Supervised methods with genomic data: a review and cautionary view

Ramón Díaz-Uriarte
Bioinformatics Unit
Spanish National Cancer Center (CNIO)
Melchor Fernández Almagro 3
Madrid, 28029
Spain.
rdiaz@cnio.es
http://ligarto.org/rdiaz

Keywords: differential expression, prediction, prognostic, microarrays, multiple testing, molecular signatures, software, statistics, machine learning, observation study

Contributed chapter to *Data analysis and visualisation in genomics and proteomics*, by F. Azuaje, and J. Dopazo, (eds.).

Abstract

We review well accepted methods to address questions about differential expression of genes and class prediction from gene expression data. We highlight some new topics that deserve more attention: testing of differential expression of specific groups of genes, intra-group heterogeneity and class prediction, gene interaction in predictors, visualisation, difficulties in the biological interpretation of predictor genes and molecular signatures, and the use of ROC[Receiver Operating Characteristic curve]-based statistics for evaluating predictors and differential expression. We end with a review of some serious problems that can limit the potential of these methods; we focus specially on inadequate assessment of the performance of new methods (due to inadequate estimation of error rates and to the use of few and "easy" data sets) and failure to recognise observational studies and include needed covariates. A final comment is made about the need for freely available source code.

1 Chapter objectives

Reviews of the analysis of gene expression data (e.g. Drăghici, 2002; Parmigiani et al., 2003; Simon et al., 2003; Slonim, 2002; Speed, 2003; Tumor Analysis Best Practices Working Group, 2004) often mention three objectives: a) class comparison, or finding/ranking of differentially expressed genes; b) class prediction or prognostic prediction; c) class discovery, also know as clustering or unsupervised analyses. We will not discuss class discovery or clustering here (it is discussed elsewhere on this book) and will concentrate on class comparison and class prediction. For the remaining two broad type of problems, this chapter has three main objectives: a) To bring a statistician, computer scientist, or computational biologist quickly up to speed by providing pointers to the literature on well accepted and standard methods [1]. b) To emphasise some topics that deserve more attention and are open to additional theoretical, empirical, and computational contributions. c) To alert editors, reviewers, and general practitioners to several serious problems that can undermine the full potential of these techniques.

2 Class prediction and class comparison

Class comparison asks if different classes of subjects (e.g., lung cancer and prostate cancer patients) differ in their gene expression; the result is often a list of genes ranked by their degree of differential expression between classes; this objective can alternatively be to examine whether other non-categorical variables (such as expression of certain proteins or survival) are associated to gene expression. **Class prediction** or prognostic prediction tries to predict the class membership (or survival or protein expression or any prognostic variable) of a set of subjects given their gene expression data. Although related, these are different objectives that answer different biological questions and require different methods (unfortunately, this difference is not always recognised in empirical work). Ranking genes often precedes trying to use genes for class prediction (see also Sackett and Haynes, 2002), but genes that show large expression differences are not necessarily good predictors (e.g., p. 299 of Whitfield et al., 2003).

3 Class comparison: finding/ranking differentially expressed genes

The most common procedures analyse each and all of the genes of the array, "asking the same question" (e.g., "is this gene differentially expressed between prostate and lung cancer patients?") for **each gene of the array**. In contrast, when there are **prespecified groups** of genes, one can ask whether that subset of genes, as a whole, shows evidence of differential expression (e.g., "are genes X, Y, Z, which are involved in cell cycle, differentially expressed between prostate and lung cancer patients?").

Specially when asking the same question for each gene of the array, there are often two different objectives: to obtain a list of genes for which "their differential expression is statistically significantly different" and to rank genes based on some measure of how distant is the expression level between conditions (and this measure can be the p-value computed before) or how likely they are to differ. These objectives are related, but measuring

[1] Lack of space precludes a full review; other lists of references can be found in `http://www.biostat.umn.edu/~weip/course/ge/syl1.html` and `http://biosun01.biostat.jhsph.edu/~gparmigi/688/readings.html`, from two well-known statisticians

the likelihood of differential expression requires additional assumptions, and obtaining p-values is more delicate than simply ranking. Even when p-values are obtained, however, they are used as informal rules of inference and to guide future experiments, rather than to provide “black or white” answers.

Asking the same question for each gene of the array

Widely accepted methods, with available software, involve the use of standard statistical tests (e.g., $t-$test for two-class comparisons, ANOVA for multi-class comparisons, Cox models for survival data, etc), where analyses are carried out gene-by-gene (reviews in Cui and Churchill, 2003; Dudoit et al., 2002 b; Reiner et al., 2003; Simon et al., 2003, ch. 7). These analyses, although conducted gene-by-gene, need to take into account that thousands of null hypotheses are being tested (one for each gene): if we were to consider any of the genes with a “rejected null” as differentially expressed, we would end up with many false rejections. Appropriate correction for **multiple testing** is often conducted using either control of the **Family Wise Error Rate** or the **False Discovery Rate**. Controlling the Family Wise Error Rate refers to controlling the probability of making one or more false discoveries, or falsely rejecting the null, over the whole family of tests; this approached was detailed in Westfall and Young (1993) and its application to microarrays was pioneered by Dudoit et al. (2002 b). In contrast, the False Discovery Rate approach controls the expected proportion of erroneously rejected nulls among the rejected hypotheses; FDR controlled has been worked on mainly by Yoav Benjamini, Daniel Yekutieli, and their collaborators (see `http://www.math.tau.ac.il/~roee/index.htm`) for lists of references and links; a recent review and applications to microarrays is Reiner et al. (2003); other approaches related to, or variations of, FDR are Storey (2002); Storey and Tibshirani (2003) and references therein; Ge et al. (2003) compare and discuss most of these different approaches. Detailed discussion of whether control of FWER or FDR is the most appropriate for a given situation is beyond the scope of this chapter; however, in many exploratory studies control of FDR is probably what most researchers need. In addition, methods for control of FDR do not require the subset pivotality assumption (Westfall and Young, 1993) to hold, and therefore are applicable to a wider range of tests; in addition, although control of FDR, as originally proposed by Benjamini and Hochberg (Benjamini and Hochberg, 1995), works only for independent (or positively regression dependent) tests statistics, the results in Reiner et al. (2003) show that violation of this assumption is generally inconsequential and there are also resampling-based FDR approaches that account for the dependence of the tests statistics.

