Gene Set Analysis Using Principal Components

Isa Kemal Pakatci
Dept. of Computer Science
University of North Carolina
Chapel Hill, NC 27599, USA
kemal@cs.unc.edu

Wei Wang
Dept. of Computer Science
University of North Carolina
Chapel Hill, NC 27599, USA
weiwang@cs.unc.edu

Leonard McMillan
Dept. of Computer Science
University of North Carolina
Chapel Hill, NC 27599, USA
mcmillan@cs.unc.edu

ABSTRACT

We present a new method for identifying gene sets associated with labeled samples, where the labels can be case versus control, or genotype differences. Existing approaches to this problem assume that variations observed within a group are due primarily to noise and they, therefore, look for significant mean shifts between groups. Biological evidence suggests variations can also result from the coordination of genes. Our method attempts to identify and assess the significance of changes in gene-gene correlation patterns. We model gene-gene correlations using principal component analysis and compare their significance to a baseline of a linear models generated by random permutations of the sample labels. Simulation results show that our method detects changes that are undetectable by Hotelling's T^2 method. Its performance on real data is comparable to existing methods with the additional capability of detecting changes in gene-interactions between sample groups.

1. INTRODUCTION

Expression microarrays provide a valuable tool for measuring the relative abundance of RNA transcripts both within and between cell and tissue types. Standard analysis methods either compare these expression levels between case and control examples and/or correlate expression levels between genes in many samples. A common simplifying assumption that is widely applied in such analyzes is that genes function independently, which is inconsistent with the underlying biology that suggests most genes interact in regulatory modules or networks.

Coordination between genes leads researchers to be more interested in the expression of gene sets rather than individual genes. Many researchers have used annotated databases of gene categories to incorporate biological knowledge into microarray expression analysis. The objective of these approaches is to determine which gene categories might be influenced by the conditions of the given experiment. Statistical methods are used both to identify and assess the significance of any changes seen among these gene categories. In this paper we evaluate predefined gene sets to determine the extent that they are associated with expression pattern differences between a given set of class labels. These class labels can be case/control or any binary phenotypic or genotypic variation.

Common statistical methods for assessing differential gene-set expression patterns between sample classes look for shifts in the vector-mean of the set and assume the observed variance is due to noise. We are interested in identifying changes in gene correlation patterns associated with the class labels. Figure 1b shows a synthetic example where the expression of two genes have a strong positive correlation in group 2 samples, and the same two genes are less and slightly negatively correlated in group 1, yet both groups have similar means. This difference in of gene interactions might be the most significant signal differentiation between the sample classes, but it is not detected by classical methods.

Our method captures significant changes in gene-set expression patterns between samples belonging to different classes. It considers more than shifts in the mean expression levels of the gene set; it also accounts for significant changes in correlation structure between genes in the set. We show the utility of our method by applying it to simulated data, and two real data examples. We also directly compare our method to [10] using a common data set.

2. RELATED WORK

One can divide existing approaches for scoring differential gene-set expression into two types based on their approach. The first type calculates a statistic for each gene individually and then calculates a summary statistic for the gene set. An example of these methods is Gene Set Enrichment Analysis (GSEA) [13], which calculates signal-to-noise ratio for each gene. GSEA ranks genes according to this score, an enrichment score (ES) for the gene set, which is calculated by comparing ranks of the genes in the gene set to the genes not in the set by using a one-sided Kolmogorov-Simirnov test. Another example is the Q2 test of Tian et al. [14] in which an association measure (either t-test or Wilcoxon rank sum) between each gene and phenotype is calculated and a gene set score is calculated by the average of individual gene association measures. Barry et al. [5] described a framework that is based on using local statistics derived for each gene and combining local statistics in a global statistic for each gene set. There are also 2x2 contingency table based methods [3,

