Chapter 13

EVOLUTIONARY COMPUTATION FOR THE INTERPRETATION OF METABOLOMIC DATA

Royston Goodacre[1,2] and Douglas B. Kell[2]
[1]Institute of Biological Sciences, University of Wales, Aberystwyth, SY23 3DD, UK [2]Department of Chemistry, University of Manchester Institute of Science and Technology, PO Box 88, Sackville St., Manchester, M60 1QD, UK

> "The fewer data needed, the better the information. And an overload of information, that is, anything much beyond what is truly needed, leads to information blackout. It does not enrich, but impoverishes."
> *Peter F. Drucker - Management: Tasks, Responsibilities, Practices*

1. INTRODUCTION

Post-genomic science is producing bounteous data floods, and as the above quotation indicates the extraction of the most meaningful parts of these data is key to the generation of useful new knowledge. Atypical metabolic fingerprint or metabolomics 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 are ideal strategies for mining such data to generate useful relationships, rules and predictions. This chapter describes these techniques and highlights their exploitation in metabolomics.

In a recent study Lyman and Varian estimated that in 2000 the world produced between 1 and 2 exabytes ($1\text{-}2.10^{18}$ bytes) of unique information (www.sims.berkeley.edu/how-much-info). This data flood is roughly 250 megabytes for every man, woman and child on earth! IBM's (www.ibm.com) estimates are that information within the life sciences doubles every 6 months, and this data explosion comes from genomic sequencing, the

G. G. Harrigan et al. (eds.), *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*

'omics' (transcriptome, proteomics, metabolomics), high-throughput screening as well as the more traditional pre-clinical and clinical trials.

Metabolomics is the third level of 'omics' analysis. The metabolome is the quantitative complement of all the low molecular weight molecules present in cells in a particular physiological or developmental state (Oliver *et al.*, 1998; Fiehn, 2002) and whilst complementary to transcriptomics and proteomics may be seen to have special advantages. In particular, we know from the theory underlying metabolic control analysis (MCA) (Kell and Westerhoff, 1986; Fell, 1996; Heinrich and Schuster, 1996; Mendes *et al.*, 1996; Kell and Mendes, 2000) as well as from experiment (Fiehn *et al.*, 2000a; Raamsdonk *et al.*, 2001), that while changes in the levels of individual enzymes may be expected to have little effect on metabolic fluxes, they can and do have significant effects on the concentrations of a variety of individual metabolites (Westerhoff and Kell, 1996). In addition, as the 'downstream' result of gene expression, changes in the metabolome are amplified relative to changes in the transcriptome and the proteome.

Currently the 'gold standard' for measuring the metabolome is gas chromatography-mass spectrometry (GC-MS) (Fiehn *et al.*, 2000a, 2000b), and whilst a single run generates the name of a metabolite (or unique designation) with its (relative) concentration, GC-MS suffers from being chemically biased because of the extraction solvents employed. It is also relatively slow both for the chromatography itself (typical run times are 30 min per sample) and for the subsequent deconvolution steps. By contrast, rather than attempting to measure every metabolite, metabolic fingerprinting methods are sufficiently rapid to enable the classification of samples according to the origin or their biological relevance (Fiehn, 2002). For high-throughput metabolic fingerprinting the methods typically employed include nuclear magnetic resonance spectroscopy (NMR) (Lindon *et al.*, 2000), direct infusion electrospray ionization-MS (Vaidyanathan *et al.*, 2001, 2002; Allen *et al.*, 2002), and Fourier transform infrared (FT-IR) spectroscopy (Winson *et al.*, 1997; Goodacre *et al.*, 1998; Oliver *et al.*, 1998). These profiling strategies generate large amounts of data, and it is obvious (Fiehn *et al.*, 2001; Mendes, 2002) that current informatic approaches need to adapt and grow in order to make the most of these data

2. MULTIVARIATE ANALYSIS

Multivariate data such as those from a metabolic fingerprint consist of the results of observations on a number of individuals (objects, or samples) of many different characters (variables) such as the spectral intensities at different mass-to-charge ratios, chemical shifts from NMR, or absorbance at

different wavenumbers from FT-IR (Martens and Næs, 1989). Each variable may be regarded as constituting a different dimension, such that if there are n variables each object may be said to reside at a unique position in an abstract entity referred to as n-dimensional hyperspace (Goodacre *et al.*, 1996). This hyperspace is necessarily difficult to visualize (Wilkinson, 1999) and the underlying theme of multivariate analysis is thus *simplification* (Chatfield and Collins, 1980) or dimensionality reduction (Tukey, 1977). This usually means that one wants to summarize 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.

Within chemometrics there are three varieties of algorithms that are used to analyze multivariate data.

2.1 The Clustering Variety

These algorithms are based on *unsupervised* learning (Duda *et al.*, 2001; Hastie *et al.*, 2001) and seek to answer the question 'How similar to one another are these samples based on the metabolite fingerprints I have collected?'

