DOI: 10.1007/s11306-005-1106-4

A metabolome pipeline: from concept to data to knowledge

Marie Brown, Warwick B. Dunn, David I. Ellis, Royston Goodacre, Julia Handl, Joshua D. Knowles, Steve O′Hagan, Irena Spasić, and Douglas B. Kell*

School of Chemistry, The University of Manchester, Faraday Building, Sackville St, PO Box 88 Manchester, M60 1QD

Received 24 August 2004; accepted 27 August 2004

Metabolomics, like other omics methods, produces huge datasets of biological variables, often accompanied by the necessary metadata. However, regardless of the form in which these are produced they are merely the ground substance for assisting us in answering biological questions. In this short tutorial review and position paper we seek to set out some of the elements of "best practice" in the optimal acquisition of such data, and in the means by which they may be turned into reliable knowledge. Many of these steps involve the solution of what amount to combinatorial optimization problems, and methods developed for these, especially those based on evolutionary computing, are proving valuable. This is done in terms of a "pipeline" that goes from the design of good experiments, through instrumental optimization, data storage and manipulation, the chemometric data processing methods in common use, and the necessary means of validation and cross-validation for giving conclusions that are credible and likely to be robust when applied in comparable circumstances to samples not used in their generation.

KEY WORDS: Metabolomics; chemometrics; data processing; databases; machine learning; genetic algorithms; genetic programming; evolutionary computing.

1. Introduction

"Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house"

Jules Henri Poincaré (1854–1912)
La Science et l'hypothése.

Since the systematic genome sequencing of the first free-living microbe (Fleischmann *et al.*, 1995), we have seen the advent of genome-wide expression profiling methods, the 'omes', in which investigators seek to understand complex biological systems on a large scale. The macromolecular omes (especially the transcriptome and proteome) were the first to gain widespread attention. However, the metabolome, the complete set of metabolites in a cell or tissue (for definitions see (Fiehn, 2001; Goodacre *et al.*, 2004)), consists of low-molecular weight chemical intermediates (Oliver *et al.*, 1998) which can be considered to be the end products of gene expression. For fundamental reasons based on metabolic control analysis http://dbk.ch.umist.ac.uk/mca_home.htm (reviews: (Kell and Westerhoff, 1986; Fell, 1996; Heinrich and Schuster, 1996)), we can expect that while change in gene (protein) expression levels will have only small effects on metabolic *fluxes*, they must have large effects on metabolite *concentrations*. Metabolomics thus represents an ideal level at which to analyse change in biological system sensitively (Harrigan and Goodacre, 2003), under conditions in which there may be negligible effects on the gross phenotype (Cornish-Bowden and Cárdenas, 2001; Raamsdonk *et al.*, 2001). Qualitative and quantitative metabolome analyses also provide a view of the biochemical status of an organism under specific conditions. For this reason increasing interest has been shown in the use of metabolomics for functional genomics, sometimes in parallel with transcriptomics and proteomics. Metabolomic data have been generated:

- For a wide variety of organisms – e.g. human (Lindon *et al.*, 2000; Fiehn and Spranger, 2003), microbial (Raamsdonk *et al.*, 2001; Kaderbhai *et al.*, 2003) and plant (Fiehn *et al.*, 2000; Roessner *et al.*, 2000),
- using a number of different approaches – metabolic profiling (Fiehn *et al.*, 2000), fingerprinting (Aharoni *et al.*, 2002; Johnson *et al.*, 2003) and footprinting (Allen *et al.*, 2003; Kaderbhai *et al.*, 2003; Allen *et al.*, 2004),
- using a number of different analytical techniques (reviews: (Sumner *et al.*, 2003; Kell, 2004))
- for many applications e.g. toxicity determination (Lindon *et al.*, 2003b; Nicholson and Wilson, 2003), diagnostics (Brindle *et al.*, 2002), gene function

*To whom correspondence should be addressed.
E-mail: dbk@umist.ac.uk; http://dbk.ch.umist.ac.uk

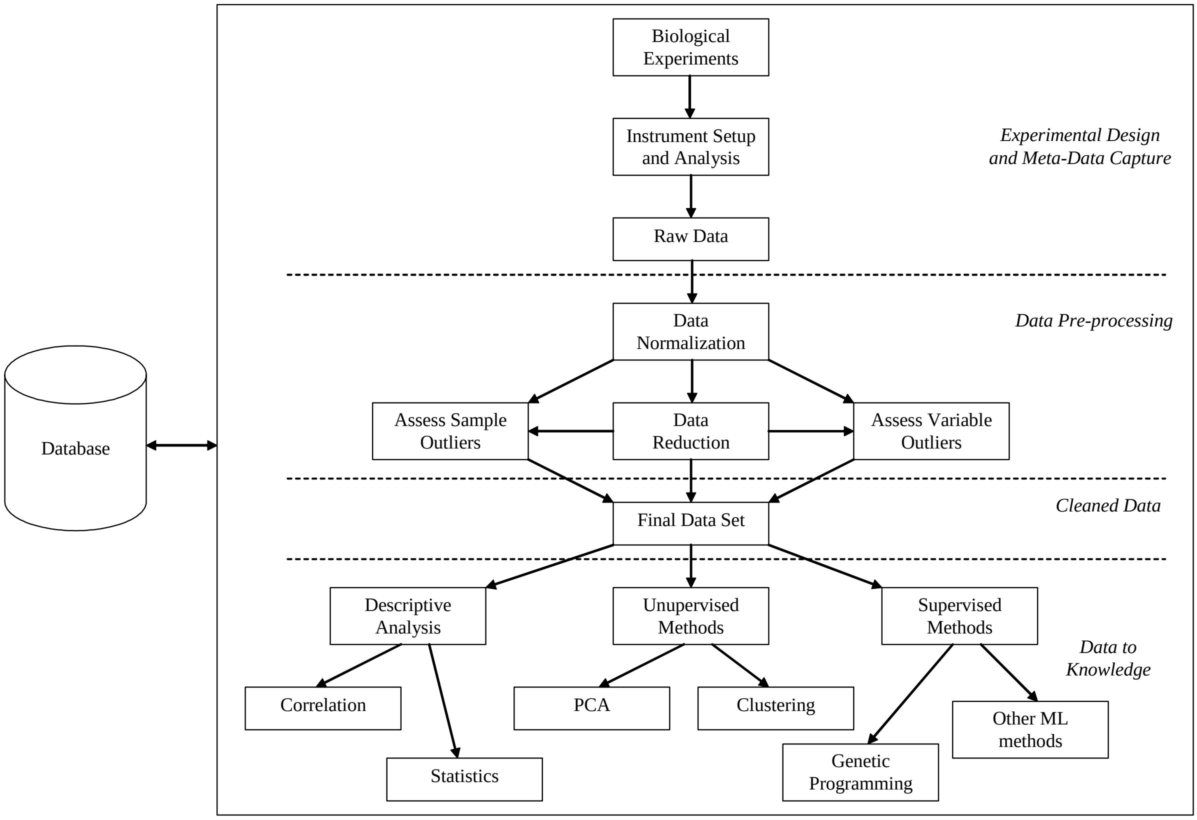


Figure 1. An overview of a pipeline for the design, performance, storage and analysis of metabolomics experiments and their attendant data.

determination (Raamsdonk *et al.*, 2001; Allen *et al.*, 2003) and in discriminating genotypes (Taylor *et al.*, 2002).

A number of reviews have discussed in detail the variety of analytical techniques and data collection and storage methods available (Fiehn, 2002; Mendes, 2002; Fiehn and Weckwerth, 2003; Fernie, 2003; Hardy and Fuell, 2003; Sumner *et al.*, 2003; Goodacre *et al.*, 2004; Kell, 2004). However, we know of no article that seeks to set out in a systematic manner the detailed way in which we might best seek to turn metabolomic data into biological knowledge. In seeking to bring together elements of best practice in this emerging discipline, we therefore offer a tutorial review, based on our own experiences over more than 10 years (Goodacre *et al.*, 1992; Goodacre *et al.*, 1993; Goodacre and Kell, 1993), of the numerical issues that one faces when conducting metabolomics experiments.

