Metabolomics, Vol. 2, No. 4, December 2006 (© 2007)
DOI: 10.1007/s11306-006-0037-z

Statistical strategies for avoiding false discoveries in metabolomics and related experiments

David I. Broadhurst,[a,b,*] and Douglas B. Kell[a,b,*]

[a]*School of Chemistry, The University of Manchester, Faraday Building, Sackville St, Manchester, M60 1QD, UK*
[b]*Manchester Centre for Integrative Systems Biology, The Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess St, Manchester, M1 7DN, UK*

Received 4 August 2006; accepted 22 September 2006

Many metabolomics, and other high-content or high-throughput, experiments are set up such that the primary aim is the discovery of biomarker metabolites that can discriminate, with a certain level of certainty, between nominally matched 'case' and 'control' samples. However, it is unfortunately very easy to find markers that are apparently persuasive but that are in fact entirely spurious, and there are well-known examples in the proteomics literature. The main types of danger are not entirely independent of each other, but include bias, inadequate sample size (especially relative to the number of metabolite variables and to the required statistical power to prove that a biomarker is discriminant), excessive false discovery rate due to multiple hypothesis testing, inappropriate choice of particular numerical methods, and overfitting (generally caused by the failure to perform adequate validation and cross-validation). Many studies fail to take these into account, and thereby fail to discover anything of true significance (despite their claims). We summarise these problems, and provide pointers to a substantial existing literature that should assist in the improved design and evaluation of metabolomics experiments, thereby allowing robust scientific conclusions to be drawn from the available data. We provide a list of some of the simpler checks that might improve one's confidence that a candidate biomarker is not simply a statistical artefact, and suggest a series of preferred tests and visualisation tools that can assist readers and authors in assessing papers. These tools can be applied to individual metabolites by using multiple univariate tests performed in parallel across all metabolite peaks. They may also be applied to the validation of multivariate models. We stress in particular that classical *p*-values such as "$p < 0.05$", that are often used in biomedicine, are far too optimistic when multiple tests are done simultaneously (as in metabolomics). Ultimately it is desirable that all data and metadata are available electronically, as this allows the entire community to assess conclusions drawn from them. These analyses apply to all high-dimensional 'omics' datasets.

KEY WORDS: statistics; machine learning; false discovery; receiver–operator characteristic; hypothesis testing; statistical power; Bonferroni correction; bias; overfitting; cross validiation; credit assignment; visualisation.

1. Introduction: binary class discrimination problems

"Thirteen years ago I moved from a department of applied mathematics and theoretical physics to a department of physiology. During these years, I've come to recognize how very difficult it is to do a good experiment" (Rapp, 1993).

"Scientific research is a process of guided learning. The object of statistical methods is to make that process as efficient as possible" (Box *et al*., 1978).

"It can be proven that most claimed research findings are false" (Ioannidis, 2005b).

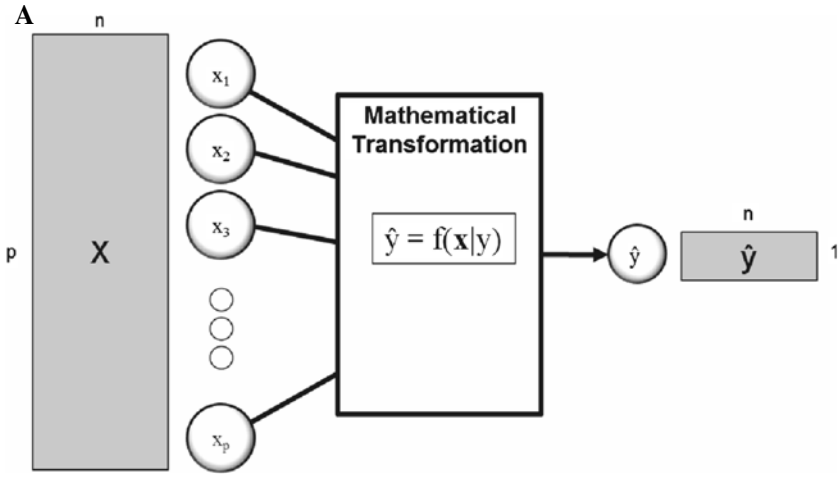
" 'what the data say' is often obscured by questionable answers to unanswerable questions (Cornfield, 1966)" (Goodman and Royall, 1988).

"Left to our own devices, ...we are all too good at picking out non-existent patterns that happen to suit our purposes" (Efron and Tibshirani, 1993)

As part of the scientific endeavour, metabolomics studies involve a search for some kind of truth, and it is appropriate therefore to start by recognising that while the renaissance of interest in metabolomics itself may be comparatively new (but cf. (Horning and Horning, 1971; Jellum *et al*., 1981; Greenaway *et al*., 1991; Tas and van der Greef, 1994)), the basics of metabolomics *studies* (as scientific studies) are not, and they can thus benefit greatly from the accumulated knowledge and standards that have become the norm in mainstream biology and medicine.

In the typical kind of experiment that is the focus of this article, the general objective (cf. (Kell and Oliver, 2004)) is the discovery of one, or more, measured variables whose values are drawn from two populations with means that differ 'significantly', these two populations (classes) being labelled 'case' and 'control'. A similar objective could be to discover a way of combining several, or all, of the measured variables in such a way that when projected into a single dimension (e.g. as a rule (Kell, 2002a)) via a mathematical transformation the predicted values are drawn from two populations with means that differ 'significantly'. This basic structure is entirely general (figure 1A), and is well described in textbooks such as (Duda *et al*., 2001) and (Hastie *et al*.,

* To whom correspondence should be addressed.
E-mail: david.broadhurst@manchester.ac.uk
E-mail: dbk@manchester.ac.uk



B

	Actual +/1	Actual -/0
Predicted +/1	a TP	b FP
Predicted -/0	c FN	d TN

Figure 1. The binary classification problem. A. The basic structure of a multivariate binary classification problem involves projecting an $n \times p$ dimensional data matrix **X** (n = number of samples, p = number of variables), via a multidimensional scaling transformation, into a one dimensional vector $\hat{\mathbf{y}}$ ($1 \times n$). This function may simply be a weighted linear sum (as in Multiple Linear Regression), a two layer function requiring estimation of a set of Latent Variables (e.g. PLS-DA) or even a multi-layer tree- based non-linear transformation (e.g. a multi-layer perceptron or a parse tree as used commonly in Genetic Programming). The optimisation/selection of this transforming function is based on *a-priori* information about class membership of each of the n samples, **y** (usually defined in terms of dummy values, 'control' = '0' and 'case' = '1'). The utility of a classifier is assessed using *a priori* class membership information of a second set of data (in the form of a test set). If the classifier produces a binary output it can be assessed in the form of a confusion matrix, while if the classifier produces a continuous output then it can be assessed in a similar fashion as is a single variable undergoing univariate significance tests. **B.** The so-called confusion matrix describing the outcome of predictive models that cross-tabulates the observed and predicted +/− or 1/0 patterns in a binary classification problem. If, in a binary prediction model we label the two classes as 1 (for cases) and 0 (for controls) under conditions in which we are treating the cases as 'positive', there are two possible prediction errors: false positives (FP) and false negatives (FN). There are also, happily, true positives and true negatives that are correctly predicted by the model. B is adapted and extended from http://asio.jde.aca.mmu.ac.uk/multivar/da4.htm, which also contains other information and is derived from (Fielding and Bell, 1997).

2001). The implication is then that knowledge of these metabolite values will be sufficient to effect the required discrimination (and subsequently, perhaps, its mechanistic basis, that might then allow external intervention). As such this is known as a binary class discrimination problem (and multi-class problems, including those involving a grade or severity of disease, can always be reduced to a series of binary class problems). We shall mainly assume that the 'input' variables are metabolite concentrations and the class referred to as 'cases' genuinely contains individuals who are in a different physiological state from those of the 'controls'. Such physiological states or individuals from whom the samples are taken may include those with disease, susceptibility to a disease, individuals suffering from a toxic effect, possessing a desirable agricultural trait, and so on) and it is taken that this class membership is thus well defined. Inaccurate labelling of individuals' class membership is known as class noise (Kell and King, 2000), and we shall largely ignore it, although we would point out that the field of semi-supervised learning is designed to help detect this and reclassify individuals as appropriate (Bennett and Demiriz, 1998; Demiriz *et al.*, 1999; Kemp *et al.*, 2003; Li *et al.*, 2003; Handl and Knowles, 2006b). Similarly, we assume that individuals have been assigned entirely to one class or the other ('crisp' membership), since so-called fuzzy class membership (Zadeh, 1965; Bezdek and Pal, 1992; Kruse *et al.*, 1994; Li and Yen, 1995) is outwith our scope here. Finally, where we touch on

multivariate analyses, we assume that the matrix of metabolites versus samples is of full rank, i.e. there are no missing values, and thus we do not deal with algorithms (such as those described elsewhere (Troyanskaya *et al.*, 2001; Zhou *et al.*, 2003; Sehgal *et al.*, 2005) for imputing them.

Since there are two possible classes, the outcome of any predictions relative to the 'true' class membership is usually set out as a binary matrix, the so-called confusion matrix (figure 1B), consisting of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Various metrics concerning the performance of a binary classifier derive from these and are given in table 1, and since they are simply based numerically on the confusion matrix they implicitly assume equal weightings on the 'cost' of misclassification. We note that in many cases of interest to this audience, a false negative is more important to avoid than is a false positive, since if the metabolite is a disease marker a false negative may be life-threatening. However, we shall again largely not pursue this issue here since our focus is on the confusion matrix itself and the means by which we may hope to maximise the Sensitivity (the conditional probability that case Y when occurring is correctly classified, i.e. $p(Y_{predicted} \mid Y_{actual})$, the probability that Y is predicted to be 1 and is 1) and the Specificity (which is the inverse, viz. $p(\overline{Y_{predicted}} \mid \overline{Y_{actual}})$, the probability that Y is predicted to be 0 and is 0) of our classifier. Note that this analysis of the confusion matrix does not explicitly take into account the prevalences of the two classes in a given population, and in consequence we do not need to be, and largely are not, either explicitly Bayesian (Berry, 1996; Bernardo and Smith, 2000; Baldi and Long, 2001) or explicitly frequentist for these present purposes. However, in terms of the interpretation of data, it is extremely important that prevalences are taken into account (e.g. (Brenner and Gefeller, 1997)), since a biomarker discovery programme for a disease with a low prevalence is likely to have many more false positives than may occur with a disease with a higher prevalence.

We note that the problems we describe are both widespread and have been widely recognised by a variety of distinguished analysts, and a summary of cautionary papers that make useful reading include (Ransohoff and Feinstein, 1978; Altman, 2001; Ioannidis *et al.*, 2001; Ein-Dor *et al.*, 2006; Ransohoff, 2004, 2005; Ioannidis, 2005a, b; Ioannidis and Trikalinos, 2005; Wacholder *et al.*, 2004). Few published papers in the metabolomics literature withstand this level of proper scrutiny (and no doubt some of ours will not, but in view of the some of the misplaced claims we have observed we still feel it useful to write this introductory review). Box 1 summarises some of the issues.

It is of interest to note that at the time of writing (July 2006), PubMed returned 578 references when for any field the search term metabolom* was used (* is a wild card, the terms thus including the terms metabolome, metabolomic and metabolomics) and increased to 762 when metabonom* was added. However, of the 762 these dropped to 82 when the term statistic* was also included. In other words, only about 10% of papers in this field appear to make mention of any statistical treatment. The percentage improved slightly when metabolom* was used

Table 1
Some metrics derived from the confusion matrix of figure 1, where N is the total number of samples and a,b,c,d refer to numbers rather than percentages. Adapted and extended from http://asio.jde.aca.mmu.ac.uk/multivar/da4.htm, which also contains other information and is derived from (Fielding and Bell, 1997). For a given value of a variable, the likelihood ratio = TP rate/FP rate = sensitivity/ (1-specificity)

Measure	Calculation
Prevalence	(a + c)/N = (TP + FN)/ total
Overall diagnostic power	(b + d)/N = (FP + TN)/ total
Correct classification rate	(a + d)/N = (TP + TN)/ total
Sensitivity	a/(a + c) = TP/(TP + FN)
Specificity	d/(b + d) = TN/(FP + TN)
False positive rate	b/(b + d) = FP/(FP + TN)
False negative rate	c/(a + c) = FN/(FN + TP)
Positive predictive power	a/(a + b) = TP/(TP + FP)
Negative predictive power	d/(c + d) = TN/(FN + TN)
Misclassification rate	(b + c)/N = (FP + FN)/total
Odds-ratio	(ad)/(cb) = (TP.TN)/(FP.FN)
Kappa	$\frac{(a+d) - (((a+c)(a+b)+(b+d)(c+d))/N)}{N - (((a+c)(a+b)+(b+d)(c+d))/N)}$
Normalised mutual information	$\frac{1 - a.\ln(a) - b.\ln(b) - c.\ln(c) - d.\ln(d) + (a+b).\ln(a+b) + (c+d).\ln(c+d)}{N.\ln N - ((a+c).\ln(a+c) + (b+d).\ln(b+d))}$

'Sensitivity' is the conditional probability that case X is correctly classified.
'Specificity' is the inverse, i.e. that case not-X is correctly classified.
'Positive predictive power' assesses the probability that a case is X if the classifier classifies the case as X.
'Negative predictive power' assesses the probability that a case is not X if the classifier does not classify the case as X.

Box 1. The main types of danger in the design and analysis of metabolomics experiments.

Inadequate sample size, in which case it is always easy to find random multivariate correlations when the number of variables greatly exceeds the number of samples

Ignorance of Type I statistical errors, i.e. high false discovery rate (FDR) due to low critical *p*-values when applying multiple significance tests in parallel across all metabolite peaks, or combinations thereof.

Overfitting, typically by failing to use independent/blind 'test' samples which are held back from model optimization and used only to test the robustness of prediction in the final phase of the study.

Inappropriate model/statistical-test selection, in which the wrong statistical test or multivariate model is chosen for a given study, either due to lack of theoretical understanding, or excessive familiarity with 'favourite' algorithms (or the software that implements them).

Bias, in which a variable of interest is differentially distributed between the classes 'case' and 'control' but happens to be correlated with another uncontrolled variable that truly underlies the variance in the metabolite of interest, such confounding variables including smoking status, gender, diet, lifestyle, pharmaceutical or recreational drug use, etc.

in the title only. Only 10 papers with metabol* included the word 'learning' (and thus 'machine learning') in any field, and those with metabonom* did not add to their number. Overall, then, only a small percentage of papers in metabolomics make much of the importance of statistics, so unsurprisingly are perhaps less than fully alert to the dangers of poor experimental design and analysis. (A referee points out that the same is true for transcriptomics and proteomics.)

2. Statistics and machine learning, and supervised versus unsupervised modelling methods

The most readily understood *modus operandi* for discussing the causes of false discovery in metabolomics is probably through the use (and misuse) of various univariate tests and visualization tools, in parallel across all metabolite peaks, and is that adopted here. However, all these arguments apply equally to the output of most multivariate methods, while some further arguments are particular to multivariate modelling and the methods of machine learning. With this in mind it may be useful, for some readers, to be aware of the many chemometric and related methods that are used in forming binary classifiers (e.g. (Mitchell, 1997; Duda *et al.*, 2001; Hastie *et al.*, 2001)). Some are principled and others are less so but let us start with two main distinctions:

(i) Some methods such as Principal Components Analysis (Jolliffe, 1986) and a variety of clustering methods (Everitt, 1993; Handl *et al.*, 2005) use only the x-data as defined in figure 1 and are fundamentally designed for what Tukey called Exploratory Data Analysis (Tukey, 1977). They are referred to as unsupervised methods and are to be contrasted with supervised methods in which knowledge of the class membership of at least some of the samples is used to guide the classifier.