Conventionally the reduction of multivariate data has been carried out using principal components analysis (PCA; (Jolliffe, 1986; Everitt, 1993)) or hierarchical cluster analysis (HCA; (Manly, 1994)). 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 can then be plotted and clusters in the data visualized; moreover this technique can be used to detect outliers. In its more conventional form, HCA calculates distances (usually Euclidean, but often Mahanalobis or Manhattan) between the objects in either the original data or a derivative thereof (*e.g.* the PCs) and uses these 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. In post-genomics such methods are sometimes referred to as 'guilt-by-association' (Altshuler *et al.*, 2000; Oliver, 2000).

2.2 The Classification/Quantification Variety

These algorithms are based on *supervised* learning (e.g. (Mitchell, 1997; Beavis *et al.*, 2000; Kell and King, 2000; Hastie *et al.*, 2001)) and seek to give answers of biological interest which have much-lower dimensionality, such as "Based on the metabolite fingerprint of this new sample I have just

collected, which class in my database does it (most likely) belong to?" and/or "what are the levels of these metabolites in my biological sample?"

The basic idea behind supervised learning is that there are some patterns (*e.g.* metabolic fingerprints) that have desired responses which are known (*i.e.* whether an animal has been challenged with a drug or placebo). 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 targets (y-data). The goal of supervised learning is to find a *model* or *mapping* that will correctly associate the inputs with the targets (Fig. 1).

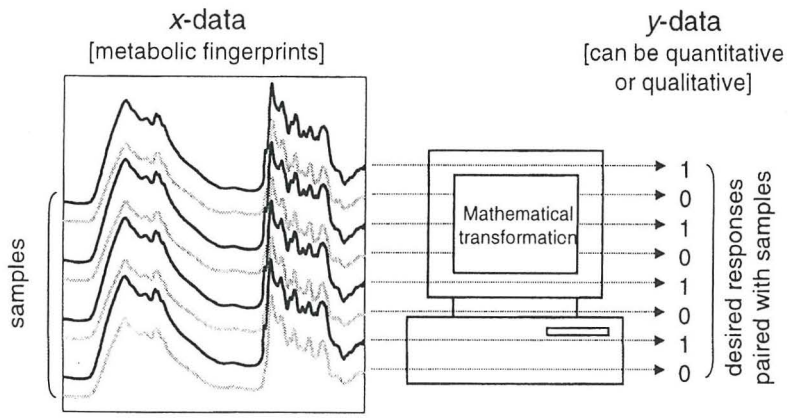

Figure 1. Supervised learning: When we know the desired responses (y-data, or targets) associated with each of the inputs (x-data, or metabolic fingerprints) then the system may be supervised. The goal is to find a mathematical transformation (model) that will correctly associate all or some of the inputs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

Many different algorithms perform supervised learning. Among the most common are (a) discriminant function analysis (DFA), which is a qualitative (categorical), cluster analysis-based method that involves projection of test data into cluster space (Manly, 1994; Radovic *et al.*, 2001), (b) partial least squares (PLS) which is a quantitative linear regression method (Martens and Næs, 1989) and (c) discriminant PLS, a qualitative (categorical) linear regression method (Martens and Næs, 1989; Alsberg *et al.*, 1998). 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 (Werbos, 1994) and radial basis functions (Broomhead and Lowe, 1988; Saha and Keller, 1990; Bishop, 1995). In these supervised learning

techniques there are minimally 4 data sets to be studied, as follows. The "training data" consist of (i) a matrix of s rows and n columns in which s is the number of objects/samples and n the number of variables (the y-data referred to above), and (ii) a second matrix, again consisting of s rows and typically 1 to i columns, in which the columns represent the variable(s) whose value(s) it is desired to know (the y-data or targets) and which for the training set have actually been previously determined by some existing "benchmark" method. The x-data (ii) are always paired with the patterns in the same row in the y-data (i). The "test data" also consist of two matrices, (iii) and (iv), corresponding to those in (i) and (ii) above, but the test set contains different samples. As the name suggests, this second pair is used to test the accuracy of the system; alternatively (and better) they may be used to cross-validate the model. That is to say, after construction of the model using the training set (i, ii) the test data (iii) are then used to challenge the calibration model so as to obtain the model's prediction of results, and these are then compared with the known or expected responses (iv). Once these are within acceptable ranges for the test data then the model is considered to be calibrated and ready to use.

2.3 The Inductive / Mining Variety

These algorithms are also based on *supervised* learning and seek to answer the question 'What have I measured in my metabolic fingerprint that makes samples in class A different from samples in class B?'

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. DFA, PLS, and ANN methods are often perceived as 'black box' approaches to modeling spectra. It is known from the statistical literature that better (*i.e.* more robust) predictions can often be obtained when only the most relevant input variables are considered (Seasholtz and Kowalski, 1993; Kell and Sonnleitner, 1995; Bø and Jonassen, 2002). Thus the best machine learning techniques should not only give the correct answer(s), but also identify a subset of the variables with maximal explanatory power. This can provide an interpretable description of what, in biological terms, is the basis for that answer. Such explanatory modeling methods do exist and are based on rule induction (Breiman *et al.*, 1984; Harrington, 1991; Quinlan, 1993; Alsberg *et al.*, 1997), inductive logic programming (Lloyd, 1987; Muggleton, 1990; King *et al.*, 1992; Lavrac and Dzeroski, 1994), and, in particular, evolutionary computation (Holland, 1992; Koza, 1992; Bäck *et al.*, 1997).

