Prior Biological Knowledge Based Approaches for the Analysis of Genome-Wide Expression Profiles Using Gene Sets and Pathways

Michael C. Wu and Xihong Lin
Department of Biostatistics
Harvard School of Public Health
Boston, MA 02115

mwu@hsph.harvard.edu and *xlin@hsph.harvard.edu*

June 22, 2009

Abstract

An increasing challenge in analysis of microarray data is how to interpret and gain biological insight of profiles of thousands of genes. This paper provides a review of statistical methods for analysis of microarray data by incorporating prior biological knowledge using gene sets and biological pathways, which consist of groups of biologically similar genes. We first discuss issues of individual gene analysis. We compare several methods for analysis of gene sets including over-representation anlaysis, gene set enrichment analysis, principal component analysis, global test, and kernel machine. We discuss the assumptions of these methods and their pros and cons. We illustrate these methods by application to a type II diabetes data set.

KEY WORDS: Gene set enrichment analysis; Global test; Hypothesis testing; Kernel machine; Over-representation analysis; Principal components; Variance component testing.

1 Introduction

Since the intial work of Schena *et al.*[1] in 1995, microarrays have become a commonly used tool in biological and medical research due to their ability to simultaneously profile the expression of thousands of genes. Initial experiments were relatively simple with no replication, one array per condition, and only crude measurements for differential expression were used—a fold change of 1.5 indicated up regulation while a fold change of 0.75 indicated down-regulation. As the complexity of these studies increased with their popularity, the need for more sophisticated tools became clear.

Currently, a standard microarray experiment consists of the simultaneous expression profiling of thousands of genes across various experimental conditions. Unless otherwise stated, we generally assume two conditions. Low-level analyses typically include image analysis (grid alignment, target detection, intensity extraction, and local background correction), normalization, and computation of a gene expression value for each probe on the chip. Significant work has been done in this area[2–6] as all futher analyses are contingent on proper low-level processing.

High level analyses typically begin by calculating a statistic (often a t-statistic) for each gene on the chip, measuring differential expression between experimental conditions. A p-value is usually generated for each gene, based on the statistic, via permutation or a parametric distribution. To account for the thousands of comparisons performed, procedures controlling the family wise error rate or the false discovery rate (FDR)[7,8] are performed. Genes that survive the correction for multiple comparisons are then considered *differentially expressed* while genes that fail to meet the criterion for significance are *non-differentially expressed*. The list of differentially expressed genes is often the final goal for the statistician and once obtained, it is the responsibility of the biological or biomedical researcher to draw further conclusions.

These traditional approaches have yielded a wealth of information regarding gene interactions, functions, and pathways. Recently, biologists have become interested in exploiting this information to facilitate and improve the analyses performed. The knowledge can be used to varying degrees,[9] but at the most basic level, it is known that most biological phenomena occur through the concerted expression of multiple genes. We can thus use our prior knowledge of what genes belong to various signalling pathways or functional groups and focus our analyses on sets of related genes, called *gene sets*. Numerous databases containing gene groupings based on various criteria have been developed. Examples include KEGG[10] and the Gene Ontology (GO) Consortium.[11]

Use of information derived from the GO consortium database is the most popular, so we briefly describe their database. The gene ontology consortium contains three principal ontologies: biological processes, cellular components, and molecular functions. Each ontology is a directed acyclic graph, creating a hierarchy of terms, called *GO terms*, that range from very broad functions, such as "physiological process", down to more specific levels, such as "microtubule depolymerization". Each ontology and GO term has a comprehensive list of genes previously demonstrated to be associated with that ontology or GO term. A number of tools have been created for mining and using the data from the GO consortium.[12]

Using groupings from the GO consorium or other annotation databases, our analyses no longer consider individual genes, but rather groups of genes. This mode of analysis overcomes a number of drawbacks, which we will explore later, associated with traditional approaches to microarray analysis and is biologically more meaningful. The goal of this article is to review methods that test for the differential expression of gene sets defined by prior knowledge.

In the next section, we will briefly review the drawbacks associated with the traditional approach and discuss the use of prior biological knowledge as a remedy for some of the problems. Section 3 introduces a few of the many popular methods using prior biological knowledge. Section 4 presents statistical issues other authors have identified regarding these methods. In Section 5, we apply these methods to a real data set and compare their performances. Finally, in Section 6, we briefly summarize the main points of the paper and discuss other practical issues regarding the use of prior biological knowledge.

2 Problems associated with traditional approaches

Advocates have suggested many different reasons for incorporating prior biological knowledge into the analysis of gene expression data.[13–17] We present some of these arguments here.

In terms of biological rationale for testing gene sets, it is well known that most pathways are not driven by a single gene, but rather by a combination of multiple genes acting in a concerted fashion. Thus, individual gene analysis may miss important pathway effects since genes that demonstrate a high level of differential expression between conditions may not be as important as a group of genes that each shows only moderate differences between conditions. In particular highly differentially expressed genes tend to be "downstream" genes. Many upstream proteins, such as transcription factors and other regulatory proteins, may only show very moderate changes, especially in contrast to high abundance proteins expressed at the end of the biological cascade. If attention is restricted to only the most highly differentially expressed genes, upstream effects are likely to be missed, despite the crucial role they play acting as activators and gatekeepers.

