Increasing interpretability of the relationship between transcriptome and phenotype through a multivariate function-based transformation of gene expression data.

Ana Conesa[1], Rasmus Bro[2], José Manuel Prats[3], Karin Kjeldahl[2], David Montaner[1] and Joaquín Dopazo[1].

1 Bioinformatics Deparment , Centro de Investigacion Principe Felipe, Valencia, Spain
3 Quality and Technology Department of Food Science, KVL, Copenhagen, Denmark
4 Department of Statistics, Polytechnic University of Valencia, Valencia, Spain

Abstract

We present here a novel approach to the analysis of transcriptomics data that integrates functional annotation of gene sets and expression values in a multivariate fashion and directly assesses the relation of functional features to a multivariate space of response phenotypical variables. Multivariate projection methods are used to obtain new correlated variables for a set of genes that share a given function. These new functional variables are then related to the response variables of interest. The analysis of the principal directions of the multivariate regression allows for the identification of gene function features correlated with the phenotype. This approach has demonstrated to be superior to the equivalent univariant approach.

Introduction

Gene expression profiling is used to study the gene regulatory basis of phenotypic or developmental characteristics. Statistical analysis of transcriptomics data is normally addressed through a two step-process: First a statistical test is performed to derive a p-value for the association of individual gene expression values to the phenotype or experimental condition(s) [1]. Secondly a number of "significant genes" are selected on the basis of a p-value threshold. These genes are further analyzed to detect the presence of significant enrichments in functional categories [2]. Such an approach presents a number of limitations: firstly, the univariate nature of the gene statistical assessments implies that many informative correlation structures within the data are ignored and that strong p-value corrections need to be applied which can hamper the identification of significant features on large datasets. Furthermore, as functional assessments (which paradoxically depend on multivariant gene activity) are performed after univariant gene selection, results are dependent on the p-value cutoff of choice, which can be problematic. Thus, too strict p-value cutoffs may lead to univariately non-significant genes (that are in fact multivariately significant but remain undetected) while too non-strict cutoffs may lead to multivariate important features getting lost among irrelevant information. Finally, when the target phenotype is not composed by a single variable but a space of different measurements (e.g. age, gender, different clinical parameters, etc.), the evaluation of differential expression under a univariate strategy can imply multiple and difficult assessments. Multivariate approaches to gene expression analysis try to overcome the limitations, e.g. using projection techniques to capture correlations patterns in gene expression data or evaluations of functionally-related gene-sets ranked by a measure of differential expression [3, 4]. While these approaches have demonstrated to be more

powerful, they still suffer from limitations regarding considering several variables in the phenotypical space. In this paper we present a novel approach to the analysis of transcriptomics data that integrates functional annotation and expression values in a multivariate fashion and directly assesses the relation of functional features to a multivariate space of response phenotypical variables.

Material and Methods

Basically, our proposal uses multivariate projection methods to obtain new correlated variables for a set of genes that share a given function. These new functional variables are then used to perform a multivariate regression on the response variables. The analysis of the principal directions of the multivariate regression allows for the identification of gene function features correlated with the phenotype. An outline of the algorithm is as follows:

1. Find the functional annotation of the genes in the transcriptomics data set
2. For each functional term, find all annotated genes and their expression values
3. Perform PCA on the expression matrix formed by the selected genes
4. Take a number of components that collect non random variation
5. Take the PCA scores for these number of components
6. Collect the scores of all function matrices in a new matrix of functional variables
7. Use this new matrix to perform PLS regression on the response variables
8. Explore PLS model and find important functional variables in that model
9. Use these functional features to explain the gene –regulation basis of the phenotype

With the above approach, two potentially critical problems can be overcome based on the assumption that important genes are correlated to similar genes. First of all, the unimportant genes are dramatically reduced in numbers which can be decisive in order to be able to detect important variations. Secondly, the important (as well as unimportant) variation is expressed in a reduced form by scores from principal component analysis. Hence, ideally, each phenomenon appears only once and therefore has a much better chance of influencing the further analysis.