3. EVOLUTIONARY COMPUTATION

Evolutionary computational-based algorithms are particularly popular inductive reasoning and optimization methods (Corne *et al.*, 1999; Michalewicz and Fogel, 2000). They are based on concepts of Darwinian selection (Bäck *et al.*, 1997) to generate and to optimize a desired computational function or mathematical expression that will yield explanatory 'rules'. These techniques include genetic algorithms (GAs; (Goldberg, 1989; Holland, 1992; Michalewicz, 1994; Mitchell, 1995)), evolution strategies (Schwefel, 1995; Beyer, 2001), evolutionary programming (Fogel, 1995, 2000) genetic programming (GP; (Koza, 1992, 1994; Banzhaf *et al.*, 1998; Koza *et al.*, 1999)) and genomic computing (GC; (Kell *et al.*, 2001; Kell, 2002A, 2002b)), and because the models are in English, and can penalize complex expressions, they may be made to be comparatively simple and easily interpreted.

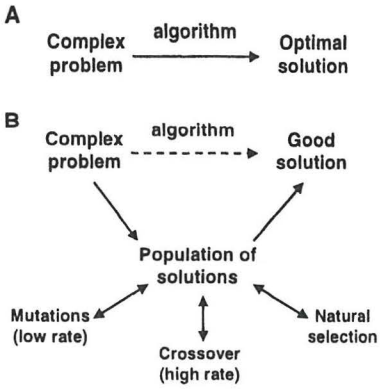

Figure 2. (A) The complex problem we wish to solve but cannot, and (B) the GA strategy.

If we consider the generic "Traveling Salesman Problem" where the object of the exercise is to find the shortest route between 20 cities, with the caveat that one may only visit each city once, we could (a) write down every possible order, (b) compute the distance for each, and (c) pick the shortest one. But is this really feasible? The number of possible orders is factorial and immense, $20! = 2.4 \times 10^{18}$, this number is so big that if your computer could check 1 million orderings every second it would still take 77,000 years to check them all! Thus even though we know how to solve the Traveling Salesman Problem we still cannot do it. This is true for identifying a subset of the variables from a metabolic fingerprint with the globally maximal explanatory power. For example, if we have measured only a modest 200 variables an exhaustive search of all possible permutations (where a variable

is either used or not) is $2^{200} = 1.6 \times 10^{60}$. These problems are NP complete (see Garey and Johnson, 1979); that is to say to find the global optimum requires exhaustive search and this is computationally impossible. Thus route A in Fig. 2 is unfeasible as no algorithm can do this and an alternative strategy needs to be found. The premise is that a 'good' solution is acceptable and so we need an alternative method to search the huge spaces of possible solutions. Importantly, however, if the search space is large but the solution space is small, *i.e.* we can solve the problem with just a small number of variables, the effective search space becomes much narrower. Thus the number of permutations of 4 variables from 200 is just 6.47×10^7. GAs offer such an approach.

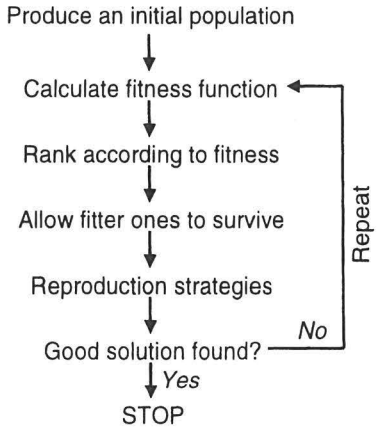

Figure 3. The overall procedure employed by GAs and GP. The criterion for a good solution will be based on setting a threshold error between the known target and the GAs' response.

In a GA a population of individuals, each representing the parameters of the problem to be optimized as a string of numbers or binary digits, undergoes a process analogous to evolution in order to derive an optimal or near-optimal solution (Fig. 2B). 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 asexual and sexual reproduction (Fig. 3).

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 selecting two parents with a frequency 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.

For variable selection prior to some supervised learning method, whether it is linear regression or ANNs, the state of each variable (in GA terminology a gene) is represented by a '1' (selected to be in the model) or a '0' (not selected) (Horchner and Kalivas, 1995; Broadhurst *et al.*, 1997). Together theses sets of variables are called a chromosome, this GA string would be of length m (where m = number of x-data input variables in the metabolic fingerprint). For example, in a variable selection problem starting with 7 variables, one possible chromosome would be 1101001. This can be translated such that variables 1, 2, 4, and 7 are to be used in the modeling process and variables 3, 5, and 6 are to be omitted. Other GA variants based on the selection of spectral windows for FT-IR and Raman spectroscopy are also popular (Williams and Paradkar, 1997; Taylor *et al.*, 1998; Roger and Bellon-Maurel, 2000; Leardi *et al.*, 2002; McGovern *et al.*, 2002).

However, whilst GAs are very successful search algorithms for tackling NP-hard problems, the disadvantage is that with the GA variable selection approach the relationship between one variable and another is not evident, only whether they contribute to a model or not. Therefore, a richer language is needed.

