Available online at www.sciencedirect.com
SCIENCE DIRECT®
Vibrational Spectroscopy 32 (2003) 33–45
VIBRATIONAL SPECTROSCOPY
www.elsevier.com/locate/vibspec

Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules

Royston Goodacre*

Department of Chemistry, UMIST, P.O. Box 88, Sackville Street, Manchester M60 1QD, UK

Abstract

Whole organism or tissue profiling by vibrational spectroscopy produces vast amounts of seemingly unintelligible data. However, the characterisation of the biological system under scrutiny is generally possible only in combination with modern supervised machine learning techniques, such as artificial neural networks (ANNs). Nevertheless, the interpretation of the calibration models from ANNs is often very difficult, and the information in terms of which vibrational modes in the infrared or Raman spectra are important is not readily available. ANNs are often perceived as 'black box' approaches to modelling spectra, and to allow the deconvolution of complex hyperspectral data it is necessary to develop a system that itself produces 'rules' that are readily comprehensible. Evolutionary computation, and in particular genetic programming (GP), is an ideal method to achieve this. An example of how GP can be used for Fourier transform infrared (FT-IR) image analysis is presented, and is compared with images produced by principal components analysis (PCA), discriminant function analysis (DFA) and partial least squares (PLS) regression.

Keywords: Artificial neural networks; Genetic programming; FT-IR

1. Introduction

Fourier transform infrared (FT-IR) and Raman are vibrational spectroscopic techniques which measures the absorbance of infrared light by molecules, or the inelastic light scattered from molecules when they are excited by a monochromatic light source [1,2]. These non-destructive techniques are not biased to any particular group of chemicals and so give 'holistic' whole organism or tissue fingerprints [3,4] of the biological sample under investigation. Due to the rapidity with which data can be collected, typified by the advent of focal plane array detectors for FT-IR [5–7] which allow infrared chemical maps of tissues to be constructed in only a few minutes, these methods are gaining considerable interest within high throughput screening programmes, disease recognition and biomarker discovery in body fluids (for excellent reviews see [8–12]).

Thus, it is possible to produce bounteous data floods, and the extraction of the most meaningful parts of these data is key to the generation of useful new knowledge about the biological system under interrogation. A typical FT-IR or Raman experiment is expected to generate thousands of data points (samples *times* variables) of which only a handful might be needed to describe the problem adequately. Evolutionary algorithms, and in particular genetic programming (GP), are ideal strategies for mining such data to generate useful relationships, rules and predictions. This paper describes GP and highlights its

* Tel.: +44-161-200-4480; fax: +44-161-200-4519.
E-mail address: r.goodacre@umist.ac.uk (R. Goodacre).

0924-2031/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0924-2031(03)00045-6

exploitation in the analysis of vibrational spectroscopic data. An example of how this would be applied to FT-IR imaging is also detailed.

2. Chemometric data analysis

Multivariate data such as those from an infrared or Raman fingerprint consist of the results of observations on a number of individuals (objects, or samples) of many different characters (variables, such as the absorbance at different wavenumbers or wavenumber shifts from a monochromatic light source for FT-IR and Raman, respectively) [13]. Each variable may be regarded as constituting a different dimension, such that if there are *n* variables (wavenumbers) each object may be said to reside at a unique position in an abstract entity referred to as *n*-dimensional hyperspace [14]. This hyperspace is obviously difficult to visualise and the underlying theme of multivariate analysis (MVA) is thus *simplification* [15] or *dimensionality reduction* [16]. In other words, one wants to summarise a large body of data by means of *relatively* few parameters, preferably the two or three which lend themselves to graphical display, with minimal loss of information, thereby allowing human interpretation.

Within chemometrics there are a variety of different algorithms that are used to analyse multivariate data, and by-and-large there are two main strategies used. The first is based on *unsupervised* learning whilst the second uses algorithms employing *supervised* learning.

2.1. Unsupervised learning

Unsupervised learning algorithms [17,18] seek to answer the question ''How similar to one another are these samples (e.g. bacteria) based on their FT-IR or Raman fingerprints I have collected?'', and are based on cluster analysis [15,19].

The reduction of the FT-IR or Raman data has typically been carried out using principal components analysis (PCA) [20] or hierarchical cluster analysis (HCA) [21]. PCA is a well-known technique for reducing the dimensionality of multivariate data whilst preserving most of the variance, and is used to identify *correlations* amongst a set of variables and to transform the original set of variables to a new set of *uncorrelated* variables called principal components (PCs). These PCs are then plotted and clusters in the data visualized; moreover this technique can be used to detect outliers. In its more conventional form, HCA employs a calculates distances (usually Euclidean, but may be Mahanalobis or Manhattan) between the objects in either the original data or a derivative thereof (e.g. the PCs) to construct a similarity matrix using a suitable similarity coefficient. These distance measures are then processed by an agglomerative clustering algorithm (although divisive algorithms are also used) to construct a dendrogram.

Provided that the data set contains ''standards'' (i.e. known things) it is evident that one can establish the closeness of any unknown samples to a standard, and thus effect the identification of the former, a technique termed 'operational fingerprinting' by Meuzelaar et al. [22] when analysing micro-organisms using pyrolysis–MS data. In post-genomics, such an approach is being referred to as 'guilt-by-association' [23,24].

2.2. Supervised learning

The unsupervised methods detailed above, although in some sense quantitative, are better seen as qualitative since their chief purpose is merely to *distinguish* objects or populations. Provided some 'gold standard' data exist on the objects being analysed, then a more powerful approach is to use *supervised* learning techniques (e.g. [18,25,26]) where one seeks to give answers of biological interest which have *much-lower dimensionality*, such as ''Based on the FT-IR fingerprint of this new sample I have just collected, which class in my database does it (most likely) belong to?'' and/or ''What is the level of toxicity this substance has on this tissue culture?''

The basic idea behind supervised learning is that there are some patterns (e.g. FT-IR fingerprints) which have desired responses which are known (i.e. the identity of this micro-organism, which has been decided by conventional approaches). These two types of data (the representation of the objects and their responses in the system) form pairs that are conventionally called inputs (*x*-data) and outputs/targets (*y*-data). The goal of supervised learning is to find a *model* or *mapping* that will correctly associate the inputs with the outputs (see Fig. 1 for a cartoon of this process).