(ii) As pointed out with great clarity by the distinguished statistician Leo Breiman (Breiman, 2001), and mirroring the philosophically reciprocal (but not reversible (Kell and Welch, 1991)) relation between the worlds of data/observation and of mental constructs (/ideas/knowledge) (Kell and Welch, 1991; Kell and Oliver, 2004), classical (Neyman-Pearson) statistics starts with an idea or hypothesis and tests the goodness of fit of the data to that hypothesis. By contrast, machine learning starts with a set of data and finds the hypothesis that best fits the data. While statisticians had long pointed out that this can lead to all sorts of problems of overfitting [often referred to in this literature as 'biased estimators' (Miller, 1990; Chatfield, 1995)], the machine learning community equally recognised that the solution to this problem is based on what are by now well-established validation and cross-validation procedures.

Indeed, while we shall later discuss the importance of validation and cross-validation, we shall largely ignore the 'data-driven' literature on machine learning (e.g. (Langley *et al.*, 1987; Weiss and Kulikowski, 1991; Anthony and Biggs, 1992; Hutchinson, 1994; Michie *et al.*, 1994; Mitchell, 1997; Vapnik, 1998; Michalski *et al.*, 1998; Michalewicz and Fogel, 2000)) and data mining (e.g. (Adriaans and Zantinge, 1996; Cabena *et al.*, 1998; Weiss and Indurkhya, 1998; Berry and Linoff, 2000; Hand *et al.*, 2001; Rud, 2001)), but provide references for those whose wish to learn more about these methods.

However, we *would* make the comment that the real aim of any of these studies is inference based on evidence, and that while the classical (and largely frequentist) approaches using statistical evidence provide one general line of reasoning, and are the main focus of the present review, there are other important approaches, often largely Bayesian, with distinguished, principled and fervent adherents (e.g. (Edwards, 1992, 2000; Royall, 1997; Pearl, 1988, 2000; Ramoni and Sabastini, 1998; Bernardo and Smith, 2000; Shipley, 2001; Jensen, 2001; Casella and Berger, 2002; Mackay, 2003; Needham *et al.*, 2006)). The proponents of likelihood in particular (e.g. (Edwards, 1992; Royall, 1997)) make an excellent case for their view of statistical inference.

3. Distributions and Normal distributions

The properties of distributions and the analysis of their variance are the focus of any number of textbooks on elementary statistics, including those aimed

at a biomedical audience (Bland, 2000; Box *et al*., 1978; Bradford Hill and Hill, 1991; Sokal and Rohlf, 1995; Rothman and Greenland, 1998; Woodward, 2000; Kirkwood and Sterne, 2003), and we are not going to reproduce them here save in terms of the ostensibly simple question of asking whether the statistical properties of a particular variable in populations from two classes differ 'significantly' or otherwise, and what that means. This basic issue is displayed in figure 2, where three examples of a single variable's potential values expressed as two hypothetical Normal distributions are given (blue = 'control'; red = 'case'). These distributions can be characterised by their mean and standard deviation (SD). As these distributions are by definition Normal, various useful and precise statements follow, for instance (i) that the proportion of observations that lie within ± 1, 2 or 3 SDs of the mean are respectively 68.27, 95.45 and 99.73, and (ii) that given the means and standard deviations we can predict with some probability p whether the two populations do indeed differ 'significantly' by more than a critical value α. A straightforward but important corollary of the normal distribution is that *by definition* nearly 5% of observations lie more than 2SDs away from the mean, while approximately 3 in a thousand lie more than 3 SDs away. Given 6000 parallel tests (the approximate number of genes in yeast (Goffeau *et al*., 1996)) more than 18 'random' genes would *appear* to be 'significant' if this thought were based on their expression levels being more than 3 SDs away from some reference expression level of genes from yeast cells measured in a nominally different condition. It is worth bearing in mind in this context that gross cholesterol levels are considered extremely important risk factors for coronary heart disease, yet the means in cholesterol levels between those with coronary artery disease and those without differ by less than 1 SD (Kannel, 1995)). While they are less predictive than LDL:HDL–cholesterol ratios (Natarajan *et al*., 2003), it should also be mentioned in terms of causality that measures such as statin treatment that also happen to lower such cholesterol levels plausibly exert their protection largely by entirely other means (e.g. (Grimes, 2006)).

Given two Normal populations (or classes), the usual procedure for testing whether the two population means differ 'significantly' is the Paired Student's t-test (Student was the pseudonym of a statistician at the Guinness brewery named W. S. Gosset, who deduced the t-distribution, that is very similar to the Normal distribution, in 1908). Another popular parametric test (i.e. one in which the parameters of the underlying Normal distribution are estimated) is one-way analysis of variance (ANOVA) which assesses the significance of the ratio of the variation within class to the variation between class.

If the distributions of the two populations are not Normal, transformations such as the logarithmic transformation can effectively turn a skewed distribution into a Normal one; however care has to be taken that by compensating for skewness in one direction one does not introduce skewness in the opposite direction. If no suitable transformations are available a variety of

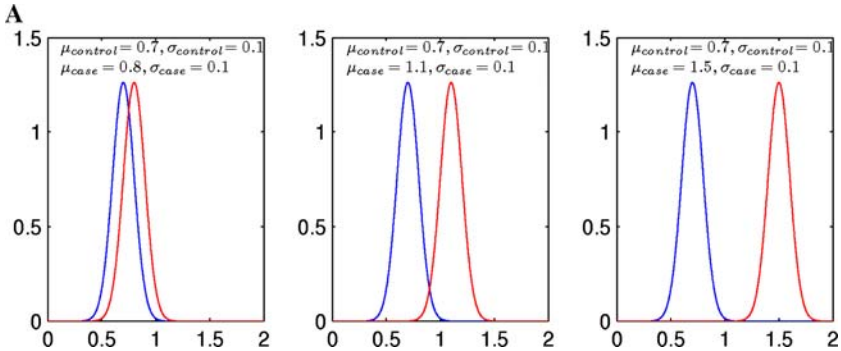


Figure 2. A. The Normal distribution curves for three very hypothetical binary discriminant examples in which the mean for the controls is always 0.7, the SD for both populations is 0.2, and the mean of the cases is respectively 0.8, 1.1 and 1.5. B. Three artificial data sets. Each consisting of 100 case samples and 100 control samples drawn from distributions as defined in A. The data are presented in the form of a scatterplot of response vs sample number, together with binned histograms, box-and-whisker plots and the respective ROC curves for the same data. The lower and upper lines of the "box" are the 25th and 75th percentiles of the sample. The distance between the top and bottom of the box is the interquartile range. The line in the middle of the box is the sample median. If the median is not centred in the box, that is an indication of skewness. The "whiskers" are lines extending above and below the box. They show the extent of the rest of the sample (unless there are outliers). Assuming no outliers, the maximum of the sample is the top of the upper whisker. The minimum of the sample is the bottom of the lower whisker. By default, an outlier is a value that is more than 1.5 times the interquartile range away from the top or bottom of the box and is marked as a red cross. The notches in the box are a graphic confidence interval about the median of a sample. A side-by-side comparison of two notched box plots provides a graphical way to determine which groups have significantly different medians. The text boxes describe the Analysis of Variance statistics; the paired t-test statistics, and the modified Z factors for each (where ‚modified' implies the more relaxed form of the equation, where a zero score implies that for two normal distributions (of equal standard deviation) their means will be 4 standard deviations apart).

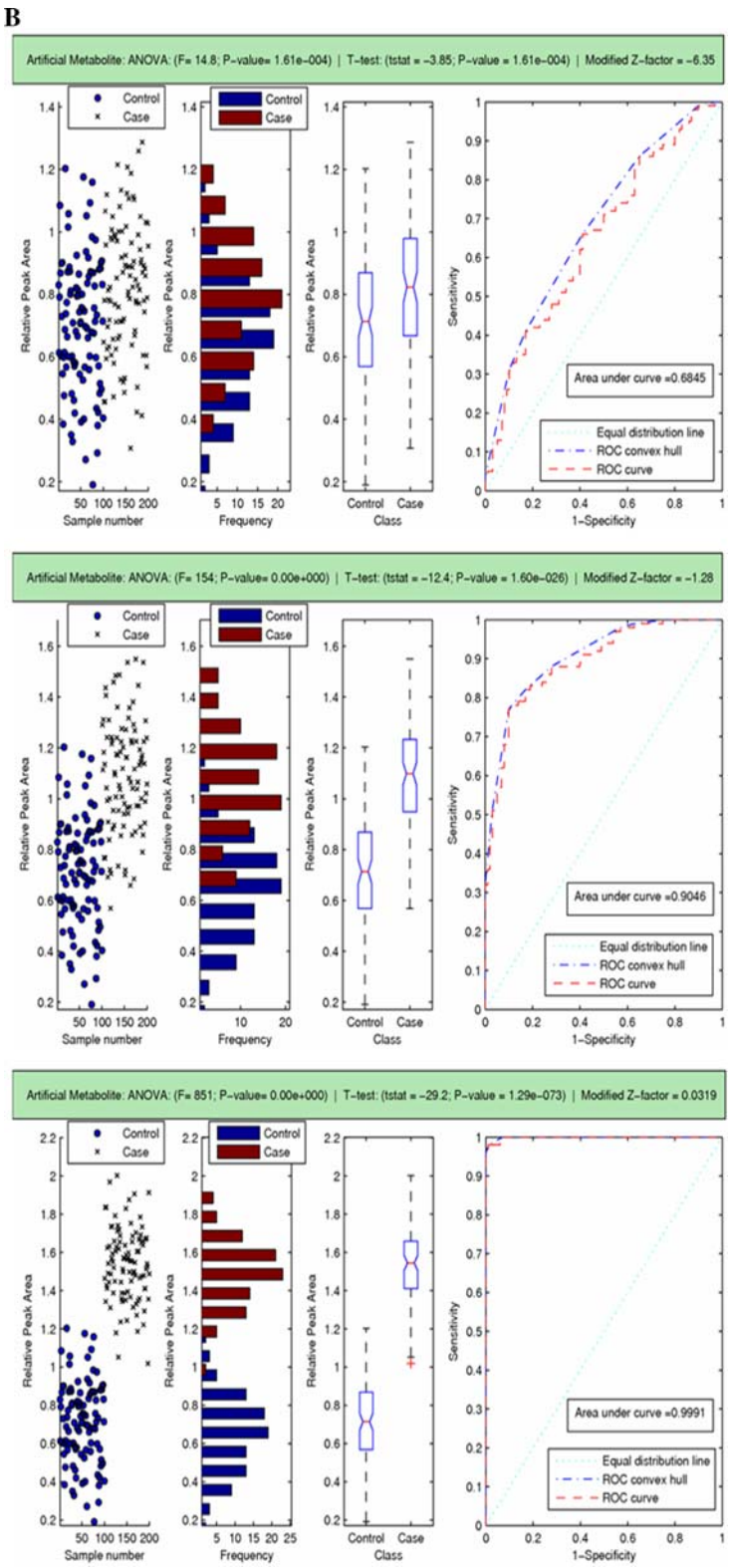


Figure 2. Continued.

non-parametric tests that do not assume a particular underlying distribution are available. There are various tests for normality. Typically, each metabolite peak may be checked for within-class kurtosis, and for within-class goodness of fit to a normal distribution using the Lilliefors test (Conover, 1980). For a given metabolite peak, if either control or case samples has kurtosis, or fails the Lilliefors test then non-parametric tests are

appropriate (Hollander and Wolfe, 1973). Of these the most common is the Mann–Whitney U test (or Wilcoxon two sample test) which by using rank orderings is largely insensitive to extreme values.

We note here that the estimates of all of these statistical metrics are themselves subject to variance, and estimates for confidence intervals for means, differences in means, and variances can be calculated (Gardner and Altman, 1989; Bland, 2000).

It is also worth noting for case-control studies that equivalent tests, and confidence intervals, can be estimated for multivariate statistical methods, such as Principal Components Analyisis (PCA, where confidence is estimated using knowledge of class membership), Canonical Variates Analysis (CVA) (see (Krzanowski, 1988)), sometimes referred to as Discriminant Function Analysis (DFA) (Manly, 1994) and Discriminant Partial Least Squares (D-PLS, or PLS-DA) (see (Eriksson *et al.*, 2001)). In this way, the actual model predictions (projections from many dimensions into one- or two-dimensional space) can be assessed for significance prior to inclusion into a confusion matrix, and suitable confidence intervals can be defined for assessing the robustness of the model using an independent test set (see later).

Another approach to assessing confidence intervals, usable for both parametric and non-parametric statistics alike, is the bootstrap family of methods (Efron and Gong, 1983; Efron and Tibshirani, 1993) that use resampling procedures (which bear relations to some types of internal cross-validation, see below). Its accessibility has increased markedly with the advent of powerful personal computers. To generate a bootstrap uncertainty estimate for a given statistic (such as a mean or median) from a set of data, we randomly generate (with replacement) p subpopulations of a size less than, or equal to, the size of the original data set. Typically, p is set to 1000, *ergo* any data point can be sampled multiple times or not at all. The desired statistic is calculated for each subpopulation, from which uncertainty estimates and confidence intervals are derived from the variances observed in and between these 'new' populations.

4. The Z- statistic

Another interesting statistic, that has assumed considerable popularity in the world of high-throughput screening but is little known outside, is the Z (and Z') statistic (Zhang *et al.*, 1999). The domain relates to determining a 'hit' in an assay in which there would be positive and negative controls (known as 'sample' and 'control') that correspondingly did or did not display activity in the assay of interest. Here, the aim is also in part to optimise the assay itself, but this is extremely pertinent to determining the quality or reliability of a method designed for binary class (patient/control) discrimination in terms of a metabolomic biomarker or biomarkers. The initial recognition of these authors was that neither the signal:noise nor the signal:background ratios provided good metrics for this, but that metrics which used both the mean *and* variance (i.e. standard deviation, SD) in both the classes were to be preferred. Specifically, the Assay Value Ratio,

$$AVR = \frac{3(SD_{signal} + SD_{background})}{|\bar{x}_{signal} - \bar{x}_{background}|}$$

Or in our particular context,

$$AVR = \frac{3(SD_{case} + SD_{control})}{|\bar{x}_{case} - \bar{x}_{control}|} \quad (1)$$

And the Z factor,

$$Z = 1 - \mathrm{AVR} \quad (2)$$

Z factors above 0.5 are considered to provide an excellent assay and hence discrimination. Alternative and less stringent forms of the Z factor that use 2 rather than 3 SDs may also be envisaged (see figures 2B, 5).