3.1 Genetic Programming

A GP is an application of the GA approach to derive mathematical equations, logical rules or program functions automatically (Koza, 1992, 1994; Gilbert *et al.*, 1997; Langdon, 1998; Koza *et al.*, 1999; Langdon and Poli, 2002). Rather than representing the solution to the problem as a string of parameters, as in a conventional GA, a GP usually (*c.f.* Banzhaf *et al.*, 1998) uses a tree structure. The leaves of the tree, or *terminals*, represent input variables or numerical constants. Their values are passed to *nodes*, at the junctions of branches in the tree, which perform some numerical or program operation before passing on the result further towards the root of the tree (Fig. 4). Genomic Computing (GC; Kell *et al.*, 2001; Kell, 2002a, 2002b) (www.abergc.com) is a variant on a GP.

The overall evolutionary procedure employed by GP is essentially identical to that of GAs. An initial (commonly 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. 5 for pictorial details).

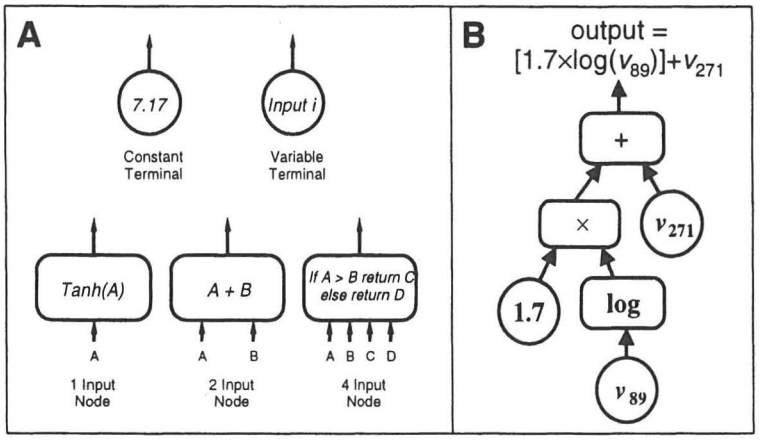

Figure 4. The richer language structure of a tree-encoded GP: (A) the building blocks and (B) a typical function tree.

(1) *Cloning:* some of the original individuals are allowed to survive unmodified.

(2) 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 an input variable.

(3) Alternatively 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 sub-trees are then swapped. At each reproduction stage because of the use of these trees to encode mathematical equations the new trees are still syntactically correct. 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 root mean squared 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 trees gets so huge that it lacks explanatory power (Langdon and Poli, 1998), 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 (Koza, 1992) 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') (Rudolph, 1997).

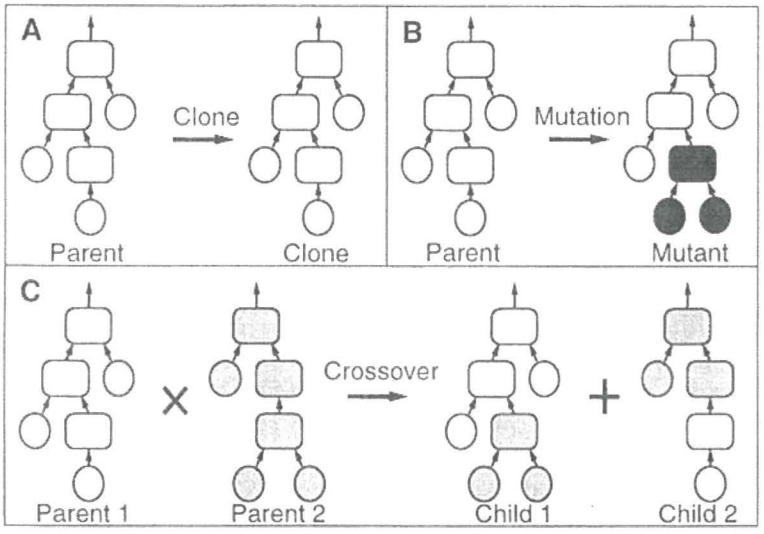

Figure 5. The GP reproduction processes, showing examples of (A) cloning, (B) mutation and (C) crossover events.

4. APPLICATION OF EVOLUTIONARY COMPUTATION-BASED METHODS TO METABOLOMICS

GAs and GPs are very efficient search algorithms and can be used to produce models that allow the deconvolution of metabolome data in chemical terms. Detailed below are five published examples illustrating this.

Example 1 (Goodacre *et al.*, 2000). Members of the genus *Bacillus* are widely distributed in soil, water, and air, and because their spores are so resistant their control is of considerable importance in the food processing industry and in the preparation of sterile products (Doyle *et al.*, 1997). In

addition, the rapid identification of *Bacillus anthracis* spores is of importance because of its potential use as a biological warfare agent (Dando, 1994; Barnaby, 1997). Therefore, there is a need for a generic characterization system that can be used to carry out large-scale and rapid detection of bacterial spores. GP was used to analyze metabolic fingerprints generated from vegetative biomass and spores using Curie-point pyrolysis-MS (Py-MS) and FT-IR. Both fingerprinting approaches could be used to differentiate successfully between vegetative biomass and spores. GP produced mathematical rules that could be interpreted in simple biochemical terms. It was found that for Py-MS, a peak at *m/z* 105 was characteristic and attributable to a pyridine ketonium ion derived from the pyrolysis of pyridine-2,6-dicarboxylic acid (dipicolinic acid), 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, dipicolinic acid, was highly characteristic of spores. Thus, although the original datasets recorded hundreds of spectral variables from whole cells simultaneously, a simple biomarker was detected that can be used for the rapid and unequivocal detection of spores of these organisms.