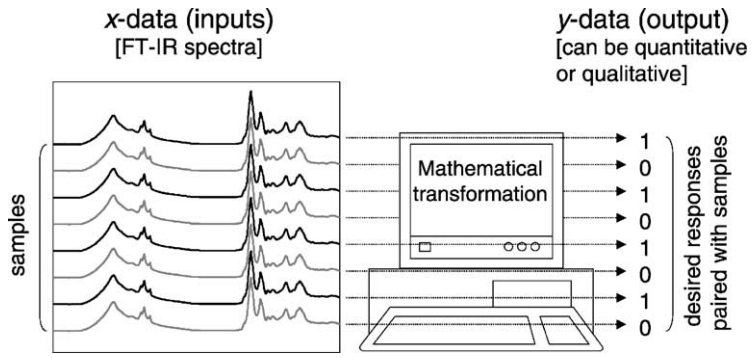

Fig. 1. *Supervised learning*. When we know the desired responses (*y*-data, or targets) associated with each of the inputs (*x*-data, or FT-IR/Raman spectra) then the system may be supervised. The goal of supervised learning is to find a mathematical transformation (model) that will correctly associate all or some of the inputs with the targets. In its conventional form this is achieved by minimising the error between the known target and the model's response (output).

Many different algorithms perform supervised learning, and they are either based on (i) discriminant algorithms, (ii) linear regression or (iii) non-linear mapping.

(i) Discriminant analysis (DA) is a qualitative (i.e. categorical), cluster analysis-based method that involves projection of test data into cluster space [21].
(ii) Although multiple linear regression (MLR) and principal components regression (PCR) are linear regression methods, the most popular approaches are based on partial least squares (PLS). PLS is a *quantitative* linear regression method [13], and can be extended to discriminant PLS, which is a qualitative (categorical) linear regression method [13,27].
(iii) However, arguably the most popular supervised learning methods are based on artificial neural networks (ANNs) which can learn non-linear as well as linear mappings. The most popular varieties are multilayer perceptrons (MLPs) [28,29] and radial basis functions (RBFs) [30–32].

The problem with the supervised learning algorithms detailed above is that the mathematical transformation from multivariate data to the target question of interest is often largely inaccessible in DA, PLS, and ANNs and these methods are often perceived as 'black box' approaches to modelling spectra, although the analysis of loading vectors for DA and regression coefficients for PLS can be informative. Indeed, for PLS there have been some interesting developments with respect to using orthogonal signal correction (OSC) [33,34] for variable selection. We know from the statistical literature that more robust predictions can often be obtained when only the most relevant input variables are considered [35,36]. Thus, the best machine learning techniques should not only give the correct answer(s), but also identify a subset of the variables with the maximal explanatory power thereby providing an interpretable description of what, in biological terms, is the basis for that answer. Such variable selection explanatory modelling methods do exist and are based on rule induction [37,38], inductive logic programming [39,40], and, in particular, evolutionary computation [41–43]. Thus, armed with these algorithms one can start to seek the answer to the question ''What have I measured in my FT-IR or Raman fingerprints that makes samples in class A (normal tissue) different from samples in class B (cancerous tissue)?''

3. Genetic programming

Evolutionary computational-based algorithms are particularly popular inductive reasoning and optimisation methods [44,45] based on the concepts of Darwinian selection [43] to generate and to optimize a desired computational function or mathematical expression to produce so called explanatory 'rules'. These techniques include genetic algorithms (GAs) [41,46,47], evolution strategies (ESs) [48], evolutionary

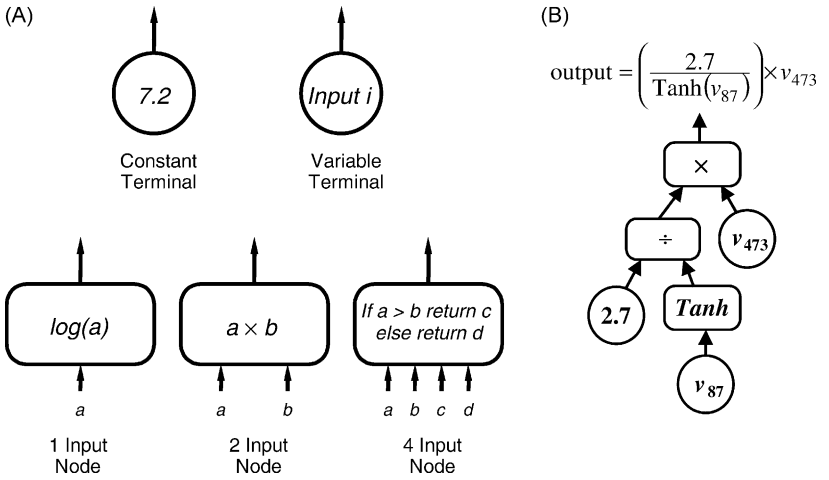

Fig. 2. The language structure of a tree-encoded GP. (A) The building blocks are represented by termini (vibrational inputs or numbers), and single input (a) and output nodes for encoding functions like $\log(a)$, $10^{(a)}$, $\tanh(a)$ and other transcendental functions; two inputs (a and b) and one output node for encoding the arithmetic functions $a \times b$, $a \div b$, $a + b$, and $a - b$; and a four input-single output note encoding a conditional 'if–then–else' statement; (B) shows a typical function tree.

programming (EP) [49] genetic programming [42,50] and genomic computing (GC) [51,52]. Of particular interest are GPs (and GCs which can be considered synonymous) because of the rich language structure used (*vide infra*), the models produced are in English, and further by reducing complex expressions, may be made to be comparatively simple thus allowing spectral interpretation and deconvolution.

A GP is an application of the GA approach to derive mathematical equations, logical rules or program functions automatically [42,53–57]. Rather than representing the solution to the problem as a string of parameters, as in a conventional GA, a GP usually [50] uses a tree structure that affords it a richer language. 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 (Fig. 2).

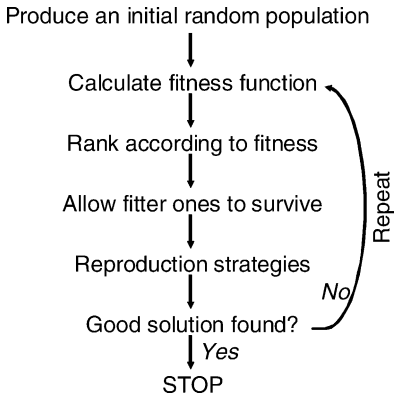

Fig. 3. The overall procedure employed by GP. The criterion for a good solution will be based on setting a threshold error between the known target and the GP's response.