Practicality also motivates the use of prior biological knowledge. Many investigators have been faced with the problem after correction for multiple comparisons, no genes meet the threshhold for statistical significance. Given that sample sizes for microarray experiments tend to be small, if the signal in the data is not strong relative to the noise, as in situations where exposures are mild, perhaps due to toxicity restrictions on patients, or where the biological response is simply weak by nature, then finding highly differentially expressed genes may be quite difficult. Multiple comparisons exacerbate the problem when high correlations exist in the data: the typical FDR and FWER controlling procedures assume that all of the hypotheses are independent, but genes are known to work together in a concerted fashion so tests may be overly conservative. Use of the empirical null hypothesis serves as a means of correcting for correlation,[18] but assumptions regarding the tail behavior of the null distribution may not hold. Regardless, when using traditional approaches, failure to detect differentially expressed genes lead to failure in drawing conclusions.

The alternative to not detecting any differentially expressed genes is to find that even after correction for multiple comparisons, a long list of "differentially expressed" genes remains. Although biological collaborators often prefer a long list of genes that meet the threshhold for statistical significance, this presents a problem in terms of interpretation. Often it is difficult to draw out a specific theme or message, or to identify potential mechanisms based on a long gene list. What conclusions are found also tend to be very subjective.

Further, comparisons of gene lists between different experiments have shown little overlap. Despite similar exposure conditions, experiments from different groups have shown dissimilar results when gene lists are compared.[19] This presents a confusing picture of what is going on biologically since each group is presenting hypotheses based on their own lists of genes called differentially expressed.

Using prior biological knowledge immediately preempts the problem that no genes are individually differentially expressed after multiple comparisons as we are no longer interested in any individual gene's significance. Furthermore, as we are no longer performing thousands of comparisons, but rather restricting attention to comparatively few pathways of specific interest, corrections for multiple comparisons need not be as extreme. Interpretation of results is facilitated as pathways are often the primary interest, and this provides a means by which the same conclusions will be drawn by different researchers presented with the same data, improving objectivity. Depending on the method applied, moderate changes can potentially be detected, and for a pathway shown to be differentially expressed, what genes are driving the difference can possibly be elucidated as well, identifying which genes are the upstream regulator genes. Finally, while a single gene is likely to show great variability in differential expression level from experiment to experiment, a pathway that contains many genes is less likely to demonstrate such variable behavior, giving more consistent results between experiments.[16]

3 Prior knowledge based methods

Numerous methods utilizing prior biological knowledge have been and are being developed. We here review a few of the many methods but emphasize that this is by no means a complete catalogue. We will compare these methods and discuss their assumptions and pros and cons in Section 4.

3.1 Over-representation Analysis

Over-representation analysis (ORA) refers to an entire class of methods which are, by far, the most commonly used as they are the earliest and the simplest developed. These methods start from a list of genes that are called differentially expressed, D, and the list of genes in the gene set of interest, S. D^c and S^c represent the set of genes not differentially expressed and the set of genes not in the gene set respectively. Based on these, the researcher looks for an over-representation of the genes in the gene set among differentially expresssed genes, or equivalently, over-representation of differentially expressed genes in the gene set. Practically speaking, this is done by creating a 2x2 contingency table based on membership in D and membership in S. Letting N be the total number of genes, and for any sets A and B, N_A denotes the cardinality of A and N_{AB} denotes the cardinality of $A \cap B$, then we can build table 1. To generate a p-value for over-representation, we test for independence between membership in D and membership in S using a Fisher's exact test,[20] specifically:

$$p = 1 - \sum_{i=0}^{N_{SD}} \frac{\left(\begin{array}{c} N_D \\ i \end{array} \right) \left(\begin{array}{c} N_{D^c} \\ N_S - i \end{array} \right)}{\left(\begin{array}{c} N \\ N_S \end{array} \right)}$$

Many variations on this method have been developed.[21] Differences focus on the construction of the list of differentially expressed genes and the test used once a 2x2 table has been constructed. Tests besides Fisher's exact test include the chi-square test, the hypergeometric test, and the binomial proportions z-test. In practice, the choice of test is unimportant. However, as will be demonstrated later, how the cutoff distinguishing differentially expressed genes from non-differentially expressed genes is constructed strongly influences whether or not a pathway is called differentially expressed. Criteria for differential expression may be based on the multiple comparisons threshold, but can be much simpler, e.g. using the 100 genes with smallest individual p-value, the top 5% most differentially expressed genes, or all genes with fold change greater than 2. For a full discussion of ORA methods, see Khatri and Draghici.[21]

3.2 Gene Set Enrichment Analysis

While ORA is attractive because of its simplicity, it relies heavily a potentially arbitrary hard cutoff. A method that remedies this is Gene Set Enrichment Analysis (GSEA). Instead of using a set cutoff, GSEA ranks all the genes on the chip based on some signed measure of differential expression from individual gene analysis and then tests the null that the genes in the gene set are uniformly distributed throughout the list of ranked genes against the alternative that the genes in the gene set tend to be closer to the top or bottom of the list. The assumption is that if a gene set is differentially expressed, then the component genes are likely to be more differentially expressed and thus clustered towards either the top or bottom of the list. This assumes that the direction of differential expression for genes in a differentially expressed gene set is the same.

The original GSEA approach was developed by Mootha *et al.*[22] Using the same notation as before, the basic algorithm is as follows:

1. Rank the N genes on the chip based on a differential expression measurement, such as t-statistic, to obtain L, the ranked gene list.