ACM-BCB 2010, Niagara Falls, NY, USA

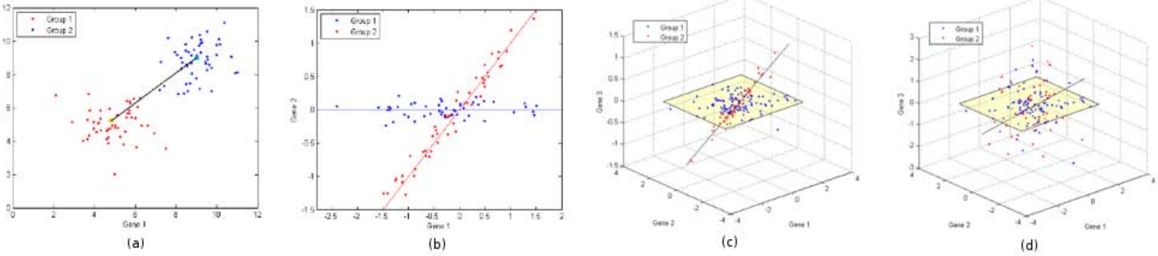

Figure 1: (a) Example where the information is on the means and all the variation is due to noise. Means are shown by squares. Line connecting the means is also shown. (b) Example where the information is on the correlation between genes. Lines indicate the fitted model (two lines) (c) Sample simulated data where number of genes is 3 ($d = 3$). Plot shows the case where there is low noise ($\sigma = 0.1$) so that the embedded pattern is evident.Group 1 lives on a plane whereas group 2 lives on a line, so the intrinsic dimensionality of group 1 and group 2 are $d_1 = 2$ and $d_2 = 1$ respectively. (d) Has the same setup as in (c) but with a high noise ($\sigma = 0.5$). It is very hard to see the differential behavior between two groups at this noise level found by PCA.

8, 11] where statistics for each gene are ranked and, based on significance cut-off, the independence of significance of genes versus existence of genes in the gene set is tested. A common assumption of these approaches is that if a gene set is associated with the class label then significant number of genes in the gene set should be similarly impacted. However, genes need not behave in unison, and class differences might reflect the extent to which subsets genes in the set are co-regulated. Methods of this type cannot capture subtle, but coordinated changes in the expression of the genes in the gene set.

A second approach assesses the significance of the gene set directly from raw expression data rather than the scores calculated for each gene [6, 7, 12, 15, 10]. Kong et al. [10] used Hotelling's T^2 statistic to test association of genes in the set with the binary phenotype. Tomfohr et al. [15] projected the expression levels of genes onto their first principal component and significance is calculated by using t-test. Goeman et al. [6, 7] scored gene sets by measuring how well expression levels can predict the class labels using logistic regression. To detect the differential expression, Mansmann and Meister [12] used an ANOVA based model which takes into account the interaction between gene and the phenotype. The assumption underlying all of these methods (t-test, Hotelling's T^2 statistic,ANOVA) is that class differences are exhibited by shifts in mean expression levels and the observed variance is due to noise. This assumption ignores the possibility of differences due to coordination between genes. Perturbations can affect gene correlation patterns rather than their mean expression. In this paper we introduce a novel method to capture more effects.

3. METHODS

Our method assumes that the samples are drawn from one or more linear manifolds according to their class label. These linear manifolds constitute our model. Formally, our method assumes that the samples are generated as follows:

$$\mathbf{x}_i = f_i(.) + \epsilon$$

where $\mathbf{x}_i$, the gene expression vector for a sample with the class label i, $f_i(.)$, is a linear function of the form $\mathbf{m}_i + \alpha_{i1}\mathbf{v}_{i1} + \alpha_{i2}\mathbf{v}_{i2} + ... + \alpha_{id_i}\mathbf{v}_{id_i}$ where d_i is the intrinsic dimension of the group i and $\epsilon \sim N(0, \sigma I)$ where I is the identity matrix and σ is the variance of the noise. The vectors, $\mathbf{v}_{ij}$, encode interactions between genes and scalars α_{ij} encode the state of each sample in group i. Our method estimates the functions, $f_i(.)$, using principal component analysis. Then we test the null hypothesis $H_0: f_1(.) = f_2(.) = \ldots = f_N(.)$ using the permutation testing over the residuals.