The overall evolutionary procedure employed by GP is depicted in Fig. 3. An initial (usually random) population of individuals, each encoding a function or expression, is generated and their fitness to produce the desired output is assessed. In the second population, three reproduction strategies are adopted (see Fig. 4 for pictorial details):

(A) *Cloning* allows some of the original individuals to survive unmodified.
(B) New individuals are generated by *mutation* where one or more random changes to a single parent individual are introduced. This can be when a node is randomly chosen, and modified either by giving it a different operator with the same number of arguments, or it may be replaced by a new random sub-tree. Terminals can be mutated by slightly perturbing their numerical values, or randomly choosing a new input variable.

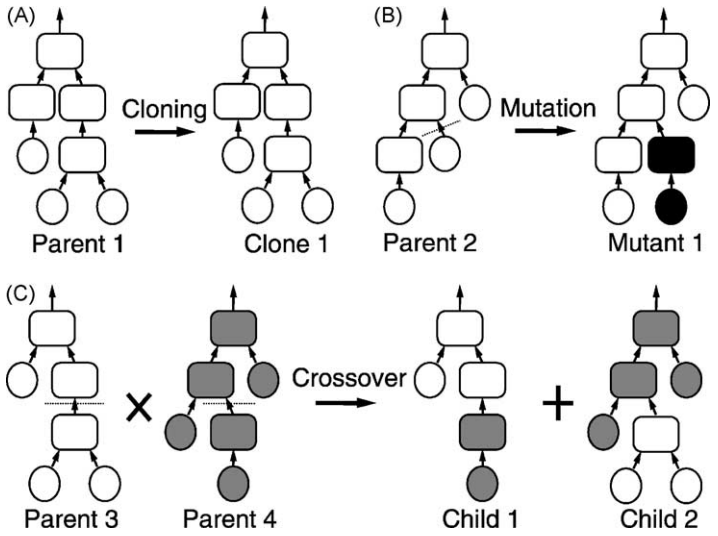

Fig. 4. The GP reproduction processes, with examples of (A) cloning, (B) mutation and (C) crossover. Dotted lines on the parents denote where the random mutation or crossover events occur.

(C) New children are generated by *crossover* where random rearrangement of functional components between two or more parent individuals takes place. Two parents are chosen with a probability related to their fitness. A node is randomly chosen on each parent tree, and the selected subtrees are then swapped.

The fitness of the new individuals in population 2 is assessed and the best individuals from the total population become the parents of the next generation. An individual's fitness is usually assessed as the error of the difference between expected values and the GP's estimated values for the training set. In order to reduce 'bloat', a phenomenon in which the GP function tree gets so huge that it lacks explanatory power [58,59], penalties to the number of nodes and depth of the tree in the individual's function tree can be applied. This overall process is repeated until either the desired result is achieved or the rate of improvement in the population becomes zero. It has been shown [42] that if the parent individuals are chosen according to their fitness values, the genetic method can approach the theoretical optimum efficiency for a search algorithm, and EAs generally are guaranteed to find the global optimum provided the best individuals are retained between generations ('elitism') [60].

4. Application of GP to the interpretation of vibrational spectroscopic data

GPs are very efficient search algorithms and because of their variable selection capabilities (that is to say, they extract the most relevant inputs and not the noise) can be used to produce models that allow the deconvolution of FT-IR and Raman data in chemical terms. Detailed below are three published examples from our laboratory illustrating this.

4.1. The detection of the dipicolinic acid biomarker in Bacillus spores [61]

The rapid identification of *Bacillus anthracis* spores is of importance because of its potential use as a biological warfare agent [62,63]. GP was used to analyse fingerprints generated from vegetative biomass and spores of various bacilli using pyrolysis–MS and FT-IR. Both fingerprinting approaches could be used to differentiate successfully between vegetative biomass and spores. GP produced mathematical rules which could be simply interpreted in biochemical terms. It was found that for pyrolysis–MS *m/z* 105 was characteristic and is a pyridine ketonium ion ($C_6H_3ON^+$) obtained from the pyrolysis of pyridine-2,6-dicarboxylic acid (dipicolinic acid, DPA), a metabolite found in spores but not in vegetative cells.

In addition, FT-IR analysis of the same system showed that a pyridine ring vibration at 1447–1439 cm^{-1} from the same metabolite, DPA, was found to be highly characteristic of spores. Thus, although the original datasets recorded hundreds of spectral variables from whole cells simultaneously, a simple biomarker can be used for the rapid and unequivocal detection of spores of these organisms.

4.2. Monitoring of complex industrial bioprocesses [64]

The ability to control industrial bioprocess is paramount for product yield optimisation, and it is imperative therefore that the concentration of the fermentation product (the determinand) is assessed accurately. Whilst infrared and Raman spectroscopies have been used for the quantitative analysis of fermentations [65–67] the transformation of spectra to determinand concentration(s) has usually been undertaken by PLS and ANNs, and one can not be sure whether the model is detecting the product itself, an increase in by-products or decrease in substrates. By contrast, GP (and GA) has recently been used to analyse IR and Raman spectra from a diverse range of unprocessed, industrial fed-batch fermentation broths containing the fungus *Gibberella fujikuroi* producing the natural product gibberellic acid. The models produced allowed the determination of those input variables that contributed most to the models formed, and it was observed that those quantitative models were predominately based on the concentration of gibberellic acid itself.

4.3. The detection of the microbial spoilage of meat [68]

The rapid detection of microbial spoilage in meats using FT-IR has only very recently been demonstrated. Attenuated total reflectance (ATR) was used for analysis where the food sample was placed in intimate contact with a crystal of high refractive index and an IR absorbance spectrum collected in just a few seconds. It was shown that FT-IR with PLS allowed accurate estimates of bacterial loads (from 10^6 to 10^9 cm^{-2}) to be determined directly from the chicken surface in the 60s, and that GP (and GA) indicated that at levels of 10^7 bacteria cm^{-2} the main biochemical indicator of spoilage as measured by FT-IR was the onset of proteolysis, a finding which is in agreement with the literature [69–71].

5. The analysis of FT-IR images using GP

5.1. The discrimination of two closely related Escherichia coli strains

Let us consider the following experiment, which demonstrates the differentiation between closely related bacteria, those of two laboratory strains of *E. coli* HB101 [72] and UB5201 [73]. It is known that these organisms are very closely related from previous whole organism fingerprinting studies using pyrolysis–MS [74], and the FT-IR spectra (*vide infra* for collection method) of *E. coli* HB101 and UB5201 do indeed look very similar (Fig. 5).