Example 2 (Johnson *et al.*, 2000). Samples from tomato fruit grown hydroponically under both high- and low-salt conditions were analyzed by FT-IR, with the aim of identifying biochemical features linked to salinity in the growth environment. Examination of the GP-derived trees showed that there were a small number of spectral regions that were consistently used. In particular, the spectral region containing absorbances potentially due to a cyanide/nitrile functional group was identified as discriminatory. Cyanide is formed in plants during ethylene biosynthesis, and ethylene production is enhanced in plants subjected to stress conditions. Therefore, one may propose that plants grown under saline conditions may therefore have enhanced levels of cyanide as a result of enhanced ethylene biosynthesis. Thus inductive reasoning *via* GP has allowed the significance of a pathway turned on under tomatoes exposed to salinity to be highlighted as potentially important. This pathway can now be subjected to conventional biochemical analysis.

Example 3 (McGovern *et al.*, 2002). The previous two examples have been qualitative (*i.e.* the outputs were categorical variables). This example now demonstrates how GA and GP can be used in a quantitative fashion. The ability to control industrial bioprocess is paramount for product yield optimization, and it is imperative therefore that the concentration of the fermentation product (the determinand) is assessed accurately. Whilst IR and Raman spectroscopies have been used for the quantitative analysis of

fermentations (McGovern *et al.*, 1999; Shaw *et al.*, 1999; Vaidyanathan *et al.*, 1999) the transformation of spectra to determinand concentration(s) has usually been undertaken by PLS and ANNs, and so one can not be sure whether the model is detecting the product itself, an increase in bi-products or decrease in substrates. By contrast, GA and GP have 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* which produces the 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.

Example 4 (Ellis *et al.*, 2002). Whilst a number of studies have applied FT-IR to the discrimination and adulteration of meats (Al-Jowder *et al.*, 1999; Downey *et al.*, 2000) its application to the rapid detection of microbial spoilage in meats has only very recently been demonstrated. A particularly robust and reproducible form of this method is attenuated total reflectance (ATR) where the food sample is placed in intimate contact with a crystal of high refractive index and an IR absorbance spectrum, a *metabolic snapshot*, collected in just a few seconds. It has been shown (Ellis *et al.*, 2002) that FT-IR with PLS allowed accurate estimates of bacterial loads (from 10^6 to 10^9 cm^{-2}) to be calculated directly from the chicken surface in 60s, and that GA and GP 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 in agreement with the literature (Dainty, 1996; Nychas and Tassou, 1997).

Example 5 (Kell *et al.*, 2001). Within functional genomics the potential power of evolutionary methods has been shown for the analysis of metabolites from transgenic tobacco plants. Tobacco is a model organism for the study of salicylate biology in plant defense, but despite a considerable amount of research, little is known regarding its synthesis, catabolism, and mode of action. Six week old control plants and a transgenic expressing a bacterial gene encoding the enzyme salicylate hydroxylase (SH-L), which is known to block salicylic acid accumulation in transgenic tobacco (Darby *et al.*, 2000) were inoculated with tobacco mosaic virus and leaf samples were analyzed by HPLC. Genomic Computing analysis of these metabolome profiles identified 3 peaks as highly discriminatory for detecting the presence of the SH-L genotype in the transgenic. One of the peaks was indeed salicylate, but the other two were unknown and are now the subject of further investigation.

5. CONCLUSION

As scientists we are all aware of the cycle of knowledge (Fig. 6) (Kell, 2002b). 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 reasoning the observations considered to be consistent or inconsistent with the hypotheses (Oldroyd, 1986). Actually, although this part is normally only implicit, by a process of induction these observations are synthesized or generalized to refine our accepted wisdom. The cycle then repeats itself until one is happy with the solution to a given problem. However, in the early stages of functional genomics programs we have a scenario where our knowledge is minute, that is to say we have no ideas about the role of an orphan open reading frame and there are few if any hypotheses to test (Brent, 1999; Brent, 2000; Kell and King, 2000). However, we can design experiments based, for example, on gene knockouts and controlled over-expression and observe the effect on the phenotype of the organism.

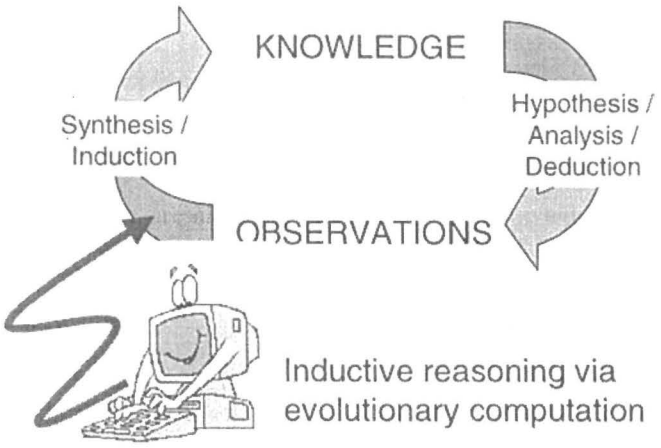

Figure 6. The cycle of knowledge showing where rule induction will play its part.