Multivariate Hypothesis Testing

There are many possible sources for gene-set candidates including gene ontology annotations (GO)[4], Kegg pathways[2], or derived gene networks[1]. Our method treats predefined gene-sets from gene ontology annotations (GO)[4], Kegg pathways[2], or derived gene networks[1] as vector variables, and solves for some lower dimensional space that captures the variation within the groups identified by class labels. After fitting a model, which consist of one linear manifold per class label, we test the significance of this fit using permutation testing over the residuals. The intuition behind the proposed method is to test whether partitioning of the samples (according to given sample labels, i.e response to a toxin) enables us to achieve a significantly lower residual than could be achieved by chance. The difference between ANOVA and our method is that ANOVA (or its high-dimensional analog, Hotelling's T^2) measure the residual by calculating the squared distance to the mean, whereas our method measure the residual by calculating squared distance to the estimated model whose form is described above. This reflects a difference in assumptions. Where ANOVA assumes that all variation is due to noise (hence only the mean contains signal) our model asssumes that variation within the groups contains some signal. Therefore, we estimate a model by extracting a linear manifold for each class label using PCA, and we measure the residual. If this residual is unlikely due to chance then this means that there is a significant association between class labels and the gene set. We measure the significance of association by using permutation testing using residuals as a test statistic.

Formally, suppose we have N groups; let X_i be the $n_i \times d$ data matrix of group i where columns are gene expression levels and rows are samples. We apply PCA on both data matrices in order to estimate the linear manifolds at our model. We do so by eigen-decomposition of covariance matrix where $\lambda_{i1}, \lambda_{i2}, ..., \lambda_{id}$ and $\mathbf{v}_{i1}, \mathbf{v}_{i2}, ..., \mathbf{v}_{id}$ are eigenvalues and eigenvectors of covariance matrix of the data matrix X_i. We estimate a linear manifold for each group i, by selecting k_i eigenvalues and corresponding eigenvectors where k_i is the number of dimensions required to explain some fraction, α, of the total variance within the group. The sum of remaining $d - k_i$ eigenvalues is the sum of the squared distances of each point to linear k_i-dimensional manifold giving a measure of the residual of the fit of this derived model to the actual data. We next randomly permute the class labels

Sim. Params.				T^2	PCA Method			
d	d_1	d_2	σ		α=.90	α=.80	α=.70	α=.50
3	2	1	0.1	0.330	**0.001**	**0.001**	**0.001**	**0.004**
3	2	2	0.1	0.492	**0.001**	**0.001**	**0.001**	0.054
10	2	2	0.1	0.368	**0.001**	**0.001**	**0.001**	**0.001**
10	8	6	0.1	0.051	**0.001**	**0.001**	**0.001**	**0.002**
3	2	1	0.5	0.300	**0.001**	**0.001**	**0.001**	**0.002**
3	2	2	0.5	0.448	**0.001**	**0.001**	**0.001**	**0.006**
10	2	2	0.5	0.311	0.639	**0.030**	**0.001**	**0.001**
10	8	6	0.5	0.046	**0.001**	**0.001**	**0.001**	**0.005**

Table 1: P-values for simulated data. P-values less than 0.05 are shown in bold. Parameterization: d is dimensionality of the whole data (number of genes). d_1, d_2 are dimensionalities of the two groups. σ is the variance of the added noise. α is the explained variance threshold that is used by our model

Net Name	Role	T^2	PCA
mmu00970	Aminoacyl-tRNA biosynth.	0.00010	0.00005
mmu04920	Adipocytokine signaling	0.00015	0.00005
mmu05222	Small cell lung cancer	0.00055	0.00005
mmu00750	Vitamin B6 metabolism	0.00330	0.00005
mmu04320	Dorso-ventral axis formation	0.08535	0.00005
mmu00750	Vitamin B6 metabolism	0.00330	0.00015
mmu03060	Protein export	0.00150	0.00010
mmu04540	Gap junction	0.01220	0.00015
mmu03050	Proteasome	0.00245	0.00020
mmu04340	Hedgehog signaling	0.00130	0.00105

Table 2: P-values as result of applying T^2 method and our method.

and search for alternative models of the same dimensions with a better fit (smaller residual). The p-value is calculated as the proportion of random group-label assignments that result in a better fit than the observed one.

4. RESULTS

We first compared our method to Hotelling's T^2 method using simulated data. We synthesized gene expression values with embedded correlation patterns that differed between class labels to which we then added Gaussian noise. Then we applied ours and the T^2 methods to detect patterns of differential expression and calculated estimated p-values based on 1000 permutations (i.e. we applied random label assignments to the same set of expression values). We also varied our method's parameter settings to explore its robustness under different conditions.