5. Receiver-Operator Characteristic (ROC) curves

An additional property of the kinds of plot shown in figure 2A (here we assume that the mean of the variable of interest is larger in the cases than in the controls) is that one can imagine taking some value, a of that variable, $\boldsymbol{x}$, (assume initially a = zero) and hypothesising that all samples above that value are 'cases' and all below it are 'controls'. One can then produce a confusion matrix for this value of $\boldsymbol{x}$ and determine the sensitivity and specificity from the formulae in table 1. This can be repeated for all values of $\boldsymbol{x}$ and a (smoothed) plot of the data made of the specificity (i.e. true positive rate) against 1-sensitivity (i.e. false positive rate). This plot will necessarily include the points 0,0 and 1,1, and is known as the Receiver–Operator Characteristic or ROC curve. It is widely considered to be one of the best means by which to describe the utility of a variable in binary classification (Egan, 1975; Metz, 1978; Hanley and McNeil, 1982; Zweig and Campbell, 1993; Raubertas *et al.*, 1994; Zhou *et al.*, 2002; Baker, 2003; Linden, 2006) (and see e.g. http://gim.unmc.edu/dxtests/ROC1.htm and http://www.anaesthetist.com/mnm/stats/roc/). If the area under the ROC curve is 0.5 (the lower limit) the variable is distributed similarly between cases and controls, such that any diagnostic test based on it is valueless for discrimination. The area under the ROC curve (the AUC) when there is complete separation of the two populations (such that

a value or set of values of the variable that lies between the ranges of the two classes is entirely diagnostic of the class) is 1 (See figure 2B). The ROC curve has its origins in signal detection theory (Egan, 1975) as applied to the radar detection of specific objects. It is especially attractive as it is insensitive to the nature of any underlying population distributions, i.e. it is non-parametric, it is independent of the prevalence of disease, two or more diagnostic tests can be compared at any or all false positive rates (FPRs), and summary measures of accuracy, such as the AUC, incorporate both components of accuracy, i.e., sensitivity and specificity, into a single measure (Obuchowski *et al.*, 2004). The AUC is an estimate of the probability that a member of one population chosen as random will exceed a member of the other population chosen at random, in the same way as does U/n_1n_2 in the Mann–Whitney U test (where U is the test score, n_1 the sample size of class1, and n_2 the sample size of class2 (Bland, 2000)). It is often considered that a value for the AUC of >0.9 is an excellent test while a value over 0.8 is still likely to be good. It does not seem to be widely used in either transcriptomics or metabolomics studies, though various simple computer programs are available (Stephan *et al.*, 2003) and we suggest that it should be adopted for omics biomarker discovery for just the reasons for which it has been widely adopted in the medical literature.

Other measures that might be considered for use as metrics of the quality or accuracy by which we can compare the 'true' and proposed class memberships of a binary partitioning include the F-measure $\left(\frac{2\cdot\text{specificity}\cdot\text{sensitivity}}{\text{specificity}+\text{sensitivity}}\right)$ (van Rijsbergen, 1979) and the adjusted Rand index (Hubert and Arabie, 1985), as well as other well-established methods that we have recently surveyed in the context of cluster validation (Handl *et al.*, 2005).

6. Hypothesis testing and statistical power.

In order to avoid ambiguity, for reasons that will hopefully soon become clear, statisticians tend to enjoy the use of double negatives and the usual (Neyman-Pearson) manner in which experiments of the present type are couched is in terms of a null hypothesis. Typically the null hypothesis is that for a given metabolite all samples are drawn from the same population (or from two populations with the same mean). A false positive (also known as a 'type I error') occurs when the hypothesis is rejected (i.e. it is claimed that the samples are drawn from two populations with significantly different means) when it should be accepted (i.e. there really is no significant difference in means). If, as is commonly the case (but see later), we allow a 5%, or one in twenty, chance that we have made a type I error, in other words if we set our criterion for a "significant difference" between two population means at the 5% level, we have a value of a parameter referred to as α of 0.05. The probability of having a false negative or making a type II error, i.e. claiming that a variable is not significant when it really is, is correspondingly known as β. The power of a test (the probability that a test will produce a significant difference at a given significance level) is $(1-\beta)$, so if β is 0.1 or 10%, then the power is 90%. Various tables (or indeed software packages – we use the nQuery Advisor http://www.statsol.ie/nquery/nquery.htm) – allow one to calculate the number of samples necessary to discriminate two populations on the basis of the distributions of a variable between them. In the case of comparing two means (μ_1 and μ_2) from two Normal distributions (standard deviation $=\sigma_1$ and σ_2 and population size $=n_1$ and n_2, respectively), where the standard error, se_{diff}, of the difference between means is $=\text{sqrt}(\sigma_1^2/n_1+\sigma_2^2/n_2)$, it can readily be shown (Bland, 2000) that the power of the test $=1-\Phi(z)$ where $z=1.96-(\mu_1+\mu_2)/\text{se}_{\text{diff}}$ (at $\alpha=0.05$; see later for why this is already too forgiving) and Φ is the equation for the standard for the Standard Normal distribution. Figure 3A shows three power curves for population sample sizes $n_1, n_2 = 10$, $n_1, n_2 = 100$ and $n_1, n_2 = 400$. The difference in means is varied from 0 to 0.5 units (where $\sigma_1, \sigma_2 = 0.2$). The plots illustrate the fact that in order to keep the probability that a test will produce a significant result at a reasonable level (for a fixed α – in this case 0.05) then the relationship between the number of samples measured and the difference in population means is critical. In other words, if the difference between 'case' and 'control' is slight then a relatively large sample set is needed when compared with a situation where the difference between 'case' and 'control' is substantial. The point is therefore, that the 'chance' – in terms of sensitivity and specificity – of being able to discriminate two populations on the basis of a candidate discriminatory variable, or model, is a function of the mean and variance in those two populations and the size of these populations, i.e. the number of samples that are assessed. So when a statistician says that 'we have failed to reject the null hypothesis', what s/he may mean is 'from the data provided we accept the null hypothesis; if the test were to be repeated with a different number, or set of samples, a different conclusion may be reached'; hence the use of the double negative. Note that these attempts to assess statistical power are related to, but often have a different emphasis from, the kind of material in the Design of Experiments literature (Box *et al.*, 1978; Deming and Morgan, 1993; Myers and Montgomery, 1995; Hicks and Turner, 1999; Montgomery, 2001) including that on the Design of Computer Experiments (Sacks *et al.*, 1989; Crary, 2002; Chen *et al.*, 2006) where we typically have control over a variety of parameters and wish to understand their effect on a number of indicator variables. A useful introduc-

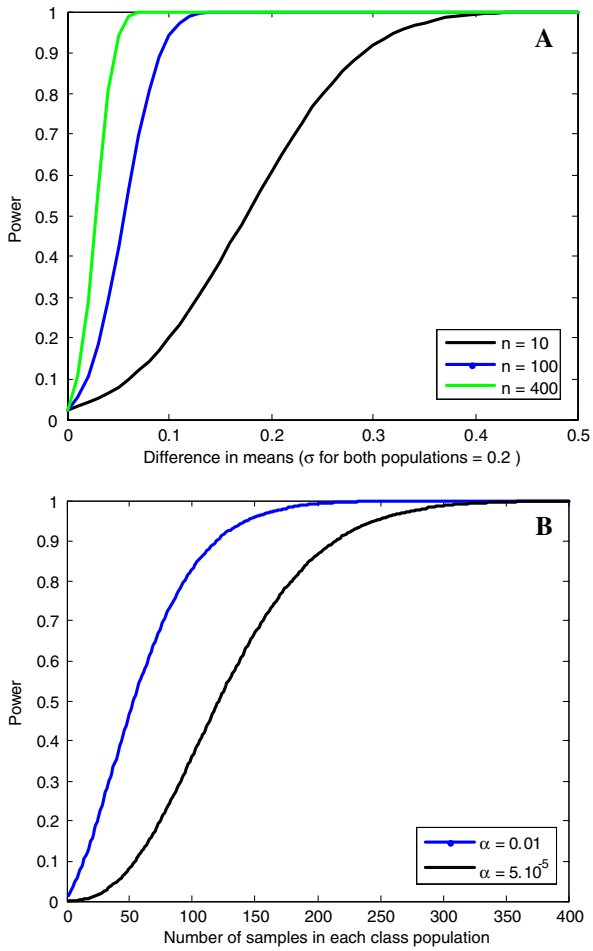


Figure 3. A. Relationship between the statistical power and the difference in means for 2 populations whose SD is 0.2 arbitrary units, for sample sizes of 10, 100 and 400. **B**. Relationship between the statistical power and the number of samples for two populations whose mean differs by 0.5 SDs for critical *p*-values of 0.01 and $5{\cdot}10^{-5}$.

tion to the calculation of statistical power is also available at Statsoft's website (http://www.statsoft.com/textbook/stpowan.html#(power_doe2).

We would stress also that the *p*-value is not universally accepted as a useful measure. "The most widely recognized practical consequence of the *p*-value's dependence on only one hypothesis is that a huge effect in a small trial or a minuscule effect in a large trial can result in identical *p*-values. To the extent that we believe the size of an effect is an essential part of the evidence relative to the hypothesis of "no effect", then the p-value is inadequate for measuring the strength of evidence (Cornfield, 1966). The move toward confidence intervals is an effort to deal with this issue by focusing on the effect size." (Goodman and Royall, 1988). Put another way, "Unlike the *p*-value, the use of evidential measures forces us to bring scientific judgment to data analysis, and shows us the difference between what the data are telling us and what we are telling ourselves." (Goodman and Royall, 1988). Finally, by contrast, and in the words of Garner and Altman (Gardner and Altman, 1989), "Small differences of no real interest can be statistically significant with large sample sizes, whereas clinically important effects may be statistically non-significant only because the number of subjects studied was small."

7. Statistical analysis using machine replication.

In order to improve the accuracy of any statistical analysis it is advised to collect machine replicates of all the biological samples in a given study. Indeed, as is widely recognised in transcriptomics, appropriate replication is one of the best ways that may be available for improving the statistical precision in many cases. Machine replicates differ from biological replicates in that in the latter case the same biological subject is re-sampled and analyzed more than once (e.g. 20 'control' patients results in 20 biological 'control' replicate measurements, whereas, 4 'control' patients repeatedly sampled 5 times results in 4 biological replicates, each containing 5 machine replicates; each strategy involves 20 measurements in total). Collection of machine replicates may be done for several reasons:

- To provide a measure of machine variance (within-sample variance), which can then be compared with between-sample and between-class variance. This provides both a measure of the sensitivity of the system used to analyze the samples and some extra information about the reproducibility of any 'discovered' biomarkers.
- Given that certain assumptions are true (see below) n machine replicates may be averaged in order to reduce the standard deviation due to random noise by a factor of $\sqrt{n}$ (for mathematical derivation see: http://zone.ni.com/devzone/conceptd.nsf/webmain/D3887A3DF70CBE0F86256A5400681ACD), thus boosting the signal to noise ratio.
- If enough machine replicates are collected semi-supervised multivariate data analysis, (such as semi-supervised PC-CVA) may be used to model the data. In this methodology the training 'class' information is the 'sample id', such that for 10 samples there are 10 classes. So in this case, the model loadings are optimized such that the ratio of between sample variance to within sample variance is maximized. This will allow clustering of samples to be revealed which would otherwise be obfuscated by noise in unsupervised methods such as PCA, without enforcing some known class membership – as in fully supervised methods.

As with any other aspect of data analysis, correctly utilizing machine replicate information requires assumptions to be made about the distribution of such

data. Thus, much care has to be taken in the way that they are used to help clarify any statistical analysis.

Univariate significance tests assume that all measurements are independently sampled from a larger population. As replicate measurements are not independent they cannot all be used in a test between 'cases' and 'controls'. Although the test score may not be unduly affected, any significance value or power calculation will be skewed as the number of degrees of freedom has been falsely increased. Instead the replicate sub-sets must be averaged to a single value. However, this can be dangerous if the within-sample variance is much greater than the between-class variance. In this case the significance test results may easily be misrepresentative of the true relationship. In multivariate analysis this principle also applies when splitting data into training and test sets. Data should be allocated to each set such that machine replicates are not separated. Again, this will ensure that the data sets are truly independent.

If machine replicate averaging is to be performed to increase the signal-to-noise ratio then there are also several assumptions that are often ignored, e.g.: that the noise is homoscedastic (Alsberg *et al.*, 1997), i.e. the mean and standard deviation remain constant across all replicate measurements, its mean is equal to zero, and the signal (without noise) is steady. If any of these assumptions is false then the averaged value may be skewed. For example, in Fourier Transform Infrared (FTIR) spectroscopy samples may be loaded using 96-well silicon sample plates (Harrigan *et al.*, 2004). If three measurements are made from a single well over a short period of time then machine replicate averaging will probably be effective. However, if a single biological sample is sub sampled 3 times into three random wells across the plate then it is more likely that the averaging of these three machine replicates will be affected by systematic errors/noise such as plate-drift or time variation. Thus, the averaged value may be skewed. Conversely, if one wishes the machine replicates to represent the variance due to machine error then, each replicate sample should be a truly independent measurement on the machine in question and should be randomly ordered amongst the experimental cohort as a whole.

8. False Discovery Rates and inadequate sample size

The term False Discovery Rate (FDRs) typically refers to the frequency of type I errors, i.e. to claims that some variable, or model, can discriminate two populations when, in fact, it cannot. As indicated in the abstract, there are a variety of reasons why this can occur, and in a certain sense we can also learn about why these may occur from the analysis of type II errors – the failure to observe an effect when one really does exist – with which they have a certain symmetry. One of the chief causes of FDRs, however, is inadequate sample size (Ioannidis *et al.*, 2001, 2003; Ransohoff, 2004; Wacholder *et al.*, 2004; Ioannidis, 2005a, b; Ioannidis and Trikalinos, 2005; Ein-Dor *et al.*, 2006; Todd, 2006), a feature that is particularly problematic in cases when the number of variables greatly exceeds the number of samples. This is, of course, a characteristic of omics studies that has long been highlighted following the development of microarrays for transcriptomics (e.g. (Benjamini and Hochberg, 1995; Efron and Tibshirani, 2002; Storey, 2002; Storey and Tibshirani, 2003; Reiner *et al.*, 2003; Jung, 2005; Xie *et al.*, 2005)). First, however, it is appropriate to discuss in outline the literature on this question of looking at multiple variables in a single study simultaneously, since the basic problems and potential solutions have also long been known.

9. Multiple hypothesis testing and the Bonferroni correction.

Much of classical statistics is concerned with univariate analyses, and the question of whether (in this context) one would 'by chance' assign a sample to a particular population on the basis of the measurement of a single variable. In a sense this can be compared to the chance of winning a lottery on the basis of purchasing a single ticket. To a good approximation (it depends on whether only a finite set of unique solutions are sold individually), this chance scales linearly with the number of lottery tickets. The same is true in statistical hypothesis testing. In other words, given various distributions, the chance of finding a discriminating variable with a p value less than a specified value (say $p < 0.01$) is increased in proportion to the number of independent tests one makes (the number of tickets, if you will). In metabolomics, if the search for discriminating biomarkers is performed using 200 metabolites and a suitable critical p-value for rejecting the null hypothesis across all the parallel tests is considered to be 0.01, then the p-value for rejecting the null hypothesis for an individual metabolite (the corrected p-value) can be approximated to $0.01/200 = 5{\cdot}10^{-5}$ (0.00005). This is, of course, a much more stringent test than $p < 0.01$ but is arguably the most appropriate way to look at this kind of multiple hypothesis testing, at least as a starting point. This said, it is conservative in the sense that it excludes type I errors at the cost of increasing the potential for type II errors (false negatives) (Bland and Altman, 1995; Cook and Farewell, 1996; Perneger, 1998), leading to the conclusion that any peak found to have a p-value below the Bonferroni-corrected level is clearly significant. Intelligent use of Bonferroni correction does not necessarily require that sample sizes be enormous (Leon, 2004), since the 'burden' can be placed on the acceptable

p-values (Storey, 2002), but it does increase substantially the stringency of our testing and the need for a great deal more skepticism that is usually observed when novel biomarkers are apparently discovered.

