BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 3 2012, pages 446–447
doi:10.1093/bioinformatics/btr656

Systems biology

Advance Access publication November 28, 2011

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

Associate Editor: Martin Bishop

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

Contact: enrico.glaab@uni.lu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 25, 2011; revised on November 9, 2011; accepted on November 24, 2011

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 Smyth (2004); Tusher *et al.* (2001)] or analyze aggregated data for gene/protein sets representing entire cellular pathways and processes (Glaab *et al.*, 2010; Guo *et al.*, 2005; Lee *et al.*, 2008). 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 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.

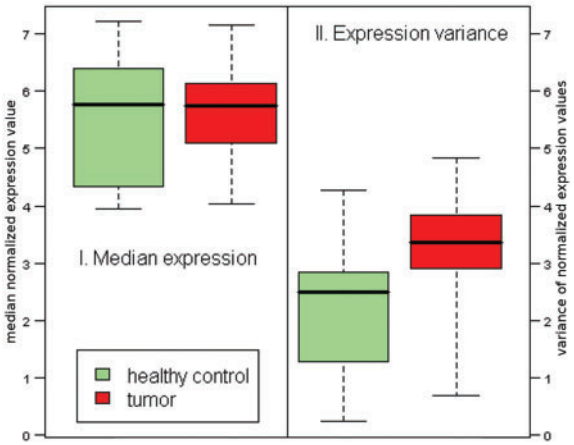

Fig. 1. Left: box plot comparing the median expression levels in the KEGG Urea cycle pathway (hsa00220) for the prostate cancer dataset by (Singh *et al.*, 2002) across 50 healthy individuals (green) and 52 tumor patients (red); right: box plot comparing the variance of expression levels in the same pathway and microarray dataset (see also Supplementary Material).

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.

(Zou and Hastie, 2008) and a recently proposed entropy score (Varshavsky *et al.*, 2006). These rankings are combined by computing the sum of ranks across the three methods and normalizing the sum-of-ranks scores by dividing by the maximum possible score. The resulting sortable ranking table of pathways contains the test statistics and significance scores, the number and identifiers of the mapped genes/proteins, and buttons to generate box plots for each pathway and forward the genes/proteins to other bioscientific web services for further analysis. Moreover, a heat-map visualization of the expression level variances is provided as output.

In the next step, the user can forward the extracted pathway variance data to a clustering module, for identifying sample groups with similar expression variance across multiple pathways, or to a classification module (for labelled data), to build models for sample classification. The clustering module provides a selection of four hierarchical clustering algorithms, three partition-based approaches and one consensus clustering approach to combine the results of the individual methods see Glaab *et al.* (2009) and Supplementary Material. In order to compare the outcome for different clustering approaches and identify a number of clusters that is optimal in terms of cluster compactness and separation between the clusters, five validity indices are computed and aggregated by computing the sum of validity score ranks across all methods and numbers of clusters. Moreover, the clustering results are visualized using both 2D plots (cluster validity score plots, principal component plots, dendrograms and silhouette plots) and interactive 3D visualizations using dimensionality reduction methods (Supplementary Material).

For a supervised analysis of the data, the classification module contains six diverse feature selection methods and six prediction algorithms, which can be combined freely by the user [see Glaab *et al.* (2009) and Supplementary Material]. To estimate the accuracy of the generated classification models, the available evaluation schemes include an external *n*-fold cross-validation as well as user-defined training/test set partitions. In addition to the average prediction accuracy and SD obtained from these evaluation methods, several other performance statistics like the sensitivity and specificity, and Cohen's Kappa statistic are computed. Additionally, a *Z*-score estimate of each gene set's utility for sample classification is determined from the frequency of its selection across different cross-validation cycles, and a heat map is generated to visualize the expression variance for the most informative gene sets. All machine learning technique implementations stem from a fully automated data analysis framework Glaab *et al.* (2009), which has previously been employed in variety of bioscientific studies (Bassel *et al.*, 2011; Glaab *et al.*, 2010; Habashy *et al.*, 2011).

To alleviate statistical limitations resulting from incomplete mappings of genes/proteins onto pathways and from multiple hypothesis testing, only pathways with a minimum of 10 mapped identifiers are considered in all analyses and *p*-values are adjusted according to Benjamini and Hochberg (1995) (see section on limitations in the Supplementary Material for details and advice).

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. 1, right; adjusted *P*-value: 2.2e-06). Interestingly, a conventional comparison of the corresponding median gene expression levels does not identify statistically significant differences between the sample groups (Fig. 1, 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 (Nishimura, 2001), Reactome (Joshi-Tope *et al.*, 2005), NCI Pathway Interaction Database (Schaefer *et al.*, 2009), WikiPathways (Pico *et al.*, 2008), InterPro (Apweiler *et al.*, 2001) and Gene Ontology [GOSlim, 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

Apweiler,R. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bassel,G.W. *et al.* (2011) A genome-wide network model capturing seed germination reveals co-ordinated regulation of plant cellular phase transitions. *Proc. Natl Acad. Sci. USA*, **108**, 9709–9714.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Glaab,E. *et al.* (2009) ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. *BMC Bioinformatics*, **10**, 358.

Glaab,E. *et al.* (2010) Learning pathway-based decision rules to classify microarray cancer samples. In Schomburg,D. and Grote,A. (eds) *German Conference on Bioinformatics 2010*, Vol. 173, Gesellschaft für Informatik, Bonn, Germany, pp. 123–134.

Guo,Z. *et al.* (2005) Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, **6**, 58.

Habashy,H.O. *et al.* (2011) RERG (Ras-related and oestrogen-regulated growth-inhibitor) expression in breast cancer: A marker of ER-positive luminal-like subtype. *Breast Cancer Res. Treat.*, **128**, 315–326.

Joshi-Tope,G. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33** (Suppl 1), D428.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480.

Lee,E. *et al.* (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, **4**, e1000217.

Nishimura,D. (2001) BioCarta. *Biotech Softwr. Internet Rep.*, **2**, 117–120.

Pico,A. *et al.* (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.

Schaefer,C. *et al.* (2009) PID: the pathway interaction database. *Nucleic Acids Res.*, **37** (Suppl 1), D674.

Singh,D. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Smyth,G. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 3.

Tusher,V. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116–5121.

Varshavsky,R. *et al.* (2006) Novel unsupervised feature filtering of biological data. *Bioinformatics,* **22**, e507.

Zou,H. and Hastie,T. (2008) *elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA*. R package version 1.0-5.