We present this here in the form of a metabolome pipeline (figure 1) which addresses the need to have a streamlined approach for data collection, storage, analysis and validation to convert the raw data into useful knowledge whilst recognizing the wide-ranging methods and approaches that are used in this area. Assuming a good experimental design and some attempt to optimize the instrumentation (which we describe briefly), the first stage in converting the data to knowledge is cleaning up the raw data and where possible relating signals to metabolites (preferably providing a chemical identity for metabolites judged to be present).

Subsequent to this is using the metabolomic data to answer biological questions, including reconstructing the metabolic networks in which they participate. The knowledge thus generated must then be available to be combined with and compared to other metabolome data, as well as integrated with transcriptome and proteome data to help build towards an understanding of the whole system operating within an organism.

2. Experimental design

The starting point in measuring the metabolome is the experimental design. This is often neglected, but the high dimensionality of omics data means that it needs especial attention. Good standard texts include (Hicks and Turner, 1999; Montgomery, 2001; Myers and Montgomery, 1995) while (Bland, 1987; Bradford Hill and Hill, 1991; Schlesselman, 1982) have a more medically oriented outlook. Textbooks of epidemiology (Rothman, 2002; Rothman and Greenland, 1998) also give valuable advice. Many researchers assume normality of statistical distributions in omics data, and this is often not found. Nowadays more or less user-friendly software allows one to effect robust experimental designs. We tend to use DesignExpert (http://www.statease.com/) for basic experimental design and response surface modelling, and nQuery Advisor (http://www.statsol.ie/nquery/nquery.htm) for statistical power calculations in case-control studies.

As well as establishing the type of variation in the independent variables that the experimenter requires, and how to optimize this within the constraints of limited experimental material, resources and time, specific consideration needs to be given to the following features:

- Biological variation, arising from variations in metabolite levels between samples of the same species grown under identical conditions. It varies from organism to organism and tissue to tissue and for plants (for instance) can be very large (Roessner *et al.*, 2000), although fortunately the machine variability is smaller (Fiehn *et al.*, 2000). In animals there are significant diurnal (Lenz *et al.*, 2003) and dietary (Solanky *et al.*, 2003) variations, which need to be appreciated when samples are collected,
- sample preparation – as standard and reproducible an approach as possible is required. This may be simple for the case of biofluids, when little or no sample preparation is needed, but may be much more complex. For high-throughput analyses simple methods need to be available,
- analytical variance in terms of the relative standard deviation of a specific experimental approach,
- the requirement for suitable controls or references,
- the type and number of internal standard(s),
- the range and sensitivity of the analytical method (overall sensitivity depends on both sample preparation/preconcentration and instrument operation),
- the number of samples and replicates to be analysed, and whether high-throughput or otherwise.

These form the basis on which the number of samples, analytical and biological replicates required are decided.

3. Instrumentation optimization

One area, related to experimental design, that in our view has not enjoyed the attention it deserves, is that of instrumental optimization. For mass spectrometers this is sometimes referred to as "tuning", and it is usually assumed that this has been done satisfactorily. In fact this is not (and cannot be) the case. If an electrospray mass spectrometer (or any other instrument) has 14 tuning parameters (e.g. the cone voltage, gas flow rates, or whatever), each of which may take just 10 values, the number of combinations of settings is then 10^{14} (and the lifetime of the Universe in seconds is 10^{17} (Barrow and Silk, 1995)). Obviously experimenters are not going to try all these combinations (this is known as "exhaustive search"), and a "heuristic" method (Reeves, 1995; RaywardSmith *et al.*, 1996; Dasgupta *et al.*, 1999; Michalewicz and Fogel, 2000), in which good but not provably optimal solutions are sought, is therefore appropriate. Evolutionary algorithms (figure 2) are especially well suited for optimization purposes (Bäck *et al.*, 1997; Corne *et al.*, 1999) and have been used in related experimental design problems such as fermentation medium optimization (Weuster-Botz and Wandrey, 1995; Davies *et al.*, 2000). Using them, we have found that the analytical performance of modern instruments such as electrospray (Vaidyanathan *et al.*, 2003; Vaidyanathan *et al.*, 2004) and GC-TOF-MS (O'Hagan *et al.*, in press) can be improved hugely. An example is given in figure 3.

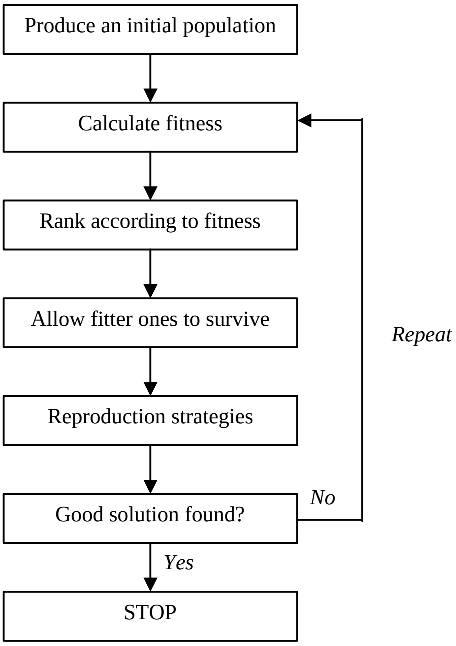

Produce an initial population
Calculate fitness
Rank according to fitness
Allow fitter ones to survive
Reproduction strategies
Good solution found?
No
Repeat
Yes
STOP

Figure 2. The basic evolutionary algorithm strategy, by which candidate solutions are evolved iteratively, using a selection step which tends to favour the better solutions in a particular generation, and mutation and recombination operators are used to create diversity within the population of candidate solutions.

4. Data gathering

The chemical complexity and range of primary and secondary metabolites present in microbial, plant and animal organisms massively exceed the comparatively limited building blocks of the transcriptome (4–5 nucleotides and derivatives), and the proteome (~20 primary amino acids). Thus, although the resulting mRNA and proteins are themselves complex the number of analytical methods used to measure them is relatively small, although the analysis of post translational modification on proteins is still problematic. By contrast, the chemical properties of metabolites range from ionic inorganic species to hydrophilic carbohydrates and sophisticated secondary natural products to hydropho-