Most gene-by-gene approaches, when computing the statistic for each gene, do not use the information contained in the rest of the genes, which could be wasteful; hierarchical Bayes or **empirical Bayes** methods allow to **“borrow information” from all of the genes** in the array when making inferences about each of the genes (see Smyth, 2004)[2]. Although not as well known as the above methods, Parmigiani and colleagues (Garrett and Parmigiani, 2003; Parmigiani et al., 2002) model gene expression using **latent categories** that are interpreted as a gene being over-expressed, under-expressed, or at baseline

[2] Another review of “moderated” or “modified” t and F statistics is Cui and Churchill (2003). The approach developed by Gordon Smyth (Smyth, 2004) is applicable to a wide range of linear models (in contrast to some earlier approaches, that were only suited for specific comparisons), and an R (`http://www.R-project.org`) package, limma, is available from Bioconductor (`http://www.bioconductor.org`), and also incorporates accounting for multiple testing. However, although applicable to linear models, borrowing strength from all other genes is not as yet implemented in an easy to use tool for problems such as censored data, often analysed with Cox models.

expression[3]; these models allows for denoising of the expression data, can enhance interpretability and help with visualisation, and ease comparisons among platforms. Finally, Bickel (2004) has argued for testing **customised null hypothesis** that redefine differential expression in a biologically meaningful way (e.g., any non-zero difference is not necessarily biologically relevant), and use ROC-based statistics[4] (see below, section 5).

Asking questions about prespecified groups of genes

Among the tens of thousands of genes in an array, there might be prespecified sets of genes (e.g., those involved in cell cycle, or those found as relevant in a previous study) about which we might want to ask whether, as a whole, these subset of genes shows evidence of differential expression between groups of patients (or whether the expression of the whole set of genes is related to some other clinical variable, such as survival). Goeman et al. (2004) have proposed a method to test whether the expression pattern of a group of genes is related to some outcome of interest (be it class membership, survival, or a non-censored continuous variable). Their approach exploits the connection between differential expression among groups and predictability of clinical outcome, and the problem of number of genes being much larger than the number of samples is overcome using penalised regression models[5]. This method constitutes a very promising way of conducting tests of differential expression of subsets of genes[6].

A different approach has been suggested by Mootha et al. (2003), who examine if the members of a set of genes are enriched (i.e., a proportion larger than expected) among the most differentially expressed genes between two classes. This method should be applicable to any other type of comparison, such as multiclass comparisons (via ANOVA) or survival data. The main differences between the approaches of Mootha et al. (2003) and Goeman et al. (2004) are listed in Table 1. Although with a different objective, a method similar to that of Mootha et al. (2003) was proposed in Díaz-Uriarte et al. (2003) (see also Al-Shahrour et al., 2004); as in Mootha et al. (2003), the approach in Díaz-Uriarte et al. (2003) only works if genes with similar ranking or order belong to the the same set but, in contrast to Mootha et al. (2003), the approach of Díaz-Uriarte et al. (2003) will detect sets of genes that are not extreme in their statistic of differential expression; however, it is a method targeted towards exploratory purposes rather than for statistical testing of prespecified hypotheses.

4 Class prediction and prognostic prediction

Overview

As explained above, the goal here is to predict the clinically relevant characteristic of a subject (be it class membership, survival, prognosis, or any other variable of interest)

[3] They use a bayesian hierarchical mixture model —with uniform distributions for abnormally high and abnormally low expression and normal distribution for baseline expression—, and the model returns, for each gene and sample, the probability that it is over-, under-, or baseline-expressed. Software —R code— is available from `http://astor.som.jhmi.edu/poe/`. See also Newton et al. (2004) who use a semiparametric hierarchical mixture model for a somewhat similar problem.

[4] R code is available from `http://www.davidbickel.com`.

[5] Penalised regression models are related to shrinkage methods, such as ridge regression, and models with random effects, and will drive many coefficients towards zero; they allow the fitting of models even when the number of samples (i.e., arrays) is smaller than the number of variables (i.e., genes).

[6] Code is available as package "globaltest" from Bioconductor.

given the genetic profile of this subject. This is also an area of extremely active research, where the disciplines of statistics and machine learning have contributed much; Table 2 shows widely accepted methods and references.

Available reviews (see Table 2) show that relatively simple and well known methods such as k-Nearest Neighbour (KNN) and Diagonal Linear Discriminant Analysis (DLDA), together with Support Vector Machines (SVM), perform very well in most classification tasks in microarray data. Because of their performance and free availability[7] in quality implementations, DLDA, KNN, and SVM should probably be used routinely as benchmarks when proposing new methods.

Five specific issues

We will discuss five issues that probably deserve more attention. First, for the user it quickly becomes evident that many methods yield non-unique solutions (see also section 6.3) or, in other words, can return different solutions of very similar quality (e.g., prediction error rate), which itself leads to the question of how to choose among solutions. A direct way of approaching this problem is via **model combination and model averaging**. Model averaging is well known among Bayesians (e.g., Hoeting et al., 1999; Wasserman, 2000), and theory shows that a (weighted) average of predictions from several models should perform better (at least no worse) than predictions from any single model. Bayesian Model Averaging approach is not without problems, however, specially selection of priors and computation, and model definition. Model averaging is also available outside the Bayesian camp; stacking was initially proposed by Wolpert (1992) in the machine learning community, and later developed by Breiman (1996) and Ting and Witten (1999) (see also Hastie et al., 2001; Ripley, 1996, for short accounts). AIC-based model averaging has been developed by Buckland et al. (1997) and Burnham and Anderson (2002). Somorjai et al. (2002) show succesfull examples of stacking applied to MR and IR spectra[8]. Finally, random forests do a kind of model averaging by using an ensemble of trees.

Regardless of which model(s) are used, two general problems can affect all models/algorithms. First, most of the available methods assume additive effects of genes. Non-additive relationships or interactions, also called synergistic (or antagonistic) effects, are present when the outcome (e.g., being of class A) depends no just on the sum of the independent contributions of X and Y, but on their combined effects. Non-additive relationships are likely both between genes (e.g., the snail [NM_005985] gene) and between genes and other factors (section 6.4). Random forests (Breiman, 2001a; Liaw and Wiener, 2002) implicitly incorporate interactions as they are an ensemble of classification trees, but the actual interactions are not easy to see. Boulesteix and Tutz (2004); Boulesteix et al. (2003) have attempted to explicitly search for **patterns of interactions and use them in predictive models**. Second, the predictive capacity of many models can be hampered by **unrecognised heterogeneity within classes** that are regarded as homogeneous. Not much work has been done in this area. This problem, for instance, was recognised in the past (e.g. Rosenwald et al., 2003) and is dealt with by Munagala et al. (2004)[9].

[7] For instance, in R, DLDA is available in package "sma", KNN in package "class" (part of the VR bundle), and SVM in package "e1071", the latter from the libsvm library of Chang and Lin (2003).

[8] However, the author has attempted, without success, both stacking and AIC-based model combination of logistic and multiresponse linear regression with genomic data.

[9] Unfortunately, their code depends on non-free software.

A final set of problems involves the **biological interpretation of class prediction models** (together with making sense of information for potentially tens of thousands of coefficients). Most methods for building predictors tend not to return models that allow for easy biological interpretation of why and how those predictors are used, and how the genes in the predictors affect and relate to the class prediction. These problems are detailed in Díaz-Uriarte (2004) and an example are methods that use dimension reduction via PCA or PLS, where all genes have loadings on all the components, making it virtually impossible to interpret the biological meaning, if any, of the components[10].