5.2. Bacterial growth and spectral acquisition

Both strains were grown separately in 100 ml liquid media (glucose (BDH),10.0 g; peptone (LAB M), 5.0 g; beef extract (LAB M), 3.0 g; H_2O, 1 l) for 16 h at 37 °C in a shaker incubator. After growth, the cells were harvested by centrifugation and washed in physiological saline (0.9% NaCl). The dry weights of the cells were then estimated gravimetrically and used to adjust the weight of the final slurries with physiological saline to approximately 40 mg ml^{-1}; this was $\sim(1-2) \times 10^7$ cells [74].

For the IR map the biomass from *E. coli* HB101 and UB5201 were applied evenly to the surface of a 7 cm × 7 cm Al plate at a concentration of $\sim$200 μg cm^{-2} (dry weight); a cartoon of a bacterial cell was drawn with the different biomass (Fig. 6A). Prior to analysis the sample carrier was oven-dried at 50 °C for 30 min. FT-IR spectra were acquired at a spatial resolution of 1 mm (therefore, this data cube was 71 pixels × 71 pixels, by 882 wavenumbers). The FT-IR instrument used was the Bruker IFS28 FT-IR spectrometer (Bruker Spectrospin Ltd., Banner Lane, Coventry, UK) equipped with an mercury–cadmium–telluride (MCT) detector (cooled with liquid N_2) and a motorised stage of a reflectance accessory, onto which the Al plate was loaded. Spectra were collected over

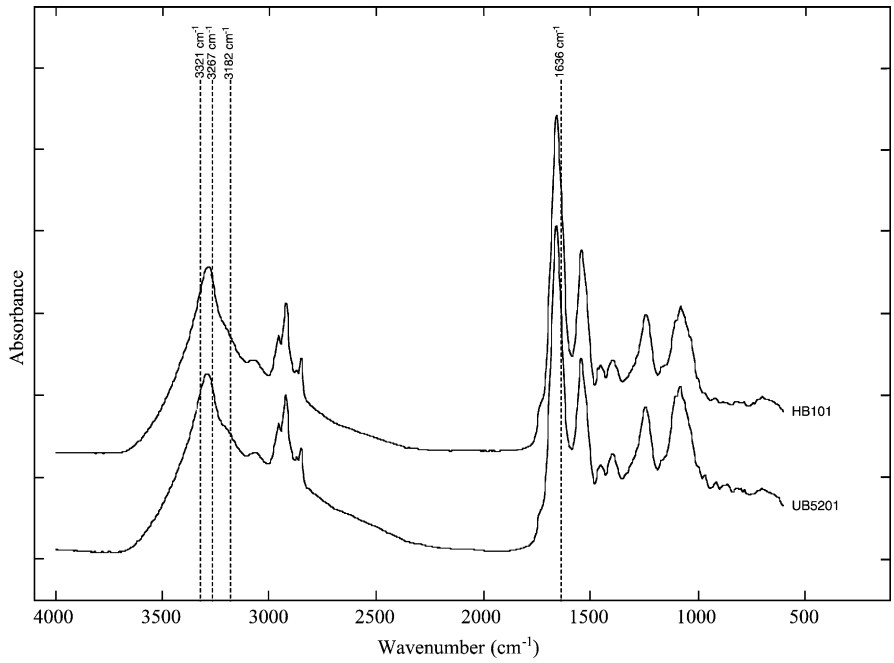

Fig. 5. Typical FT-IR spectra of *Escherichia coli* UB5201 and HB101 strains. Also highlighted are the wavenumbers selected by GP analysis.

the wavenumber range 4000–600 cm^{-1}. Spectra were acquired at a rate of 20 s^{-1}, the spectral resolution used was 4 cm^{-1}, and 16 spectra were co-added and averaged.

5.3. 'Classical' analysis of IR chemical images

The traditional approach used to analyse FT-IR (and indeed Raman) images from tissues or other materials [5] has been to plot the area under specific peaks for protein, lipid and polysaccharides. In order to compensate for sample thickness effects it is often prudent to plot the lipid-to-protein ratio. This was performed for the *E. coli* map where the lipid-to-protein ratio was calculated as the integration of the CH_2 stretch between 2912 and 2936 cm^{-1} divided by the integration of C=O vibration between 1651 and 1674 cm^{-1}. The resultant image is shown in Fig. 6B which bears no resemblance to the real cartoon image shown in Fig. 6A.

The approach of using spectral windows is a valid one but presumes that one already knows which are the important discriminative vibrations. Without this *a priori* knowledge one can use a full spectral approach and compress the data via PCA as demonstrated by, e.g. [75]. This was performed on the *E. coli* map and whilst there are some features in the first principal component score (Fig. 7A), the two bacterial strains cannot be differentiated, and neither were they in any of the other PCs scores extracted (data not shown). Therefore, since we have also collected spectra of the *E. coli* HB101 and UB5201 it would seem sensible to use a supervised method, and calibrate it with these reference spectra. DA and ANNs have been used to create chemical images from tissues [7,75]. Thus, projection of the spectra from the *E. coli* image into discriminant function analysis (DFA) space was performed. Briefly, PCA followed by DFA was carried out on 14 spectra from *E. coli* HB101 coded as one group and 14 of *E. coli* UB5201 coded as a separate class (8 PCs were extracted which explained 99.9% of the variance). The spectra from the image were first projected into the PCA space and then the resultant PCs projected into the DFA space, and the first PC-DF score was plotted (Fig. 7B). In the PC-DFA image it is easier to differentiate between the two *E. coli* strains,