2. An Enrichment Score (ES) is then calculated for the date set. For gene G_i (the i-th gene in L), let:

   $$E_i = \begin{cases} \sqrt{\frac{N_{S^c}}{N_S}} & \text{if } G_i \text{ is in } S \\ -\sqrt{\frac{N_S}{N_{S^c}}} & \text{if } G_i \text{ is NOT in } S \end{cases},$$

   where N_S is the number of genes in Set S and N_{S^c} is the number of genes not in set S. Define the enrichment score $\text{ES}(S) = max_{1 \le j \le N} \sum_{i=1}^{j} E_i$.

3. To determine significance, permutation is used to generate the null distribution:

   (a) Randomly permute the class labels.

   (b) Re-rank the genes to generate a new ranked gene list L^*.

   (c) Calculate $\text{ES}(S)^*$, the enrichment score based on L^*

   (d) Repeat the above for 1000 permuations.

4. A p-value is generated by comparing our original $\text{ES}(S)$ to the distribution of the $\text{ES}(S)^*$.

The Enrichment Score is essentially a modified Kolmogorov-Smirnov statistic. Several improvements have been made to the method. Sweet-cordero *et al.*[23] extended GSEA to multiple gene sets and multiple data sets and the Subramanian *et al.*[13] modified the enrichment score so that each gene's contribution is weighted by its correlation with the phenotypic outcome.

Many methods similar to GSEA have been developed. The Gene Set Analysis (GSA) of Efron and Tibshirani[24] is based on the GSEA method, but uses a "maxmean statistic", M, instead of the Kolmogorov-Smirnov statistic for the enrichment score, potentially leading to greater power. If t_i is the differential expression measurement (t-statistic) for the i^{th} gene in the gene set, then the max mean statistic is given by:

$$M = \max\left\{\left|\frac{\sum_{i=1}^{N_S} I(t_i > 0)t_i}{N_S}\right|, \left|\frac{\sum_{i=1}^{N_S} I(t_i < 0)t_i}{N_S}\right|\right\}$$

Note that the denominator is N_S. For evaluation of significance, GSA argues for permutation of genes in addition to the permuation of class labels. A method by Smythe[25] and Tian *et al.*[26] uses the averaged t-statistic for the enrichment score. Tian *et al.* also makes further modifications to GSEA in the case where one wishes to compare differential expression of one gene set to differential expression of another. Other methods by Pavlidis *et al.*[27] and Rhanenfhrer *et al.*[28] are similar in flavor.

3.3 Global Test Method using Generalized Linear Models

The global test[14] does not rely on the potentially unstable individual gene analyses. This method exploits the duality beween association and prediction: if a gene set can be used to predict the clinical outcome, its expression pattern must differ for different outcomes. If Y is the outcome of interest (possibly continuous or possibly 1/0 for case/control status), and letting X be the $n \times N_S$ matrix of gene expression values for the gene set (where n is the number of samples) so that x_{ij} is the gene expression value of the j^{th} gene of the i^{th} sample, the global test is motivated by a regression model to predict the outcome based on gene set expression:

$$g\{\mathcal{E}(Y_i|\boldsymbol{\beta})\} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{N_S} x_{iN_S}, \tag{1}$$

where $g(\cdot)$ is a link function in generalized linear models,[29] such as the logit link for the two group comparison, and α is an intercept. Then testing for an overall predictive effect for the gene set is equivalent to testing:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{N_S} = 0$$

In most cases, the number of genes in the gene set is large relative to the sample size, so an additional assumption that the β's are iid with mean 0 and variance τ^2 is made. Under this assumption, our null hypothesis is simply:

$$H_0 : \tau^2 = 0$$

An alternative interpretation is to rewrite the earlier model by setting $r_i = \sum_{j=1}^{N_S} x_{ij}\beta_j$. Then under the null, $\mathbf{r} = (r_1, \ldots, r_n)$ where n is the number of samples, has mean $\mathbf{0}$ and covariance $\tau^2 XX'$. We thus rewrite our model as:

$$g\{\mathcal{E}(Y_i|\boldsymbol{\beta})\} = \alpha + r_i, \tag{2}$$

which corresponds to a random effects model. Assuming α is known, a score statistic for testing $H_0 : \tau^2 = 0$ is given as

$$T = \frac{(Y-\mu)'R(Y-\mu) - \mu_2 \text{trace}(R)}{\sqrt{2\mu_2^2 \text{trace}(R^2) + (\mu_4 - 3\mu_2^2)\sum_i R_{ii}^2}}$$

where $R = \frac{1}{N_S}XX'$, $\mu = g^{-1}(\alpha)$, and μ_2 and μ_4 are the second and fourth central moments of Y under the null.[14,30] T can be approximated by:

$$Q = \frac{(Y-\mu)'R(Y-\mu)}{\mu_2}$$

which is also assymptotically normal under H_0. However, since the sample size is likely to be low, Goeman *et al.* suggest that significance be evaluated by permuting the class labels to obtain a null distribution for Q and then comparing the original statistic to the permuted distribution. Since α is never known in real situations, some adjustments are necessary to estimate μ and μ_2.

A nice by-product of the global test statistic is that i^{th} gene's contribution to Q is simply:

$$Q_i = \frac{1}{\mu^2}[X_i^t(Y-\mu)]^2$$

as $Q = \frac{1}{N_S}\sum_i Q_i$. Thus, if a pathway is determined to be significantly differentially expressed, by estimating the contribution (influence score) for each gene, researchers can determine which genes are driving the difference, giving more information as to the biological mechanisms involved.