We have tested the method on the dataset by Heijne et al. [5] of a toxicogenomics study on rats. In this experiment, rats are administrated the drug bromobencen at three different doses (High, Medium and Low) and blood/urine samples are taken after 6, 24, and 48 hours of treatment. There are control (no administration) and placebo (only drug vehicle administration) rat groups. For each experimental condition one to three rats are taken for gene expression profiling and microarray experiments are done with a dye-swap design. There is gene expression information for 2665 genes. Data was normalized by lowess and centered for each dyeswap set. All computations were performed in R, using Limma [6] and plsr packages.

Additionally, there are measurements for the same rats of physiological and morphological variables (Body Weight (g), Kidneys weight(g), Kidney/BW (g/kg),Liver (g),Liver/BW , Bilirubin tot, ASAT, ALAT, LDH, Albumin g/l, ALP (U/l), Creatin umol/l, Cholesterol (mmol/l), Glucose (mmol/l), Phospholipids (mmol/l), Triglycerides (mmol/l), Tot.Protein (g/l), Urea (mmol/l, A/G ratio, GSH corr.(M).

Gene ontology functional annotation was assigned to these genes and the annotation score [7] was computed for each annotation term using the software blast2go (inclusive analysis). GO terms with an annotation score and at least four annotated genes were selected as functional data.

Results

PCA analysis of both gene expression data (Figure 1) and clinical variables (Figure 2) revealed a first component of variability that basically differentiates the High bromobenze doses treatment at 24 and 48 hours from the rest of the conditions. A pretty similar PCA score plot was found for gene expressions and clinical variables, indicating that the major pattern of the variability in both datasets had similar structures and related to the intensive administration of the drug.

The GO selection procedure obtained a total of 1140 GO terms from the 3 ontologies were selected. After PCA-based transformations 343 functional components were obtained corresponding to 117 different GO terms. Most GO terms generated one or two PCA components (functional variables) and only a few (high hierarchy GO levels) generated between 10 and 20 variables. PCA analysis of the new matrix of functional variables shows a projected space similar to those obtained previously with gene expression and clinical data (Figure 3), but explained variance for the first , most discriminating, component was clearly higher (38% with functional variables versus 21% with gene expression data).

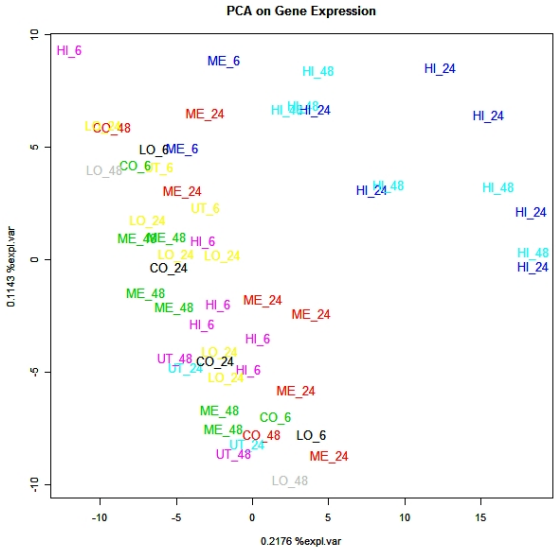

Figure 1. PCA analysis of gene expression data

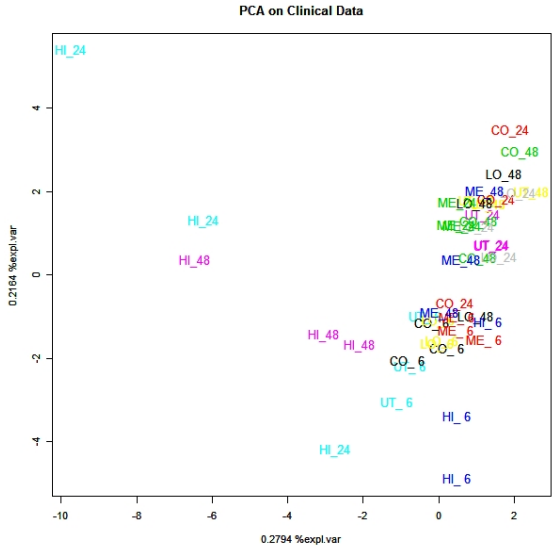

Figure 2. PCA analysis of clinical variables.