Actually, Bonferroni correction assumes that each of the variables is independent (of the others), and while this is clearly not entirely true these correlations are known to apply to only a comparatively small number of metabolites (Kose *et al.*, 2001; Steuer *et al.*, 2003; Weckwerth and Morgenthal, 2005; Camacho *et al.*, 2005; Steuer, 2006); a divisor of 100 rather than 200 in the above example would make little difference to the number of metabolites found to be false positives (type I errors) by Bonferroni analysis in metabolomics studies. An illustration of the effects of Bonferroni correction is given in figure 3B.

Although there are other methods (e.g. (Benjamini and Hochberg, 1995; Cook and Farewell, 1996; Reiner *et al.*, 2003; Storey and Tibshirani, 2003; Jung, 2005; Xie *et al.*, 2005), and see also http://www.tufts.edu/~gdallal/mc.htm), we would comment that as well as being the most conservative, a pure Bonferroni analysis is both conceptually easy to understand and numerically easy to implement.

10. Stratification and subpopulations

A related error goes as follows: "although I could find no statistically significant marker for discriminating between the cases and controls when I compared the entire (matched) population from a carefully designed experiment, perhaps a subpopulation – say women between 40 and 50 – would show that there is in fact a marker with a significance value such as $p < 0.05$ that I might get away with publishing". Actually this kind of *post hoc* stratification or 'data dredging' (Todd, 2006) also amounts to multiple hypothesis testing, and while the numbers in the subpopulations are even lower than in the whole population, it too must be subjected to a Bonferroni correction if tested in this way. Note that this has nothing to do with the real stratification or differences in populations that are, for example, the focus of pharmacogenomics (Evans and Johnson, 2001; Evans and Relling, 1999, 2004) and analyses designed to reduce confounder effects (Rothman and Greenland, 1998).

11. An example from gene associations in cancer studies

Since the early days of transcriptome analysis (Golub *et al.*, 1999), many workers have looked to detect different gene expression in cancerous versus normal tissues. Partly because of the expense of transcriptomics (and the inherent noise in such data (Schena, 2000; Tu *et al.*, 2002; Cui and Churchill, 2003; Liang and Kelemen, 2006)), the numbers of samples and their replicates is often small while the number of candidate genes is typically in the thousands. Given the above, there is clearly a great danger that most of these will not in practice withstand scrutiny on deeper analysis (despite the ease with which one can create beautiful heat maps and any number of 'just-so' stories to explain the biological relevance of anything that is found in preliminary studies!). This turns out to be the case, and we review a recent analysis (Ein-Dor *et al.*, 2006) of a variety of such studies.

Ein-Dor and colleagues (Ein-Dor *et al.*, 2006) recognised these problems, and reasoned that if the FDRs in 'comparable' studies were very great the genes observed would, by definition, be different from each other. (Equivalently, repeated findings of the same thing would give one confidence in their significance, as was also found in protein-protein associated studies (von Mering *et al.*, 2002), arguably for similar reasons.) This was easy to test simply by comparing the gene lists found in different studies of the 'same' cancers. While cancer poses special problems, as its development has a stochastic part caused by increasing mutation rates and DNA and other cellular damage (Duesberg *et al.*, 2000), if the markers are to be any use we must find ones that are statistically reproducible.

Probably the most widely accepted theory in machine learning (Mitchell, 1997; Vapnik, 1998) is due to Valiant (1984), and is known as PAC ('probably approximately correct') learning. Ein-Dor et al. (2006) developed this theory to assess the likely FDR in these kinds of studies. A comparison of two studies using hundreds of samples found respectively 76 (Wang *et al.*, 2005) and 70 (van 't Veer *et al.*, 2002; van de Vijver *et al.*, 2002) 'significant' genes, but only 3 of them were in fact the same (Ein-Dor *et al.*, 2006)! Equally, permuting subsets of the samples also 'discovered' entirely different lists of genes. Thus the figure of merit Ein-Dor *et al.* (2006) introduced, f, "is the overlap between two Prospective Gene Lists (PGLs), obtained from two different training sets of n samples in each. That is, $0 \le f \le 1$ is the fraction of shared genes that appear on both PGLs; the closer f is to 1, the more robust and stable are the PGLs obtained from an experiment". For the typical sample sizes used in the studies surveyed by Ein-Dor *et al.* (2006), the overlap between two PGLs, obtained from two training sets using the PAC analysis, was of the order of only a few percent.

12. Meta-analysis

Analyses (such as that of Ein-dor and colleagues above) that combine data from different trials of nominally the 'same' experiment are known as meta-analyses in the literature of medical statistics. Overall, there is a

feeling that such datasets should not be combined into a single large dataset as this produces bias, and instead should be analysed and compared separately (Sharp *et al.*, 1996; Altman and Deeks, 2002). Thus, meta-analysis is treated as a two-stage process in which the data for each study are first summarised, and then those summaries are combined statistically. We note, however, that *publication* bias – the greater likelihood of publishing a positive than a negative result – can create optimistic inaccuracies in such analyses (Williamson *et al.*, 2005).

13. Bias

We have mentioned bias several times. The problem is often referred to as a problem of 'confounding variables or confounding factors', although the latter phrase has a slightly different emphasis and meaning in the epidemiological literature (''confounding is a distortion in the estimated exposure effects that result from differences in risk between the exposed and unexposed that are not due to exposure'' (Rothman and Greenland, 1998)). Imagine a study in which we wished to measure biomarkers for ethnicity, and compared the serum or urine metabolome of samples taken from Japanese or Russian people. No doubt we would find differences, but it would be quite wrong to ascribe these to ethnicity as the differences are just as likely to be due to something else that co-varies with ethnicity. Diet is likely the most important co-varying difference here. Of course this is an 'obvious' example, though such studies may well equally be confounded by the difference in time it takes the samples to get to the airport, i.e. differences in transport and storage conditions, which may in consequence be extremely well coupled with the thing one is trying to measure. This is almost inevitable in multi-country studies without great care being taken. Ransohoff (2005), whose outstanding review should be read by every researcher, refers to bias as ''the most important 'threat to validity' that must be addressed in the design, conduct and interpretation of such (i.e. biomarker) research'', and he comments that ''Bias can be so powerful in non-experimental observational research that a study should be presumed 'guilty' – or biased – until proven innocent''. In contrast to the small sample size problems that exacerbate false discovery rates, bias *cannot* be compensated by large sample numbers – in fact this can even make things worse by persuading readers of the validity of spurious differences that are actually due simply to confounding factors that happen to correlate with the class discrimination of interest. Naturally the correlation improves with sample size, as does the bias.

Bias can be exceptionally difficult to remove in case-control biomarker studies, although careful age and gender matching of the two classes is a good start. Having a gender bias (in which say males are more common in the case than in the control cohort) means that there is a danger of learning a model that is actually discriminating on gender. Similarly, it is likely that more cases will be taking drugs against the disease that they are known to have than are taking the same drugs in the control population. A recent study (Kirschenlohr *et al.*, 2006) suggested that both of these problems were probably a major feature of an earlier study on the purported detection of coronary artery disease using NMR. In this case (Kirschenlohr *et al.*, 2006) it was argued that only 6% of the model could be attributed to the coronary artery disease, compared with the 50% that might be achieved, ignoring prevalence, by random guessing of the class membership.

The only way to know that one has avoided bias, and for the readers of publications to know that authors have avoided important bias that might have affected the conclusion that should be drawn form the data, is to publish all the metadata (data about the samples) along with the metabolomic data. In this way, readers can establish that the models are not being made on confounding factors by comparing closely the distributions of the claimed biomarkers not only between the different classes (case and control) but with respect to the metadata. The fact that this was probably true of the famous ovarian cancer proteomics data soon led to the discovery (Baggerly *et al.*, 2004; Diamandis, 2004) that the original data (Petricoin *et al.*, 2002) were almost certainly not sound. Several years on, ''there are no clear success stories in which discovery proteomics has led to a deployed protein biomarker'' (Rifai *et al.*, 2006). Including the properties of the instrument in the analysis is probably key to getting good data here (Ressom *et al.*, 2005).

14. Overfitting/overtraining and proper (cross and external) validation

As mentioned above, there is a nowadays a trend towards data-driven models of biology (Brent, 1999, 2000; Brent and Lok, 2005; Kell and Oliver, 2004). In particular, one important area is that in which multiple combinations of variables are tested to see if they make a functional classifier. Classical statistics rightly recognised that this is really a version of the multiple-hypothesis-testing problem given above, and effectively provides a huge number of lottery tickets with which, one might, by chance, 'win', i.e. find a combination of markers that effectively discriminated the two classes. However, it may of course be the case that such a combination of markers really is discriminatory. How then is one to find out? The answer adopted by the machine learning community is to use a subset of the

data – the hold-out set – that is not used in the generation of the model *in any way at all* (Ransohoff, 2004). Thus the set used in producing the model is called the training set. Many of the powerful methods available today, such as neural networks and Evolutionary Algorithms, can usually learn the training set (White, 1992), even if the inputs consist of random numbers. Such models will not of course give reliable data when exposed to new examples, i.e. they will not generalize well. The initial model is said to have overfitted the data. The same kinds of problems arise in process modelling (Kell and Sonnleitner, 1995).

In order to address this issue of overfitting the machine learning community has developed several validation techniques. If during modeling some sort iterative parameter optimization is required, such as in the case with PLS-DA, PC-DFA, or multilayer perceptrons and other kinds of Neural Networks, etc., then internal model validation is commonplace. The simplest is to split the training data into two sets, typically with a ratio of 2:1 training:validation, and where the validation set is matched and completely representative of the training set. The validation set is used to assess the goodness of prediction statistic (Q^2) (Eriksson *et al.*, 2001) by projecting the data through a model previously built using the training set (assessed using the goodness of fit statistic R^2). As the model parameters are optimised (e.g. by increasing the number of latent variables), R^2 and Q^2 initially follow the same upward trend from 0 to 1. However as the models start to overfit, the trajectories diverge, R^2 tending toward 1 and Q^2 falling back toward 0. It is assumed that the model will have achieved its optimal predictive powers, and thus generalize well, at the initial point of divergence. Finally, one uses a completely separate set, the test set, to assess how accurate the model one has selected really is. In many published studies this very important third step is left out. There seems to be a failure to understand that as the validation set *is* used in the optimization of the model it is to all intents and purposes part of the training set, thus it *cannot* also be used to assess the general predictive powers of the final model. This can only be done with an independent test set.

Another popular validation method is n-fold cross-validation (Martens and Næs, 1989; Eriksson *et al.*, 2001; Brereton, 2003), where the dataset is randomly spilit into n mutually exclusive subsets (the folds) of approximately equal size (the special case where n = the number of samples is known as leave-one-out cross-validation – though Golbraikh and Tropsha, in a wonderfully titled article, make a very clear case against the use of leave-one-out Q^2 in multivariate problems (Golbraikh and Tropsha, 2002)). The model is trained n times each time holding out one of the folds as an internal validation set. The Q^2 values for the n validation steps are then averaged to give an overall Q^2 value for that particular model. This process is repeated as the model parameters are optimised and the optimisation halted by looking for a maximum in the Q^2 curve. This method is considered useful because all the training data are at some point used to both train and validate the model, and in this way no data are 'wasted' in a holdout validation set. Unfortunately, even though this algorithm has been shown to be extremely effective in many circumstances it can potentially be misused when the number of variables $p \gg$ than the number of samples s as is generally the case in omic studies. A useful and lucid comparison between n-fold cross-validation and other model selection methods is provided by Kohavi (Kohavi, 1995). He points out that "*if a {machine learning classifier} is unstable for a particular dataset under a set of perturbations introduced by cross−validation, the accuracy estimate is likely to be unreliable*".

When $p \gg s$ there is a great danger that machine learning models will fall foul of the *curse of dimensionality* (Bellman, 1961). For example if one took 100 observations along one-dimension, one could draw a histogram of the results, and draw clear and statistically valid inferences (as illustrated in figure 2B). If one now considers the corresponding 200-dimensional hypercube, the 100 observations are now isolated points in a vast empty space (given a unit hypercube and random sample distribution the relationship between dimension and average distance to nearest neighbour is linear, such that at $p = 10$, 99% of the samples are greater than 0.5 from the origin (Hastie *et al.*, 2001)). To get similar coverage to the one-dimensional space would now require 100^{300} (or 10^{600}) observations (Bellman, 1961), a number that may be compared to 10^{17} which is the lifetime of the known Universe in seconds (Barrow and Silk, 1995). Now consider that we wish to construct a linear decision plane in order to discriminate between cases and controls in this multidimensional space. Also consider that the orientation of this decision plane will be optimised using the 'goodness of fit' criterion described above. Now, if $p \gg s$, and thus the hypercube (or X-space) is sparsely populated then the influence of each sample on the *goodness of fit* and therefore on the orientation of the decision plane can potentially be huge. This is of course dependent on the distribution of the samples in the X-space. If each class is tight and occupies a small and separate volume in space then the leverage effect of each individual may be small. However, if for example the multidimensional within-class variance > between class variance (i.e. the samples are more widely dispersed in X-space) then leverage could be substantial. So if we now use n fold cross-validation to optimise the orientation of the decision plane one can deduce that removing each validation set could also easily have a huge effect on the orientation of this plane for each of the n training models.

In linear algorithms such as PLS-DA or PC-DFA it is possible to check for instability by means of comparing

the loadings vector for each of the n training models. If the values differ considerably between models then in effect n completely different models have been created to explain the same discrimination; which one is then correct (and is this even a valid question)? This phenomenon is termed the *Rashomon Effect* by Breiman (2001). Rashomon is a Japanese movie in which four people, from different vantage points, witness an incident in which one person dies and another is supposedly raped. When they come to testify in court, they all report the same facts, but their stories of what happened are very different. In this paper Breiman states "*This effect is closely connected to what I call instability* (Breiman, 1966) *that occurs when there are many different models crowded together that have about the same training or test set error. Then a slight perturbation of the data or in the model construction will cause a skip from one model to another. The two models are close to each other in terms of error, but can be distant in terms of the form of the model.*" In situations such as these (in our hypothetical case where the n training models differ significantly) the leverage effect of each holdout set must be considerable, thus the assumption that the classifier is stable with respect to the data set is false and therefore n-fold validation will potentially produce a badly trained final model. Add into this the fact that if s is small then s/n will be even smaller, so it is very unlikely that each of the n validation sets will be fully representative of the whole data, and will probably be biased in relation to the metadata, adding to the probability of a bad model. Great care must therefore be taken when using this method of validation and again the only way of proving the model's unquestionable robustness is to use an independent test set as described above, or by using more sophisticated methodologies such as bootstrapping.