3.4 Global Test Method Using Kernel Machines

The global test of Goeman *et al.*[14] assume in model (1) the effects of genes within a gene set are additive. Genes within a pathway are often correlated and interact to each other. The additive assumption hence might be too strong in practice. Liu *et al.*[17] proposed modeling the gene set effects using kernel machines, which allow joint flexible nonlinear effects of genes within a pathway on a phenotype. Specifically, we replace the generalized linear model (1) by the generalized nonlinear model

$$g\{\mathcal{E}(Y_i|h)\} = \alpha_0 + h(x_{i1}, \cdots, x_{ip}), \tag{3}$$

where $h(\cdot)$ is a linear or nonlinear smooth function and its functional form can be estimated from the data. When $h(\cdot)$ is linear in x's, (3) reduces to the generalized linear model (1).

Liu *et al.* proposed to estimate $h(\cdot)$ using kernel machines, which can proceed by estimating $h(\cdot)$ using (2) assuming the r are random effects with mean 0 and covariance $\tau^2 K$, where K is a kernel matrix whose (i, i')th element can be viewed as a measure of similarity of the gene profile of the ith subject and that of the i'th subject. If $h(\cdot)$ is a linear function in x's, then the (i, i')th element of K is $(x_i^T x_{i'} + c)$ and K reduces to XX' if $c = 0$. If $h(\cdot)$ is a qudratic fucntion of x's including two-way interactions of the x's, the (i, i')th component of K is $(x_i^T x_{i'} + c)^2$. If $h(\cdot)$ is a smooth function of x's expanded by radius basis, the (i, i')th element of K is $exp\{(x_i - x_{i'})^T (x_i - x_{i'})/c\}$. A variance component test for $H_0 : \tau^2 = 0$ under (3) corresponds to a global test for no gene set effect by allowing for nonlinear effects when an appropriate kernel function is assumed.

3.5 Principal Components Analysis (PCA)

Another set of approaches that do not rely on individual gene analyses are based on the principle of dimension reduction. Although gene sets already contain far fewer genes than the total number on a chip, the dimensionality of the gene set still often exceeds the sample size. By sufficiently reducing the dimentionality of the data, standard univariate or multivariate methods can be applied. The most commonly used means of dimension reduction is principal components analysis.

Principal components analysis seeks to identify the b directions of greatest variability in the data and then project the data onto the space spanned by these directions. Mathematically, these directions are given by the eigenvectors of the sample covariance matrix ($\widehat{\Sigma}$) corresponding to the b largest eigenvalues of $\widehat{\Sigma}$. Suppose that the expression value for each gene has been centered by their respective sample means, then letting $V = [v_1, v_2, \ldots v_p]$, $E = diag(e_1, e_2, \ldots, e_p)$, and $e_1 \geq e_2 \geq \ldots \geq e_p$, where v_j is the eigenvector corresponding to the j^{th} largest eigenvalue, e_j, V and E can be found by the singular value decomposition of X:

$$X = UE^{\frac{1}{2}}V^t$$

Projecting the data into a smaller subspace reduces the dimensions of the data while keeping the most information since the directions of greatest variability are retained.

To our knowledge, the first method of this type to apply these ideas to detection of differentially expressed gene sets is the method proposed by Tomfohr *et al.*,[15] which we will refer to as PCA_T. The idea is to reduce the gene set to its first principal component, so that we have a single "super-gene" or "metagene". The supergene's expression value for the i^{th} subject is the first component score:

$$X_i^{\text{new}} = v_1^t X_i$$

Since X^{new} is now one dimensional, we can then use a standard two-sample univariate test, e.g. t-test or Wilcoxon test, to evaluate the significance of the supergene. If the supergene is found to be differentially expressed, then the entire gene set is considered to be differentially expressed.

Often times, the first principal component may be insufficient for summarizing a gene set's activity or it may capture variability not associated with differences resulting from clinical outcomes. For instance, it has been demonstrated in the genome wide association testing literature that the first principal component identifies variability resulting from differences in ancestry among subjects.[31] Thus, a natural extension of PCA_T is to consider additional higher order components and reduce the gene set to the first b principal components. This approach was first published by Kong *et al.*[32] and we refer to it as PCA_K. Instead of a single supergene, b supergenes summarize the gene set. Choices for b are briefly discussed below, but b is necessarily less than the number of positive eigenvalues, $d = rank(\widehat{\Sigma})$. If $V_{(b)} = [v_1, v_2, \ldots, v_b]$ then the new component scores, expression values for the super genes, for the i^{th} subject are:

$$X_i^{\text{new}} = V_{(b)}^t X_i$$

Since X is now an $n \times b$ matrix, one can use Hotelling's T^2 test to evaluate significance. For completeness, the Hotelling's T^2 statistic is found by:

$$T^2 = \frac{n_1 n_2}{n}(\bar{X}_1^{\text{new}} - \bar{X}_2^{\text{new}})^t \widehat{\Sigma}_p^{-1} (\bar{X}_1^{\text{new}} - \bar{X}_2^{\text{new}})$$

where n_j is the number of subjects with clinical outcome j, $\bar{X}_j^{\text{new}}$ is the vector of mean expression values for the supergenes among subjects with clinical outcome j, and $\widehat{\Sigma}_p = ((n_1 - 1)\widehat{\Sigma}_1 + (n_2 - 1)\widehat{\Sigma}_2)/(n-2)$ is the pooled covariance matrix ($\widehat{\Sigma}_j$ is the covariance matrix of the supergenes among subjects with outcome j). To generate a p-value, one can either permute the class labels and generate a permutation distribution for T^2, or alternatively, under the commonly used assumption of normality, it is know that $\frac{(n-b-1)}{(n-2)b}T^2 \sim F_{b,n-b-1}$.