Visualisation methods can help with biological interpretation in this task. For microarray data the **biplot**, as extended by Pittelkow and Wislon (2003)[11], is particularly useful, specially use of the GE-biplot both before and after selecting genes according to different criteria of relevance.

In addition, **"molecular signatures" or "gene expression signatures"** are key features in many studies in cancer research (Alizadeh et al., 2000; Golub et al., 1999; Pomeroy et al., 2002; Rosenwald et al., 2002; Shaffer et al., 2001; Shipp et al., 2002) and seem to imply the idea of coordinate expression of subsets of genes, so that some of these sets of coordinate expression would be related to some criterion of interest (e.g., cancer type, or survival) (for an almost definition of a signature see p. 375 in Shaffer et al., 2001). Recently Stegmaier et al. (2004) provide a very interesting example of a high-throughput, generic, method for screening of compounds that induce differentiation of leukaemia cells, based on gene expression signature of five genes; so gene expression signatures work as a surrogate for a biological state. In spite of their apparent relevance, however, there seems to be no approach for identifying molecular signatures. Recently, we proposed a method that is explicitly designed to try to identify molecular signatures: it finds sets of genes that are tightly coexpressed and that can be used as successful predictors (Díaz-Uriarte, 2004). This method could also help uncover situations that are inconsistent with the assumptions underlying the existence of a few, easily interpretable, signature components of coexpressed genes. However, there are several unsolved issues. On the one hand, the implicit model underlying Díaz-Uriarte (2004) is one where most of the genes are not relevant for prediction, relevant genes are involved in one and only one "signature component" (i.e., non-overlapping signature components), and the signature components are common, and behave similarly, in different groups; there are, however, richer biological models for biological signatures. In addition, there are related issues regarding differences in patterns of gene coexpression within and among-groups and potential instability concerns (see also section 6.3) about some results (see sections 3.2 and 3.3 in Díaz-Uriarte, 2004). Some of these issues might be solved with extensions to the method, and some might require completely different approaches. For example, modifications of the Plaid model of Lazzeroni and Owen (Lazzeroni and Owen, 2002) (see also Turner et al., 2004), which might allow a more principled, model-based, approach to the problem, within a richer class of models; or an extension of the simultaneous clustering and classification approach in Jörnsten and Yu (2003), where we could add normal mixture models with restrictions on the covariance matrix for clustering; or an approach based on the latent class methods of Parmigiani and colleagues (Garrett and Parmigiani, 2003; Parmigiani et al., 2002), where signature components are based on under-, over- or baseline expression (instead of expression levels), and potentially non-overlapping sets of genes for different classes. Work along these lines is currently in progress in our group. In any case, regardless of the exact method

[10] Naively interpreting components using loadings or eliminating genes with small loadings is often not justified and can lead to unexpectedly suboptimal solutions (Cadima and Jolliffe, 2001; Jolliffe, 2002)

[11] R code is available from Y. Pittelkow on request (see `http://cbis.anu.edu.au/software.html`).

used, it is also relevant that the search for molecular signatures highlights that finding a few sets of genes with biological interpretability can be worth even if it leads to small loses in predictive performance (see also Somorjai et al., 2003) because good classification performance, per se, does not shed any light into the underlying biological or clinical phenomena.

5 ROC curves for evaluating predictors and differential expression

Specially for the two-class setting, common measures of performance (e.g. Baker et al., 2002; Hastie et al., 2001; Pepe, 2003) are **Sensitivity**, or True Positive Rate, the probability of predicting a positive outcome when the true state is positive (i.e., $\frac{TP}{TP+FN}$ in Table 3) and **Specificity**, the probability of predicting a negative outcome when the true state of a case is negative (i.e., $\frac{TN}{TN+FP}$)[12].

Sensitivity and Specificity are often used to construct a Receiver Operating Characteristic (**ROC**) curve[13]. A ROC curve (see, e.g., figure 1) (e.g. Pepe, 2003; Pepe et al., 2001; van Belle, 2002, ch. 4) is a plot of Sensitivity in the ordinate against one minus Specificity or the **False Positive Rate** (i.e., $= \frac{FP}{TN+FP}$) in the abscissa. In other words, a plot of the probability of a hit against the probability of false alarm (Duda et al., 2001). This shows us how the sensitivity and the false positive rate change as we modify the threshold that classifies a subject as a member of one class or the other. In addition, we can use as a statistic the “Area under the curve” for a ROC curve, which is “(…) an overall measure of classification accuracy over all possible decision thresholds” (Bickel, 2004; Pepe, 2003).

ROC curves and ROC-based statistics are widely (and successfully) used to evaluate the diagnostic utility of medical tests (e.g., X-rays, ultrasounds, biochemical tests, etc, as reviewed in the excellent book by Pepe, 2003). It seems reasonable that similar approaches could be used with microarray data, specially since ROC-based statistics are very flexible devices that allow us, for example, to model covariate effects on the ROC curves, and to combine multiple test results (see Pepe, 2003, for review). As mentioned above (section 3), Bickel (2004) and Pepe et al. (2003) have argued for the use of ROC-based statistics to rank genes. These authors (see also Xu and Li, 2003) argue that ranking genes using ROC-based statistics is more meaningful than using t- and F-based statistics or p-values. Using the area under the ROC curve for two groups is a measure of differential expression that also provides information on the discriminatory capacities of genes: the empirical area under the ROC curve is equal to the probability that a randomly selected patient from one of the groups will have a larger expression value than a randomly selected patient from the other group (Bickel, 2004; Pepe, 2003), and this summary, from the clinical or biological perspective, is often much more meaningful than a t-statistic or a p-value. In addition, the area under the ROC curve is equivalent to the Wilcoxon rank sum statistic ($\equiv$ Mann-Whitney U statistic), and thus it is a distribution-free rank statistic (Pepe, 2003; Pepe et al., 2003). Besides the area under the whole curve, Pepe et al. (2003) suggest

[12] Lemon et al. (2003) have argued that the **Positive-predictive value** (PPV), “(…) the likelihood that a positive test result indicates a true positive” (i.e., $\frac{TP}{TP+FP}$) can be more relevant than sensitivity and specificity; however, this needs to be done carefully. In fact, for cancer screening the **Predictive Value Positve** (PVP) (similar in spirit to the PPV) and the **Predictive Value Negative** (PVN) are probably more important than the sensitivity and specificity, but they must be computed taking into account the prevalence, and not just the entries from the Table 3, as explained in Baker et al. (2002); Pepe (2003); van Belle (2002). This caveat is particularly important for very low prevalence diseases.