It has been implied that PLS-DA is capable of coping with this curse of dimensionality through the use of latent variables (Eriksson *et al.*, 2001). However as these latent variables are also optimized using goodness of fit, the problem remains dependent on the number and distribution of the sample population. The same source notes that "A necessary condition for PLS-DA to work reliably is that each class is tight and occupies a small and separate volume in X-space...Moreover, when some of the classes are not homogeneous and spread significantly in X-space, the discriminant analysis does not work" (Eriksson *et al.*, 2001).

Note that in some literature the terminology for the validation and test sets is exchanged (we prefer the version above with the final held-out set being the 'test set'). The key point is that the set finally used for assessing the model must not have had any part in its generation whatsoever. Any model that is not validated using samples that had absolutely *no* part in the production of the model is prone to overfitting. Period Ransohoff (2005) comments that only about 10% of transcriptomics studies have used independent (external) validation (Ntzani and Ioannidis, 2003).

The situation is not helped by the vast range in sophistication of modeling software available to users of metabolomic instrumentation. Users can often be encouraged to plot spurious correlations, which do not in isolation fully explain the significance of the model's predictive ability. For example in Discriminant Partial Least Squares if 10 latent variables are needed to produce a 'robust' and 'validated' model, simply plotting the scores values of the first two latent variables is not sufficient proof of class separation in X-space (especially if class confidence regions are not shown, and test set values are not overlaid on this plot). The final overall model prediction scores for the training, validation, and test sets are required. Only then can the model be correctly assessed using univariate statistics and graphical tools of the type described here. In the case of PLS this is quite straightforward as any model created can be reduced to the form of $\hat{\mathbf{y}} = \mathbf{bX}$ (where, $\hat{\mathbf{y}}$ is the predicted numerical response, $\mathbf{b}$ is the loadings vector if there is a single response variable, and $\mathbf{X}$ is the input data vector) (Alsberg *et al.*, 1998; Wold *et al.*, 2001) and many good visualization and statistics tools are available.

Rowland (2003) gives an excellent example of what may be required, in the context of genetic programming, a technique (e.g. (Koza, 1992; Langdon, 1998; Koza *et al.*, 2003)) that we favour since it gives easily interpretable rules that relate to actual metabolites rather than to latent variables and that also generalize well (Kell *et al.*, 2001; Kell, 2002, b; Goodacre and Kell, 2003; Kenny *et al.*, 2005).

It is also worth stressing that the distributions of samples in the training, validation and test sets should be such that they effectively come from the same populations, since if they do not failure is almost certainly guaranteed – and assessing whether this is the case can nevertheless expose bias. The most famous example is perhaps a tale (Goodacre *et al.*, 1996) about an experiment designed to train a neural network-based classifier involving the discrimination of images of landscapes from images that were otherwise similar but also contained battle tanks. Unfortunately the images without the tanks had been taken on a sunny day and those with the tanks on an overcast day, so the classifier had only learnt whether the sun was shining or not! Of course this came to light, as it were, when a second test set was used that did not have such a correlation between brightness and tankness.

Finally it is also worth noting that simply quoting R^2 and Q^2 values as a measure of quality for a given model provides a very vague description of predictive ability. Generally a $Q^2 > 0.5$ is regarded as good; however, this is very much application-dependent (Eriksson *et al.*, 2001). A simple scatter, box, or ROC plot (e.g. see figure 2B) of both training and overlaid test predictions give readers, and reviewers, far more confidence in any claims of model utility than a single conglomerate score with many assumed characteristics.

15. Credit assignment in multivariate calibration

Assuming that one is not simply looking for a model that will classify biological samples without explaining how (a so called 'black box' model), the primary aim of any multivariate machine learning of the type of present interest must be the discovery of important discriminatory biomarkers for a given metabolomics experiment. We have shown how it is possible to evaluate each metabolite univariately (i.e. in isolation). However, if no single metabolite is deemed a singularly good biomarker, one must expand the search to subsets of two or more metabolites that in combination provide good discrimination. Of course, as with parallel single metabolite significance tests, the possibility of False Discovery increases with the number of metabolites measured in parallel. However, whereas in the univariate case the number of parallel tests increases linearly, in the multivariate case the number of parallel tests is equal to $p!/(s!(p-s)!)$ where p is the total number of measured metabolites and, s, is the number of metabolites that appear in a candidate subset. For example, if $p = 100$ and $s = 5$ there are 7.52×10^7 possible combinations (and thus parallel tests).

The discovery of *significant biomarker subsets* (*SBSs*) can be achieved in one of two ways. First in what we will term 'the bottom up approach', where we hypothetically look at all the possible combinations of s in p, whilst varying s from 2 to p (The curve in figure 4. shows the relationship between the number of possible combinations of s in p, as s is varied between 2 and p. The area under the curve $= \sum_{s=2}^{p} p!/(s!(p-s)!)$ is the total number of possible combination of any subset size). For each subset a particular model is built (for example, the simplest being a weighted linear sum, of the form $\hat{\mathbf{y}} = \mathbf{bX_{sub}}$) and the predictions from this model are then evaluated by either univariate significance tests or model validation. Any model above a certain quality threshold can then have its constituent metabolites listed as an SBS. Of course using this strategy the amount of parallel tests will quickly become huge and computationally intractable. There are simple ways of limiting the number of tests, such as disallowing any models to be built that already contain SBSs of a lower dimensionality (i.e. if subset $\{w,x,y\}$ is an SBS then one may choose not to test $\{w,x,y,z\}$), or strictly limiting the maximum number of metabolites in a subset. A more heuristic solution is to use a search algorithm to traverse the metabolite subset space. In particular methods involving the use of evolutionary algorithms to 'evolve' SBSs or their equivalent have proved very successful for chemical problems (Lucasius and Kateman, 1994; Lucasius *et al.*, 1994; Horchner and Kalivas, 1995; Broadhurst *et al.*, 1997; Judson, 1997; Shaffer and Small, 1997; Gillet *et al.*, 2002; Jarvis and Goodacre, 2005). Alternatively, both the model and the subset to be tested can be evolved in the form of a Genetic Program (see above).

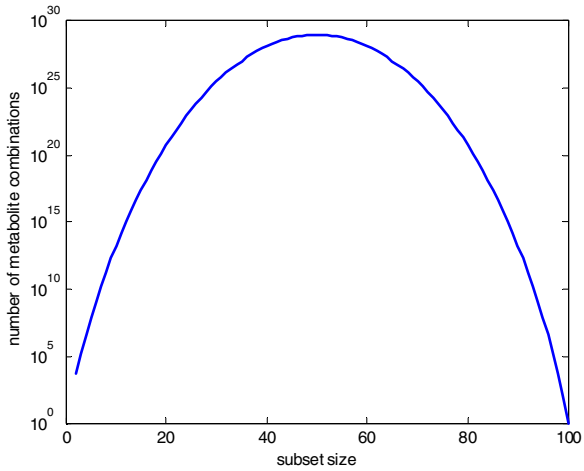


Figure 4. The combinatorial explosion of possible models that occurs when one tries to makes models using subsets of a total (here) of 100 metabolites. The abscissa is the total number of metabolites allowed to contribute to the model.

The second approach is what we will term 'the top down approach'. This method involves constructing a multivariate model using all of the available measured metabolites at once. Ignoring the many non-linear methods for the moment, most multivariate methods depend on parameter estimation using the covariance matrix of the raw data. Unfortunately when there are more variables than observations this can easily lead to a numerically unstable inverse matrix (or one incalculable due to singularity) (Seber and Wild, 1989). Therefore, the most popular linear multivariate models use latent variables (variables that serve to reduce the dimensionality of the raw data, such that stable covariance-based parameter estimation can be achieved). Either way, the final model can be expressed in terms of a multivariate classifier as defined in figure 1A. In the case of linear models such as MLR, PLS-DA, and PC-DFA this model can simply be reduced to the following single equation: $\hat{\mathbf{y}} = \mathbf{bX}$ (where $\mathbf{b}$ is the loadings, or weights vector, of length p). The predictions from these models can again be evaluated by either univariate significance tests, or model validation. Assuming proper validation has been performed then the problem is now how to relate individual loadings values to the significance of the model's performance. Many publications assume that if a metabolite has a relatively high loading (arbitrarily set, for instance, to > 2SD of the overall loadings distribution) then it is automatically deemed significant. Thus if q out of p loadings are greater than 2SD then the q corresponding metabolites are reported as being significant biomarkers. This of course may be true, but judicious use of Occam's Razor (i.e. when multiple competing theories have equal predictive powers, the principle recommends selecting those that introduce the fewest assumptions and postulate the fewest hypothetical entities; see also (Seasholtz and Kowalski, 1993)) implies that these assumptions be

tested. One method would be to use some sort of sensitivity analysis (Frey and Patil, 2002; Oakley and O'Hagan, 2004; Saltelli *et al*., 2004; White and Kell, 2004; Kell and Knowles, 2006). Another more straightforward method is to remove the q metabolites from the raw data and build two new models in parallel, one model with the *q* metabolites in isolation, the other with the remaining data. If the first model proves successful (and the loadings are relatively consistent) then the assumption that these *q* metabolites are significant biomarkers holds true; if not then the assumption is false. If the first model is a success but the second model fails then we can assume that the *q* metabolites are the sole important biomarkers. However, if both the models are a success then there are more significant biomarkers than originally thought and the process of isolation and model building must continue. Eventually, a list of robust metabolite subsets will be produced. An example of where this approach has been successfully used can be seen in (Catchpole *et al*., 2005). Breiman (2001) is also quite scathing about directly interpreting loadings plots and provides an example where variable isolation and remodeling proves that seemingly important variables are in fact not so influential after all.

16. Multi-objective optimisation

Related to this, as well as many input variables one may also have many objectives. This is an important field that is not really the subject of this review, but we do consider it useful to mention a few works on the subject (Ringuest, 1992; Dasgupta *et al*., 1999; Zitzler, 1999; Van Veldhuizen and Lamont, 2000; Deb, 2001; Coello Coello *et al*., 2002; Handl and Knowles, 2004, 2006a; Knowles and Hughes, 2005; Handl *et al*., 2006), noting especially that there are many conditions in which even single-objective problems can be given multiple objectives in order to aid in their solution (Knowles *et al*., 2001).

17. Visualisation issues

As stated above, we would argue that the best solution for allowing a full understanding of the significance or otherwise of a particular piece of work is to make available *all* the data and metadata. Given the existence of the Web, this is nowadays straightforward. However, data alone are but the ground substance for understanding, and there are a variety of tools that can help readers of papers evaluate the data using different views. This topic is usually referred to as Data Visualisation (Cleveland, 1993, 1994; Fortner, 1995; Wilkinson, 1999; Friendly, 2000; Tufte, 2001), and is especially significant in the analysis of multivariate omics data. For reasons of cost we can effect many more analyses in metabolomics than are likely to be done in transcriptomics or proteomics (we have already published elements of a 750-sample study (Kell *et al*., 2005) and are presently

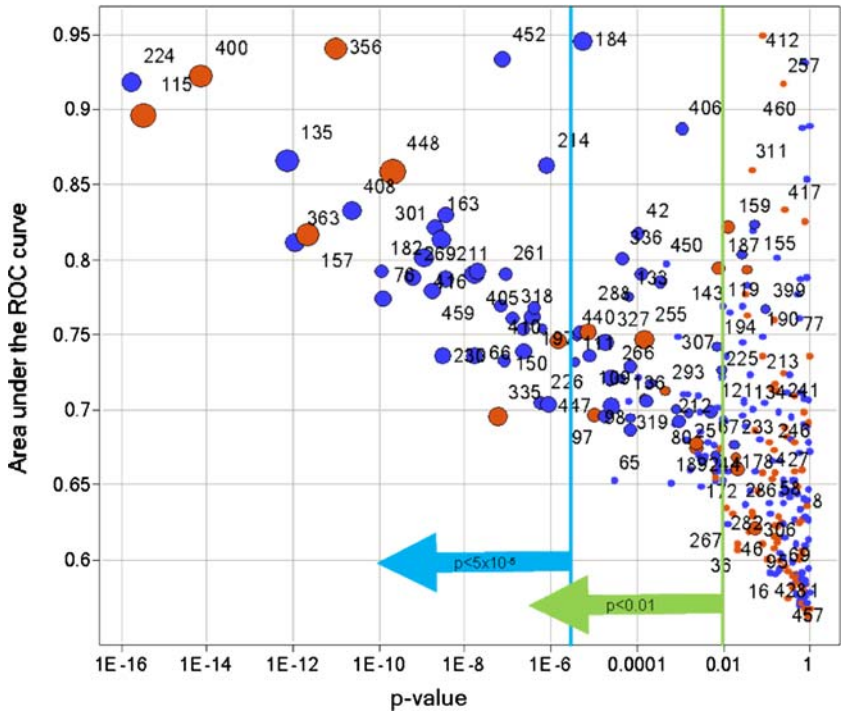


Figure 5. A plot of the area under the ROC curve vs the *p*-value for 286 metabolites from a dataset of 87 pre-eclamptic cases and 87 carefully matched controls (Kenny *et al*., 2005) analysed using an optimised GC-tof method (O'Hagan *et al*., 2005). The colour of the symbols encodes whether the metabolite is raised (blue) or lowered (red) in the cases while the size encodes the modified Z factor.

writing up one concerning some 1500 yeast gene knockouts), while FDRs are decreased by the fact that the number of metabolites is somewhat less than the number of transcripts or proteins (Oliver *et al*., 1998). Further, for fundamental reasons explained by metabolic control analysis (Kell and Westerhoff, 1986; Fell, 1996; Heinrich and Schuster, 1996; Cornish-Bowden and Cárdenas, 2000; Cascante *et al*., 2002), the variation in the metabolome is expected to be much greater (Raamsdonk *et al*., 2001; Urbanczyk-Wochniak *et al*., 2003; Kell, 2004), and thus we have the opportunity to achieve levels of statistical significance that the other omes are much less likely to achieve without very large sample numbers. Todd (2006) gives equivalent arguments for SNPs. The area under the ROC curve and the statistical p-value are not entirely unrelated, but they do nevertheless measure different properties; ROC curves are more sensitive to the actual class distributions across