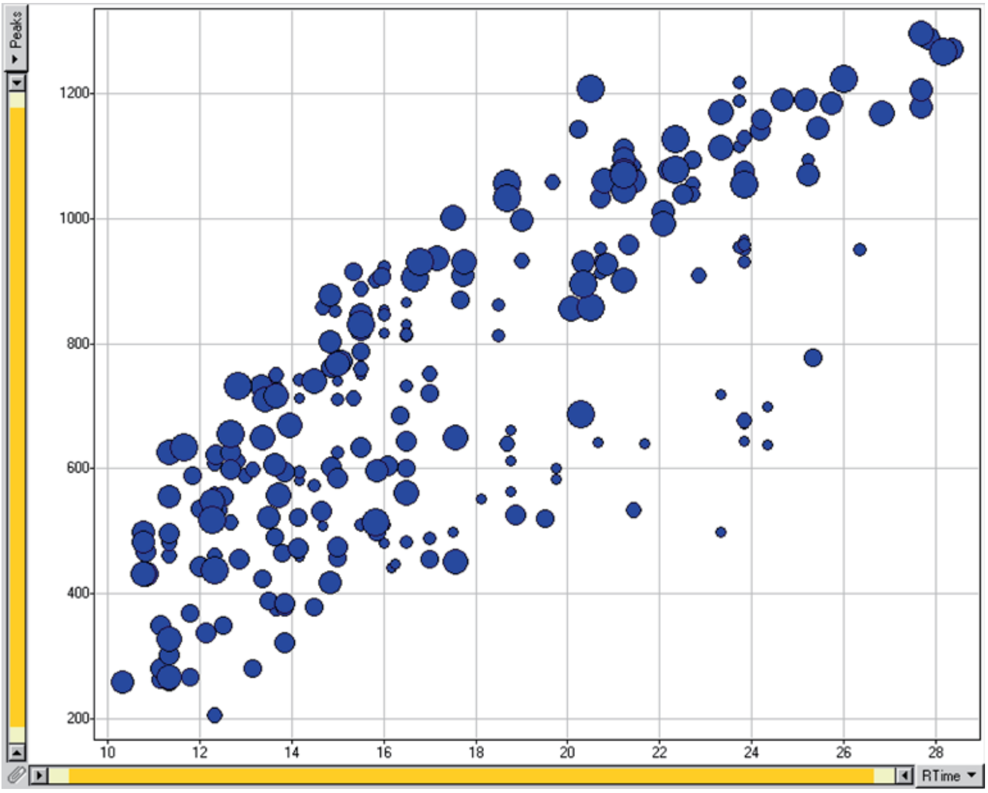
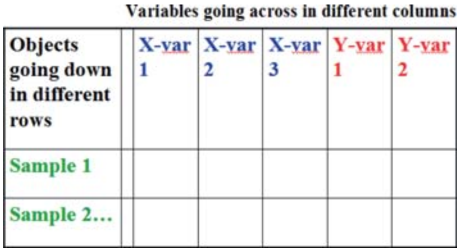


Figure 3. Optimisation of a GC-TOF instrument using closed loop evolutionary computing (O'Hagan *et al.*, in press). The figure shows 120 experiments in which both the number of peaks detected (maximized) and the run time (minimized) (both shown), as well as the signal:noise (not shown) were optimised. The generation is encoded via the size of the symbol. There is an obvious trade-off, and we settled on the conditions giving a run time of just over 20 min and just over 1200 peaks.

bic lipids – with a wide range of polarity, solubility and volatility. It is thought that currently no analytical method can fully identify the metabolome and thus the method chosen in each experiment essentially only targets a subset of the full complement of metabolites. Fingerprinting methods such as NMR (Lindon *et al.*, 2000; Raamsdonk *et al.*, 2001; Nicholson *et al.*, 2002; Lindon *et al.*, 2003a; Nicholson and Wilson, 2003), FTIR (Oliver *et al.*, 1998; Ellis *et al.*, 2003; Goodacre, 2003; Kaderbhai *et al.*, 2003) and pyrolysis (Goodacre *et al.*, 1992; Goodacre and Kell, 1993; Goodacre and Kell, 1996) or direct-injection mass spectrometry (Goodacre *et al.*, 2002; Allen *et al.*, 2003; Kaderbhai *et al.*, 2003; Allen *et al.*, 2004) provide high-dimensional inputs for classification methods, but rarely allow one to identify the chemical bases for their classifications. Thus, the wide chemical complexity of the metabolome means that extraction and separation methods of some kind are normally required, usually involving selective detection (so-called hyphenated (Wilson and Brinkman, 2003) techniques such as GC-MS, LC-MS and so on) and since chromatographic separations are often imperfect there is necessarily a data-deconvolution step. Another issue, which follows from the "amplification" of metabolite concentration changes relative to those of proteins (see above) is the very large dynamic range of metabolites (e.g. pM to mM) that may be important; no available instrument can presently cover such a range without differential dilution, although four orders of magnitude are possible in favourable circumstances.



Variables going across in different columns

Objects going down in different rows	X-var 1	X-var 2	X-var 3	Y-var 1	Y-var 2
Sample 1					
Sample 2…					

Figure 4. A propositional approach to describing and using metabolomics data (the x-data) for analyzing complex systems. These may have other specific properties (the y-data) which one may also wish to 'explain' in terms of the x-data.

The outcome of these stages is a full-rank matrix of peak number(/identity), wavenumber, chemical shift, m/z and so on (the so-called x-variables) versus sample number, together with associated metadata such as gender, age, physiological traits, disease status and so on, in the form given in figure 4. Some of the metadata will typically be characters that one is interested in predicting, and these are known as the y-data. In some cases the y-data class membership of the samples is known, and in this case one may exploit supervised learning methods (see below).

5. Data analysis

Irrespective of the analytical technique used, the analysis of the data is essentially performed in three stages. Initially the raw data need to be preprocessed to convert them to a suitable form. Secondly it may be useful to subject these modified data to data reduction so that only the most relevant input variables are used in the subsequent data analysis (Seasholtz and Kowalski, 1993). Some methods used in these two stages are listed in table 1 below.

Each of these processes requires very careful thought, since when we are dealing with data containing hundreds of variables (dimensions) the "knock-on" effects of each numerical data processing step are simply not transparent, and the conclusions drawn should preferably be robust to the type of data pre-processing steps used. Thus normalizing to a constant total signal (to take into account varying sample sizes) introduces dependencies between the variables that would not exist without this step. Similarly how one treats missing variables can have significant effects on the position of individual samples in clustering diagrams. Missing values may arise because they are below the limit of detection (in which case it is reasonable to assign a value of zero), or because they were not collected. Deconvolution and further processing of hyphenated data to establish the contribution of each eluting component is a very difficult and active area, which needs to begin by "registering" (Woodward *et al.*, 2004) or aligning datasets (Duran *et al.*, 2003). Automating this reliably is a high priority for metabolomics.

The objective of the third stage of the data analysis is to find patterns within the data which give useful biological information that can be used to *generate* hypotheses that can be further tested and refined. For the metabolome because the biological differences between samples sometimes arises from comparatively small differences in many metabolite concentrations, recognizing the pattern and interpreting it is not straightforward. The methods available for metabolome analysis can be placed in four main (and partly overlapping) categories – univariate and multivariate statistical, unsupervised learning (which looks at the overall pattern or structure of the data), supervised learning (which uses known information to help guide the classification of the data (Duda *et al.*, 2001; Hastie *et al.*, 2001)), and system-based analyses which use theories such as MCA (Fell, 1996) to help interpret the data in terms of the biological networks that generated them (Kell, 2004). Many unsupervised learning methods are equivalent to clustering methods and are often statistically based, while supervised methods come in many varieties (Weiss and Kulikowski 1991; Michie *et al.*, 1994; Mitchell, 1997), including statistical, neural, rule-based, evolutionary and so on. Listed in table 2 are examples of a variety of methods, together with selected references.

6. Univariate and multivariate statistical methods

Before looking into the more complex multivariate methods, it is always desirable to look at the statistical properties (mean and variance) of individual metabolites and the relations between them (and each other) and the other measured properties of interest. Since there are n^2 correlations for n metabolites these can be quite difficult to analyse if the process is not automated. Unusual variances may be due to specific outliers, and these need to be assessed and if necessary removed before sensible conclusions can be drawn.

Although classical multivariate statistics based on the analysis of variance (ANOVA) continues to be the method of choice in many fields, especially experimental medicine, its philosophy is really quite different from that which underpins the omics revolution (Kell and Oliver, 2004). This is very nicely set out in a paper (entitled "Statistical modeling: the two clusters") by the distinguished statistician Leo Breiman (Breiman, 2001b), in which he points out that statistics assume an underlying model, including distributions of properties, and assesses the goodness of fit to the model, while machine learning methods make no such assumptions and use the data to *determine* the best models – so-called non-parametric approaches to modelling. The utility of the models is then assessed using cross-validation methods (see below). Though these "two cultures" sound similar, they are in fact profoundly different in their basis, purpose, implementation and performance, our prejudice being for the unbiased nature of the latter.

