bar{x}_{(i,1)} - \bar{x}_i)^2}{\sigma_{(i,0)}{}^{2} + \sigma_{(i,1)}{}^{2}}$$

Equation 3 **Characteristicity**

percentage of the total ion count for that sample. Each spectrum contained 150 m_n values, ranging from m_{51} to m_{200}.

Each GP was trained using an 84-sample training set of 42 vegetative and 42 sporulated samples, with the fitness function returning the RMS error for the rule's output as compared with the desired output (which was set to 0.0 for vegetative samples and 1.0 for sporulated samples). To classify a sample, an output value < 0.5 was taken to indicate a vegetative sample, and one ≥ 0.5 to indicate a sporulated sample.

The 10 example GP-derived rules use 34 of the 150 possible variables in the spectra but, because of the size constraints on the rules, each individual rule uses only about 6 to 9 variables. Surprisingly, the GP frequently selects variables other than those indicated by the statistical metrics to be the most characteristic for classification purposes (Table 2). In fact, in addition to those that are most characteristic, the GP consistently selects variables that are *least* characteristic. Inspection of the rules show that these non-characteristic variables are often being used as internal standard reference points: the fact that they are uncorrelated with the outputs (*i.e.* are independent or even constant) means that they are ideal variables to use in operations such as normalisation and baseline correction between samples. The use of low characteristicity variables in this way suggests that, although the data have already been normalised using the total ion count for the whole data set, the process of selecting a subset of the variables may necessitate a further normalisation of the selected variables to enable the best predictive models to be derived.

An example of the use of internal reference points is seen in the rule from example run 3 (Table 1), where three highly characteristic variables m_{105}, m_{64} and m_{76} (with average characteristicity rankings of 4.67, 2.67 and 1 respectively) are all found as ratios with m_{134} (average rank 140.33). This rule also contains a scaling factor of m_{97} (average rank 103), again as a ratio with m_{134} and additionally uses m_{178} (average rank 107.67) as a baseline correction factor.

The observation that the GP models use variables at both extremes of the characteristicity scale has obvious implications for the conventional approach to variable selection, which is to choose only those variables with high statistical characteristicity.

It was noted that m_{105} appeared in the rules much more frequently than any other variable, despite several other variables having a higher average characteristicity ranking. On closer examination, it was found that the data ranges for either sporulated or vegetative classes for m_{105} overlapped by just 9 samples, which was the lowest such overlap in the whole data set. The next lowest, m_{76}, had 20 overlapping samples. Thus m_{105}, although having a lower statistical characteristicity, actually provided the greatest discrimination between the sporulation classes of any single input variable. The statistical metrics do not take into account the actual distribution of points both within and between the classes, and so may give a somewhat inaccurate measure for the suitability of variables for use in a classification problem such as this.

The consistent selection by the GP of a small number of *m/z* peaks with high characteristicity, specifically peaks m_{105}, m_{64} and m_{76}, strongly suggests that molecules with these masses (105, 64 and 76 atomic mass units) are involved in the physico-chemical processes underlying the analysis of bacterial sporulation by PyMS. It is possible, therefore, to design a directed chemical analysis in order to identify molecules with these masses in the pyrolysate, and thereby to improve the understanding of the biochemistry of sporulation. Such a directed analysis would not have been so readily possible without the variable selection provided by the GP.

Table 2 Characteristicity rankings (relative positions in a sorted list) of the spectral peaks used by the rules in Table 1.

Peak	Ranking		
(*m/z*)	F-test	*t*-test	Characteristicity
51	11	7	15
55	107	109	126
58	14	14	20
64	2	4	2
72	79	68	114
76	1	1	1
79	12	13	14
80	63	112	78
84	8	10	5
89	10	9	11
94	72	32	86
96	131	137	129
97	112	67	130
105	7	3	4
114	16	15	27
117	35	26	28
118	29	21	32
125	148	148	150
134	142	140	139
135	37	40	66
142	78	72	110
146	126	101	82
151	145	147	133
152	129	142	140
154	26	17	34
156	46	76	58
157	70	98	71
162	124	102	76
164	123	150	127
167	143	149	136
168	125	133	131
178	128	106	89
179	94	116	116
198	116	111	90

4. Conclusions

The vegetative and sporulated biomass from 36 *Bacillus* species were analysed by Curie-point pyrolysis mass spectrometry (PyMS). Direct visual analysis of these 150-variable spectra was not possible, and so a GP approach was used to reduce the data dimensionality by deriving predictive models based on constrained-length rules.

The GP was able to derive rules capable of modeling the data using just 6-9 variables, which may be a reflection of the intrinsic dimensionality of this data set, *i.e.* an indication of the minimum number of variables within the data set which are able to provide all the information necessary to derive the best predictive models.

Inspection of the GP-derived rules showed that the models used not only the most characteristic variables (as measured by standard statistical metrics) but also the *least* characteristic. This enabled the GPs to derive good predictive models by performing normalisation and baseline correction between different samples, an ability likely to be lost if only the most characteristic variables are used.

The GP was also able to identify variables which are actually more characteristic for classification purposes than the statistical metrics would suggest. Unlike the statistical methods, the GP is able to take into account the precise distribution of data values for any given variable, and so the GP was able to select variables by assessing their actual discriminatory performance, rather than estimating their discriminatory value based on statistical theory.

The identification of characteristic ions by the GP allows a directed chemical analysis to be designed, and therefore this approach has the potential to lead to a better understanding of the physico-chemical processes underlying the classification of bacterial sporulation using PyMS.