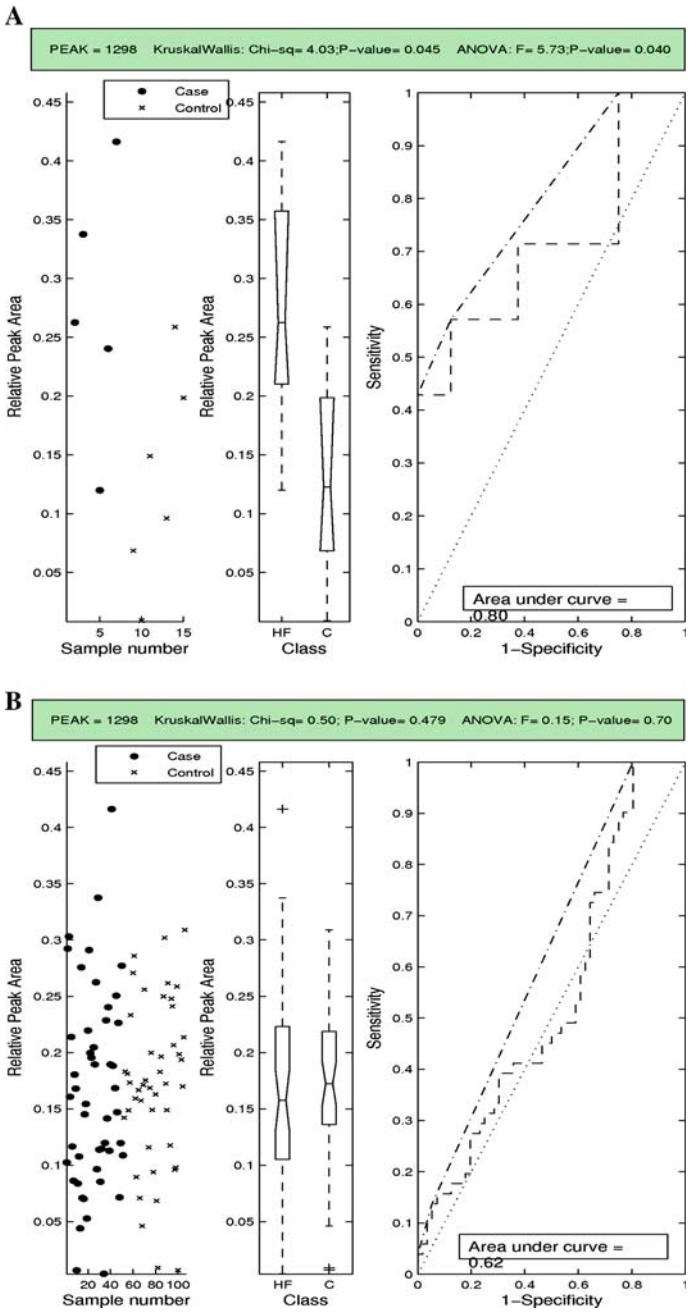


Figure 6. Effect of sample size on the apparent significance of a particular marker (metabolite 1298). A ten per cent random sample was taken from a study of 50 cases + 50 controls. A randomly selected 'significant' peak from this subset (where the critical p-value is 0.05) is shown in the form of a box-whisker plot and ROC curve A. This peak could easily be considered a significant biomarker (if no correction to the critical p-value is made to account for parallel statistical tests). If this metabolite is examined using the whole data set (B) the *p*-value increases to 0.48 and thus it is therefore not really a significant biomarker.

the total range, rather than being encoded as an estimate of this distribution as calculated by variance-based significance tests. Consequently, we find that a plot of the AUC (Area under the ROC curve) vs the p-value (for a particular significance test) is an extremely convenient way of comparing all the measured metabolites in a single graph. Figure 5 shows one for data acquired as part of the study described by Kenny *et al.* (2005). The figure also shows the conventional cutoff of $p < 0.01$ and that after applying a Bonferroni correction for a data set containing 200 metabolites to make it $5{\cdot}10^{-5}$. This decreases the number of 'significant' metabolites from 148 to 49.

As already discussed, small data sets can lead to false discoveries. To illustrate this circumstance we took a dataset (not yet published) of human serum metabolites measured using a GC-tof-MS assay (O'Hagan *et al.*, 2005). 50 'cases' with matched 'controls' were processed resulting in 272 statistically usable peaks. These data were split into two sets: *Set A* is the original 50 'case', 50 'control'; *Set B* simulates a very small sample size study by randomly sampling a matched subset of 5 'case', 5 'control' from **A**. AUC vs *p*-value plots were constructed for both sets (not shown). A randomly selected 'significant' peak from data set B (where the critical *p*-value is 0.05) is shown in figure 6A. This peak could easily be considered a significant biomarker (if no correction to the critical *p*-value is made to account for parallel statistical tests); however if this metabolite is examined using the whole data set (figure 6B) the *p*-value increases to 0.48. Thus one can see how easy it is to select a peak falsely as 'a possible marker' when there are so few samples.

Plots such as those of Figs 2B and 6 allow one to see the distribution of *individual* data points for a single variable in a manner that summary statistics such as mean and SD do not. We also prefer (Kell *et al.*, 2001) plots such as that of figure 7 in which individual samples are displayed as a function of a small number of variables (rather than latent variables).

The principal focus of this review has been the effective discovery of true metabolite biomarkers for binary classification experiments. Whether the biomarkers are discovered univariately or multivariately, the result is a set of important metabolities. As mentioned earlier, there can be significant correlations between different metabolites as a series of samples are measured as part of a study, whatever the mechanism (Kose *et al.*, 2001; Steuer *et al.*, 2003; Camacho *et al.*, 2005; Weckwerth and Morgenthal, 2005; Steuer, 2006). (This said, we note also that a combination of 'omic data and attendant GO terms shows that most interventions in pathways have their greatest effect on pathway elements that are nearest to the site of intervention, such that expression profiles may indeed be used to infer them (di Bernardo *et al.*, 2005), although this is not a watertight principle (Westerhoff and Kell, 1987).) Describing and subsequently understanding the

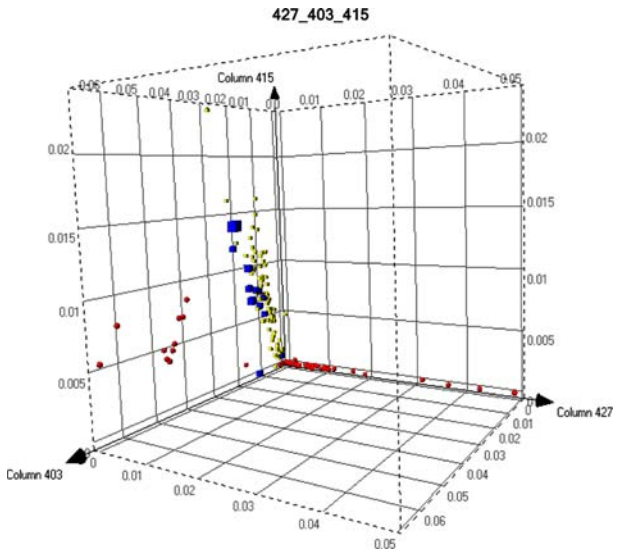


Figure 7. Visualisation of metabolomic data from individual samples. The plot shows the values of 3 metabolite peaks for cases (red) and control (blue and yellow) taken from a study of metabolic markers for pre-eclampsia (Kenny *et al.*, 2005).

cause of these correlations is sometimes difficult with many measured variables (particularly true with omic data). A convenient means for summarising such relationships for significant correlation is the so-called spring-embedding plot (figure 8), in which nodes (metabolites) are displayed as an undirected graph by constructing a virtual physical model and running an iterative solver to find a low-energy configuration. Following an approach proposed by Kamada and Kawai (1989), an ideal spring is placed between every pair of nodes such that its length is set to the shortest path distance between the endpoints, and the spring constant is proportional to the correlation between nodes. The springs push the nodes so their geometric distance in the layout approximates their path distance in the graph. (In statistics, this algorithm is also known as multidimensional scaling. Its application to graph drawing was noted by Kruskal and Seery (1980), and has since been used by many others (e.g., Eades, 1984; Ebbels *et al.*, 2006; Fruchterman and Reingold, 1991; Kim *et al.*, 2001)). Figure 8 was produced using *Graphviz* open source graph visualization software (Gansner and North, 2000), where the nodes are the Bonferonni-corrected significant metabolites, with their size inversely proportional to their *p*-value, and edges only exist between two nodes if the Spearman's rank correlation coefficient is > 0.6. It is worth noting that if the sample size is small the reliability of such coefficients is questionable.

A variety of other multidimensional scaling algorithms such as the Sammon mapping (1969) convert complex high-dimensional relationships into 'maps' that may be visualized in 2 and 3 dimensions. Kohonen's self-organising maps are one such (Kohonen, 1989;

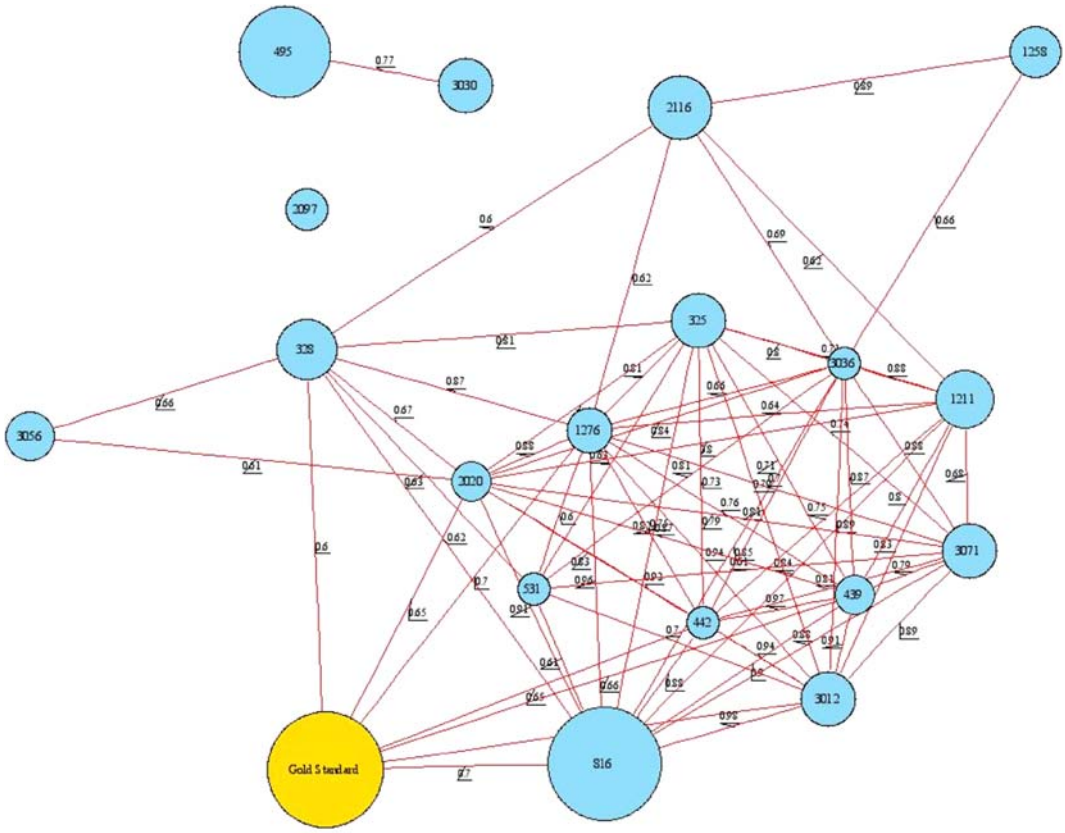


Figure 8. A spring-embedded correlation plot showing the correlations between metabolites from a case:control study where the nodes (whose size is inversely proportional to the p-value) are the significant metabolites after Bonferonni correction, while edges only exist between two nodes if the Spearman's rank correlation coefficient is >0.6. Also any isolated node is a significant biomarker but has no significant correlation to any other peak and its location is arbitrary. Gold standard = current standard measurement for classification.

Zupan and Gasteiger, 1993), while Agrafiotis and colleagues (Farnum *et al.*, 2003) describe a variety of visualization tools within the framework of molecular diversity, including, in consequence, those based on illustrating degrees of similarity.

18. Recommendations

We summarise our recommendations in the form of a Box. It is usefully read in conjunction with Ransohoff's paper (2005).

Potential problem	Recommendation
Bias	Make (and publish) a table that includes for each binary class the numbers (for categorical data) or mean and variance in the distribution of samples between cases and controls for each of the metadata classes (such as gender, age, pharmaceutical and recreational drug use, dietary information, ethnicity, etc.). For those that are unequal check that the same model that claims to discriminate cases from controls cannot discriminate well solely on the basis of these co-variates.
Inadequate sample size	The study must be powered correctly, recognising that for multiple (parallel) measurements/tests the necessary powering differs from that for a single measure.
Excessive false discovery rate due to multiple hypothesis testing	Apply suitable corrections, the best being the most conservative Bonferrroni correction that decreases the statistical p-value in rough proportion to the number of variables (metabolites) being modelled. Report only those that survive this test.
Inappropriate use of particular numerical methods	Check that the methods used are fit for purpose. For instance, methods like PLS can not deal sensibly with non-monotonic data (these are common in metabolomics studies, for instance when a drug is beneficial at low doses and toxic at high ones). Many other methods fail on disjoint populations.
Overfitting	Ensure that one is not simply learning what amounts to a training set. This is only really checked by external validation.

Potential problem	Recommendation
Failure to perform adequate validation and cross-validation	True validation in supervised learning systems requires testing the model on samples that have not been used in its construction at all. It is good practice to try and discriminate the training, (validation) and test samples for readers.

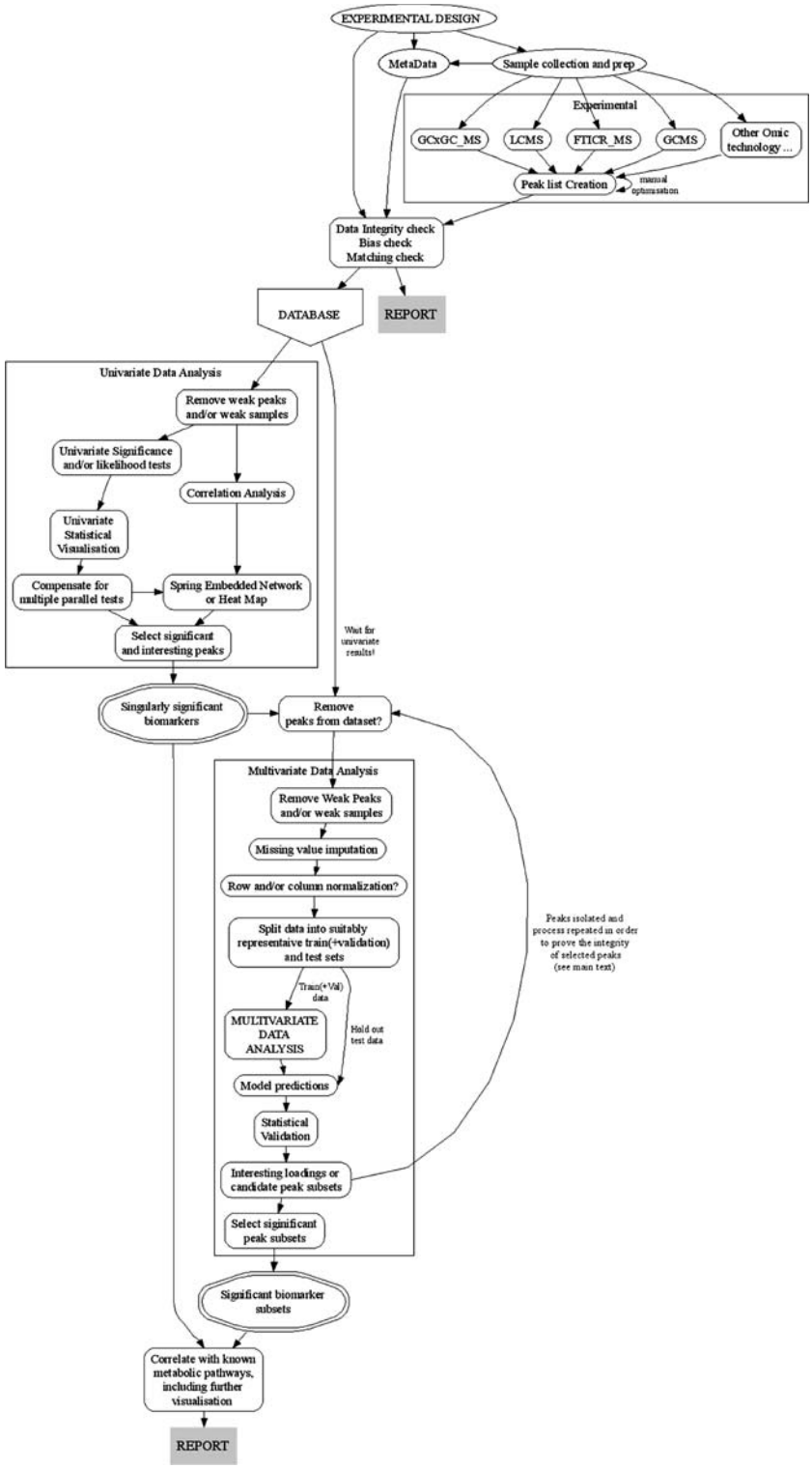


Figure 9. A workflow describing some of the useful analyses that may be performed during the generation and analysis of metabolomic biomarker data designed for binary class discrimination.