Table 1
Methods of data preprocessing and reduction

Data preprocessing	Data reduction
Normalization of data – data transforms	Limiting the data analysis to a specified range of the processed data
Normalization of data – using internal standard(s)	Excluding variables or samples whose variation within replicates is outside the allowable analytical limits
Deconvolution of peaks	Excluding sample outliers, identified e.g. by PCA
Addressing baseline shifts and machine drift	
Dealing with missing values	

Table 2
An overview of data analytical methods

Univariate and multivariate Statistical	Unsupervised Learning	Supervised Learning	"Theory-based"
Mean	PCA (Jolliffe, 1986)	Discriminant Analysis (Fisher, 1951)	Metabolic Control Analysis (MCA) (Fell, 1996)
Standard Deviation	Clustering (Everitt, 1993; Duda et al., 2001; Hastie et al., 2001)	(Discriminant) Partial Least squares (Martens and Næs, 1989)	Bayesian belief networks (Bernardo and Smith, 2000; Berry, 1996; Leonard and Hsu, 1999; Ramoni and Sabastini, 1998)
% Coefficient of Variation	Self-organising maps (Kohonen, 1989)	Artificial Neural Nets (Bishop, 1995; Ripley, 1996)	
Correlation and regression (Flury and Riedwyl, 1988)	Auto-associative neural networks to effect non-linear PCA (Kramer, 1991)	Rule Induction (Breiman, 2001a; Brieman et al., 1984; Quinlan, 1993)	
Mutual information (Shannon and Weaver, 1949; Battiti, 1994; Gilbert et al., 1997)		Inductive logic programming (Muggleton, 1990)	
		Evolutionary computation (Bäck et al, 1997)	

7. Clustering methods

In the absence of sufficient training data for supervised methods, the application of unsupervised techniques, in particular clustering methods, becomes necessary. Clustering algorithms take as their input a set of objects typically represented as feature vectors, where each vector describes some measured property (e.g. the intensity in a spectrum sampled at N points), and aim to assign each of these vectors to a group, such that those placed in the same group are more similar to each other than to those placed in different groups. "Similarity" here, essentially means proximity in the multidimensional feature space. To measure proximity, clustering algorithms make use of one of a number of distance functions, e.g. Euclidean distance, Mahalanobis distance, cosine distance, etc. (Jain *et al.*, 1999).

This loose but intuitive concept of clustering can be quite difficult to realize in practice. One reason for this is the difficulty, even for humans, to establish unambiguously the clusters that exist within a data set (see figure 5). Secondly, even in cases where an umambiguous partitioning of the data would be possible, clustering algorithms can fail drastically. This is because most existing clustering techniques rely on estimating the quality of a particular partitioning by means of just one internal evaluation function (an objective function that measures intrinsic data properties such as the spatial separation between clusters or their compactness). Hence, the internal evaluation function is assumed to reflect the quality of the partitioning reliably, an assumption that may be violated for certain data sets (Estivill-Castro, 2002). However, given that many objective functions for clustering are complementary, the simultaneous optimisation (e.g. by means of multi-objective evolutionary algorithms) of several of these objectives can help to overcome this problem and ensure a robust algorithm performance (Handl and Knowles, 2004).

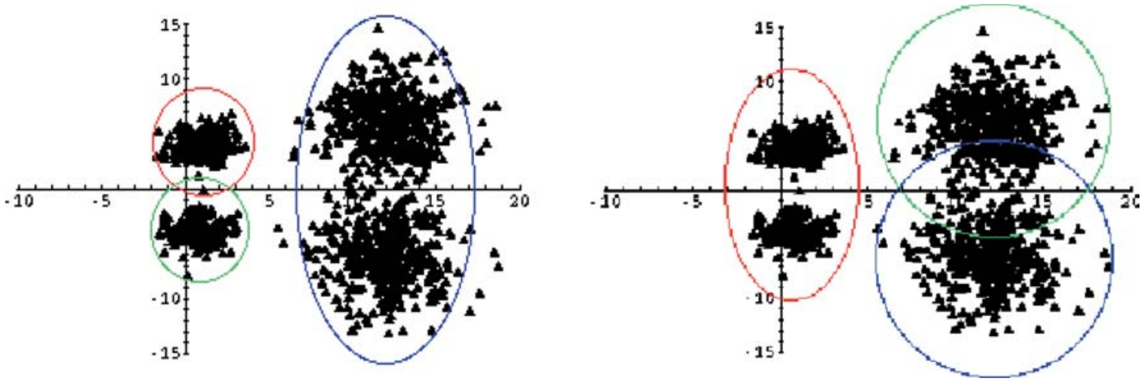


Figure 5. Two possible partitions of the same data set. Dependent on the optimisation criterion either one could be considered of better quality. Different clustering algorithms will produce differing results.

A third fundamental difficulty of clustering is the determination of the number of clusters in the data set. Most existing algorithms require this parameter to be provided, which is a major problem in a setting where the structure of the data is completely unknown. Whilst there have been recent attempts to determine the number of clusters automatically (including the Gap statistic (Tibshirani *et al.*, 2001), Resampling (Dudoit and Fridlyand, 2002) and others (De Smet *et al.*, 2002)), no entirely reliable method exists to date.

In general, the application of clustering algorithms for data analysis requires a careful analysis of the results produced. A major problem with traditional clustering algorithms (e.g. *k*-means, hierarchical algorithms and self-organizing maps) is the fact that *they return a partitioning without any estimate of the reliability of this result*. Indeed many force all points to be in at least one cluster, whether this is justified or not. In order to obtain an acceptable degree of confidence in a clustering result, it is useful to cluster the data repeatedly, using a randomized algorithm (like *k*-means with different initialization), different algorithms, or resampling of the data (in which a subset of the data only is used as input), and observe the stability of the partitioning with respect to these changes (this is commonly used in phylogenetics, and known as bootstrapping (Page and Holmes, 1998)). Alternatively, clustering results can be assessed using internal or external measures of clustering quality. Internal measures (e.g. F-measure, Rand Index (Halkidi *et al.*, 2001)) provide an estimate of the degree of structure in the data and can help us to determine whether the data exhibits sufficient structure (i.e. compact, well-separated clusters) or whether it seems to be essentially uniform (in which case the data-preprocessing and the distance function employed may have to be reconsidered, as these can have a crucial impact on the discernability of clusters). Different to internal measures, external measures (e.g. F-Measure, Rand Index (e.g., F-Measure, Rand Index Halkidi *et al.*, 2001)) require a reference partitioning or knowledge of the true class labels respectively. Hence, they are useful to establish the difference between two partitionings of a data set, or for the evaluation of a clustering algorithm on data sets where the correct classification is known.

8. Supervised learning methods

Supervised learning methods are used when we have information on both the inputs and the outputs that one is desiring to understand or to classify. Typically these come as paired data sets (as in figure 4). This allows us to "train" a model using some kind of a teacher. A typical example in metabolomics would be where we have two classes of sample, from patients with a disease and from healthy controls. In this case we wish to determine a biomarker or set of biomarkers from the inputs that can be used to classify the samples into disease or control. For dealing with multivariate data this class assignment problem is shown below in its simplest form in figure 6 (Kell and King, 2000). The input consists of a large number of data-points that can represent a wide range of variables, which may be categorical, binary (e.g. true/false), or numerical (severity of disease/grade of cancer). A series of mathematical transformations when applied to the inputs are used to generate the outputs. Metabolomics data are thus the inputs and represent the *x*-data of figure 4.