This study highlights some of the ways in which genetic programming can aid the practice of science in ways not easily achieveable using more conventional data modelling methods. With its ability to select those variables which provide the maximum information for the minimum of model complexity, and to derive simple, interpretable expressions using those variables, GP can provide insights into the mechanisms underlying scientific problems which would otherwise remain undiscovered. Genetic programming-based data modelling is therefore a uniquely powerful tool with the potential to advance scientific knowledge in a wide variety of experimental fields.

Acknowledgments

R.J.G., J.T., J.J.R, and D.B.K. thank the UK EPSRC and BBSRC for financial support. R.J.G. and J.T. thank Dr. Andrew Edmonds of Science in Finance, Ltd. for financial and intellectual support. R.G. thanks the Wellcome Trust for financial support (grant number 042615/Z/94/Z). R.G. and B.S. thank the Society for General Microbiology for a vacation studentship.

References

1. Magee, J.T., *Whole-organism fingerprinting*, in *Handbook of New Bacterial Systematics*, Goodfellow, M. and O'Donnell, A.G., Editors. 1993, Academic Press: London. p. 383-427.

2. Goodacre, R. and Kell, D.B., *Pyrolysis mass spectrometry and its applications in biotechnology.* Cur. Opin. Biotechnol., 1996. **7**: p. 20-28.

3. Meuzelaar, H.L.C., Haverkamp, J., and Hileman, F.D., *Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials*. 1982, Amsterdam: Elsevier.

4. Goodacre, R., Neal, M.J., Kell, D.B., Greenham, L.W., Noble, W.C., and Harvey, R.G., *Rapid identification using pyrolysis mass spectrometry and artificial neural networks of Propionibacterium acnes isolated from dogs.* J. Appl. Bacteriol., 1994. **76**(2): p. 124-134.

5. Freeman, R., Goodacre, R., Sisson, P.R., Magee, J.G., Ward, A.C., and Lightfoot, N.F., *Rapid identification of species within the Mycobacterium tuberculosis complex by artificial neural network analysis of pyrolysis mass spectra.* J. Med. Microbiol., 1994. **40**(3): p. 170-173.

6. Sisson, P.R., Freeman, R., Law, D., Ward, A.C., and Lightfoot, N.F., *Rapid detection of verocytotoxin production status in E. coli by artificial neural network analysis of pyrolysis mass spectra.* J. Anal. Appl. Pyrol., 1995. **32**: p. 179-185.

7. Goodacre, R., Hiom, S.J., Cheeseman, S.L., Murdoch, D., Weightman, A.J., and Wade, W.G., *Identification and discrimination of oral asaccharolytic Eubacterium spp. using pyrolysis mass spectrometry and artificial neural networks.* Cur. Microbiol., 1996. **32**: p. 77-84.

8. Goodacre, R., Edmonds, A.N., and Kell, D.B., *Quantitative analysis of the pyrolysis mass spectra of complex mixtures using artificial neural networks - application to amino acids in glycogen.* J. Anal. Appl. Pyrol., 1993. **26**(2): p. 93-114.

9. Goodacre, R., Trew, S., Wrigley-Jones, C., Saunders, G., Neal, M.J., Porter, N., and Kell, D.B., *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, 1995. **313**(1-2): p. 25-43.

10. Timmins, É.M., Howell, S.A., Alsberg, B.K., Noble, W.C., and Goodacre, R., *Rapid differentiation of closely related Candida species and strains by pyrolysis mass spectrometry and fourier transform infrared spectroscopy.* Journal of Clinical Microbiology, 1998. **36**, p. 367-374.

11. Goodacre, R., Hammond, D., and Kell, D.B., *Quantitative analysis of the adulteration of orange juice with sucrose using pyrolysis mass spectrometry and chemometrics.* J. Anal. Appl. Pyrol., 1997. **40/41**: p. 135-158.

12. Alsberg, B.K., Goodacre, R., Rowland, J.J., and Kell, D.B., *Classification of pyrolysis mass spectra by fuzzy multivariate rule induction - comparison with regression, K-nearest neighbour, neural and decision-tree methods.* Anal. Chim. Acta, 1997. **348**: p. 389-407.

13. Goodacre, R., Shann, B., Gilbert, R.J., Timmins, É.M., McGovern, A.C., Alsberg, B.K., Logan, N.A. and Kell, D.B. *The characterisation of Bacillus species from PyMS and FT-IR data.* In *Proc. 1997 ERDEC Scientific Conference on Chemical and Biological Defense Research.*, Aberdeen Proving Ground. (in press).

14. Taylor, J., Goodacre, R., Wade, W., Rowland, J.J., and Kell, D.B., *The deconvolution of pyrolysis mass spectra using genetic programming: application to the identification of some Eubacterium species*. FEMS Microbiology Letters, 1998. **160**: p.237-246.

15. Gilbert, R.J., Goodacre, R., Woodward, A.M., and Kell, D.B., *Genetic programming: a novel method for the quantitative analysis of pyrolysis mass spectral data*. Anal. Chem., 1997. **69**(21): p.4381-4389.

16. Causton, D.R., *A Biologist's Advanced Mathematics.* 1987, London: Allen and Unwin.

17. Manly, B.F.J., *Multivariate Statistical Methods : A Primer*. 1994, London: Chapman & Hall. 215.

18. Holland, J.H., *Adaption in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. 1992: MIT Press.

19. Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*. 1989: Addison-Wesley.

20. Mitchell, M., *An Introduction to Genetic Algorithms*. 1995, Boston: MIT Press.

21. Koza, J.R., *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. 1992, Cambridge, MA: MIT Press. 819.

22. Koza, J.R., *Genetic Programming II: Automatic Discovery of Reusable Programs*. 1994, Cambridge, MA: MIT Press.

23. Singleton, A., *Genetic Programming in C*, Byte, 1994. **19**: p. 1-28.