[13] The package ROC in Bioconductor offers several utilities for building and using ROC curves.

using the empirical estimates of the ROC at a given False Positive Rate, t_0, $ROC(t_0)$, and the partial area under ROC at t_0, $pAUC(t_0)$, as measures of differential expression. These statistics do depend on t_0, and a reasonable t_0 could be the False Positive Rate that is acceptable in practice: when screening asymptomatic people, where prevalence of cancer is very low in average risk populations, it is important to keep the False Positive Rate extremely low because otherwise there would be large numbers of people undergoing expensive and invasive procedures (Baker et al., 2002; Pepe et al., 2003).

6 Caveats and admonitions

6.1 Estimating the error rate of the predictor

To evaluate the performance of a predictor, it is common to provide the error rate of the predictions. However, many papers, including "high-profile" ones, report error rates that are severely biased, leading to overoptimistic claims about the performance of different methods. This is a most unfortunate situation because lack of appropriate rigour in the application and adherence to appropriate rules of evidence undermines trust in the promises of these technologies. These severe problems were addressed in the bioinformatics literature in Ambroise and McLachlan (2002) and Simon et al. (2003). In spite of the seriousness of the problem, the practice of reporting severely biased error rates is still common, and this has prompted a recent review (Ransohoff, 2004) that tries, once again, to alert users, reviewers, and editors against computing, reporting, and accepting overly optimistic error rates. We will review here the two most common problems, remembering that our objective when providing an estimate of the error rate is to provide an estimate of the likely error rate we will make when we apply our classifier to new data sets from the same population.

On possible problem is reporting the **"resubstitution rate"**, the error rate computed from the very same observations that were used to build the classifier, because the resubstitution error rate is severely biased-down due to overfitting: if we fit a classifier to a data set, we can expect it to "adapt to" some peculiarities of the data, which will make it work well with those data, but might lead it to work poorly with data not yet seen by the classifier or learner. This problem is even more serious with microarray data, where there are thousands of genes that can be part of a predictor. With so many variables, and so few samples, it is very easy to find a predictor that works perfectly in a completely random data set (see, for example, Fig. 8.4 in Simon et al., 2003). To solve this problem either cross-validation or bootstrap have been used; both methods build the predictor using a subset of the data, and then predict the values for the remaining data, thus insuring that the predictions are from data not used for the training.

A second common problem is to carry out the cross-validation *after* the gene selection: all samples are used for gene selection, and the cross-validation process does not include gene selection. This leads to very optimistic estimates of the error rate, as shown in Ambroise and McLachlan (2002) and Simon et al. (2003) because we incur in a problem similar to overfitting when the gene selection is carried out. The solution is to perform cross-validation or bootstrap so that all steps of the analysis (including gene selection, but also other potential steps such as imputation) are included in the cross-validation[14].

[14] Of course, all these comments apply to other approaches, such as stepwise, forward, and backward selection methods in linear or logistic regression; in addition, these selection methods are well known for their instability and their leading to biased p-values (e.g., section 4.3 in Harrell, 2001). Anyway, these variable selection methods ought to be subject, too, to cross-validation or bootstrap.

Whether cross-validation (and what size of folds) or bootstrap (and what type of bootstrap) should be used is beyond the scope of this review (see Ambroise and McLachlan, 2002; Braga-Neto and Dougherty, 2004; Davison and Hinkley, 1997; Efron and Gong, 1983; Efron and Tibshirani, 1993, 1997; Simon et al., 2003).

6.2 Reinventions of the wheel and comparisons among methods

There are two related problems that slow the development of the field just simply by overwhelming researchers with new publications and algorithms. On the one hand, there is a fair amount of "repeated reinventions of the wheel", or ignorance of previously dealt with problems (many of them, with solutions by now). In addition, many new methods that are published are not evaluated against "standard" competing methods (see also section 4), or are evaluated using only data sets regarded as "easy" (e.g., the Leukaemia data set of Golub et al., 1999), making it hard to asses how new methods really perform (in sharp contrast, for example, Dettling and Bühlmann, 2004, use six different data sets and three competing predictors). Hopefully, more strict standards for evaluation of proposed methods (together with the requirements of a freely available "reference implementation" —section 7) will decrease the amount of new proposed methods, will shorten the "to-read" pile, and will allow researchers to carry out wider and more exhaustive searches for more mature solutions to similar problems from other fields.

6.3 Stability of results or which set of candidate genes is biologically relevant?

Suppose a predictor has been built that includes 20 genes. How far can we take biological interpretation on the relevance of these genes? A paper by Somorjai et al. (2003) suggests that often not very far; the problem is the instability or non-uniqueness of results, a phenomenon called the "Rashomon effect" by L. Breiman (Breiman, 2001b). It is very common that, if we re-run a given procedure with only minor changes or using bootstrap samples, we end up with very different sets of models, suggesting that there are many different "optimal" subsets of genes (because there are many different descriptions that give approximately the same minimum error rate Breiman, 2001b). Somorjai et al. (2003) show how this can arise because of small sample sizes and an extremely small sample per feature ratio (i.e., very small number of arrays relative to the number of genes). Somorjai et al. (2003) suggest using a variety of classifiers or predictors and finding whether the same features are selected; if the same set of genes is repeatedly selected, we would be more confident that the set is reasonably robust. Of course, this way of examining robustness to selection methods cannot be used if feature selection is carried out using the same filter method for different classifiers (e.g., finding the 200 genes with largest F-ratio, and then using those 200 genes with DLDA, KNN, and SVM). Additionally, the bootstrap can be used to examine variation in solutions achieved. The multiplicity problem deserves much more careful attention and prompts for cautious interpretation of results.

6.4 Recognising observational studies and the need of including covariates

Although microarray studies are often referred to as "experiments" they are frequently observational studies. The differences between observational and experimental studies are well known in statistics and epidemiology, and affect both analyses and interpretation of results. Observational studies present several potential problems, specially:

- Background differences between groups and presence of potential confounding variables; confounding is a pervasive problem. Potter (2003) illustrates it with examples of the relation between vegetable consumption and cancer being confounded by differences in smoking associated with vegetable consumption (smokers also tend to eat fewer vegetables) and differences in expression profiles between cancer types being related to the unmeasured confounding of age and sex. A related problem is interaction, such as when the degree of association between an exposure factor (e.g., expression of gene A) and the disease is different for different levels of the confounding variable, such as sex (Collett, 2003); there is evidence that this might be the case in lung cancer (Patel et al., 2004). The problems of confounding and interaction are discussed in more detail below.

- Biases arising from handling of units (e.g., case samples are frozen several hours after collection whereas control samples are frozen immediately; Potter, 2003)) or from biases during the selection of subjects for the study or from informative patterns of missingness.

- Samples too small to allow for generalisations to the populations of interest, and problems of reproducibility.