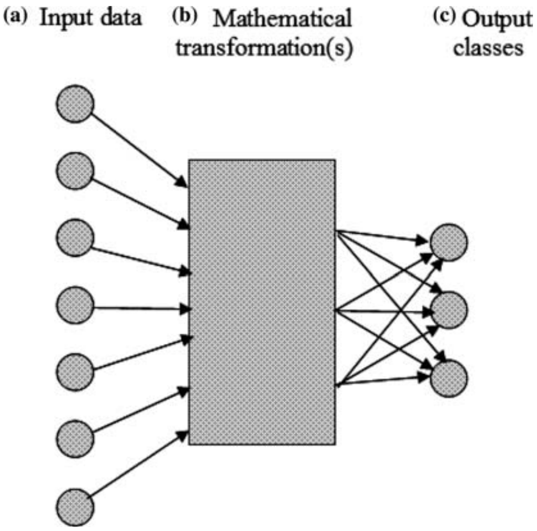


Figure 6. The class assignment problem. The inputs can be considered, and are referred to, as the "explanatory variables" or "x-data" whereas the functional or the other classes of interest, which are still variables associated with the samples, are referred to as "dependent variables" or "y-data" and are to be obtained as the outputs.

Supervised methods are much more powerful than unsupervised ones such as the widely used principal components analysis (PCA) and clustering methods because they concentrate on the variance that matters for the question of interest (e.g. (Goodacre *et al.*, 1992; Goodacre *et al.*, 1993)). In machine learning, methods that use only the *x*-data, unsupervised methods, are distinguished from supervised learning methods which are trained using both the *x*-data and the *y*-data (Jain and Dubes, 1988; Weiss and Kulikowski 1991; Michie *et al.*, 1994; Bishop, 1995; Livingstone, 1995; Ripley, 1996; Mitchell, 1997; Duda *et al.*, 2001).

The ideal method for supervised learning not only gives the correct answer, but explains how it got there ("credit assignment"). Some methods such as (artificial) neural networks are good at performing classification but poor at explaining the basis for it, while deterministic rule-based methods such as CART gives rules that are easy to understand but may not be as accurate as one would wish. A particularly powerful form of supervised learning is evolutionary computing (figure 2), which is based on Darwinian principles of natural selection and (here) is used to generate and to optimize a mapping between the input and the output variables. Genetic programming (GP) is a subset of evolutionary computing and was largely developed and popularized by John Koza (Koza, 1992, 1994; Banzhaf *et al.*, 1998; Langdon, 1998; Koza *et al.*, 1999; Foster, 2001; Langdon and Poli, 2002; Koza *et al.*, 2003). This method involves an arrangement in which the rules are arrayed in a tree-like structure that is read from the bottom and a subset of variables are passed through appropriate operators or functions to provide the output. So-called parse-trees can be mutated and recombined to provide variants that remain syntactically correct. They help evolve solutions to complex problems that are simple and intelligible, generating equations essentially in the form of rules, thereby having both desirable properties (accuracy and intelligibility) mentioned above. GP has been used successfully by us in identifying metabolites in terms of their involvement in particular processes (Gilbert *et al.*, 1999; Johnson *et al.*, 2000; Kell *et al.*, 2001; Kell, 2002; Allen *et al.*, 2003; Goodacre, 2003; Goodacre and Kell, 2003; Allen *et al.*, 2004).

A particular trend is towards voting methods of various kinds (Bauer and Kohavi, 1999; Dietterich, 2000; Breiman, 2001a), in which ensembles of "weak" learners contribute to more robust classifications via a committee voting approach (Bishop, 1995) than is possible with single classifiers alone (Hastie *et al.*, 2001).

Correlation analyses can be used to investigate the dependency of metabolites on one another (Kose *et al.*, 2001; Fernie, 2003; Steuer *et al.*, 2003; Urbanczyk-Wochniak *et al.*, 2003). Most frequently pairs of metabolites show no clear relationship to each other but when they do occur they are commonly of two types. The first is when there is high correlation indicating two-closely linked metabolites and the second is when there is non-linear correlation between metabolites. This may suggest that one of the metabolites is more "constrained" than another and that they are connected in some manner through a feedforward or feedback mechanism (Fernie, 2003), although other mechanisms are possible. As correlation methods however, (Pearl, 2000), no distinction can be made between relations of the substrate-product variety and those based on regulatory interactions. However, such correlations can be of value in seeking biosynthetic precursors of metabolites whose structures are unknown, as a vehicle to assist in structure determination.

Finally, Pattern recognition analysis of the metabolome can also be achieved using co-response analysis (Raamsdonk *et al.*, 2001) based on MCA (Hofmeyr and Carnish Bowden, 1996; Raamsdonk *et al.*, 2001) where the co-variation of pairs of metabolites under different conditions can provide useful information of their "connectedness" (Kose *et al.*, 2001; Steuer *et al.*, 2003).

9. Data storage

Driven in part by the needs of transcriptomics (Brazma *et al.*, 2001), there is an increasing recognition that we need standards and interoperable databases for storing proteome (Orchard *et al.*, 2003; Taylor *et al.*, 2003) and metabolome data (Hardy and Fuell, 2003), as well as for the metabolic models (Hucka *et al.*, 2003) that are a substantial part of the systems biology agenda (Kell, 2004). The raw data generally do not support flexible access and its structure, as discussed above, may vary greatly from experiment to experiment depending on the analytical technique used (e.g. m/z peaks for MS, peak retention times and mass spectrum for GC/MS and LC/MS, chemical shifts for NMR, wave number for FT-IR, etc.). Since large amounts of data need to be stored, handled and disseminated efficiently, databases are used to store the raw or processed data in a structured form and to provide fast and modular access to such data. Further advantages of using a database to store the experimental data include enforcing consistency and integrity of the data (Hardy and Fuell, 2003).

Another question to decide is the choice of the types of data to be stored in a metabolomics database. It is generally useful to store information about the whole range of wet experiments including growth, sample preparation and analytical experiments. From the metabolomics point of view, it is important to store the so-called meta-data (the data about the data, a term which refers to specific conditions, protocols and parameters used for growth/cultivation and sample preparation experiments) in order to support reproducibility of results and to analyse the effects that cultivation conditions and sample preparation have on the quality of chemical analysis. Wet experiments are generally performed in relation to some background knowledge, with the ultimate aim of enriching that knowledge. Once such a database is sufficiently populated, some types of biochemical knowledge can be acquired automatically by applying machine learning to the content of the database. In this context, it is practical to store the background knowledge in the database explicitly and in a machine-readable form.

The roles of the background knowledge stored in the database cover the provision of the biological context for genetic strains being examined (e.g. functions of specific genes), the interpretation of the results of analytical experiments (e.g. mapping a mass spectrum to specific compounds and their chemical properties), support for the reasoning process of data mining (e.g. annotation for supervised learning methods), etc. Apart from the experimental data and the relevant background knowledge, it is important to cover yet another aspect of metabolomics in the post-genomic era, which is concerned with the results of machine learning methods used to turn the metabolomics data into information.

Namely, the conclusions produced by statistical and machine learning methods (e.g. correlation, clustering methods, genetic programming, etc.) also need to be stored for future use in a suitable form, since some of these methods are computationally intensive. Numerical analyses of these types that one might wish to store include:

– pair-wise correlation or mutual information data between variables, either in full or in an ordered list.
– clustering or multivariate statistical information.
– a derived rule in a variety of possible formats, etc.

For example, rules uncovered by GP can be stored in the eXtensible Markup Language (XML) format, which is platform independent and can be converted automatically (assuming the provision of appropriate scripts) into appropriate code depending on users' preferences for specific platforms. Other data (e.g. computational times, evaluation results, etc.) related to specific uses of machine learning methods should be stored as well, in order to support the comparison of different methods. This also provides a convenient means of assessing the suitability of specific methods (and their parameters) for specific types of data (e.g. a naive Bayes classifier may work well for FT-IR spectra but not mass spectra).