A fundamental issue always present when using principal components is the choice of b, the number of components to use. Kong *et al.* simply use a hard threshold on the eigenvalues but admit that this may not be optimal. This problem has been studied in various applied settings by many authors.[33–37] Suggested rules for choosing b are:

1. **First Component Only**: $b = 1$ as in Tomfohr *et al.*'s method.

2. **Proportion of Variability Explained**: The proportion of variability explained by the first q principal components is given by: $r_q = \frac{\sum_{k=1}^{q} e_k}{\sum_{k=1}^{d} e_k} = \frac{\sum_{k=1}^{q} e_k}{\text{trace}(\widehat{\Sigma})}$. Typical cutoffs range between 70% to 90%. The number of components to that explain 70% of the total variability is found by: $b = argmin_q\{r_q > 0.70\}$.

3. **Zhu's Method**: A commonly used method of estimating the number of components is to generate a Scree plot (a barplot of the eigenvalues) and then look for an "elbow" or "big gap" in the graph. An elbow between the q^{th} and $(q+1)^{th}$ eigenvalue suggests that there is a rapid decrease in the relative importance of the components. In the past, this method tended to be subjective and not practical in many situations because it was not automated, but Zhu and Ghodsi propose a simple algorithm for identifying elbows.

   Suppose we want to see if there is a gap between the q^{th} and $(q+1)^{th}$ eigenvalues. Let $\mathcal{A}_1 = \{e_1, e_2, \ldots, e_q\}$ and $\mathcal{A}_2 = \{e_{q+1}, e_{q+2}, \ldots, e_d\}$. If there truly is a gap at the q^{th} position $\mathcal{A}_1$ and $\mathcal{A}_2$ can be considered as samples from two different distributions, $f(e; \theta_1)$ and $f(e; \theta_2)$ respectively. If we assume the two samples are independent, then the log-likelihood of our data is given by:

   $$\ell(q, \theta_1, \theta_2) = \sum_{k=1}^{q} \log f(e_k; \theta_1) + \sum_{k=q+1}^{d} \log f(e_k; \theta_2)$$

   For convenience, we use the normal density for f and we can obtain a profile log-likelihood by plugging in: $\widehat{\theta}_1 = [\bar{e}_1, s^2]$ and $\widehat{\theta}_2 = [\bar{e}_2, s^2]$, where $\bar{e}_1 = \sum_{k=1}^{q} e_k/q$, $\bar{e}_2 = \sum_{k=q+1}^{d} e_k/(d-q)$, and $s^2 = \frac{(q-1)s_1^2 + (d-q-1)s_2^2}{d-2}$ with s_1^2 and s_2^2 equaling the variances of $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively. b is then set to the value of q that maximizes the profile likelihood.

   Despite the naive, but convenient, assumptions of normality and independence, empirical results suggest that the overall algorithm is still effective.

4. **Guttman-Kaiser's Average Eigenvalue Rule**: All eigenvalues greater in magnitude than the average of the eigenvalues are retained. The method was initially designed for PCA based on the correlation matrix. If all of the genes were independent, then the principal components would be identical to the original data and have unit variance. Thus, any eigenvalue less than 1 in magnitude carries less information than one of the original variables and is not worth

keeping. Noting that 1 is the mean of the eigenvalues from the correlation matrix, we instead compare the eigenvalues from the covariance matrix to the mean.

5. **Jolliffe's Modified Average Eigenvalue Rule:** All eigenvalues greater in magnitude than 0.7 times the average of the eigenvalues are retained. The constant 0.7 was chosen based on simulation.

6. **Bartlett's test**: This method sequentially tests for equality of eigenvalues starting from d down to 1. If the last $d-q$ eigenvalues are equal, then they contain equally little information and should be discarded. To determine b, test whether the last $d-q$ values are all equal against the alternative that there are at least two that are different. If we reject the null, then b is set to $d-q+1$, otherwise we increase q by 1 and test again.

   To actually test for the equality of the last $d-q$ eigenvalues, we use the statistic:

   $$n'\left[(d-q)\log\bar{e} - \sum_{k=q+1}^{d}\log e_k\right]$$

   which approximately follows a χ^2_ν distribution with $\nu = 0.5(d-q+2)(d-q-1)$. $n' = n - \frac{2d+11}{6}$ improves the approximation.

The authors do not agree as to which rules are optimal. This may depend on the individual setting and correlation structure of the gene set.

If a pathway is determined to be significantly differentially expressed by PCA_T, genes driving the difference are identified as the genes with the greatest loadings. Tomfohr *et al.* refer to these as "activity levels". If PCA_K is used, one can find activity levels by identifying which supergenes are most differentially expressed and examining the loadings for generating those supergenes. Multiple differentially expressed supergenes may suggest differing mechanisms for differential expression.