These issues are well known in epidemiology, which studies patterns of disease and possible factors that affect these patterns of disease by using mainly observational data (Collett, 2003; Potter, 2003). However, as indicated by Potter (2003) concerns related to microarrays being often observational studies are mostly absent from standard papers and textbooks on microarray design and analysis (Churchill, 2002; Drăghici, 2002; Simon and Dobbin, 2003; Simon et al., 2003; Speed, 2003; Tumor Analysis Best Practices Working Group, 2004; Yang and Speed, 2002). In particular, it is surprising that confounding and interaction have not been given more consideration (see also Ntzani and Ioannidis, 2003, who show that an alarmingly large number of predictive studies with DNA arrays do not include adjustments for other known, and potentially competing, predictors). Confounding and interaction can be addressed, at least partially, by appropriately using relevant covariates in the statistical models[15].

How is this relevant for microarray data? As Potter (2003) illustrates, many of the differences seen in expression profiles between different types of cancers can be the result of confounding by age and sex. Another example is provided by Patel et al. (2004), who have reviewed evidence that clearly indicates that there are sex-specific differences in susceptibility to, and biology and progression of, lung cancer. Some of these sex-specific differences could be related to differential expression of certain genes, decreased DNA repair capacity in women, increased incidence of certain mutations, and estrogen signalling. All of these factors and differences make it extremely likely that both confounding and interaction will occur related to sex in studies of the relationship between gene expression and cancer[16], and in the development of predictive models. However, the good news is that sex and age of patients are often known for each microarray sample; these two variables,

[15] Harrell (2001, pp. 3 and 390) emphasises the importance of multivariable modelling in observational studies because they allow us to control (hold constant mathematically the effect of) variables that might differ between groups because the study is observational

[16] Interactions are very likely given the complex mappings between transcript levels and protein levels (O'Neill et al., 2003)

thus, should routinely be included in the analysis as covariates and to examine possible interactions. (Interestingly, Patel et al., 2004, call for undertaking sex-specific research in lung cancer). Of course, comments regarding sex and age are extensive to other potential confounders (e.g., diet, exercise, region of origin, etc), for which information might be available. Controlling for the effect of confounders with strong effects (and, from the biology we know, sex and age are likely to be confounders with strong effects in many cases), can lead to increases in statistical power, because a source of variation is being taken into account rather than being thrown into the error term[17]. Thus, by controlling the effects of covariates we can be more likely to detect differential expression between conditions. On the other hand, if differences between groups are mainly due to confounders (e.g., because of a disproportionate presence of one sex in one of the groups), only after controlling for the confounder can we trust that differential expression of certain genes or the predictive ability of our model is not due to confounding. With respect to interactions (e.g., that the effects of changes in the expression of certain genes depend strongly on, say, sex), their presence can be an important finding in itself, as is the case of sex-differences and lung cancer biology (Patel et al., 2004). Finally, if there are interactions with, say, sex, we will obtain lower error rates if we develop different predictive models for men and women than if we use a model that makes predictions independently of sex.

6.5 Collaboration between statisticians and biologists and the use of software "magic bullets"

Successful use of microarrays to answer biologically relevant questions will require close collaboration between biologists and statisticians during the complete process of the study. The need for statisticians' advice during the experimental design has been discussed before (Churchill, 2002; Simon and Dobbin, 2003; Yang and Speed, 2002) and is not the subject of this chapter; however, it should be remembered that full details of the experimental set up are necessary for the use of appropriate statistical methods. In the context of this chapter, statisticians need to realize that there are often many subtleties in the interpretation of microarray results that preclude simple mappings from RNA expression data to phenotypes (O'Neill et al., 2003). At the same time, statistical help is needed to insure that the statistical model and test being used is addressing the biological questions of interest. What in any case is unrealistic is to expect that if the biologist sends a file with 15000 rows by 200 columns (genes by subject) to the statistician, the statistician will return to the biologist the list of, say, 30 genes that are the answer to the biological question. But that is, in fact, what some users often expect from software tools or statistical consulting, and what some statisticians might believe is possible/desirable. And this also means that the questions asked are sometimes reformulated to accommodate the available software.

The problem of those expectations and procedures is that they lack key ingredients often needed to provide an answer to the underlying biological question. Table 4 lists some typical questions that a statistician might ask[18]. Only after these (and other) questions have been answered, it is time to search for the appropriate tool, which might be a web tool, a GUI-based stats program, or might require the competent use of command-driven programs or the development of new programs to carry out the customised required analyses.

[17] This is the idea behind blocking in experimental design: controlling a know source of variation.

[18] van Belle (2002) provides a very accessible account for the reasons behind those, and many other, questions statisticians ask.

7 Final note: source code should be available

Many new methods papers are published every month, and biologists and applied statisticians do not have the time to implement each and every idea that is published, nor to deal with the complications associated with patented algorithms. Sometimes, however, when researchers ask for software from authors of methods paper they face answers such as "…my method is straightforward to implement from the explanations in my paper", "…the method will soon be available as part of program XYZ (which is proprietary)", or " …I am not in the business of providing software to anyone".

In the opening lecture of the Royal Statistical Society meeting of 2002, titled "Statistical methods *need* software", Brian Ripley (Ripley, 2002) proposed "(…) a reference implementation, some code which is warranted to give the authors intended answers in a moderately-sized problem. It need not be efficient, but it should be available to anyone and everyone." Calls for availability of software, including source code, in bioinformatics research have also been made in other settings (e.g. Dudoit et al., 2003; Marshall, 2003), and the Open Bioinformatics Foundation (`http://www.open-bio.org/`) is "focused on supporting open source programming in bioinformatics." The Free Software Foundation (`http://www.fsf.org`) and the Open Source Initiative (`http://www.opensource.org/`) explain free and open source software. The reasons for making source code available in bioinformatics and microarray research are summarised by Dudoit et al. (2003, p. 46) and are reproduced in Table 5.

In this review, and following the above spirit, we have been highly biased towards methods for which software, including source code, is available; besides the philosophical issues involved, this is also a pragmatic decision.

8 Acknowledgements

J. Goeman and V. Mootha for answers about the workings of their procedures. C. Lázaro-Perea provided detailed and careful comments on the ms. that forced me to rewrite it and greatly improved it. The editors and reviewers for very helpful comments on the ms. The author was partially supported by the Ramón y Cajal program of the Spanish MCyT (Ministry of Science and Technology); funding partially provided by project TIC2003-09331-C02-02 of the Spanish MCyT.

References

Al-Shahrour, F., J. Herrero, Á. Mateos, J. Santoyo, R. Díaz-Uriarte, and J. Dopazo (2004). Using gene ontology on genome-scale studies to find significant associations of biologically relevant terms to groups of genes. *Neural Networks for Signal Processing XIII*.

Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, and L. M. Staudt (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature 403*, 503–511.

Ambroise, C. and G. J. McLachlan (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA 99*(10), 6562–6566.

Baker, S. G., B. S. Kramer, and S. Srivastava (2002). Markers for early detection of cancer: statistical guidelines for nested case-control studies. *BMC Med Res Methodol 2*, 4.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statistical Society, Series B 57*, 289–300.

Bickel, D. R. (2004). Degrees of differential gene expression: detecting biologically significant expression differences and estimating their magnitudes. *Bioinformatics 20*, 682–688.