Metabolomics is one 'omics approach with which one can generate data floods from these genetic manipulations (as indeed are transcriptomics and proteomics, and the same general conclusions given here apply equally to these methods). Thus we are then positioned at the bottom of Fig. 6 where we have collected a great many observations and the trick is to drive the cycle round *via* inductive reasoning to generate new hypotheses. Evolutionary computing methods can be considered to be rule induction

methods that are entirely data-driven and are thus especially appropriate for problems that are data-rich but hypothesis/information-poor. Rule induction by GP and GC can be used to generate rules and hence hypotheses from suitable examples. Of course these new theories will not necessarily be correct, but by testing them new knowledge will be generated which will lead to an increased understanding of the function of the orphan gene. In the new post-genomic biology, then, we shall need good databases (Mendes, 2002), very good data, and even better algorithms, with which to turn our data into knowledge.

ACKNOWLEDGEMENTS

The authors are indebted to the UK BBSRC (Engineering and Biological Systems Committee), the UK EPSRC and the Royal Society of Chemistry for financial support.

REFERENCES

Al-Jowder O, Defernez M, Kemsley EK, Wilson RH. Mid-infrared spectroscopy and chemometrics for the authentication of meat products. *J Agric Food Chem* 47: 3210-3218 (1999).

Allen JK, Davey HM, Broadhurst D *et al.* Metabolic footprinting: a high-throughput, high-information approach to cellular characterisation and functional genomics. *Nature Biotechnol* submitted (2002).

Alsberg BK, Goodacre R, Rowland JJ, Kell DB. Classification of pyrolysis mass spectra by fuzzy multivariate rule induction - comparison with regression, k-nearest neighbour, neural and decision-tree methods. *Anal Chim Acta* 348: 389-407 (1997).

Alsberg BK, Kell DB, Goodacre R. Variable selection in discriminant partial least squares analysis. *Anal Chem* 70: 4126-4133 (1998).

Altshuler D, Daly M, Kruglyak L. Guilt by association. *Nature Genet* 26: 135-137 (2000).

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

Banzhaf W, Nordin P, Keller RE, Francone FD. *Genetic Programming: An Introduction.* Morgan Kaufmann, San Francisco (1998).

Barnaby W. *The Plague Makers: The Secret World of Biolgoical Warfare.* Vision Paperbacks, London (1997).

Beavis RC, Colby SM, Goodacre R *et al.* Artificial intelligence and expert systems in mass spectrometry. In *Encyclopedia of Analytical Chemistry.* Meyers RA (Ed) pp. 11558-11597, John Wiley and Son, Chichester (2000).

Beyer H-G. *The Theory of Evolution Strategies.* Springer, Berlin (2001)

Bishop CM. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford (1995).

Bø TH, Jonassen I. New feature subset selection procedures for classification of expression profiles. http://genomebiologycom/2002/3/4/research/00171 3: research0017.1-0017.11 (2002).

Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees.* Wadsworth Inc, Pacific Grove (1984).

Brent R. Functional genomics: learning to think about gene expression data. *Curr Biol* 9: R338-R341 (1999).

Brent R. Genomic biology. *Cell* 100: 169-183 (2000).

Broadhurst D, Goodacre R, Jones A *et al.* Genetic algorithms as a method for variable selection in PLS regression, with application to pyrolysis mass spectra. *Anal Chim Acta* 348: 71-86 (1997).

Broomhead DS, Lowe D. Multivariable function interpolation and adaptive networks. *Complex Sys* 2: 321-355 (1988).

Chatfield C, Collins AJ. *Introduction to Multivariate Analysis.* Chapman and Hall, London (1980).

Corne D, Dorigo M, Glover F (Ed). *New Ideas in Optimization.* McGraw Hill, London (1999).

Dainty RH. Chemical/biochemical detection of spoilage. *Int J Food Microbiol* 33: 19-33 (1996).

Dando M. *Biological Warfare in the 21st Century.* Brassey's Ltd., London (1994).

Darby RM, Maddison A, Mur LAJ *et al.* Cell specific expression of salicylate hydroxylase in an attempt to separate localised HR and systemic signalling establishing SAR in tobacco. *Plant Mol Pathol* 1: 115-124 (2000).

Downey G, McElhinney J, Fearn T. Species identification in selected raw homogenized meats by reflectance spectroscopy in the mid-infrared, near-infrared, and visible ranges. *Appl Spectr* 54: 894-899 (2000).

Doyle MP, Beuchat LR, Montville TJ (Ed) *Food Microbiology: Fundamentals and Frontiers.* American Society of Microbiology Press, Washington DC (1997).

Duda RO, Hart PE, Stork DE. *Pattern Classification.* 2nd Edn. John Wiley and Sons, London (2001).

Ellis DI, Broadhurst D, Kell DB *et al.* Rapid and quantitative detection of the microbial spoilage of meat using FT-IR spectroscopy and machine learning. *Appl Env Microbiol* 68: 2822-2828 (2002).

Everitt BS. *Cluster Analysis.* Edward Arnold, London (1993).