Table 1 shows results on simulated data. Notice that the T^2 method does not detect any of the embedded patterns. The T^2 method tests for significance of the mean shift, but in our simulated data set the means of the two groups were very close. On the other hand, our method gives significant p-values (< 0.05) especially when noise level is low ($\sigma < 0.5$). At low noise levels our method detects correlated differential expression levels for any variance thresholds, $\alpha > 0.5$. When we look at the high noise cases ($\sigma = 0.5$), our method still detects most patterns. This result is impressive because a noise variance of 0.5 makes the two groups nearly indistinguishable from each other as shown in Figure 1d.

In general, p-values are insensitive to the threshold setting, α, however, for data with unknown noise levels one can search for an optimal setting of α by searching the range between 0 and 1 for the smallest p-value.

4.1 Gene Regulation Pathways

We used the publicly available dataset cited in [10] to compare our method the Hotelling's T^2 method used by the same paper. In this dataset there are 23695 genes and 19 samples, 10 were from mice treated with a chemical, RAD001, which prevents the activation of TOR (Target of Rapamycin). The remaining 9 samples are from the mice were treated with a placebo. The mouse gene pathway sets are taken from KEGG database [9]. We used 201 pathways obtained from KEGG database on September 2009

Gene Name	Chromosome	Position
Hadhsc	3-	131,816,213-131,854,921
Ppt1	4+	121,206570-121,228,697
Hadha	5-	28,524982-28,561,657
Hadha	5-	28,524982-28,561,657
Hadhb	5+	28,561999-28,591,338
Hadhb	5+	28,561999-28,591,338
Echs1	7-	127,763417-127,774,118
Ppt2	17-	33,112,563-33,122,931
Acaa2	18+	75,308,994-75,337,099
Hsd17b4	18+	50,574,867-50,642,794

Table 3: Positions of the genes in the mmu00062 network. Positions are given in base pairs.

We varied the threshold values and observed more significant results for lower threshold values. We hypothesize that this is due to a high-level of noise in the expression data, thus, causing small gene-correlation signals to be lost if the threshold is too selective, which is consistent with the trend seen in the simulated data. Table 2 shows top 10 networks based on our method with threshold α=0.75, which coresponds to linear-group models that explain 75% or more of the observed variance.

The most significant network is mmu00970 is also among the top ten networks found by T^2 method. This is an example of an overlap between our method and the T^2 method. In general, when either method gives a significant p-value (i.e. <0.05) the other method also gives a significant p-value, which means our method is comparable to the existing approaches. However there are notable exceptions, such as, mmu05410, which was detected as significant by our method but was not detected by T^2. This network is related to heart disease which is related to the experiment under investigation, and might warrant addtional exploration.

4.2 eQTL study

Expression Quantitative Trait Locus (eQTL) analysis studies gene expression as a phenotypic variation relative to differences in genetic background. For genetic background, we considered DNA sequence variants called Single Nucleotide Polymorphisms (SNPs).

The data used are from a selection strain of mice produced to investigate fat metabolism. The mice were selected over several generations to have a high ratio of body fat for a fixed diet. Gene expression data was extracted from hypothalamus which is responsible for metabolic processes and controlling the hormonal activity. Expression and genotype data was provided for both founder and derived mice. We expect to observe a perturbation in expression levels of the genes that are related to fat metabolism. Thus, we fixed the set of genes to be in KEGG pathway mmu00062 - Fatty Acid Elongation in mitochondria, which is related to fat metabolism in cells. The group labels are defined by 1813 genome-wide SNPs. In this experiment we fixed the set of genes and searched for group labels (genotypes) that are associated with the selected set of genes,whereas in the previous experiment group labels were given apriori and we searched through gene-sets. The variance threshold was set

SNP Name	Chr.	Pos.	P-value	Heterozygosity
rs3704486	4	125.09	0.001	Dominant Minority
mCV22554962	5	117.83	0.002	Dominant Minority
rs3719258	7	134.52	0.002	Dominant Minority
rs8253487	1	87.49	0.002	Dominant Majority
rs3700831	1	177.85	0.002	Additive
rs3683699	18	30.65	0.003	Additive
rs6286913	18	30.96	0.003	Additive
mCV24690992	12	63.87	0.004	Additive
rs3677302	10	20.8	0.004	Additive
rs3708441	1	177.49	0.004	Additive

Table 4: Results of eQTL analysis applied on hypothalamus tissue. Top 10 SNPs that are found to be significantly attached to genes in mmu00062 pathway are shown

to $\alpha = 0.75$. Table 3 shows the genes in the selected network and their genomic positions.