Before implementing a specific database, the structure of the data needs to be described by developing a suitable model transcribed into a database schema. Depending on the specific purpose of a database, a suitable balance needs to be struck between its generality and commitment to specific organisms or analytical methods (Mendes, 2002; Hardy and Fuell, 2003; Li *et al.*, 2003). Entity-relationship diagrams (Chen, 1976) traditionally used to model relational databases are nowadays typically being replaced by Unified Modelling Language (UML) models (Booch *et al.*, 1999). For example, UML has been used to model the databases described in (Paton *et al.*, 2000; Cornell *et al.*, 2003; Taylor *et al.*, 2003; Jones *et al.*, 2004). UML is an object-oriented modelling language which uses classes and relations as its main structuring mechanism. Classes are used to describe structural aspects of homogenous sets of objects by means of their attributes, operations and relations. XML is also becoming increasingly used for modelling and integration of biochemical subdomains, e.g. Chemical Markup Language (CML), Systems Biology Markup Language (SBML) (Hucka *et al.*, 2003), Generalized Analytical Markup Language (GAML), etc. (Achard *et al.*, 2001). Many database models supply an XML schema of the database (e.g. (Taylor *et al.*, 2003; Jenkins *et al.*, 2004; Jones *et al.*, 2004)). Through the use of elements and their hierarchical organisation, XML allows specification of data integrated together with its structure. The syntax of XML documents also makes them self-descriptive and thus largely self-documented.

Once a suitable schema is developed, it needs to be implemented as a database. A UML model can be translated straightforwardly into an object-oriented database. For example, such an approach has been taken in the development of the object-oriented GIMS database used to store genomic and functional data (Paton *et al.*, 2000; Cornell *et al.*, 2003). Also, it can be relatively easily translated into a relational or XML database. For instance, the object-oriented model for functional genomics described in (Jones *et al.*, 2004) has been implemented as a relational database. The PEDRo model for proteomics experimental data, on the other hand, is used to convert data into the corresponding XML format, and the XML files so produced can be stored in a database repository of the user's choice. Further, an XML model can be manipulated automatically by XML-aware applications to produce an object-oriented or relational schema (or any other kind for that matter) or directly translated into an XML-native database. Features used to determine a specific choice of database type may include flexibility in terms of generality, extensibility, ease of access and portability. Further, speed of access may be important from the data mining point of view having in mind the sheer volume of data that needs to be processed. In addition, interoperability with the existing databases sometimes needs to be supported as well.

XML databases are particularly suitable for highly variable data (see above) that cannot be easily represented by fixed table-like structures. The highly variable structures could generally be retained in relational databases as well. However, a new table needs to be created for each XML element type that can contain other nested elements. This could dramatically increase the number of tables required, and, therefore, negatively affect the transparency of the database and its performance. XML has been suggested as the most appropriate basis for creating a standard for the exchange of metabolomics data (Li *et al.*, 2003). However, relational databases still offer the fastest access and if the data to be accessed can be easily described by fixed table-like structures (e.g. mass spectrum), then a relational database is a natural choice for some metabolomics applications.

Flexibility in database design is important in this area because as new analytical techniques or data analysis methods become available then it is important that this information can be stored and readily accessed.

10. Validation including cross validation

Behind all stages of the data processing and analysis, statistical and other numerical validation methods need to be used to ensure that the quality of the data is high and that conclusions and interpretations drawn from the data can be justified. While this statement is an obvious statement of good intentions, it is surprisingly hard to be rigorous when very high-dimensional data are involved. For example, we discussed above the general lack of validation of clusters.

Machine learning methods are extremely powerful, and such power can sometimes be dangerous. It is always possible, for instance using neural networks, to learn associations between inputs and outputs of the form given in figure 6, *even if all the values used are random numbers*! This is because such systems possess what is referred to as a "content addressable memory", so that once they have "learned" something it can be retrieved. Of course when other inputs are used the outputs are equally nonsensical, and we wish our models to have the ability to generalize, i.e. to produce "correct" outputs when presented with "new" inputs.

The essential strategy used to avoid this pitfall is to control the training in such a way that the model is tested using samples that are not used in the training phase but for which the "correct" answer is known (Chatfield, 1995; Mitchell, 1997; Duda *et al.*, 2001; Hastie *et al.*, 2001). Resampling methods of this type include bootstrapping schemes (Efron and Tibshirani, 1993), or leave-*k*-out cross-validation where the data for all except *k* samples are used serially to predict the *k* samples omitted; $k > 1$ is considered to be more robust. A common method, and one we usually use, is to split the data into three sets. One ("training set") is used for training the learning system, another ("validation set") is used to tune the method (in iterative algorithms such as regression, neural or evolutionary methods, this means to determine when training is stopped so as to avoid overfitting), and a third set ("test set") is used as a final test of the ability of the model to generalise. Note that in some works the meanings of the terms "validation set" and 'test set' are interchanged, and the final set is also commonly referred to as a hold-out set. Some of the issues used to determine which samples one would assign to each set, and the use of these methods in GP, are discussed well and in detail by Rowland (Rowland, 2003).

11. Concluding remarks

"Errors using inadequate data are much less than those using no data at all"

Charles Babbage (1792–1871).

In conclusion, it is important whilst aiming to produce useful data to recognize the limitations of all high-volume high-throughput methods currently used in measuring and analyzing the metabolome. The reproducibility will vary considerably from organism to organism, from tissue to tissue and between analytical and extraction methods used. In tandem with this are the problems that the analysis of high-dimensional data presents. However, a well-defined approach can be used to maximize the potential of the raw data to ensure that coupled to the chemometric data processing that is necessarily required, the data may be used to give meaningful and useful results.

Acknowledgments

We thank the BBSRC, EPSRC, NERC, The Gottlieb Daimler and Karl Benz Foundation and the RSC for financial support, and Nigel Hardy and Helen Fuell for useful discussions.

References

Achard, F., Vaysseix, G. and Barillot, E. (2001). XML, bioinformatics and data integration. *Bioinformatics* **17**, 115–125.

Aharoni, A., Ric de Vos, C.H., Verhoeven, H.A., *et al.* (2002). Non-targeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *Omics* **6**, 217–234.

Allen, J.K., Davey, H.M., Broadhurst, D., *et al.* (2003). High-throughput characterisation of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.* **21**, 692–696.

Allen, J.K., Davey, H.M., Broadhurst, D., Rowland, J.J., Oliver, S.G. and Kell, D.B. (2004). Discrimination of the modes of action of antifungal substances using metabolic footprinting. *Appl. Environ. Micorbiol.*, **70**, 6157–6165.

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

Banzhaf, W., Nordin, P., Keller, R.E. and Francone, F.D. (1998). *Genetic Programming: An Introduction*. Morgan Kaufmann, San Francisco.

Barrow, J.D. and Silk, J. (1995). *The Left Hand of Creation: The Origin and Evolution of the Expanding Universe*. Penguin, London.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Networks* **5**, 537–550.

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning* **36**, 105–139.

Bernardo, J.M. and Smith, A.F.M. (2000). *Bayesian Theory*. Wiley, Chichester.

Berry, D.A. (1996). *Statistics: A Bayesian Perspective*. Duxbury Press, Belmont.

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

Bland, M. (1987). *An Introduction to Medical Statistics*. Oxford University Press, Oxford.

Booch, G., Raumbaugh, J. and Jacobson, I. (1999). *Unified Modelling Language User Guide*. Addison–Wesley, .

Bradford Hill, A. and Hill, I.D. (1991). *Bradford Hill's Principles of Medical Statistics, 12th edn*. Edward Arnold, London.

Brazma, A., Hingamp, P., Quackenbush, J., *et al.* (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371.

Breiman, L. (2001a). Random forests. *Machine Learning* **45**, 5–32.

Breiman, L. (2001b). Statistical modeling: the two cultures. *Stat. Sci.* **16**, 199–215.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth International, Belmont.