Boulesteix, A. and G. Tutz (2004). Identification of interaction patterns and classification with applications to microarray data. *SFB386 Discussion paper 369.*

Boulesteix, A. L., G. Tutz, and K. Strimmer (2003). A CART-based approach to discover emerging patterns in microarray data. *Bioinformatics 19*, 2465–2472.

Braga-Neto, U. M. and E. R. Dougherty (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics 20*, 374–380.

Breiman, L. (1996). Stacked regressions. *Machine Learning 24*, 49–64.

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

Breiman, L. (2001b). Statistical modeling: the two cultures (with discussion). *Statistical Science 16*, 199–231.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997). Model selection: an integral part of inference. *Biometrics 53*, 603–618.

Bureau, A., J. Dupuis, B. Hayward, K. Falls, and P. Van Eerdewegh (2003). Mapping complex traits using Random Forests. *BMC Genet 4 Suppl 1*, S64.

Burnham, K. P. and D. R. Anderson (2002). *Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed.* New York: Springer.

Cadima, J. F. C. L. and I. T. Jolliffe (2001). Variable selection and the interpretation of principal subspaces. *Journal of Agricultural, Biological, and Environmental Statistics 6*, 62–79.

Chang, C.-C. and C.-J. Lin (2003). Libsvm: a library for support vector machines. Technical report, URL: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nat Genet 32 Suppl*, 490–495.

Collett, D. (2003). *Modelling binary data, 2nd ed.* London: Chapman and Hall.

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

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap methods and their application.* Cambridge: Cambridge University Press.

Dettling, M. and P. Bühlmann (2004). Finding predictive gene groups from microarray data. *J. Multivariate Anal.*, in press.

Díaz-Uriarte, R. (2004). A simple method for finding molecular signatures from gene expression data. Technical report, URL:http://www.arxiv.org/abs/q-bio.QM/0401043.

Díaz-Uriarte, R., F. Al-Shahrour, and J.Dopazo (2003). *Methods of Microarray Data Analysis III, papers from Camda '02*, Chapter The Use of Go Terms to Understand the Biological Significance of Microarray Differential Gene Expression Data, pp. 233–247. Kluwer.

Drăghici, S. (2002). *Data analysis for DNA microarrays.* London: Chapman and Hall.

Duda, R. O., P. E. Hart, and D. G. Stork (2001). *Pattern classification, 2nd ed.* New York: John Wiley.

Dudoit, S., J. Fridlyand, and T. P. Speed (2002 a). Comparison of discrimination methods for the classification of tumors suing gene expression data. *J Am Stat Assoc 97*(457), 77–87.

Dudoit, S., Y. Yang, M. Callow, and T. Speed (2002 b). Statistical methods for identifying differentially expressed genes in replicated cdna experiments. *Statistica Sinica 12*, 111–139.

Dudoit, S., R. C. Gentleman, and J. Quackenbush (2003). Open source software for the analysis of microarray data. *Biotechniques Suppl*, 45–51.

Efron, B. and G. Gong (1983). A leisurely look at the bootstrap, the jacknife, and cross-validation. *Am Stat 37*(1), 36–48.

Efron, B. and R. J. Tibshirani (1993). *An introduction to the bootstrap.* London: Chapman and Hall.

Efron, B. and R. J. Tibshirani (1997). Improvements on cross-validation: the .632+ bootstrap method. *J. American Statistical Association 92*, 548–560.

Frank, I. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Garrett, E. and G. Parmigiani (2003). *The analysis of gene expression data: methods and software*, Chapter POE: Statistical methods for qualitative analysis of gene expression, pp. 362–387. Springer.

Garthwaite, P. H. (1994). An intepretation of partial least squares. *J Am Stat Assoc 89*(425), 122–127.

Ge, Y., S. Dudoit, and T. Speed (2003). Resampling-based multiple testing for microarray data analysis (with discussion). *TEST 12*, 1–77.

Ghosh, D. (2003). Penalized discriminant methods for the classificatin of tumors from gene expression data. *Biometrics In press*.

Goeman, J. J., S. A. van de Geer, F. de Kort, and H. C. van Houwelingen (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics 20*, 93–99.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Gunther, E. C., D. J. Stone, R. W. Gerwien, P. Bento, and M. P. Heyes (2003). Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proc Natl Acad Sci U S A 100*, 9608–9613.

Gusnanto, A., Y. Pawitan, and A. Ploner (2003). Variable selection in gene and protein expression data. Technical report, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm.

Harrell, J. F. E. (2001). *Regression modeling strategies*. New York: Springer.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The elements of statistical learning*. New York: Springer.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: a tutorial. *Statistical Science 14*, 382–417.

Huang, X. and W. Pan (2003). Linear regression and two-class classification with gene expression data. *Bioinformatics 19*, 2072–2078.

Jolliffe, I. T. (2002). *Principal component analysis, 2nd ed.* New York: Springer.

Jörnsten, R. and B. Yu (2003). Simultaneous gene clustering and subset selection for sample classification via MDL. *Bioinformatics 19*, 1100–1109.

Lazzeroni, L. and A. Owen (2002). Plaid models for gene expression data. *Statistica Sinica 12*, 61–86.

Lemon, W. J., S. Liyanarachchi, and M. You (2003). A high performance test of differential gene expression for oligonucleotide arrays. *Genome Biol 4*, R67.

Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *Rnews 2*, 18–22.

Marshall, E. (2003). The upside of good behavior: make your data freely available. *Science 299*, 990.

Mootha, V. K., C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet 34*, 267–273.

Munagala, K., R. Tibshirani, and P. O. Brown (2004). Cancer characterization and feature set extraction by discriminative margin clustering*. *BMC Bioinformatics 5*, 21.

Newton, M. A., A. Noueiry, D. Sarkar, and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics 5*, 155–176.

Nguyen, D. V. and D. M. Rocke (2002). Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics 18*(9), 1216–1226.

Ntzani, E. E. and J. P. Ioannidis (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet 362*, 1439–1444.

O'Neill, G. M., D. R. Catchpoole, and E. A. Golemis (2003). From correlation to causality: microarrays, cancer, and cancer treatment. *Biotechniques Suppl*, 64–71.

Park, P. J., L. Tian, and I. S. Kohane (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics 18, S1*(S1), S120–S127.

Parmigiani, G., E. Garrett, R. Anbazhaghan, and E. Gabrielson (2002). A statistical framework for expression-based molecular classification in cancer. *J. Royal Statistical Society, Series B 64*, 717–736.

Parmigiani, G., E. Garrett, R. Irizarry, and Z. SL. (2003). *The analysis of gene expression data: methods and software*, Chapter The analysis of gene expression data: an overview of methods and software, pp. 1–45. Springer.

Patel, J. D., P. B. Bach, and M. G. Kris (2004). Lung cancer in US women: a contemporary epidemic. *JAMA 291*, 1763–1768.

