PathVar: analysis of gene and protein expression variance in cellular pathways using microarray data

Enrico Glaab 1,2* and Reinhard Schneider 1,2
1 Structural and Computational Biology Unit, EMBL, Meyerhofstrasse 1, 69117, Heidelberg and 2 Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg, Germany

ABSTRACT

Summary: Finding significant differences between the expression levels of genes or proteins across diverse biological conditions is one of the primary goals in the analysis of functional genomics data. However, existing methods for identifying differentially expressed genes or sets of genes by comparing measures of the average expression across predefined sample groups do not detect differential variance in the expression levels across genes in cellular pathways. Since corresponding pathway deregulations occur frequently in microarray gene or protein expression data, we present a new dedicated web application, PathVar, to analyze these data sources. The software ranks pathway-representing gene/protein sets in terms of the differences of the variance in the within-pathway expression levels across different biological conditions. Apart from identifying new pathway deregulation patterns, the tool exploits these patterns by combining different machine learning methods to find clusters of similar samples and build sample classification models.

Availability: freely available at http://pathvar.embl.de

1 INTRODUCTION

In the search for new diagnostic biomarkers, one of the first steps is often the identification of significant differences in the expression levels of genes or proteins across different biological conditions. Commonly used statistical methods for this purpose quantify the extent and significance of changes in measures of the average expression levels of single genes/proteins [see for example [1,2,3,4,5]] or analyze aggregated data for gene/protein sets representing entire cellular pathways and processes [6,7,8,9]. However, since these approaches compare measures of averaged expression levels, they cannot study how the variance of expression levels across the genes/proteins of a cellular pathway (termed ‘pathway expression variance’ here) changes under different biological conditions. In this article, we present a web application for microarray data analysis to identify and prioritize pathways with changes in the pathway expression variance across samples (unsupervised setting) or predefined sample groups (supervised setting). In particular, we show example cases on cancer data in which significant pathway deregulations manifest themselves in terms of changes in the variance of the gene/protein expression levels in pathways, while no significant changes can be detected in the median pathway expression levels (see section ‘Results on Cancer Microarray Data’ and Fig. 1). Finally, we discuss how the software enables automated sample clustering and classification using the extracted pathway expression variances.

2 WORKFLOW AND METHODS

PathVar identifies and analyzes deregulation patterns in pathway expression using two possible analysis modes, a supervised and an unsupervised mode, chosen automatically depending on the availability of sample class labels.

In the first step, the user uploads a pre-normalized, tab-delimited microarray dataset and chooses an annotation database to map genes/proteins onto cellular pathways and processes (see Section 4). Next, in the supervised analysis mode, the software computes two gene/protein set rankings in terms of differential pathway expression variance using a parametric T-test and a non-parametric Mann-Whitney U-test (or respectively, an F-test and Kruskal-Wallis test for multi-class data). Alternatively, in the unsupervised analysis mode, three feature rankings are obtained from the pathway expression variance matrix (rows = pathways, columns = samples) by computing the absolute variances across the columns/samples, the magnitude of the loadings in a sparse principal component analysis

*To whom correspondence should be addressed.

Fig. 1. Left: box plot comparing the median expression levels in the KEGG Urea cycle pathway (hsa00220) for the prostate cancer dataset by Singh et al. [20] across 50 healthy individuals (green) and 52 tumor patients (red); right: box plot comparing the variance of expression levels in the same pathway and microarray dataset (see also Supplementary Material).
be combined freely by the user [see Glaab score plots, principal component plots, dendrograms and silhouette plots) in terms of cluster compactness and separation between the clusters, five clustering approaches and identify a number of clusters that is optimal Material]. To estimate the accuracy of the generated classification models, employed in variety of bioscientific studies (Bassel data analysis framework Glaab machine learning technique implementations stem from a fully automated selection across different cross-validation cycles, and a heat map is generated Kappa statistic are computed. Additionally, a other performance statistics like the sensitivity and specificity, and Cohen’s as well as user-defined training/test set partitions. In addition to the average (1995) (see section on limitations in the Supplementary Material for details)

3 RESULTS ON CANCER MICROARRAY DATA

The microarray prostate cancer dataset by Singh et al. [2002], containing 52 tumor samples and 50 healthy control samples, is a typical example for a cancer-related high-throughput dataset with gene expression deregulations across many cellular pathways. When analyzing this data using both a comparison of median gene expression levels in KEGG pathways across the sample classes, and a comparison of the expression level variances with PathVar, the top-ranked pathway in terms of differential expression variance, Urea cycle and metabolism of amino groups (hsa00220), showed a significant increase of the variance in the tumor samples (see Fig. right; adjusted P-value: 2.26e-06). Interestingly, a conventional comparison of the corresponding median gene expression levels does not identify statistically significant differences between the sample groups (Fig. left). Similar results were obtained for other cancer-associated KEGG pathways, including the angiogenesis-related VEGF signaling pathway (hsa04370) and the inflammation-related Natural killer cell mediated cytotoxicity (hsa04650) process. Corresponding statistics and box plots are provided in the Supplementary Material, which also contains results from the clustering module and the classification module, similar outputs for a further microarray study, as well as details on the used data and normalization procedures. In summary, PathVar identifies statistically significant pathway deregulations, different from those detected by methods for comparing averaged expression levels, and provides pathway-based clustering and classification models that enable a new interpretation of microarray data.

4 IMPLEMENTATION

All data analysis procedures were implemented in the R statistical programming language and made accessible via a web interface written in PHP on an Apache web server. Gene and protein sets representing cellular pathways and processes were retrieved from the databases KEGG (Kanehisa et al., 2008), BioCarta (Schaefer et al., 2005), Reactome (Jo市教育) et al., 2009), NCI Pathway Interaction Database (Schaefer et al., 2008), WikiPathways (Picó et al., 2009) InterPro (Kuwata et al., 2007) and Gene Ontology (GO) (Ashburner et al., 2000) and will be updated on a regular basis. A detailed tutorial for the software is provided on the web page.

Funding: German Academic Exchange Service (DAAD) short-term fellowship to (E.G.).

Conflict of Interest: none declared.

REFERENCES