In eQTL experiments, it is common to compare the genomic positions of genes to the positions of those genetic markers that are found to be most significant. If the most significant markers are close to the genes in the pathway, it is perhaps due to regulation behavior modifications in the gene set corresponding to a potential cis-regulatory module. Table 4 shows most significant SNPs found using our algorithm. As can be seen from the table the most significant SNP is rs3704486 which is located on the chromosome 4 at approximately 125 megabases. Table 3 shows that the Ppt1 gene (palmitoyl-protein thioesterase 1) is also located on chromosome 4 at the 121th megabase. A 4 megabase distance might be considered to be quite large, but note that our SNP data is sparse (it does not cover the DNA in high resolution), and that the identified SNP is the second closest SNP to this gene. We also applied Hotelling's T_2 test to this SNP and got a p-value of 0.76, implying that our method indentified a potential association where Hotelling's T_2 test does not. The proximity of SNP position to one of the genes of the pathway signifies a likely true association. This result shows the utility of our method, but any inferences are subject to further experimental investigations.

5. CONCLUSION AND FUTURE WORK

Analyzing gene expression experiments gives us insights about biological mechanisms. Existing computational tools concentrate on detecting shifts in gene expression due to changes in the mean-level of expression. Our PCA-based method goes beyond merely detecting mean shifts by modeling the correlation patterns among coordinating genes. Using simulated and real data, we have shown that this method offers a tool for detecting differentially expressed gene patterns under different conditions.

We demonstrated our method's use for eQTL analysis where the categories are based on genotypic variations. Current approaches to eQTL study analyze the data on gene-by-gene basis, ignoring gene-gene interactions) whereas our method treats gene-sets as a unit. This approach enables the detection the genetically influenced gene-network variations. We showed an example of the utility of our approach on eQTL data, by detecting potential cis-regulatory modules.

In this work, we focused on cases where gene sets are provided a priori. Our long-term goal is to discover gene-sets that work together to regulate a phenotype of interest. This would enable biologists to focus on gene-networks that are supported by computational evidence. Considering the fact that there are tens of thousands of genes, the space of potential gene-sets is huge, and efficiently searching through this space is a daunting computational problem. Furthermore, when applied to eQTL analysis, the number of potential group-labels considered, which scales with the size of the marker set, futher adds to the computational challenge.

6. ACKNOWLEDGEMENTS

This work was supported by EPA STAR grant RD832720.

7. REFERENCES

[1] Biocarta. `http://www.biocarta.com`.

[2] Kegg: Kyoto encyclopedia of genes and genomes. `http://www.genome.jp/kegg/`.

[3] F. Al-Shahrour et al. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics (Oxford, England)*, 20(4):578–580, March 2004.

[4] M. Ashburner et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature genetics*, 25(1):25–29, May 2000.

[5] W. T. Barry, A. B. Nobel, and F. A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, May 2005.

[6] J. J. Goeman et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, January 2004.

[7] J. J. Goeman et al. Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21(9):1950–1957, May 2005.

[8] D. A. Hosack et al. Identifying biological themes within lists of genes with ease. *Genome biology*, 4(10), 2003.

[9] M. Kanehisa et al. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucl. Acids Res.*, pages gkp896+, October 2009.

[10] S. W. Kong, W. T. Pu, and P. J. Park. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, 22(19):2373–2380, October 2006.

[11] H. K. Lee, W. Braynen, K. Keshav, and P. Pavlidis. Erminej: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, 6, 2005.

[12] U. Mansmann and R. Meister. Testing differential gene expression in functional groups. goeman's global test versus an ancova approach. *Methods of information in medicine*, 44(3):449–453, 2005.

[13] V. K. Mootha et al. Pgc-1Îś-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, June 2003.

[14] L. Tian et al. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13544–13549, September 2005.

[15] J. Tomfohr, J. Lu, and T. B. Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6, 2005.