Pawitan, Y., J. Bjöhle, S. Wedren, K. Humphreys, L. Skoog, F. Huang, L. Amler, P. Shaw, P. Hall, and J. Bergh (2004). Gene expression profiling for prognosis using cox regression. *Statist Med In press.*

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction.* Oxford: Oxford University Press.

Pepe, M. S., R. Etzioni, Z. Feng, J. D. Potter, M. L. Thompson, M. Thornquist, M. Winget, and Y. Yasui (2001). Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst 93*, 1054–1061.

Pepe, M. S., G. Longton, G. L. Anderson, and M. Schummer (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics 59*, 133–142.

Pittelkow, Y. E. and S. R. Wislon (2003). Visualisation of gene expression data —the ge-biplot, the chip-plot and the gene-plot. *Statistical Applications in Genetics and Molecular Biology 2*, Article 6.

Pomeroy, S., P. Tamayo, M. Gaasenbeek, L. Sturla, M. Angelo, M. McLaughlin, J. Kim, L. Goumnerova, P. Black, C. Lau, J. Allen, D. Zagzag, J. Olson, T. Curran, C. Wetmore, J. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. Louis, J. Mesirov, E. Lander, and T. Golub (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature 415*, 436–442.

Potter, J. D. (2003). Epidemiology, cancer genetics and microarrays: making correct inferences, using appropriate designs. *Trends Genet 19*, 690–695.

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

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

Ripley, B. D. (1996). *Pattern recognition and neural networks.* Cambridge: Cambridge University Press.

Ripley, B. D. (2002). Statistical methods *Need* software: a view of statistical computing. Opening lecture, RSS 2002. http://www.stats.ox.ac.uk/ ripley/RSS2002.pdf.

Romualdi, C., S. Campanaro, D. Campagna, B. Celegato, N. Cannata, S. Toppo, G. Valle, and G. Lanfranchi (2003). Pattern recognition in gene expression profiling using dna array: a comparative study of different statistical methods applied to cancer classification. *Hum. Mol. Genet. 12*(8), 823–836.

Rosenwald, A., G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane, E. M. Hurt, H. Zhao, L. Averett, L. Yang, W. H. Wilson, E. S. Jaffe, R. Simon, R. D. Klausner, J. Powell, P. L. Duffey, D. L. Longo, T. C. Greiner, D. D. Weisenburger, W. G. Sanger, B. J. Dave, J. C. Lynch, J. Vose, J. O. Armitage, E. Montserrat, A. Lopez-Guillermo, T. M. Grogan, T. P. Miller, M. LeBlanc, G. Ott, S. Kvaloy, J. Delabie, H. Holte, P. Krajci, T. Stokke, L. M. Staudt, and the Lymphoma/Leukemia Molecular Profiling Project (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *N Engl J Med 346*(25), 1937–1947.

Rosenwald, A., G. Wright, K. Leroy, X. Yu, P. Gaulard, R. D. Gascoyne, W. C. Chan, T. Zhao, C. Haioun, T. C. Greiner, D. D. Weisenburger, J. C. Lynch, J. Vose, J. O. Armitage, E. B. Smeland, S. Kvaloy, H. Holte, J. Delabie, E. Campo, E. Montserrat, A. Lopez-Guillermo, G. Ott, H. K. Muller-Hermelink, J. M. Connors, R. Braziel, T. M. Grogan, R. I. Fisher, T. P. Miller, M. LeBlanc, M. Chiorazzi, H. Zhao, L. Yang, J. Powell, W. H. Wilson, E. S. Jaffe, R. Simon, R. D. Klausner, and L. M. Staudt (2003). Molecular diagnosis of primary mediastinal B cell lymphoma identifies a clinically favorable subgroup of diffuse large B cell lymphoma related to Hodgkin lymphoma. *J Exp Med 198*, 851–862.

Sackett, D. L. and R. B. Haynes (2002). The architecture of diagnostic research. *BMJ 324*, 539–541.

Shaffer, A., A. Rosenwald, E. Hurt, J. Giltnane, L. Lam, O. Pickeral, and L. Staudt (2001). Signatures of the immune response. *Immunity 15*, 375–385.

Shipp, M. A., K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub (2002). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine 8*(1), 68–74.

Simon, R., M. D. Radmacher, K. Dobbin, and L. M. McShane (2003). Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute 95*(1), 14–18.

Simon, R. M. and K. Dobbin (2003). Experimental design of DNA microarray experiments. *Biotechniques Suppl*, 16–21.

Simon, R. M., E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, and Y. Zhao (2003). *Design and analysis of DNA microarray investigations.* New York: Springer.

Slonim, D. K. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nat Genet 32 Suppl*, 502–508.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology 3*, Article 3.

Somorjai, R. L., B. Dolenko, and R. Baumgartner (2003). Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics 19*, 1484–1491.

Somorjai, R. L., B. Dolenko, A. Nikulin, P. Nickerson, D. Rush, A. Shaw, M. Glogowski, J. Rendell, and R. Deslauriers (2002). Distinguishing normal from rejecting renal allografts: application of a three–stage classification strategy to mr and ir spectra of urine. *Vibrational Spectroscopy 28*, 97–102. Stacking, with individual classifier's coefficient weighted using bootstrap.

Speed, T. e. (2003). *Statistical analysis of gene expression microarray data.* London: Chapman and Hall.

Stegmaier, K., K. N. Ross, S. A. Colavito, S. O'Malley, B. R. Stockwell, and T. R. Golub (2004). Gene expression-based high-throughput screening(GE-HTS) and application to leukemia differentiation. *Nat Genet 36*, 257–263.

Stone, M. and R. J. Brooks (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *J. R. Stat. Soc. B 52*(2), 237–269.

Storey, J. (2002). A direct approach to false discovery rates. *J. Royal Statistical Society, Series B 64*, 479–498.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A 100*, 9440–9445.

Ting, K. M. and I. H. Witten (1999). Issues in stacked generalization. *Journal of Artificial Intelligence Research 10*, 271–289.

Tumor Analysis Best Practices Working Group, T. (2004). Guidelines: Expression profiling - best practices for data generation and interpretation in clinical trials. *Nat Rev Genet 5*, 229–237.

Turner, H., T. Bailey, and W. Krzanowski (2004). Improved biclustering of microarray data demonstrated through systematic performance tests. *Comput. Statist. Data Anal.*, In press.

van Belle, G. (2002). *Statistical rules of thumb.* New York: Wiley.

Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. *J Math Psychol 44*, 92–107.

Westfall, P. H. and S. S. Young (1993). *Resampling-based multiple testing. Examples and methods for p-value adjustment.* New York: Wiley.

Whitfield, C. W., A. M. Cziko, and G. E. Robinson (2003). Gene expression profiles in the brain predict behavior in individual honey bees. *Science 302*, 296–299.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks 5*, 241–259.

Xu, R. and X. Li (2003). A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics 19*, 1284–1289.