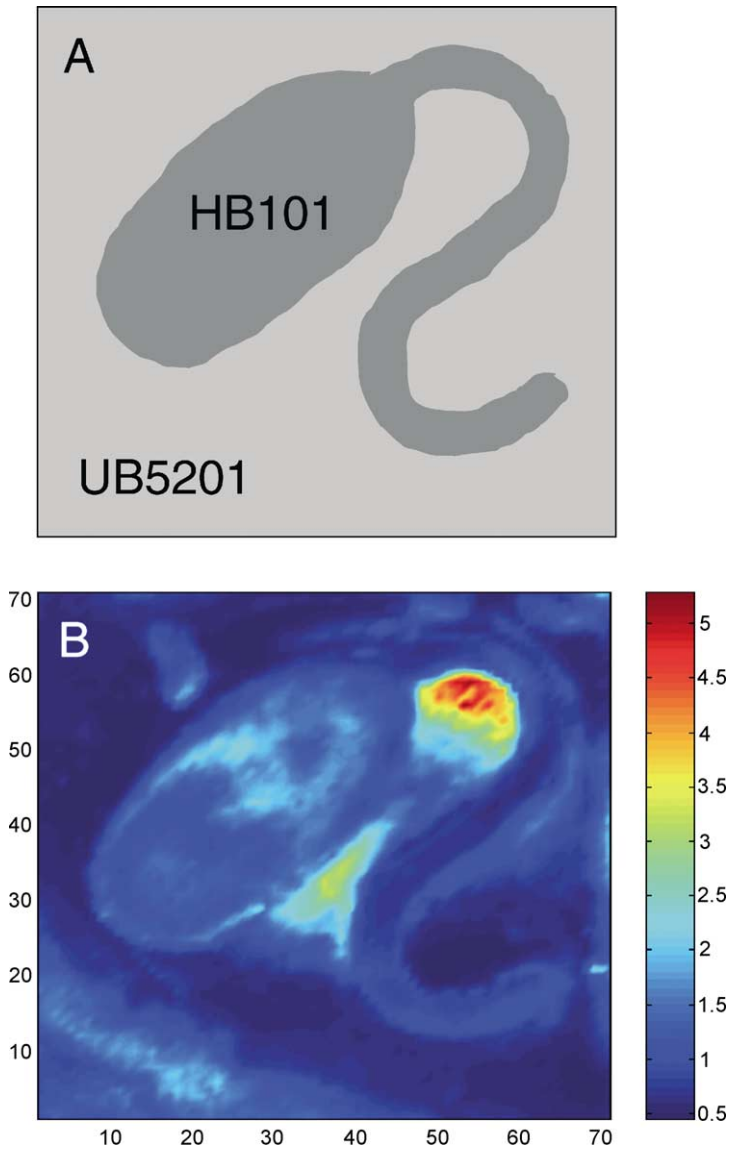

Fig. 6. (A) Cartoon of the original image and (B) the protein:lipid ratio of the 71 pixels × 71 pixels hypercube. The lipid-to-protein ratio was calculated as the integration of the CH_2 stretch between 2912 and 2936 cm^{-1} divided by the integration of C=O vibration between 1651 and 1674 cm^{-1}.

however, the PC-DFA loadings are complex (Fig. 8A) and no obvious spectral features were found to be discriminating; although on closer inspection the amide I band was found to be discriminating. PLS was also performed on the 14 spectra from *E. coli* HB101 coded as 1 and 14 of *E. coli* UB5201 coded as 0. The PLS model was calibrated with a single latent variable and challenged with the spectra from the image and the resultant predictions were plotted (Fig. 7C). The PLS generated image was not as clear as that produced from PC-DFA, which might be because PLS uses a linear regression algorithm rather than a discriminatory-based one. Inspection of the regression coefficients from the PLS model (Fig. 8B) were highly complex, with many input variables being selected.

5.4. GP analysis of IR chemical images

The GP employed the Genomic Computing software Gmax-bio™ (Aber Genomic Computing, Aberystwyth, UK) which runs under Microsoft Windows NT on an IBM-compatible PC. An introduction to Gmax-bio™ is given elsewhere [51,52], and the default parameter settings for population size (1000), mutation and recombination rates were used throughout. The operators that were used were +, −, /, *, $\log_{10}(x)$, 10^x, and $\tanh(x)$. The fitness calculation used is $F = 1/(0.01 + S/B)$ where the values of S and B are determined by the FITNESS setting. In this expression, S is a statistic derived from the model, which ranges between 0 and infinity and B is a normalising quantity. The value of B is chosen such that a perfect model

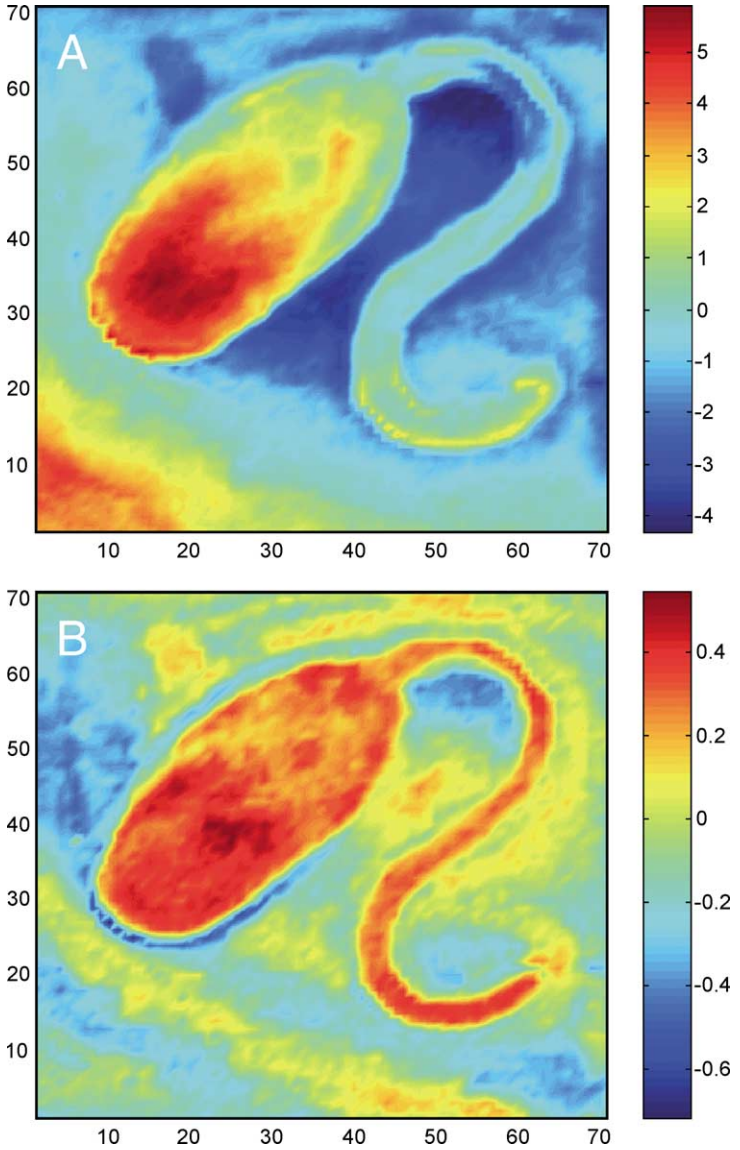

Fig. 7. Images produced from (A) PC scores showing PC 1 which explained 86.2% of the total variance, (B) PC-DFA projection analysis showing PC-DF 1, (C) PLS predictions (UB5201 coded as 0 and HB101 as 1) and (D) GP analysis where the tree output = $(3267\,\mathrm{cm}^{-1} - 3321\,\mathrm{cm}^{-1}) \times (1636\,\mathrm{cm}^{-1}/3182\,\mathrm{cm}^{-1})$; for the GP UB5201 was also coded as 0 and HB101 as 1.