A generally useful concept in bioinformatics (Goble *et al.*, 2001) is the concept of the pipeline (Potter *et al.*, 2004; Brown *et al.*, 2005) or workflow (Peleg *et al.*, 2002; Oinn *et al.*, 2004, 2006; Stevens *et al.*, 2004; Romano *et al.*, 2005), especially in the present context for the analysis of omics data. Here, the tools involved in the data analysis are stitched together using standardised environments or interfaces to form a workflow, after which they may then be enacted in a more or less automated manner. Distributed environments using systems such as Taverna (Oinn *et al.*, 2004, 2006; Stevens *et al.*, 2004) to enact the necessary workflows provides an attractive way forward (Chen and Hofestädt, 2006; Kell, 2006). A suggested and useful pipeline for metabolomics data analysis is given in figure 9.

19. Concluding comments

The world of science is littered with examples of false conclusions being drawn from ostensibly well designed experiments (http://www.ems.psu.edu/~fraser/BadScience.html), and the bad design of experiments will usually ensure such an outcome. In many areas of post-genomic discovery, the proper methods of statistical analysis are not entirely clear, for instance how best to treat correlated variables in terms of a Bonferroni-type correction for significance when doing multiple hypothesis-testing. However, there are many elements of good practice that are well established in biomedicine and we in metabolomics have no reason not to follow them in our own experiments. Even well-established principles such as single- and preferably double-blinding of samples are not made explicit, and could as well be. But the most important issue is the recognition that with very many variables, potentially with significant noise, the false discovery rates and premature claims of significance are likely to be major problems. It is hoped that this review will lower their frequency.

Acknowledgments

We thank the BBSRC, MRC and BHF for financial support and many colleagues for useful discussions and examples.

References

Adriaans, P. and Zantinge, D. (1996). *Data Mining,* Addison-Wesley, Harlow, Essex.

Alsberg, B.K., Kell, D.B. and Goodacre, R. (1998). Variable selection in discriminant partial least-squares analysis. *Anal. Chem.* **70**, 4126–4133.

Alsberg, B.K., Woodward, A.M., Winson, M.K., Rowland, J. and Kell, D.B. (1997). Wavelet denoising of infrared spectra. *Analyst* **122**, 645–652.

Altman, D.G. (2001). Systematic reviews of evaluations of prognostic variables. *BMJ* **323**, 224–228.

Altman, D.G. and Deeks, J.J. (2002). Meta-analysis, Simpson's paradox, and the number needed to treat. *BMC Med. Res. Methodol.* **2**, 3.

Anthony, M. and Biggs, N. (1992). *Computational Learning Theory,* Cambridge University Press, Cambridge.

Baggerly, K.A., Morris, J.S. and Coombes, K.R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* **20**, 777–785.

Baker, S.G. (2003). The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *J. Natl. Cancer Inst.* **95**, 511–515.

Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

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

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour,* Princeton University Press, Princeton, NJ.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.* **57**, 289–300.

Bennett, K. and Demiriz, A. (1998). Semi-supervised support vector machines. *Adv. Neural Inf. Proc. Syst.* **12**, 368–374.

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

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

Berry, M.J.A. and Linoff, G.S. (2000). *Mastering the Art of Data Mining,* Wiley, New York.

Bezdek J.C. and Pal, S.K. (Eds) (1992). *Fuzzy Models for Pattern recognition: Methods That Search for Structures In Data.* IEEE Press., New York.

Bland, J.M. and Altman, D.G. (1995). Multiple significance tests: the Bonferroni method. *BMJ* **310**, 170.

Bland, M. (2000). *An Introduction to Medical Statistics,* Oxford University Press, Oxford.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). *Statistics for Experimenters,* Wiley, New York.

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

Breiman, L. (1966). The heuristics of instability in model selection. *Ann. Statist.* **24**, 2350–2381.

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

Brenner, H. and Gefeller, O. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat. Med.* **16**, 981–91.

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

Brent, R. (2000). Genomic biology. *Cell* **100**, 169–183.

Brent, R. and Lok, L. (2005). A fishing buddy for hypothesis generators. *Science* **308**, 504–506.

Brereton, R.G. (2003). *Chemometrics: Data Analysis for the Laboratory and Chemical Plant,* Wiley, New York.

Broadhurst, D., Goodacre, R., Jones, A. and Rowland Kell, J.J. D.B. (1997). Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Anal. Chim. Acta.* **348**, 71–86.

Brown, M., Dunn, W.B., Ellis, D.I., Goodacre, R., Handl, J., Knowles, J.D., O'Hagan, S., Spasic, I. and Kell, D.B. (2005). A metabolome pipeline: from concept to data to knowledge. *Metabolomics* **1**, 35–46.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. (1998). *Discovering Data Mining: From Concept to Implementation,* Prentice Hall, Englewood Cliffs, NJ.

Camacho, D., de la Fuente, A. and Mendes, P. (2005). The origins of correlations in metabolomics data. *Metabolomics* **1**, 53–63.

Cascante, M., Boros, L.G., Comin-Anduix, B., de Atauri, P., Centelles, J.J. and Lee, P.W. (2002). Metabolic control analysis in drug discovery and disease. *Nat. Biotechnol.* **20**, 243–249.

Casella, G. and Berger, R.L. (2002). *Statistical Inference, (2 edn)*. Duxbury, Pacific Grove, CA.

Catchpole, G.S., Beckmann, M., Enot, D.P., Mondhe, M., Zywicki, B., Taylor, J., Hardy, N., Smith, A., King, R.D., Kell, D.B., Fiehn, O. and Draper, J. (2005). Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc. Natl. Acad. Sci.* **102**, 14458–14462.

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

Chen, M. and Hofestädt, R. (2006). A medical bioinformatics approach for metabolic disorders: biomedical data prediction, modeling, and systematic analysis. *J. Biomed. Inform.* **39**, 147–159.

Chen, V.C.P., Tsui, K.L., Barton, R.R. and Meckesheimer, M. (2006). A review on design, modeling and applications of computer experiments. *IIE Trans.* **38**, 273–291.

Cleveland, W.S. (1993). *Visualizing Data,* Hobart Press, Summit, NJ.

Cleveland, W.S. (1994). *The Elements of Graphing Data,* Hobart Press, Summit, NJ.

Coello, C.A., van Veldhuizen, D.A. and Lamont, G.B. (2002). *Evolutionary Algorithms for Solving Multi-Objective Problems,* Kluwer Academic Publishers, New York.

Conover, W.J. (1980). *Practical Nonparametric Statistics,* Wiley, New York.

Cook, R.J. and Farewell, V.T. (1996). Multiplicity considerations in the design and analysis of clinical trials. *J. Roy. Stat. Soc. A* **159**, 93–110.

Cornfield, J. (1966). Sequential trials, sequential analysis and likelihood rinciple. *Am. Stat.* **20**, 18–23.

Cornish-Bowden, A. and Cárdenas, M.L. (2000). From genome to cellular phenotype-a role for metabolic flux analysis?. *Nat. Biotechnol.* **18**, 267–269.

Crary, S.B. (2002). Design of computer experiments for metamodel generation. *Analog. Integr. Circ. Sig. Proc.* **32**, 7–16.

Cui, X. and Churchill, G.A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **4**, 210.

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

Deb, K. (2001). *Multi-Objective Optimization Using Evolutionary Algorithms,* Wiley, New York.

Deming, S.N. and Morgan, S.L. (1993). *Experimental Design: A Chemometric Approach,* Elsevier, Amsterdam.

Demiriz, A., Bennett, K. and Embrechts, M.J. (1999). Semi-supervised clustering using genetic algorithms in Dagli, C.H., Buczak, A.L., Ghosh, J., Embrechts, M.J. and Ersoy, O. (Eds), *Intelligent Engineering Systems Through Artificial Neural Networks*. ASME Press, New York, pp. 809–814.

di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E. and Collins, J.J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* **23**, 377–383.

Diamandis, E.P. (2004). Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J. Natl. Cancer Inst.* **96**, 353–356.

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

Duesberg, P., Stindl, R. and Hehlmann, R. (2000). Explaining the high mutation rates of cancer cells to drug and multidrug resistance by chromosome reassortments that are catalyzed by aneuploidy. *Proc. Natl. Acad. Sci. USA* **97**, 14295–14300.

Eades, P. (1984). A heuristic for graph drawing. *Congressus Numerantium* **42**, 149–160.

Ebbels, T.M.D., Buxton, B.F. and Jones, D.T. (2006). springScape: visualisation of microarray and contextual bioinformatic data using spring embedding an 'information landscape'. *Bioinformatics* **22**, e99–e108.

Edwards, A.W.F. (1992). *Likelihood,* Johns Hopkins University Press, Baltimore.

Edwards, D. (2000). *Introduction to Graphical Modeling, (2 edn)*. Springer, Berlin.

Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *Am. Stat.* **37**, 36–48.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23**, 70–86.

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

Egan, J.P. (1975). *Signal Detection Theory and ROC Analysis,* Academic Press, New York.

Ein-Dor, L., Zuk, O. and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci USA* **103**, 5923–5928.

Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold, S. (2001). *Multi- and Megavariate Data Analysis: Principles and Applications,* Umetrics Academy, Umeå.

Evans, W.E. and Johnson, J.A. (2001). Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Annu. Rev. Genomics. Hum. Genet.* **2**, 9–39.

Evans, W.E. and Relling, M.V. (1999). Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**, 487–491.

Evans, W.E. and Relling, M.V. (2004). Moving towards individualized medicine with pharmacogenomics. *Nature* **429**, 464–468.

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

Farnum M.A., DesJarlais, R. and Agrafiotis, D.K. (2003). Molecular diversity in Gasteiger, J. (Ed.), *Handbook of Cheminformatics: vol 4 From Data to Knowledge*. Wiley/VCH, Weinheim, pp. 1640–1686.

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

Fielding, A.H. and Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.* **24**, 38–49.

Fortner, B. (1995). *The Data Handbook, (2 edn)*. Springer, New York.

Frey, H.C. and Patil, S.R. (2002). Identification and review of sensitivity analysis methods. *Risk Anal.* **22**, 553–578.

Friendly, M. (2000). *Visualising Categorical Data,* SAS Institute, Cary, NC.

Fruchterman, T.M.J. and Reingold, E.M. (1991). Graph Drawing by Force-Directed Placement. *Software –practice & experience* **21**, 1129–1164.

Gansner, E.R. and North, S.C. (2000). An open graph visualization system and its applications to software engineering. *Software: Practice and Experience* **30**, 1203–1233.

Gardner, M.J. and Altman, D.G. (1989). *Statistics with Confidence: Confidence Intervals And Statistical Guidelines,* BMJ, London.

Gillet, V.J., Khatib, W., Willett, P., Fleming, P.J. and Green, D.V.S. (2002). Combinatorial library design using a multiobjective genetic algorithm. *J. Chem. Inf. Comput. Sci.* **42**, 375–385.

Goble, C.A., Stevens, R., Ng, G., Bechhofer, S., Paton, N.W., Baker, P.G., Peim, M. and Brass, A. (2001). Transparent access to multiple bioinformatics information sources. *IBM. Syst. J.* **40**, 532–551.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996). Life With 6000 Genes. *Science* **274**, 546–567.

Golbraikh, A. and Tropsha, A. (2002). Beware of q2!. *J. Mol. Graph Model* **20**, 269–276.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999).

Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

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., Neal, M.J. and Kell, D.B. (1996). Quantitative analysis of multivariate data using artificial neural networks: a tutorial review and applications to the deconvolution of pyrolysis mass spectra. *Z. Bakteriol.* **284**, 516–539.

Goodman, S.N. and Royall, R. (1988). Evidence and scientific research. *Am. J. Publ. Health* **78**, 1568–1574.

Greenaway, W., May, J., Scaysbrook, T. and Whatley, F.R. (1991). Identification by gas chromatography-mass spectrometry of 150 compounds in propolis. *Z. Naturforsch. C* **46**, 111–121.

Grimes, D.S. (2006). Are statins analogues of vitamin D? *Lancet* **368**, 83–6.

Hand, D., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining,* MIT Press, Cambridge, MA.

Handl, J., Kell, D.B. and Knowles, J. (2006). Multiobjective optimization in bioinformatics and computational biology. *IEEE Trans Comput Biol Bioinformatics* (in the press).

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).

Handl, J. and Knowles, J. (2006a) An evolutionary approach to multiobjective clustering. *IEEE Trans Evol Comput* (in press).

Handl, J. and Knowles, J. (2006b). Semi-supervised feature selection via multiobjective optimization. *International Joint Conference on Neural Networks (IJCNN 2006)*. Proc WCCI 2006, IEEE Press, pp. 6351–6358.

Handl, J., Knowles, J. and Kell, D.B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics* **21**, 3201–3212.

Hanley, J.A. and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

Harrigan, G.G., LaPlante, R.H., Cosma, G.N., Cockerell, G., Goodacre, R., Maddox, J.F., Luyendyk, J.P., Ganey, P.E. and Roth, R.A. (2004). Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity. *Toxicol. Lett.* **146**, 197–205.

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, (5 edn)*. Oxford University Press, Oxford.

Hollander, M. and Wolfe, D.A. (1973). *Nonparametric Statistical Methods,* Wiley, New York.

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

Horning, E.C. and Horning, M.G. (1971). Metabolic profiles: gas-phase methods for analysis of metabolites. *Clin Chem* **17**, 802–809.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *J. Classif.* **2**, 193–218.

Hutchinson, A. (1994). *Algorithmic Learning,* Clarendon Press, Oxford.

Ioannidis, J.P. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *JAMA* **294**, 218–228.

Ioannidis, J.P. (2005b). Why most published research findings are false. *PLoS Med.* **2**, e124.

Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A. and Contopoulos-Ioannidis, D.G. (2001). Replication validity of genetic association studies. *Nat. Genet.* **29**, 306–309.

Ioannidis, J.P. and Trikalinos, T.A. (2005). Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *J. Clin. Epidemiol.* **58**, 543–549.

Ioannidis, J.P., Trikalinos, T.A., Ntzani, E.E. and Contopoulos-Ioannidis, D.G. (2003). Genetic associations in large versus small studies: an empirical assessment. *Lancet* **361**, 567–571.

Jarvis, R.M. and Goodacre, R. (2005). Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. *Bioinformatics* **21**, 860–868.