Fell DA. *Understanding the Control of Metabolism.* Portland Press, London (1996).

Fiehn O. Metabolomics – the link between genotypes and phenotypes. *Plant Mol Biol* 48: 155–171 (2002).

Fiehn O, Kloska S, Altmann T. Integrated studies on plant biology using multiparallel techniques. *Curr Opin Biotechnol* 12: 82-86 (2001).

Fiehn O, Kopka J, Dörmann P *et al.* Metabolite profiling for plant functional genomics. *Nature Biotechnol* 18: 1157-1161 (2000a).

Fiehn O, Kopka J, Trethewey RN, Willmitzer L. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal Chem* 72: 3573-3580 (2000b).

Fogel DB. A comparison of evolutionary programming and genetic algorithms on selected constrained optimization problems. *Simulation* 64: 397-404 (1995).

Fogel DB. *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence.* IEEE Press, Piscataway (2000).

Garey M, Johnson D. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* Freeman, San Francisco (1979).

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

Goldberg DE. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading (1989).

Goodacre R, Neal MJ, Kell DB. Quantitative analysis of multivariate data using artificial neural networks: a tutorial review and applications to the deconvolution of pyrolysis mass spectrtra. *Z Bakteriol* 284: 516-539 (1996).

Goodacre R, Shann B, Gilbert R *et al.* The detection of the dipicolinic acid biomarker in *Bacillus* spores using Curie-point pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. *Anal Chem* 72: 119-127 (2000).

Goodacre R, Timmins ÉM, Burton R *et al.* Rapid identification of urinary tract infection bacteria using hyperspectral, whole organism fingerprinting and artificial neural networks. *Microbiol* 144: 1157-1170 (1998).

Harrington PB. Fuzzy rule-building expert systems: minimal neural networks. *J Chemometrics* 5: 467-486 (1991).

Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer-Verlag, Berlin (2001).

Heinrich R, Schuster S. *The Regulation of Cellular Systems*. Chapman and Hall, New York (1996).

Holland JH. *Adaption in Natural and Artificial Systems*. MIT Press, Cambridge (1992).

Horchner U, Kalivas JH. Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection. *Anal Chim Acta* 311: 1-13 (1995).

Johnson HE, Gilbert RJ, Winson MK *et al.* Explanatory analysis of the metabolome using genetic programming of simple, interpretable rules. *Genet Program Evolv Mach* 1: 243-258 (2000).

Jolliffe IT. *Principal Component Analysis*. Springer-Verlag, New York (1986).

Kell DB. Defence against the flood: a solution to the data mining and predictive modelling challenges of today. *Bioinformatics World* (part of Scientific Computing News) Issue 1: 16-18 (2002a) http://www.abergc.com/biwpp16-18_as_publ.pdf.

Kell DB. Genotype-phenotype mapping: genes as computer programs. *Trends Genet* in press (2002b).

Kell DB, Darby RM, Draper J. Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. *Plant Phys* 126: 943-951 (2001).

Kell DB, King RD. On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. *Trends Biotechnol* 18: 93-98 (2000).

Kell DB, Mendes P. Snapshots of systems: metabolic control analysis and biotechnology in the post-genomic era. In *Technological and Medical Implications of Metabolic Control Analysis*. Cornish-Bowden A, Cárdenas ML (Ed) pp. 3-25, Kluwer Academic Publishers, Dordrecht (2000) (see http://qbab.aber.ac.uk/dbk/mca99.htm).

Kell DB, Sonnleitner B. GMP - Good Modelling Practice: an essential component of Good Manafacturing Practice. *Trends Biotechnol* 13: 481-492 (1995).

Kell DB, Westerhoff HV. Towards a rational approach to the optimization of flux in microbial biotransformations. *Trends Biotechnol* 4: 137-142 (1986).

King RD, Muggleton S, Lewis RA, Sternberg MJE. Drug design by machine learning - the use of inductive logic programming to model the structure-activity-relationships of trimethoprim analogs binding to dihydrofolate-reductase. *Proc Natl Acad Sci USA* 89: 11322-11326 (1992).

Koza JR. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, Cambridge (1992).

Koza JR. *Genetic Programming II: Automatic Discovery of Reusable Programs*. MIT Press, Cambridge (1994).

Koza JR, Bennett FH, Keane MA, Andre D. *Genetic Programming III: Darwinian Invention and Problem Solving*. Morgan Kaufmann, San Francisco (1999).

Langdon WB. *Genetic Programming and Data Structures: Genetic Programming + Data Structures = Automatic Programming*! Kluwer Academic Publishers, Boston (1998).

Langdon WB, Poli R. Fitness causes bloat: mutation. In *Proc First European Workshop on Genetic Programming*. Vol. 1391. Banzhaf W, Poli R, Schoenauer M, Fogarty TC (Ed) pp. 37-48, Springer-Verlag, Berlin (1998).

ftp://ftp.cwi.nl/pub/W.B.Langdon/papers/WBL.euro98_bloatm.ps.gz.

Langdon WB, Poli R. *Foundations of Genetic Programming*. Springer-Verlag, Berlin (2002).

Lavrac N, Dzeroski S. *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, Chichester (1994).