4 Statistical Issues

Numerous methods other than those described exist. In the next section we will compare the described methods on a real data set, but here we first introduce two statistical considerations that have recently been identified. This work is largely the result of Goeman and Buhlmann.[38]

4.1 The Null Hypothesis

Each of the methods seeks to test for differential expression of the gene set between experimental conditions. However, as Goeman and Buhlmann[38] and Tian *et al.*[26] point out, there are two ways of actually formulating the null hypothesis:

1. Formulation 1: H_0^{comp}: The genes in the gene set S are at most as differentially expressed as the genes not in S.

2. Formulation 2: H_0^{self}: The genes in the gene set S are not differentially expressed.

Methods that use formulation 1 are the most prevalent as they include all of the currently used over-representation analysis methods and the original GSEA method. The methods described testing formulation 2 are the global test, PCA, and some later variants of GSEA. While both null's seem similar, they are actually quite different. Goeman and Buhlmann call formulation 1 a *competitive null hypothesis* while formulation 2 is a *self-contained null hypothesis*. Fundamentally, a competitive null pits the genes in S against all other genes in the experiment, while a self-contained null looks only at the genes in the gene set and ignores all of the other genes on the microarray. Both Tian *et al.* and Goeman and Buhlmann favor self-contained tests.

Criticism of tests using H_0^{comp}revolve primarily around issues of power. Generally, a self-contained test will tend to have more power than a competitive test, as H_0^{self}tends to imply H_0^{comp}. Under the competitive setup, a gene set's significance is penalized in experiments where many genes are differentially expressed as the standard for significance has been raised. Allison *et al.*[39] and subsequently Goeman and Buhlmann describe the competitive framework as a "zero-sum game". Asside from the power considerations, this creates negative correlation between p-values and is problematic as the standard false discovery rate corrections may not be valid under negative correlation. As self-contained nulls completely ignore other gene sets, this is not an issue.

Goeman and Buhlmann also note that self-contained tests are direct generalizations of single gene tests. This is a nice property as testing a gene set with a single gene is equivalent to testing the gene individually. This property does not hold for competitive tests. A related result is that a self-contained test can directly test the null that there are no differentially expressed genes on the chip, potentially serving as a quality check. A competitive null cannot treat the entire data set as a gene set as there would be no complement to compare with.

The main criticism of self contained tests is that they may be overly powerful in settings where many genes appear to be highly differentially expressed. In particular, large gene sets may contain a few genes that appear to be differentially expressed leading to a significant p-value for that gene set despite it's biological irrelevance. In such cases, Goeman and Buhlmann suggest using self-contained tests as an initial screening and then following up with a competitive test in the second stage.

4.2 The Sampling Unit

Goeman and Buhlmann identify another important statistical issue that raises the most fundamental question: what is the sampling unit? Classical tests are based on experiments that have a population of subjects, and then *subject sampling* is performed: we consider our data a sample of subjects drawn from the population. Each subject has the same set of fixed attributes (in this case genes). In contrast, the over-representation methods described above perform *gene sampling.* The tests used still assume the samples are drawn from the population, but in this case the population considered is the set of genes. This reverses the classical setup: we now consider our data as a sample of genes coming from a fixed set of subjects. In the first set up, our sample size is the number of subjects (arrays) while in the latter it is actually the number of genes. GSEA, GSA, global test, and PCA all sample subjects.

This has a dramatic impact on the interpretation of results obtained. Specifically, a significant p-value gives confidence that if we were to draw a new sample, a gene set would again be differentially expressed, i.e. results generalize to the population from which we sampled. Under the subject sampling scheme, this means that we are confident that the association between genes and the experimental conditions will be found for a new group of samples. In contrast, under the gene sampling scheme, a significant p-value gives confidence that for a new set of genes from the same subjects, a similar association between being in the gene set and being called "differentially expressed" will be found.

The gene sampling scheme is usually not the preferred set up. Generally, experiments are performed with the intention of finding results that generalize beyond the sample of subjects. Indeed, a new experiment would likely take a new sample of subjects rather than a new sample of genes. Also, tests for both sampling schemes assume that sampling units are independent and identically distributed. Assuming genes are independent is extremely unrealistic: the entire purpose of using prior biological knowledge is to exploit the information that genes work together.

5 Comparison on Type II Diabetes Data

This data set consists of gene expression profiles of muscle tissue from 17 subjects with type 2 diabetes and 17 subjects with normal glucose tolerance (a third group with impaired glucose tolerance was omitted from our analysis). It was first analyzed in Mootha *et al.*[22] using GSEA and is available form the Broad Institute Website (http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi). Details on low-level processing are available in the original manuscript. After removing all genes with no single measure greater than 100 (genes not expressed in the data), 10983 genes remained. To compare the performance of the described methods, to each of 133 pathways (the original 149 considered by Mootha *et al.* less the pathways containing 0 or 1 gene only) we applied:

1. ORA considering the top 100 genes as "differentially expressed" and testing for association using Fisher's exact test.

2. ORA considering the top 100 genes as "differentially expressed" and testing for association using the χ^2-test.

3. ORA considering the top 5% most differentially expressed genes (as determined by two-sample t-test) as "differentially expressed" and testing for association using Fisher's exact test.

4. ORA considering the top 5% most differentially expressed genes as "differentially expressed" and testing for association using the χ^2-test.