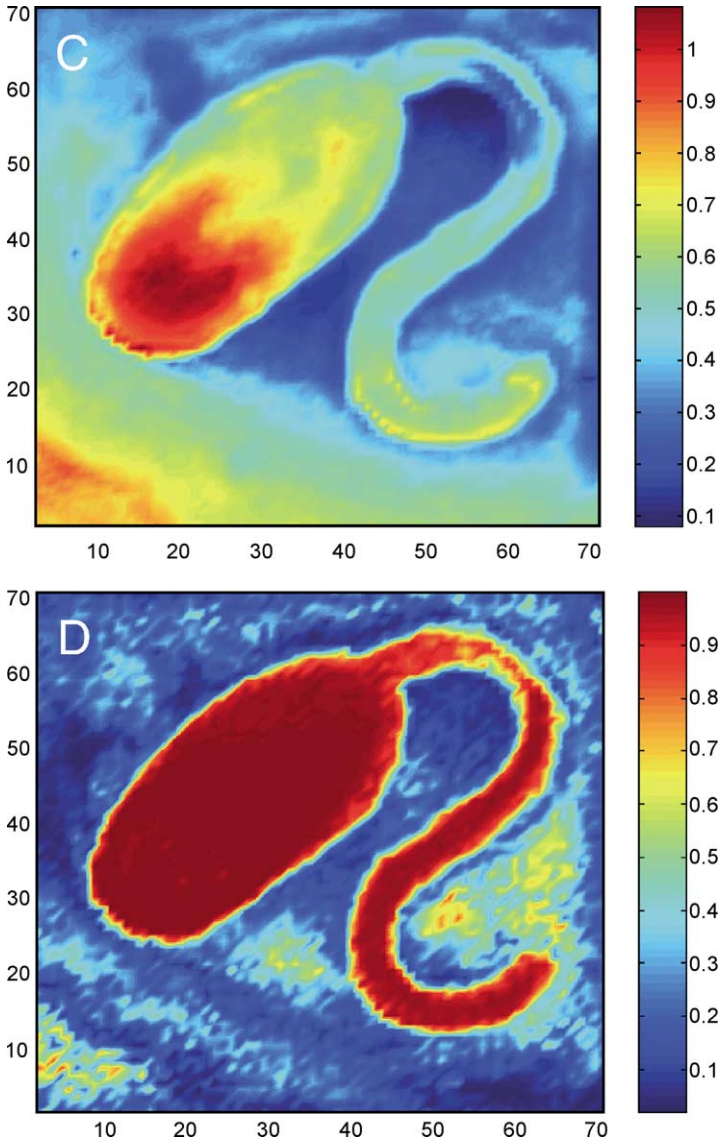

Fig. 7. (*Continued*).

yields $F = 100$, and a model which performs no better than random chance yields $F = 1$.

The GP was calibrated with the same 28 spectra from *E. coli* HB101 and UB5201 that were used for PC-DFA, and the output for the GP for spectra from UB5201 was coded as 0 and those from HB101 as 1. After evolution the GP's output was ($3267\,cm^{-1} - 3321\,cm^{-1}$) × ($1636\,cm^{-1}/3182\,cm^{-1}$), the outputs for all 5041 spectra (71 pixels × 71 pixels) in the *E. coli* map were calculated and plotted spatially (Fig. 7D). It is obvious that the two strains are clearly differentiated by the algebraic combination of these four vibrations. The vibrations selected are highlighted in Fig. 5 and it is likely that these all arise from proteins: $1636\,cm^{-1}$ is a C=O stretching on the side of the amide I band, 3267 and $3321\,cm^{-1}$ are on the O–H vibration, and $3182\,cm^{-1}$ from N–H. It is known that proteins rich in α-helix have an amide I maximum at $\sim 1650\,cm^{-1}$ and this shifts to $\sim 1620\,cm^{-1}$ for proteins rich in β-sheets [76].

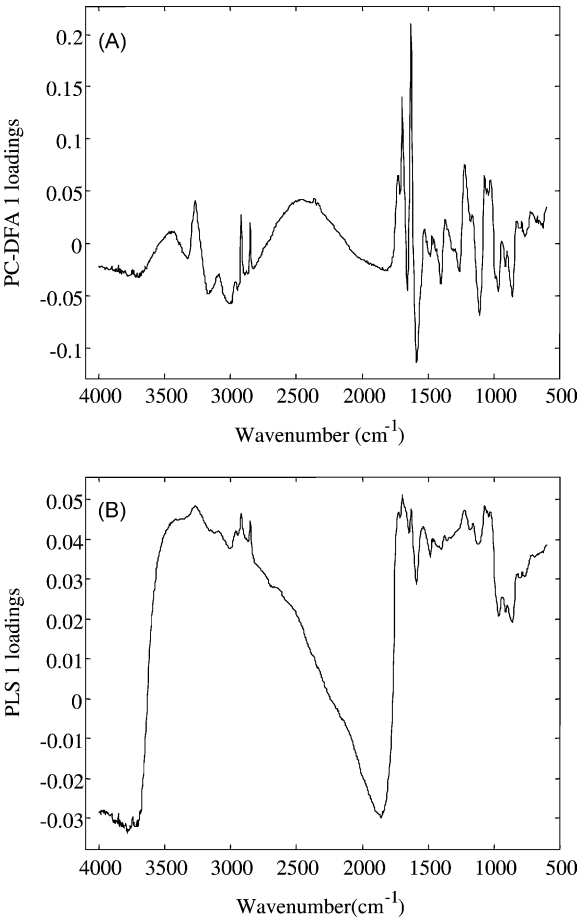

Fig. 8. Loadings plots showing (A) PC-DF 1 loadings and (B) PLS first latent variable loading.