Leardi R, Seasholtz MB, Pell RJ. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. *Anal Chim Acta* 461: 189-200 (2002).

Lindon JC, Nicholson JK, Holmes E, Everett JR. Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids. *Concepts Magn Reson* 12: 289-320 (2000).

Lloyd JW. *Foundations of Logic Programming*. Springer-Verlag, Berlin (1987).

Manly BFJ. *Multivariate Statistical Methods: A Primer*. Chapman and Hall, London (1994).

Martens H, Næs T. *Multivariate Calibration*. John Wiley and Sons, Chichester (1989).

McGovern AC, Broadhurst D, Taylor J *et al.* Monitoring of complex industrial bioprocesses for metabolite concentrations using modern spectroscopies and machine learning: application to gibberellic acid production. *Biotechnol Bioeng* 78: 527-538 (2002).

McGovern AC, Ernill R, Kara BV *et al.* Rapid analysis of the expression of heterologous proteins in *Escherichia coli* using pyrolysis mass spectrometry and Fourier transform infrared spectroscopy with chemometrics: application to α2-interferon production. *J Biotechnol* 72: 157-167 (1999).

Mendes P. Emerging bioinformatics for the metabolome. *Briefings Bioinformat* 3: 134-45 (2002).

Mendes P, Kell DB, Westerhoff HV. Why and when channeling can decrease pool size at constant net flux in a simple dynamic channel. *Biochim Biophys Acta* 1289: 175-186 (1996).

Michalewicz Z. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, Berlin (1994).

Michalewicz Z, Fogel DB. *How to Solve It: Modern Heuristics*. Springer-Verlag, Heidelberg (2000).

Mitchell M. *An Introduction to Genetic Algorithms*. MIT Press, Boston (1995).

Mitchell TM. *Machine Learning*. McGraw Hill, New York (1997).

Muggleton SH. Inductive logic programming. *New Generation Comput* 8: 295-318 (1990).

Nychas GJE, Tassou CC. Spoilage processes and proteolysis in chicken as detected by HPLC. *J Sci Food Agric* 74: 199-208 (1997).

Oldroyd D. *The Arch of Knowledge: An Introduction to the History of the Philosophy and Methodology of Science*. Methuen, New York (1986).

Oliver SG. Proteomics: guilt-by-association goes global. *Nature* 403: 601-603 (2000).

Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. *Trends Biotechnol* 16: 373-378 (1998).

Quinlan JR. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo (1993).

Raamsdonk LM, Teusink B, Broadhurst D *et al.* A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnol* 19: 45-50 (2001).

Radovic BS, Goodacre R, Anklam E. Contribution of pyrolysis mass spectrtrometry (Py-MS) to authenticity testing of honey. *J Anal Appl Pyrolysis* 60: 79-87 (2001).

Roger JM, Bellon-Maurel V. Using genetic algorithms to select wavelengths in near-infrared spectra: application to sugar content prediction in cherries. *Appl Spectr* 54: 1313-1320 (2000).

Rudolph G. *Convergence Properties of Evolutionary Algorithms.* Verlag Dr Kovac, Hamburg (1997).

Saha A, Keller JD. Algorithms for better representation and faster learning in radial basis functions. In *Advances in Neural Information Processing Sytems.* Vol. 2. Touretzky D (Ed) pp. 482-489, Morgan Kaufmann, San Mateo (1990).

Schwefel H-P. *Evolution and Optimum Seeking.* John Wiley and Sons, New York (1995).

Seasholtz MB, Kowalski B. The parsimony principle applied to multivariate calibration. *Anal Chim Acta* 277: 165-177 (1993).

Shaw AD, Kaderbhai N, Jones A *et al.* Non-invasive, on-line monitoring of the biotransformation by yeast of glucose to ethanol using dispersive Raman spectroscopy and chemometrics. *Appl Spectr* 53: 1419-1428 (1999).

Tukey JW. *Exploratory Data Analysis.* Addison-Wesley, Reading (1977).

Vaidyanathan S, Kell DB, Goodacre R. Flow-injection electrospray ionization mass spectrometry of crude cell extracts for high-throughput bacterial identification. *J Am Soc Mass Spectrom* 13: 118-128 (2002).

Vaidyanathan S, Macaloney G, McNeill B. Fundamental investigations on the near-infrared spectra of microbial biomass as applicable to bioprocess monitoring. *Analyst* 124: 157-162 (1999).

Vaidyanathan S, Rowland JJ, Kell DB, Goodacre R. Rapid discrimination of aerobic endospore-forming bacteria via electrospray-ionisation mass spectrometry of whole cell suspensions. *Anal Chem* 73: 4134-4144 (2001).

Werbos PJ. *The Roots of Back-Propagation: From Ordered Derivatives to Neural Networks and Political Forecasting.* John Wiley and Sons, Chichester (1994).

Westerhoff HV, Kell DB. What BioTechnologists knew all along...? *J Theor Biol* 182: 411-420 (1996).

Wilkinson L. *The Grammar of Graphics.* Springer-Verlag, New York (1999).

Williams RR, Paradkar RP. Correcting fluctuating baselines and spectral overlap with genetic regression. *Appl Spectr* 51: 92-100 (1997).

Winson MK, Goodacre R, Woodward AM *et al.* 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).