5. The original GSEA method, with genes ranked by two-sample t-statistics.

6. Gene Set Analysis, with genes ranked by two-sample t-statistics (software available from the authors' website: http://www-stat.stanford.edu/ tibs/GSA/index.html).

7. Global test (software available from Bioconductor: http://www.bioconductor.org/).

8. PCA_T

9. PCA_K with the number of components to use determined by taking the maximum of 3 or the number of components necessary to account for 70% of the variability. For gene sets with fewer than three genes, the number of components used was equal to the number of genes in the gene set.

Results comparing the number of gene sets identified as differentially expressed, at the nominal $\alpha = 0.05$ level, by each method are given in the diagonal of Table 2. Non-diagonal entries contain the number of gene sets simultaneously identified by the two methods. As an example, 18 pathways were identified as differentially expressed by GSEA and 9 pathways were indentified as differentially expressed by both GSEA and GSA.

Over-representation analysis is clearly very sensitive to the cut-off used. If the top 100 genes, as ranked by t-statistic, were called "differentially expressed", only seven pathways were identified by Fisher's exact test. In contrast, if the cutoff is lowered so that the top 5% genes are called "differentially expressed", then 60 pathways are identified. When comparing the use of Fisher's exact test to use of the χ^2-test, it initially appears that the χ^2 test is able to identify many more pathways at the same cutoff. However, the pathways identified by using the χ^2-test and not identified by the fisher's exact test all contain very few genes (less than 5 genes). In such situations, the χ^2 test may not be appropriate. For pathways comprised of more genes, results were essentially the same.

GSA identifies only two more pathways as differentially expressed than GSEA, but the pathways identified by each overlap by only 9 pathways despite GSA being developed based on GSEA.

Though theoretical justifications suggest self-contained tests are more powerful, yet the global test and PCA_T identified only 4 and 5 pathways, respectively. The global test may not perform optimally in identifying pathways consisting of many genes with no predictive ability and only a few with predictive ability. Under the assumptions of the test, the global test should identify such pathways, but inclusion of many other genes may dilute the signal and introduce extraneous noise. A method employing variable selection may perform better.

Using PCA_T assumes that the first principal component sufficiently summarizes the entire pathway's activity. Given that signal is low and noise often high in expression profiling experiments and that we have not accounted for other baseline effects such as ancestry, it is not surprising that the first principal component does not provide good separation of diabetics from normal patients. Indeed, Bair *et al.*[40] suggest some filtering of genes is necessary in order for the first component to capture the difference of interest.

PCA_K identifies more (16) pathways than PCA_T, suggesting that use of more principal components provides a better summary of a gene set's expression. On the otherhand, of the 5 pathways identified as being differentially expressed by PCA_T, only one was identified by PCA_K. Truth of the PCA_K null implies truth of the PCA_T null, so rejection of the PCA_T null should imply rejection of the PCA_K null. However, in practice though the first component is different, this difference is diluted by the additional components used, suggesting that the number of components used is not optimal. A better method for determining the number of components to test needs to be developed. Of the 16 pathways identified by PCA_K, 6 were also identified by the original GSEA program. These 6 were also identified as being significantly differentially expressed by GSA. It is important to note, that GSEA and GSA assume that all of the genes in a significantly differentially expressed pathway will be differentially expressed in the same direction, i.e. the genes will tend to all be closer to the top of the ranked gene list or all be closer to the bottom of the ranked gene list. In contrast, PCA_K (as well as PCA_T and the global test) does not consider the directionality. This better matches biological assumptions: in a given pathway, certain genes will be turned on and others turned off in response to stimuli. Pathways identified by PCA_K but not by GSEA or GSA may be such pathways in which direction of differential expression is different.

6 Discussion

When applied to the type 2 diabetes data, we see that using prior biological knowledge can potentially identify pathways of interest. If the traditional approach had been taken, no conclusions could have been drawn as no genes met the criteria for differential expression after controlling for the false discovery rate. The number of pathways identified by ORA tends to be somewhat unstable depending on the number of genes called "differentially expressed", though if this method is used, the specific test used does not appear to make a difference as long as the sizes of the gene sets are not small. The interpretation of the ORA results is difficult, however, as they treat all genes as the sampling unit. These methods should be used very cautiously, if at all. Global test and PCA_T may not perform well for gene sets that include many irrelevant genes. The enrichment scoring methods appear to function well as does PCA_K. All three methods produce results that are biologically reasonable, but it is not clear which method is preferable in practice, despite the competitive nature of the enrichment scoring methods.

A major weakness that all prior biological knowledge based described suffer is the quality of the prior biological knowledge incorporated. The methods we described here all deal with analyzing gene sets which are grouped based on some biological principle and the assumption is that all of the genes in the biological grouping are associated with each other in a biologically meaningful fashion. However, this assumption is not always true: the quality of the groupings is not always gauranteed. Databases such as KEGG are of good quality, as are other databases curated by humans. Databases curated by algorithms tend to contain the most inaccuracies and errors. For instance, data from the gene ontology consortium is the most common source of gene set groupings, but the data also tend to include information from weaker sources. In particular, annotations based on IAE (Inferred from Electronic Annotation) are viewed quite skeptically, but according to the GO annotation website (http://www.geneontology.org) as of May, 2009, only 64,568 of 160,498 human GO annotations are from non-IAE sources. Mistakes also tend to arrise from inconsistencies and abmbiguity in gene names/symbols. Use of high quality groupings is absolutely essential.