Since the *side* of the amide I band was chosen (rather than the centre of the peak), in addition to the other protein vibrations, suggests that the major difference between *E. coli* HB101 and UB5201 is due to the protein complement of these cells, rather than a change in the polysaccharide or lipid components.

6. Concluding remarks

We are all hopefully aware of the cycle of knowledge (Fig. 9) [52,77]. One has some preconceived notions about the problem domain, experiments are designed to test these hypotheses, the observations from these experiments are recorded and by deductive

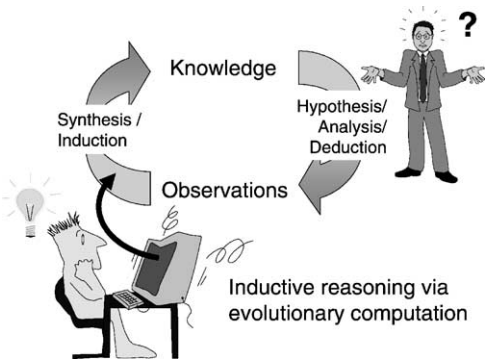

Fig. 9. The cycle of knowledge, showing where rule induction will play its part.

reasoning the observations considered to be consistent or inconsistent with the hypotheses [78]. Actually, although this part is normally only implicit, by a process of induction these observations are synthesised or generalised to refine our accepted wisdom. The cycle then repeats itself until one is happy with the solution to a given problem.

However, if one is in a scenario where our knowledge is minute, e.g. that is to say we have no idea about the biochemical or physiological differences between two organisms. What are we to do? It is unreasonable to go to the Sigma–Aldrich catalogue and collect the spectra of every known metabolite (and although a rather 'stamp collecting' exercise it is fair to say that this would be a useful resource). Nevertheless, we can design experiments based, for example, on whether *Bacillus* is sporulated or exists as vegetative biomass, collect 'holistic' whole organism fingerprints by FT-IR or Raman spectroscopies and then use rule induction via evolutionary computation to elucidate what are the key bond vibrations of the vibrational spectra are important for the discrimination. Coupled to MS and NMR, this will then give an insight into the sorts of biochemical species (metabolites) that are important, and help understand more fully the biological system under investigation.

Acknowledgements

I am indebted to the UK BBSRC (Engineering and Biological Systems Committee) and the UK EPSRC for financial support.

References

[1] P.R. Griffiths, J.A. de Haseth, Fourier Transform Infrared Spectrometry, Wiley, New York, 1986.
[2] N.B. Colthup, L.H. Daly, S.E. Wiberly, Introduction to Infrared and Raman Spectroscopy, Academic Press, New York, 1990.
[3] J.T. Magee, in: M. Goodfellow, A.G. O'Donnell (Eds.), Handbook of New Bacterial Systematics, Academic Press, London, 1993, pp. 383–427.
[4] R. Goodacre, É.M. Timmins, R. Burton, N. Kaderbhai, A.M. Woodward, D.B. Kell, P.J. Rooney, Microbiology 144 (1998) 1157–1170.
[5] E.N. Lewis, P.J. Treado, R.C. Reeder, G.M. Story, A.E. Dowrey, C. Marcott, I.W. Levin, Anal. Chem. 67 (1995) 3377–3781.
[6] P. Colarusso, L.H. Kidder, I.W. Levin, J.C. Fraser, J.F. Arens, E.N. Lewis, Appl. Spectrosc. 52 (1998) 106A–120A.
[7] P. Lasch, W. Haensch, E.N. Lewis, L.H. Kidder, D. Naumann, Appl. Spectrosc. 56 (2002) 1–9.
[8] J. Schmitt, H.C. Flemming, Int. Biodeterior. Biodegrad. 41 (1998) 1–11.
[9] L. Mariey, J.P. Signolle, C. Amiel, J. Travert, Vib. Spectrosc. 26 (2001) 151–159.
[10] D. Naumann, Appl. Spectrosc. Rev. 36 (2001) 239–298.
[11] W. Petrich, Appl. Spectrosc. Rev. 36 (2001) 181–237.
[12] K. Maquelin, C. Kirschner, L.-P. Choo-Smith, N. van den Braak, H.P. Endtz, D. Naumann, G.J. Puppels, J. Microbiol. Methods 51 (2002) 255–271.
[13] H. Martens, T. Næs, Multivariate Calibration, Wiley, Chichester, 1989.
[14] R. Goodacre, M.J. Neal, D.B. Kell, Z. Bakteriol. 284 (1996) 516–539.
[15] C. Chatfield, A.J. Collins, Introduction to Multivariate Analysis, Chapman & Hall, London, 1980.
[16] J.W. Tukey, Exploratory Data Analysis, Addison-Wesley, Reading, MA, 1977.
[17] R.O. Duda, P.E. Hart, D.E. Stork, Pattern Classification, 2nd edition, Wiley, London, 2001.
[18] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer-Verlag, Berlin, 2001.
[19] B.S. Everitt, Cluster Analysis, Edward Arnold, London, 1993.
[20] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1986.
[21] B.F.J. Manly, Multivariate Statistical Methods: A Primer, Chapman & Hall, London, 1994.
[22] H.L.C. Meuzelaar, J. Haverkamp, F.D. Hileman, Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials, Elsevier, Amsterdam, 1982.
[23] D. Altshuler, M. Daly, L. Kruglyak, Nat. Genet. 26 (2000) 135–137.
[24] S.G. Oliver, Nature 403 (2000) 601–603.
[25] T.M. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.
[26] R.C. Beavis, S.M. Colby, R. Goodacre, P.B. Harrington, J.P. Reilly, S. Sokolow, C.W. Wilkerson, in: R.A. Meyers (Eds.), Encyclopaedia of Analytical Chemistry, Wiley, Chichester, 2000, pp. 11558–11597.
[27] B.K. Alsberg, D.B. Kell, R. Goodacre, Anal. Chem. 70 (1998) 4126–4133.
[28] D.E. Rumelhart, J.L. McClelland, The PDP Research Group Parallel Distributed Processing, Experiments in the Microstructure of Cognition, vols. I and II, MIT Press, Cambridge, MA, 1986.
[29] P.J. Werbos, The Roots of Back-Propagation: From Ordered Derivatives to Neural Networks and Political Forecasting, Wiley, Chichester, 1994.
[30] D.S. Broomhead, D. Lowe, Complex Syst. 2 (1988) 321–355.
[31] A. Saha, J.D. Keller, in: D. Touretzky (Eds.), Advances in Neural Information Processing Systems, Morgan Kaufmann, San Mateo, CA, 1990, pp. 482–489.
[32] C.M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.
[33] S. Wold, H. Antti, F. Lindgren, J. Ohman, Chemom. Intell. Lab. Syst. 44 (1998) 175–185.
[34] A. Hoskuldsson, Chemom. Intell. Lab. Syst. 55 (2001) 23–38.
[35] M.B. Seasholtz, B. Kowalski, Anal. Chim. Acta 277 (1993) 165–177.
[36] D.B. Kell, B. Sonnleitner, Tibtech 13 (1995) 481–492.
[37] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, Pacific Grove, CA, 1984.
[38] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[39] R.D. King, S. Muggleton, R.A. Lewis, M.J.E. Sternberg, Proc. Natl. Acad. Sci. 89 (1992) 11322–11326.
[40] N. Lavrac, S. Dzeroski, Inductive Logic Programming: Techniques and Applications, Ellis Horwood, Chichester, 1994.
[41] J.H. Holland, Adaption in Natural and Artificial Systems, MIT Press, Cambridge, MA, 1992.
[42] J.R. Koza, Genetic Programming: On the Programming of computers by Means of Natural Selection, MIT Press, Cambridge, MA, 1992.
[43] T. Bäck, D.B. Fogel, Z. Michalewicz, Handbook of Evolutionary Computation, IOP Publishing/Oxford University Press, Oxford, 1997.
[44] D. Corne, M. Dorigo, F. Glover (Eds.), New Ideas in Optimization. McGraw-Hill, London, 1999.
[45] Z. Michalewicz, D.B. Fogel, How to Solve it: Modern Heuristics, Springer-Verlag, Heidelberg, 2000.
[46] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1989.
[47] M. Mitchell, An Introduction to Genetic Algorithms, MIT Press, Boston, 1995.
[48] H.-G. Beyer, The Theory of Evolution Strategies, Springer-Verlag, Berlin, 2001.
[49] D.B. Fogel, Evolutionary Computation: Towards a New Philosophy of Machine Intelligence, IEEE Press, Piscataway, 2000.
[50] W. Banzhaf, P. Nordin, R.E. Keller, F.D. Francone, Genetic Programming: An Introduction, Morgan Kaufmann, San Francisco, CA, 1998.