Jellum, E., Bjornson, I., Nesbakken, R., Johansson, E. and Wold, S. (1981). Classification of human cancer cells by means of capillary gas chromatography and pattern recognition analysis. *J. Chromatogr.* **217**, 231–237.

Jensen, F.V. (2001). *Bayesian Networks and Decision Graphs,* Springer, Berlin.

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

Judson, R. (1997). Genetic algorithms and their use in chemistry. *Rev. Comput. Chem.* **10**, 1–73.

Jung, S.H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics* **21**, 3097–104.

Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Inf .Proc. Lett.* **31**, 7–15.

Kannel, W.B. (1995). Range of serum cholesterol values in the population developing coronary artery disease. *Am. J. Cardiol.* **76**, 69C–77C.

Kell, D.B. (2002a). Genotype:phenotype mapping: genes as computer programs. *Trends. Genet.* **18**, 555–559.

Kell, D.B. (2002b). Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules. *Mol. Biol. Rep.* **29**, 237–41.

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

Kell, D.B. (2006). Metabolomics, modelling and machine learning in systems biology: towards an understanding of the languages of cells . The 2005 Theodor Bücher lecture. *FEBS J.* **273**, 873–894.

Kell, D.B., Brown, M., Davey, H.M., Dunn, W.B., Spasic, I. and Oliver, S.G. (2005). Metabolic footprinting and Systems Biology: the medium is the message. *Nat. Rev. Microbiol.* **3**, 557–565.

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 Knowles, J.D. (2006). The role of modeling in systems biology in Szallasi, Z., Stelling, J. and Periwal, V. (Eds), *System Modeling in Cellular Biology: From Concepts to Nuts and Bolts*. MIT Press, Cambridge, pp. 3–18.

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 Sonnleitner, B. (1995). GMP - Good Modelling Practice: an essential component of good manufacturing practice. *Trends Biotechnol.* **13**, 481–492.

Kell, D.B. and Welch, G.R. (1991). No turning back, Reductonism and Biological Complexity. *Times Higher Educational Supplement* **9th August**, 15.

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

Kemp, C., Griffiths, T., Stromsten, S. and Tenenbaum, J.B. (2003) Semi-supervised learning with trees. Adv. Neural Inf Proc Syst 16.

Kenny, L.C., Dunn, W.B., Ellis, D.I., Myers, J., Baker, P.N., The GOPEC Consortium and Kell, D.B. (2005) Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. Metabolomics 1, 227–234 - online DOI: 10.1007/s11306–005–0003–1.

Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N. and Davidson, G.S. (2001). A gene expression map for Caenorhabditis elegans. *Science* **293**, 2087–2092.

Kirkwood, B.R. and Sterne, J.A.C. (2003). *Essential Medical Statistics,* Blackwell, Oxford.

Kirschenlohr, H.L., Griffin, J.L., Clarke, S.C., Rhydwen, R., Grace, A.A., Schofield, P.M., Brindle, K.M. and Metcalfe, J.C. (2006). Proton NMR analysis of plasma is a weak predictor of coronary artery disease. *Nat. Med.* **12**, 705–710.

Knowles, J.D. and Hughes, E.J. (2005). Multiobjective optimization on a budget of 250 evaluations. Evolutionary Multi-Criterion Optimization (EMO 2005), LNCS 3410, 176–190 http://dbk.ch.umist.ac.uk/knowles/pubs.html.

Knowles, J.D., Watson, R.A. and Corne, D.W. (2001). Reducing local optima in single-objective problems by multi-objectivization in E. Zitzler *et al*., (ed.), Proc. 1st Int. Conf. on Evolutionary Multi-criterion Optimization (EMO'01), Springer, Berlin, pp. 269–283.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, pp. 1137–1143.

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, Mass.

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.

Kruse, R., Gebhardt, J. and Klawonn, F. (1994). *Foundations of Fuzzy Systems,* John Wiley, Chichester.

Kruskal, J.B. and Seery, J.B. (1980). Designing network diagrams. Proc. 1st General Conf. on Social Graphics, pp. 22–50.

Krzanowski, W.J. (1988). *Principles of Multivariate Analysis: A User's Perspective,* Oxford Univeristy Press, Oxford.

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

Langley, P., Simon, H.A., Bradshaw, G.L. and Zytkow, J.M. (1987). *Scientific Discovery: Computational Exploration Of The Creative Processes,* MIT Press, Cambridge, MA.

Leon, A.C. (2004). Multiplicity-adjusted sample size requirements: a strategy to maintain statistical power with Bonferroni adjustments. *J. Clin. Psychiatry* **65**, 1511–1514.

Li, H.-X. and Yen, V.C. (1995). *Fuzzy Sets And Fuzzy Decision-Making,* CRC Press, Boca Raton, Florida.

Li, T., Zhu, S., Li, Q., and Ogihara, M. (2003). Gene functional classification by semi-supervised learning from heterogeneous data. *Proc ACM Symp. Appl. Computing*. pp. 78–82.

Liang, Y. and Kelemen, A. (2006). Associating phenotypes with molecular events: recent statistical advances and challenges underpinning microarray experiments. *Funct .Integr Genomics* **6**, 1–13.

Linden, A. (2006). Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *J. Eval. Clin. Pract.* **12**, 132–139.

Lucasius, C.B., Beckers, M.L.M. and Kateman, G. (1994). Genetic algorithms in wavelength selection – a comparative-study. *Analytica Chimica Acta* **286**, 135–153.

Lucasius, C.B. and Kateman, G. (1994). Understanding and using genetic algorithms .2. Representation, configuration and hybridization. *Chemometrics and Intelligent Laboratory Systems* **25**, 99–145.

Mackay, D.J.C. (2003). *Information Theory, Inference and Learning Algorithms,* Cambridge University Press, Cambridge.

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

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

Metz, C.E. (1978). Basic principles of ROC analysis. *Semin Nucl Med* **8**, 283–98.

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

Michalski, R.S., Bratko, I. and Kubat, M. (Eds) (1998). *Machine Learning and Data Mining,* Methods and applications, Wiley, Chichester.

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

Miller, A.J. (1990). *Subset Selection in Regression,* Chapman and Hall, London.

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

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

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

Natarajan, S., Glick, H., Criqui, M., Horowitz, D., Lipsitz, S.R. and Kinosian, B. (2003). Cholesterol measures to identify and treat individuals at risk for coronary heart disease. *Am. J. Prev. Med.* **25**, 50–7.

Needham, C.J., Bradford, J.R., Bulpitt, A.J. and Westhead, D.R. (2006). Inference in Bayesian networks. *Nat. Biotechnol.* **24**, 51–53.

Ntzani, E.E. and Ioannidis, J.P. (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* **362**, 1439–44.

O'Hagan, S., Dunn, W.B., Brown, M., Knowles, J.D. and Kell, D.B. (2005). 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.* **77**, 290–303.

Oakley, J.E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *JR Stat. Soc. A* **66**, 751–769.

Obuchowski, N.A., Lieber, M.L. and Wians, F.H. Jr. (2004). ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clin. Chem.* **50**, 1118–25.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A. and Li, P. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 3045–3054.

Oinn, T., Li, P., Kell, D., Goble, C., Goderis, A., Greenwood, M., Hull, D., Stevens, R., Turi, D. and Zhao, J. (2006). *Taverna/Mygrid: Aligning a Workflow System with the Life Sciences Community Workflows for eScience,* Springer, Guildford 299–318.

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.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,* Morgan Kaufmann, San Francisco.

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

Peleg, M., Yeh, I. and Altman, R.B. (2002). Modelling biological processes using workflow and Petri Net models. *Bioinformatics* **18**, 825–37.

Perneger, T.V. (1998). What's wrong with Bonferroni adjustments. *BMJ* **316**, 1236–8.

Petricoin, E.F. III, Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C. and Liotta, L.A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577.

Potter, S.C., Clarke, L., Curwen, V., Keenan, S., Mongin, E., Searle, S.M., Stabenau, A., Storey, R. and Clamp, M. (2004). The Ensembl analysis pipeline. *Genome Res*. **14**, 934–941.

Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M., Berden, J.A., Brindle, K.M., Kell, D.B., Rowland, J.J., Westerhoff, H.V., van Dam, K. and Oliver, S.G. (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.

Ransohoff, D.F. (2004). Rules of evidence for cancer molecular-marker discovery and validation. *Nat*. *Rev*. *Cancer* **4**, 309–314.

Ransohoff, D.F. (2005). Bias as a threat to the validity of cancer molecular-marker research. *Nat*. *Rev*. *Cancer* **5**, 142–149.

Ransohoff, D.F. and Feinstein, A.R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N*. *Engl*. *J*. *Med*. **299**, 926–930.

Rapp, P.E. (1993). Chaos in the neurosciences: cautionary tales from the frontier. *Biologist* **40**, 89–94.

Raubertas, R.F., Rodewald, L.E., Humiston, S.G. and Szilagyi, P.G. (1994). ROC curves for classification trees. *Med*. *Decis*. *Making* **14**, 169–174.

Reiner, A., Yekutieli, D. and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368–375.

Ressom, H.W., Varghese, R.S., Abdel-Hamid, M., Eissa, S.A., Saha, D., Goldman, L., Petricoin, E.F., Conrads, T.P., Veenstra, T.D., Loffredo, C.A. and Goldman, R. (2005). Analysis of mass spectral serum profiles for biomarker selection. *Bioinformatics* **21**, 4039–4045.

Rifai, N., Gillette, M.A. and Carr, S.A. (2006). Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat*. *Biotechnol*. **24**, 971–983.

Ringuest, J.L. (1992). *Multiobjective Optimization: Behavioral and Computational Considerations,* Kluwer Academic Publishers, Dordrecht.

Romano, P., Marra, D. and Milanesi, L. (2005). Web services and workflow management for biological resources. *BMC Bioinformatics* **6**(Suppl 4), S24.

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

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

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm,* Chapman and Hall/CRC, London.

Rud, O.P. (2001). *Data Mining Cookbook,* Wiley, New York.

Sacks, J., Welch, W., Mitchell, T. and Wynn, H. (1989). Design and analysis of computer experiments (with discussion). *Statist Sci* **4**, 409–435.

Saltelli, A., Tarantola, S., Campolongo, F. and Ratt, M. (2004). *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models,* Wiley, New York.

Sammon, J.W. Jr. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans*. *Computers* **C-18**, 401–409.

Schena, M. (Eds) (2000). *Microarray Biochip Technology,* Eaton Publishing, Natick, MA.

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

Seber, G.A.F. and Wild, C.J. (1989). *Nonlinear Regression,* Wiley, New York.

Sehgal, M.S., Gondal, I. and Dooley, L.S. (2005). Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics* **21**, 2417–2423.

Shaffer, R.E. and Small, G.W. (1997). Learning optimization from nature – genetic algorithms and simulated annealing. *Anal*. *Chem*. **69**, A236–A242.

Sharp, S.J., Thompson, S.G. and Altman, D.G. (1996). The relation between treatment benefit and underlying risk in meta-analysis. *BMJ* **313**, 735–738.

Shipley, B. (2001). *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference,* Cambridge University Press, Cambridge.

Sokal, R.R. and Rohlf, F.J. (1995). *Biometry, (3 edn)*. Freeman, New York.

Stephan, C., Wesseling, S., Schink, T. and Jung, K. (2003). Comparison of eight computer programs for receiver-operating characteristic analysis. *Clin*. *Chem*. **49**, 433–439.

Steuer, R. (2006). On the analysis and interpretation of correlations in metabolomic data. *Brief Bioinform*. **7**, 151–158.

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

Stevens, R., McEntire, R., Goble, C., Greenwood, M., Zhao, J., Wipat, A. and Li, P. (2004). myGrid and the drug discovery process. *DDT Biosilico*. **4**, 140–148.

Storey, J.D. (2002). A direct approach to false discovery rates. *J*. *Roy*. *Stat*. *Soc*. *B* **64**, 479–498.

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **100**, 9440–5.

Tas, A.C. and van der Greef, J. (1994). Mass spectrometric profiling and pattern recognition. *Mass Spectrum Rev*. **13**, 155–181.

Todd, J.A. (2006). Statistical false positive or true disease pathway? *Nat*. *Genet*. **38**, 731–733.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525.

Tu, Y., Stolovitzky, G. and Klein, U. (2002). Quantitative noise analysis for gene expression microarray experiments. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **99**, 14031–14036.

Tufte, E.R. (2001). *The Visual Display of Quantitative Information, (2 edn)*. Graphics Press, Cheshire, CT.

Tukey, J.W. (1977). *Exploratory Data Analysis,* Addison-Wesley, Reading, MA.

Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L. and Fernie, A.R. (2003). Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep* **4**, 989–993.

Valiant, L.G. (1984). A theory of the learnable. *Comm ACM* **27**, 1134–1142.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H. and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009.

van Rijsbergen, C. (1979). *Information Retrieval,* Butterworth, London.

Van Veldhuizen, D.A. and Lamont, G.B. (2000). Multiobjective evolutionary algorithms: analyzing the state-of-the-art. *Evol Comput* **8**, 125–147.

Vapnik, V.N. (1998). *Statistical Learning Theory,* Wiley, New York.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. and Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl. Cancer Inst.* **96**, 434–442.

Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., Jatkoe, T., Berns, E.M., Atkins, D. and Foekens, J.A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679.

Weckwerth, W. and Morgenthal, K. (2005). Metabolomics: from pattern recognition to biological interpretation. *Drug Discov. Today* **10**, 1551–1558.

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.

Weiss, S.M. and Indurkhya, N. (1998). *Predictive Data Mining,* Morgan Kaufmann, San Francisco.

Westerhoff, H.V. and Kell, D.B. (1987). Matrix method for determining the steps most rate-limiting to metabolic fluxes in biotechnological processes. *Biotechnol. Bioeng.* **30**, 101–107.

White, H. (1992). *Artificial Neural Networks: Approximation and Learning Theory,* Blackwell, Oxford.

White, T.A. and Kell, D.B. (2004). Comparative genomic assessment of novel broad-spectrum targets for antibacterial drugs. *Comp. Func. Genomics* **5**, 304–327.

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

Williamson, P.R., Gamble, C., Altman, D.G. and Hutton, J.L. (2005). Outcome selection bias in meta-analysis. *Stat. Methods Med. Res.* **14**, 515–524.

Wold, S., Trygg, J., Berglund, A. and Antti, H. (2001). Some recent developments in PLS modeling. *Chemometr. Intell. Lab Syst.* **58**, 131–150.

Woodward, M. (2000). *Epidemiology: Study Design and Data analysis,* Chapman and Hall/CRC, London.

Xie, Y., Pan, W. and Khodursky, A.B. (2005). A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* **21**, 4280–4288.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control* **8**, 338–353.

Zhang, J.H., Chung, T.D.Y. and Oldenburg, K.R. (1999). A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* **4**, 67–73.

Zhou, X., Wang, X. and Dougherty, E.R. (2003). Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics* **19**, 2302–2307.

Zhou, X.H., Obuchowski, N.A. and McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine,* Wiley, New York.

Zitzler, E. (1999). *Evolutionary Algorithms for Multiobjective Optimization: Methods And Applications,* Shaker Verlag, Aachen.

Zupan, J. and Gasteiger, J. (1993). *Neural Networks for Chemists,* Verlag Chemie, Weinheim.

Zweig, M.H. and Campbell, G. (1993). Receiver-Operating Characteristic (ROC) plots - a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **39**, 561–577.