Use of prior biological knowledge can alleviate some of the problems associated with analysis of gene expression profiles. Use of these methods has led to a better understanding of the biological mechanisms underlying phenotypic responses. These methods do, however, have problems of their own: in addition to relying heavily on the quality of the information used, we have seen that methods seeking to accomplish the same task provide differing results, all of which may be reasonable. Issues regarding multiple comparisons are also of concern since gene sets may be very highly correlated, differing by only a few genes. Clearly, more research is necessary to deal with the unresolved statistical issues and the problem of inconsistent results.

Acknowledgement

This research is supported by a grant from the National Cancer Institute (CA–76404).

References

[1] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235):467.

[2] Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proceedings of the National Academy of Sciences. 2001;98(1):31–36.

[3] Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Research. 2001;29(12):2549.

[4] Bolstad B, Collin F, Simpson K, Irizarry R, Speed T. Experimental design and low-level analysis of microarray data. International review of neurobiology. 2004;60:25.

[5] Speed T. Statistical analysis of gene expression microarray data. CRC Press; 2003.

[6] Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association. 2004;99(468):909–917.

[7] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;p. 289–300.

[8] Storey J. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002;64(3):479–498.

[9] Nacu S, Critchley-Thorne R, Lee P, Holmes S. Gene expression network analysis and applications to immunology. Bioinformatics. 2007;23(7):850–858.

[10] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27.

[11] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000;25(1):25–29.

[12] Gentleman R. Using GO for statistical analyses. In: COMPSTAT 2004. Springer; 2004. p. 171.

[13] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545–15550.

[14] Goeman JJ, Van De Geer SA, De Kort F, Van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. 2004;20(1):93–99.

[15] Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. BMC bioinformatics. 2005;6(1):225.

[16] Manoli T, Gretz N, Groene HJ, Marc K, Eils R, Brors B. Group testing for pathway analysis improves comparability of different microarray datasets. Bioinformatics. 2006;22(20):2500–2506.

[17] Liu D, Lin X, Ghosh D. Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics. 2007;63(4):1079–1088.

[18] Efron B. Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association. 2007;102(477):93–103.

[19] Fortunel N, Otu H, Ng H, Chen J, Mu X, Chevassut T, et al. Comment on "'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". Science. 2003;302(5644):393.

[20] Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. Global functional profiling of gene expression. Genomics. 2003;81(2):98–104.

[21] Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics. 2005;21(18):3587–3595.

[22] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics. 2003;34(3):267–273.

[23] Sweet-Cordero A, Mukherjee S, Subramanian A, You H, Roix JJ, Ladd-Acosta C, et al. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nature genetics. 2004;37:48–55.

[24] Efron B, Tibshirani R. On testing the significance of sets of genes. Annals of Applied Statistics. 2007;1(1):107–129.

[25] Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology. 2004;3(1).

[26] Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proceedings of the National Academy of Sciences. 2005;102(38):13544–13549.

[27] Pavlidis P, Lewis DP, Noble WS. Exploring gene expression data with class scores. In: Pacific Symposium on Biocomputing; 2002. p. 474–485.

[28] Rahnenfuhrer J, Domingues FS, Maydt J, Lengauer T. Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data. Statistical Applications in Genetics and Molecular Biology. 2004;3(1).

[29] Mccullagh P, Nelder JA. Generalized linear models Monographs on statistics and applied probability. Chapman and Hall London; 1989.

[30] Lin X. Variance component testing in generalised linear models with random effects. Biometrika. 1997;84(2):309–326.

[31] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38(8):904–909.

[32] Kong SW, Pu WT, Park PJ. A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics. 2006;22(19):2373.

[33] Cangelosi R, Goriely A. Component retention in principal component analysis with application to cDNA microarray data. Biology Direct. 2007;2(1):2.

[34] Peres-Neto PR, Jackson DA, Somers KM. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Computational Statistics and Data Analysis. 2005;49(4):974–997.

[35] Valle S, Li W, Qin SJ. Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. Ind Eng Chem Res. 1999;38(11):4389–4401.

[36] Jolliffe I. Principal component analysis. Springer; 2002.

[37] Zhu M, Ghodsi A. Automatic dimensionality selection from the scree plot via the use of profile likelihood. Computational Statistics and Data Analysis. 2006;51(2):918–930.

[38] Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007;23(8):980.

[39] Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics. 2006;7(1):55–65.

[40] Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. Journal of the American Statistical Association. 2006;101(473):119–137.

Table 1: 2x2 ORA Contigency table based on membership in the list of differentially expressed genes (D) and the list of genes in the gene set (S)

	Diff. Expressed	Not Diff. Expressed	
In gene set	N_{SD}	N_{SD^c}	N_S
Not in gene set	N_{S^cD}	$N_{S^cD^c}$	N_{S^c}
total	N_D	N_{D^c}	N

Table 2: Diabetes Dataset Results

	Top 100 Genes		Top 5%						
	Fisher	χ^2-test	Fisher	χ^2-test	GSEA	GSA	Global	PCA_T	PCA_K
100 (Fisher)	7	7	7	7	2	5	2	2	2
100 (χ^2-test)		32	10	32	3	5	3	3	2
5% (Fisher)			60	60	8	17	4	3	14
5% (χ^2-test)				82	9	17	4	4	14
GSEA					18	9	1	0	6
GSA						20	2	2	10
Global							4	1	3
PCA_T								5	1
PCA_K									16