[51] D.B. Kell, R.M. Darby, J. Draper, Plant Physiol. 126 (2001) 943–951.
[52] D.B. Kell, Trends Genet. 18 (2002) 555–559.
[53] J.R. Koza, Genetic Programming II: Automatic Discovery of Reusable Programs, MIT Press, Cambridge, MA, 1994.
[54] R.J. Gilbert, R. Goodacre, A.M. Woodward, D.B. Kell, Anal. Chem. 69 (1997) 4381–4389.
[55] W.B. Langdon, Genetic Programming and Data Structures: Genetic Programming + Data Structures = Automatic Programming!, Kluwer Academic Publishers, Boston, 1998.
[56] J.R. Koza, F.H. Bennett, M.A. Keane, D. Andre, Genetic Programming III: Darwinian Invention and Problem Solving, Morgan Kaufmann, San Francisco, CA, 1999.
[57] W.B. Langdon, R. Poli, Foundations of Genetic Programming, Springer-Verlag, Berlin, 2002.
[58] W.B. Langdon, R. Poli, in: W. Banzhaf, R. Poli, M. Schoenauer, T.C. Fogarty (Eds.), Proceedings of the First European Workshop on Genetic Programming, Springer-Verlag, Berlin, 1998, pp. 37–48.
[59] V. Podgorelec, P. Kokol, Genet. Program. Proc. 1802 (2000) 326–337.
[60] G. Rudolph, Convergence Properties of Evolutionary Algorithms, Verlag Dr Kovac, Hamburg, 1997.
[61] R. Goodacre, B. Shann, R.J. Gilbert, É.M. Timmins, A.C. McGovern, B.K. Alsberg, D.B. Kell, N.A. Logan, Anal. Chem. 72 (2000) 119–127.
[62] M. Dando, Biological Warfare in the 21st Century, Brassey's Ltd., London, 1994.
[63] W. Barnaby, The Plague Makers: The Secret World of Biolgoical Warfare, Vision Paperbacks, London, 1997.
[64] A.C. McGovern, D. Broadhurst, J. Taylor, N. Kaderbhai, M.K. Winson, D.A. Small, J.J. Rowland, D.B. Kell, R. Goodacre, Biotechnol. Bioeng. 78 (2002) 527–538.
[65] A.C. McGovern, R. Ernill, B.V. Kara, D.B. Kell, R. Goodacre, J. Biotechnol. 72 (1999) 157–167.
[66] A.D. Shaw, N. Kaderbhai, A. Jones, A.M. Woodward, R. Goodacre, J.J. Rowland, D.B. Kell, Appl. Spectrosc. 53 (1999) 1419–1428.
[67] S. Vaidyanathan, G. Macaloney, B. McNeill, Analyst 124 (1999) 157–162.
[68] D.I. Ellis, D. Broadhurst, D.B. Kell, J.J. Rowland, R. Goodacre, Appl. Environ. Microbiol. 68 (2002) 2822–2828.
[69] R.H. Dainty, J. Food Microbiol. 33 (1996) 19–33.
[70] G.J.E. Nychas, C.C. Tassou, J. Food Microbiol. 74 (1997) 199–208.
[71] D.I. Ellis, R. Goodacre, Trends Food Sci. Technol. 12 (2002) 413–423.
[72] T. Maniatis, F. Fritsch, J. Shambrook, Molecular Cloning: A Laboratory Manual, Cold Spring Harbour Laboratory, New York, 1982.
[73] F. de la Cruz, J. Grinsted, J. Bacteriol. 151 (1982) 222–228.
[74] É.M. Timmins, R. Goodacre, J. Appl. Microbiol. 83 (1997) 208–218.
[75] P. Lasch, D. Naumann, Cell. Mol. Biol. 44 (1998) 189–202.
[76] M. Jackson, H.H. Mantsch, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 95–120.
[77] R. Goodacre, D.B. Kell, in: G.G. Harrigan, R. Goodacre (Eds.), Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis, Kluwer Academic Publishers, Dordrecht, 2003.
[78] D. Oldroyd, The Arch of Knowledge: An Introduction to the History of the Philosophy and Methodology of Science, Methuen, New York, 1986.