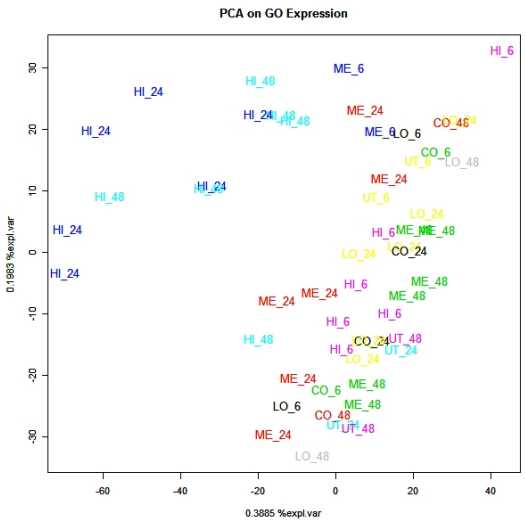

Figure 3. PCA analysis of the new matrix of functional variables

Figure 4. Loading plot of the clinical variables on the two first components of the PLS model

PLS regression was performed predicting one of the clinical variables from the functional data matrix and the number of components was selected by leave-one-out cross validation, indicating 5 components which had a Q2 (cross-validation-prediction error) of 0.58. Figure 4 shows the loading plot of the clinical variables on the two first components of the PLS model, and Figure 5 shows the score plot for the functional data. Table 1 gives the importance value (VIP) of each clinical variable in the model, showing a most determinant LDH, ALAT, GSH and ASA and ALP. The score plot of the functional variables shows a slight modification of the distribution of samples, being the 24 hours HI dose treatment the condition that best correlates with the clinical variables. The multivariate approach captures a large number of GO functional terms that are displayed in DAG structure as plotted by the Blast2GO program [7] in figure 6. Selected GO terms include response to stimulus, heme binding, oxidoreductase activity, glutathione transferase, apoptosis, terms like apotosis, ribosomal unit and cytoskeleton.

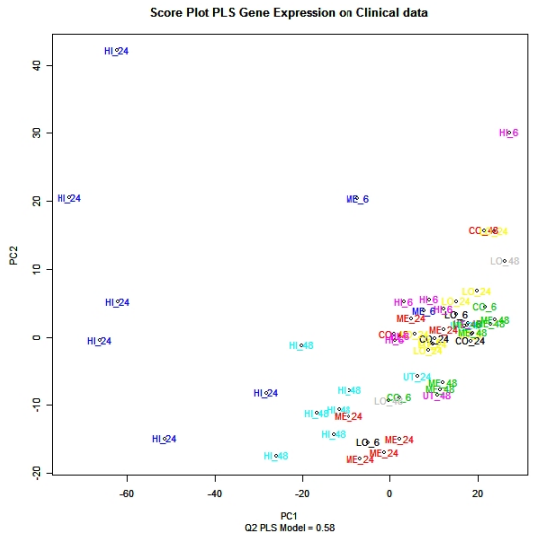

Figure 5. Score plot for the functional data

Clinical Varibles	Importance in the PLS Model
LDH	5.60E+05
ASAT	2.20E+04
ALAT	1459
GSH.corr..M.	711.8
ALP..U.l.	6.539
Body.Weight..g.	0.05506
Liver.BW..g.kg.	0.0201
Bilirubin.tot	0.01915
Creatin.umol.l	0.003769
Glucose..mmol.l.	0.002205
Tot.Protein..g.l.	0.001464
Liver..g.	0.0007265
Urea..mmol.l.	0.0004417
Phospholipids..mmol.l.	0.0001061
Kidney.BW..g.kg.	0.0001051
Albumin.g.l	9.93E-05
Cholesterol..mmol.l.	9.66E-05
Triglycerides..mmol.l.	4.48E-05
A.G.ratio	6.25E-06
Kidneys..g.	1.21E-06