Brindle, J.T., Antti, H., Holmes, E., *et al.* (2002). Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nat. Med.* **8**, 1439–1444.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *J. R. Stat. Soc. Ser. A* **158**, 419–466.

Chen, P. (1976). The entity-relationship model – toward a unified view of data. *ACM Trans. Database Syst.* **1**, 9–36.

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

Cornell, M., Paton, N.W., Hedeler, C., *et al.* (2003). GIMS: an integrated data storage and analysis environment for genomic and functional data. *Yeast* **20**, 1291–306.

Cornish-Bowden, A. and Cárdenas, M.L. (2001). Silent genes given voice. *Nature* **409**, 571–572.

Dasgupta, P., Chakrabarti, P.P. and DeSarkar, S.C. (1999). *Multi-objective Heuristic Search*. Vieweg, Braunschweig.

Davies, Z.S., Gilbert, R.J., Merry, R.J., Kell, D.B., Theodorou, M.K. and Griffith, G.W. (2000). Efficient improvement of silage additives using genetic algorithms. *Appl. Environ. Microbiol.* **66**, 1435–1443.

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B. and Moreau, Y. (2002). Adaptive quality-based clustering of gene expression profiles. *Bioinformatics* **18**, 735–746.

Dietterich, T.G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems*, pp. 1–15.

Duda, R.O., Hart, P.E. and Stork, D.E. (2001). *Pattern Classification, 2nd ed.* John Wiley, London.

Dudoit, S., Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.* **3**, RESEARCH0036.

Duran, A.L., Yang, J., Wang, L. and Sumner, L.W. (2003). Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics* **19**, 2283–2293.

Efron, B. and Tibshirani, R.J. (1993). *Introduction to the Bootstrap*. Chapman and Hall, London.

Ellis, D.I., Harrigan, G.G. and Goodacre, R. (2003). Metabolic fingerprinting with Fourier transform infrared spectroscopy in Harrigan, G.G., Goodacre, R. (Eds), *Metabolic profiling: its role in biomarker discovery and gene function analysis*. Kluwer, Boston. pp. 111–124.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newslett. Arch.* **4**, 65–75.

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

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

Fernie, A.R. (2003). Metabolome characterisation in plant system analysis. *Funct. Plant Biol.* **30**, 111–120.

Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Func. Genomics*. **2**, 155–168.

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

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N. and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**, 1157–1161.

Fiehn, O. and Spranger, J. (2003). Use of metabolomics to discover metabolic patterns associated with human disease in Harrigan, G.G. and Goodacre, R. (Eds), *Metabolic profiling: its role in biomarker discovery and gene function analysis.* Kluwer Academic Publishers, Boston, pp. 199–215.

Fiehn, O. and Weckwerth, W. (2003). Deciphering metabolic networks. *Eur. J. Biochem.* **270**, 579–588.

Fisher, R.A. (1951). *The Design of Experiments, 6th ed.* Oliver & Boyd, Edinburgh.

Fleischmann, R.D., Adams, M.D., White, O., *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach*. Chapman and Hall, Londonc.

Foster, J.A. (2001). Evolutionary computation. *Nat. Rev. Genet.* **2**, 428–436.

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

Gilbert, R.J., Johnson, H.E., Rowland, J.J., *et al.* (1999). Genetic programming as an analytical tool for metabolome data in Langdon, W.B., Poli, R., Nodin, P. and Fogarty, T. (Eds), *Late-breaking papers of EuroGP-99, Software Engineering*. CWI, Amsterdam, pp. 23–33.

Goodacre, R. (2003). Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. *Vib. Spectrosc.* **32**, 33–45.

Goodacre, R. and Kell, D.B. (1993). Rapid and quantitative analysis of bioprocesses using pyrolysis mass spectrometry and neural networks. Application to indole production. *Anal. Chim. Acta.* **279**, 17–26.

Goodacre, R. and Kell, D.B. (1996). Pyrolysis mass spectrometry and its applications in biotechnology. *Curr. Opin. Biotechnol.* **7**, 20–28.

Goodacre, R. and Kell, D.B. (2003). Evolutionary computation for the interpretation of metabolome data in Harrigan, G.G. and Goodacre, R. (Eds), *Metabolic profiling: its role in biomarker discovery and gene function analysis*. Kluwer Academic Publishers, Boston, pp. 239–256.

Goodacre, R., Kell, D.B. and Bianchi, G. (1992). Neural networks and olive oil. *Nature* **359**, 594.

Goodacre, R., Kell, D.B. and Bianchi, G. (1993). Rapid assessment of the adulteration of virgin olive oils by other seed oils using pyrolysis mass spectrometry and artificial neural networks. *J. Sci. Food Agric.* **63**, 297–307.

Goodacre, R., Vaidyanathan, S., Bianchi, G. and Kell, D.B. (2002). Metabolic profiling using direct infusion electrospray ionisation mass spectrometry for the characterisation of olive oils. *Analyst* **127**, 1457–1462.

Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. and Kell, D.B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* **22**, 245–252.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On clustering validation techniques. *J. Intell. Inf. Syst.* **17**, 107–145.

Handl, J. and Knowles, J. (2004) Evolutionary Multiobjective Clustering. PPSN VIII, LNCS 3242, 1081–1091 (see http://dbk.ch.umist.ac.uk/Papers/HandlKnowlesPPSN-webversion.pdf).

Hardy, N. and Fuell, H. (2003). Databases, data modeling and schemas: database development in metabolomics in Harrigan, G.G. and Goodacre, R. (Eds), *Metabolic profiling: its role in biomarker discovery and gene function analysis*. Kluwer Academic Publishers, Boston, pp. 277–291.

Harrigan, G.G. and Goodacre, R. (Eds) (2003). *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis.* Kluwer Academic Publishers, Boston.

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

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

Hicks, C.R. and Turner, K.V. Jr (1999). *Fundamental Concepts in the Design of Experiments, 5th edn.* Oxford University Press, Oxford.

Hofmeyr, J.H. and Cornish-Bowden, A. (1996). Co-response analysis: a new experimental strategy for metabolic control analysis. *J. Theor. Biol.* **182**, 371–380.

Hucka, M., Finney, A., Sauro, H.M., *et al.* (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999). Data clustering: a review. *ACM Comput. Surveys* **31**, 264–323.

Jenkins, H., Hardy, N., Beckmann, M., Draper, J., Smith, A.R., Taylor, J. *et al.* (2004). A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnol.* **22**, 1601–1606.

Johnson, H.E., Broadhurst, D., Goodacre, R. and Smith, A.R. (2003). Metabolic fingerprinting of salt-stressed tomatoes. *Phytochemistry* **62**, 919–928.

Johnson, H.E., Gilbert, R.J., Winson, M.K., *et al.* (2000). Explanatory analysis of the metabolome using genetic programming of simple, interpretable rules. *Genet. Progr. Evolvable Machines* **1**, 243–258.

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

Jones, A., Hunt, E., Wastling, J.M., Pizarro, A. and Stoeckert, C.J. Jr (2004). An object model and database for functional genomics. *Bioinformatics* **20**, 1583–1590.

Kaderbhai, N.N., Broadhurst, D.I., Ellis, D.I., Goodacre, R. and Kell, D.B. (2003). Functional genomics via metabolic footprinting: monitoring metabolite secretion by *Escherichia coli* tryptophan metabolism mutants using FT-IR and direct injection electrospray mass spectrometry. *Comp. Funct. Genom.* **4**, 376–391.

Kell, D.B. (2002). Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules. *Mol. Biol. Rep.* **29**, 237–241.

Kell, D.B. (2004). Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* **7**, 296–307.

Kell, D.B., Darby, R.M. and Draper, J. (2001). Genomic computing: explanatory analysis of plant expression profiling data using machine learning. *Plant Physiol.* **126**, 943–951.

Kell, D.B. and King, R.D. (2000). 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.

Kell, D.B. and Oliver, S.G. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* **26**, 99–105.