Yang, Y. H. and T. Speed (2002). Design issues for cDNA microarray experiments. *Nat Rev Genet 3*, 579–588.

	Goeman et al., 2004	Mootha et al., 2003
Testing	If the set of genes that belongs to set S shows differential expression between classes A and B.	If the "most differentially expressed" genes are mainly of one of the sets.
Statistic	Multivariate: all genes in the set fitted simultaneously using a generalised linear model[1].	Univariate (gene-by-gene).
Ease of application	Requires development of math for different cases (already done for two-class, multiclass, and censored data).	Only needs ordering of genes with criteria of our choice.
Assumes equal behaviour of genes in set	No.	Genes in the set(s) of interest must have a similar ranking of the statistic[2].
Application to different sets	Need to carry out different tests for each of different sets of genes.	Can be applied at once over different sets, and a permutation test carried out to test the single null hypothesis that no gene set is associated with the class distinction.

[1] In general, for multivariate hypotheses ("are the genes of set S differentially expressed between groups A and B?") we should prefer procedures that are fully multivariate (Krzanowski, 1988, pp. 235 and ff.).

[2] Requiring the set of genes to have a similar ranking of the statistic does not by itself guarantee that the set of genes will be made of genes that are co-expressed.

Table 1: Comparison of methods in Goeman et al., 2004 and Mootha et al., 2003 for testing hypotheses about pre-specified sets of genes.

Method	References
Classification	
Diagonal Linear Discriminant Analysis (DLDA)	Dudoit et al. (2002 a), Simon et al. (2003), Romualdi et al. (2003), Huang & Pan (2003), Duda et al. (2001) and Hastie et al. (2001)[1]
K-Nearest Neighbour	Dudoit et al. (2002 a), Simon et al. (2003), Romualdi et al. (2003), Duda et al. (2001) and Hastie et al. (2001)
Support Vector Machines (SVM)	Guyon et al. (2002), Lee & Lee (2003), Simon et al. (2003), Romualdi et al. (2003), Duda et al. (2001) and Hastie et al. (2001)
Partial Least Squares	Stone & Brooks (1990), Garthwaite (1994), Ghosh (2003), Gusnanto et al. (2003), Huang & Pan (2003), Nguyen & Rocke (2002)
Random forests	Breiman (2001), Liaw & Wiener (2002), Bureau et al. (2003), Gunther et al. (2003)
Survival data	
Partial Least Squares	Park et al. (2002)
Penalised Cox regression	Pawitan et al. (2004)

[1] Dudoit et al. (2002 a), Simon et al. (2003), Romualdi et al. (2003) are general reviews that include reviews and results from different data sets. Huang & Pan (2003) show the relationships between several of these (and other) methods. Duda et al. (2001) and Hastie et al. (2001) are general overviews, with additional background material in statistics and machine learning.

Table 2: Well known and good-performing class prediction methods. Because classification has been much more studied than prediction of survival, the methods listed for survival data are not as well known.

	Predicted	
True	Diseased	Healthy
Diseased	True Positive (TP)	False Negative (FN)
Healthy	False Positive (FP)	True Negative (TN)

Table 3: Confusion matrix for a two-class classification problem, with indication of the usual labels for the four types of outcome.

Are genes grouped in families, and are we interested in the overall responses of groups of genes, or should we look at individual genes?

Are certain genes or spots in the array more relevant biologically, maybe because they are easier to measure reliably with other assays?

Is there additional information on which genes are likely to be differentially expressed?

Do you really need the best possible predictor that statistical computing will get you, or do you want a small list of genes very likely to be differentially expressed?

In what stage of the scientific discovery process is this study, and how tight control do you require over the Type I error rate?

What other information and variables about the patients, besides the microarray data, do you have available?

What population do you expect the results of these studies to be relevant for?

Are these the original, complete data, and are these the original biological questions, or have the data and questions gone through an already long run of analyses which has already filtered data and reoriented hypotheses?

What is the next stage of this study, or what do you want to do with these results?

What additional studies could be done to confirm the results from these analyses?

Table 4: Some relevant questions statisticians and biologists should engage in a dialog about.

- full access to the algorithms and their implementation, which allows users to understand what they are doing when they run a particular analysis
- the ability to fix bugs and extend and improve the supplied software
- encouraging good scientific computing and statistical practice by providing appropriate tools, instruction, and documentation
- providing a workbench of tools that allow researchers to explore and expand the methods used to analyse biological data
- ensuring that the international scientific community is the owner of the software tools needed to carry out research
- promoting reproducible research by providing open and accessible tools with which to carry out that research (reproducible research as distinct from independent verification)

Table 5: Reasons why source code should be available in bioinformatics, from p. 46 of Dudoit et al. (2003).

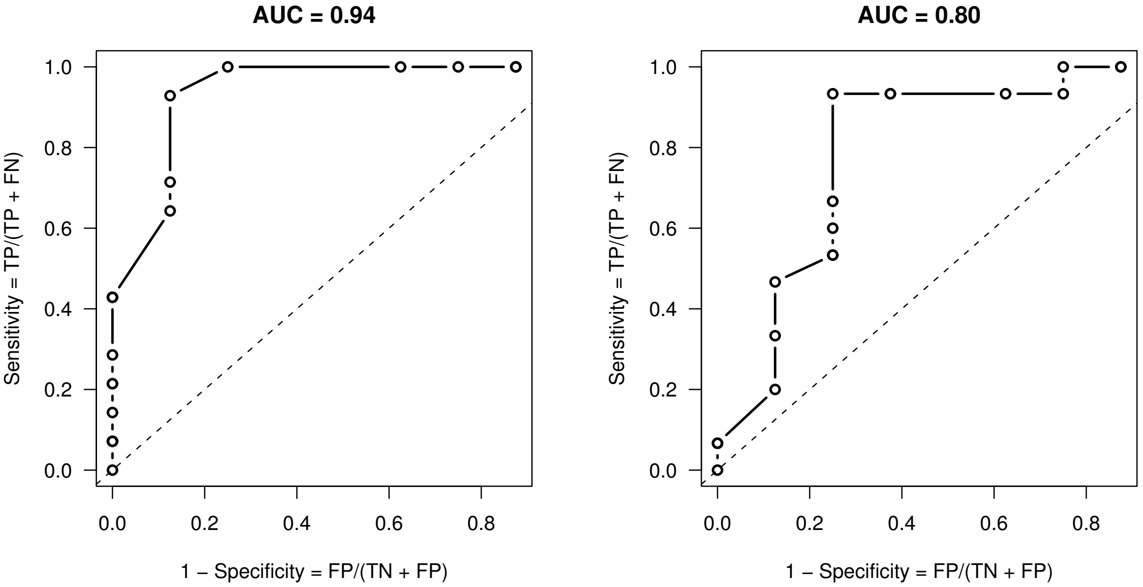

Figure 1: Two ROC curves from real microarray data; on top of each we indicate the Area Under the ROC Curve.