Table 1. Importance value of each clinical variable in the model

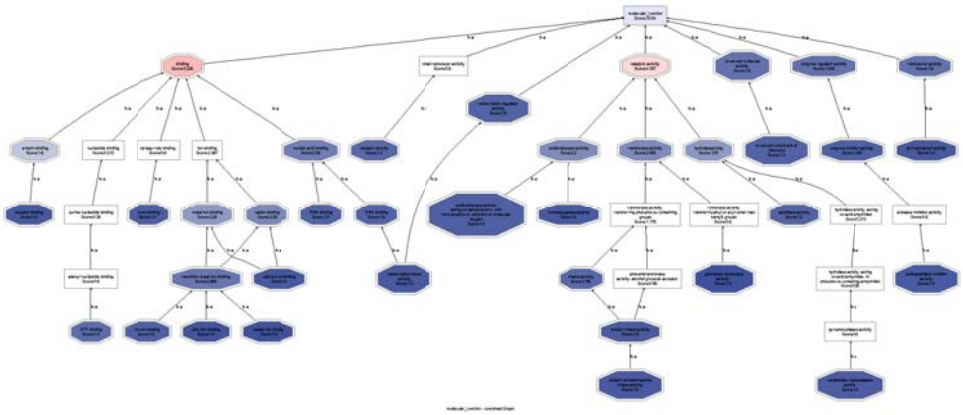

Figure 6. DAG representation of the selected terms

Discussion

The approach presented integrates in one analysis three basic elements of transcriptomic analysis: gene expression data, functional annotation and phenotype characteristics, providing a direct correlation of gene function to response variables.

The important clinical and functional variables identified in the study are very much in agreement with previous analysis of the liver toxicology response: GSH (glutathione) is a principal player of the detox response by conjugating xenobiotics to be targeted for degradation. ALA, ASAT and ALP (alkaline phosphatase) are typical indicators of oxidative stress [5]. At the side of functional terms, the set of functionalities obtained represent the general molecular response to drug administration: glutathione transferase and oxidoreductase activities play a role in detoxification. Heme binding is indicative of the activity of heme oxigenase and cytochromes in the process. Additionally, protein synthesis and ribosome proliferation are also affected by the cellular stress. Other processes such as changes in cytoskeleton organization have also been reported and are highlighted by the functional analysis.

Comparison of the functional analysis approach proposed here to a similar univariate analysis on the same dataset previously performed by us [8] shows that multivariate methods are more effective in highlighting relevant gene functions. While the more traditional approach identified terms belonging to a few GO branches (basically, ribosome and oxidoreductase/heme activities) (data not shown), the proposed approach showed a more diverse set of relevant functions. The DAG representation of the selected terms (Figure 6) shows functional terms distributed in numerous braches of the gene ontology and a various levels, expanding from high specific (6-7 level) to more general (level 2) that collect selected terms at higher specificities.

References

1. Speed T: **Statistical Analysis of Gene Expression Microarray Data**. London: Chapman & Hall/CRC; 2003.
2. Dopazo J: **Functional interpretation of microarray experiments**. *Omics* 2006, **10**(3):398-410.
3. Al-Shahrour F, Arbiza L, Dopazo H, Huerta-Cepas J, Minguez P, Montaner D, Dopazo J: **From genes to functional classes in the study of biological systems**. *BMC Bioinformatics* 2007, **8**:114.
4. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information**. *Bioinformatics* 2005, **21**(13):2988-2993.
5. Heijne WH, Stierum RH, Slijper M, van Bladeren PJ, van Ommen B: **Toxicogenomics of bromobenzene hepatotoxicity: a combined transcriptomics and proteomics approach**. *Biochem Pharmacol* 2003, **65**(5):857-875.
6. Smyth GK: **Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments**. In: *Statistical Applications in Genetics and Molecular Biology.* vol. 3; 2004: 3.
7. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.
8. Conesa A, Nueda MJ, Ferrer A, Talon M: **maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments**. *Bioinformatics* 2006, **22**(9):1096-1102.