Kell, D.B. and Westerhoff, H.V. (1986). Metabolic control theory – its role in microbiology and biotechnology. *FEMS Microbiol. Rev.* **39**, 305–320.

Kohonen, T. (1989). *Self-Organization and Associative Memory*. Springer-Verlag, Berlin.

Kose, F., Weckwerth, W., Linke, T. and Fiehn, O. (2001). Visualizing plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics* **17**, 1198–1208.

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

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

Koza, J.R., Bennett, F.H., Keane, M.A. and Andre, D. (1999). *Genetic Programming III: Darwinian Invention and Problem Solving*. Morgan Kaufmann, San Francisco.

Koza, J.R., Keane, M.A., Streeter, M.J., Mydlowec, W., Yu, J. and Lanza, G. (2003). *Genetic Programming: Routine Human-Competitive Machine Intelligence*. Kluwer, New York.

Kramer, M.A. (1991). Nonlinear principal components analysis using auto-associative neural networks. *AIChE J* **37**, 233–243.

Langdon, W.B. (1998). *Genetic Programming and Data Structures: Genetic Programming + Data Structures = Automatic Programming!* Kluwer, Boston.

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

Lenz, E.M., Bright, J., Wilson, I.D., Morgan, S.R. and Nash, A.F.P. (2003). A 1H-NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects. *J. Pharm. Biomed. Anal.* **33**, 1103–1115.

Leonard, T. and Hsu, J.S.J. (1999). *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*. Cambridge University Press, Cambridge.

Li, X.J., Brazhnik, O., Kamal, A., *et al.* (2003). Databases and visualization for metabolomics in Harrigan, G.G. and Goodacre, R. (Eds), *Metabolic profiling: its role in biomarker discovery and gene function analysis*. Kluwer Academic Publishers.

Lindon, J.C., Holmes, E. and Nicholson, J.K. (2003). So whats the deal with metabonomics? Metabonomics measures the fingerprint of biochemical perturbations caused by disease, drugs, and toxins. *Anal. Chem.* **75**, 384A–391A.

Lindon, J.C., Nicholson, J.K., Holmes, E., *et al.* (2003b). Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicol. Appl. Pharmacol.* **187**, 137–46.

Lindon, J.C., Nicholson, J.K., Holmes, E. and Everett, J.R. (2000). Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids. *Concepts Magn. Reson.* **12**, 289–320.

Livingstone, D. (1995). *Data Analysis for Chemists*. Oxford University Press, Oxford.

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

Mendes, P. (2002). Emerging bioinformatics for the metabolome. *Brief Bioinform.* **3**, 134–145.

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

Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (Eds) (1994). *Machine Learning: Neural and Statistical Classification*. Ellis Horwood, Chichester.

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

Montgomery, D.C. (2001). *Design and Analysis of Experiments, 5th edn*. Wiley, Chichester.

Muggleton, S.H. (1990). Inductive logic programming. *New Gen. Comput.* **8**, 295–318.

Myers, R.H. and Montgomery, D.C. (1995). *Response Surface Methodology: Process and Product Optimization using Designed Experiments*. Wiley, New York.

Nicholson, J.K., Connelly, J., Lindon, J.C. and Holmes, E. (2002). Metabonomics: a platform for studying drug toxicity and gene function. *Nat Rev. Drug Discov.* **1**, 153–161.

Nicholson, J.K. and Wilson, I.D. (2003). Understanding ‘global’ systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Disc.* **2**, 668–676.

O’Hagan, S., Dunn, W.B., Brown, M., Knowles, J.D., Kell, D.B. (2004). Closed-loop, multiobjective optimisation of analytical instrumentation: gas-chromatography-time-of-flight mass spectrometry of the metabolomes of human serum and of yeast fermentations. *Anal. Chem.*, In press.

Oliver, S.G., Winson, M.K., Kell, D.B. and Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends Biotechnol.* **16**, 373–378.

Orchard, S., Hermjakob, H. and Apweiler, R. (2003). The proteomics standards initiative. *Proteomics* **3**, 1374–1376.

Page, R.D.M. and Holmes, E.C. (1998). *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, Oxford.

Paton, N.W., Khan, S.A., Hayes, A., *et al.* (2000). Conceptual modelling of genomic information. *Bioinformatics* **16**, 548–557.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge.

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

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

Ramoni, M. and Sabastini, P. (1998). *Theory and Practice of Bayesian Belief Networks.* Edward Arnold, London.

RaywardSmith, V.J., Osman, I.H., Reeves, C.R. and Smith, G.D. (Eds) (1996). *Modern heuristic search methods.* Wiley, Chichester.

Reeves, C.R. (Eds) (1995). *Modern heuristic techniques for combinatorial problems.* McGraw Hill, London.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.

Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. and Willmitzer, L. (2000). Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry. *Plant J.* **23**, 131–142.

Rothman, K.J. (2002). *Epidemiology: An Introduction.* Oxford University Press, Oxford.

Rothman, K.J. and Greenland, S. (1998). *Modern Epidemiology, 2nd edn.* Lippincott, Williams & Wilkins, Philadelphia.

Rowland, J.J. (2003). Model selection methodology in supervised learning with evolutionary computation. *Biosystems* **72**, 187–196.

Schlesselman, J.J. (1982). *Case–Control Studies – Design, Conduct, Analysis.* Oxford University Press, Oxford.

Seasholtz, M.B. and Kowalski, B. (1993). The parsimony principle applied to multivariate calibration. *Anal. Chim. Acta* **277**, 165–177.

Shannon, C.E. and Weaver, W. (1949). *The Mathematical Theory of Communication.* University of Illinois Press, Urbana.

Solanky, K.S., Bailey, N.J.C., Beckwith-Hall, B.M., *et al.* (2003). Application of biofluid 1H nuclear magnetic resonance-based metabonomic techniques for the analysis of the biochemical effects of dietary isflavones on human plasma profile. *Anal. Biochem* **323**, 197–204.

Steuer, R., Kurths, J., Fiehn, O. and Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics* **19**, 1019–1026.

Sumner, L.W., Mendes, P. and Dixon, R.A. (2003). Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817–836.

Taylor, C.F., Paton, N.W., Garwood, K.L., *et al.* (2003). A systematic approach to modelling capturing and disseminating proteomics experimental data. *Nat. Biotechnol* **21**, 247–254.

Taylor, J., King, R.D., Altmann, T. and Fiehn, O. (2002). Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics* **18**(Suppl 2), S241–S248.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J Roy Stat Soc B* **63**, 411–423.

Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., *et al.* (2003). Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep* **4**, 989–993.

Vaidyanathan, S., Broadhurst, D.I., Kell, D.B. and Goodacre, R. (2003). Explanatory optimisation of protein mass spectrometry via genetic search. *Anal. Chem* **75**, 6679–6686.

Vaidyanathan, S., Kell, D.B. and Goodacre, R. (2004). Selective detection of proteins in mixtures using electrospray ionization mass spectrometry: influence of instrumental settings and implications for proteomics. *Anal. Chem.*, **76**, 5024–5032.

Weiss, S.H. and Kulikowski, C.A. (1991). *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural networks, Machine Learning, and Expert Systems.* Morgan Kaufmann Publishers, San Mateo, CA.

Weuster-Botz, D. and Wandrey, C. (1995). Medium optimization by genetic algorithm for continuous production of formate dehydrogenase. *Proc. Biochem* **30**, 563–571.

Wilson, I.D. and Brinkman, U.A. (2003). Hyphenation and hypernation the practice and prospects of multiple hyphenation. *J. Chromatogr. A* **1000**, 325–356.

Woodward, A.M., Rowland, J.J. and Kell, D.B. (2004). Fast automatic registration of images using the phase of a complex wavelet transform: application to proteome gels. *Analyst* **129**, 542–552.