BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 20 2012, pages 2701–2702
doi:10.1093/bioinformatics/bts513

Gene expression

Advance Access publication August 24, 2012

COMPADRE: an R and web resource for pathway activity analysis by component decompositions

Roberto-Rafael Ramos-Rodriguez[1], Raquel Cuevas-Diaz-Duran[1], Francesco Falciani[2], Jose-Gerardo Tamez-Peña[1] and Victor Trevino[1,*]

[1]Cátedra of Bioinformática and Department of Computer Sciences, Tecnológico de Monterrey, Campus Monterrey, Monterrey, Nuevo León, México and [2]School of Biosciences and IBR, University of Birmingham, Edgbaston, UK

Associate Editor: Trey Ideker

ABSTRACT

Summary: The analysis of biological networks has become essential to study functional genomic data. Compadre is a tool to estimate pathway/gene sets activity indexes using sub-matrix decompositions for biological networks analyses. The Compadre pipeline also includes one of the direct uses of activity indexes to detect altered gene sets. For this, the gene expression sub-matrix of a gene set is decomposed into components, which are used to test differences between groups of samples. This procedure is performed with and without differentially expressed genes to decrease false calls. During this process, Compadre also performs an over-representation test. Compadre already implements four decomposition methods [principal component analysis (PCA), Isomaps, independent component analysis (ICA) and non-negative matrix factorization (NMF)], six statistical tests (*t*- and *f*-test, SAM, Kruskal–Wallis, Welch and Brown–Forsythe), several gene sets (KEGG, BioCarta, Reactome, GO and MsigDB) and can be easily expanded. Our simulation results shown in Supplementary Information suggest that Compadre detects more pathways than over-representation tools like David, Babelomics and Webgestalt and less false positives than PLAGE. The output is composed of results from decomposition and over-representation analyses providing a more complete biological picture. Examples provided in Supplementary Information show the utility, versatility and simplicity of Compadre for analyses of biological networks.

Availability and implementation: Compadre is freely available at http://bioinformatica.mty.itesm.mx:8080/compadre. The R package is also available at https://sourceforge.net/p/compadre.

Contact: vtrevino@itesm.mx

Supplementary information: Supplementary Data are available at *Bioinformatics* online

Received on May 8, 2012; revised on July 25, 2012; accepted on August 13, 2012

1. INTRODUCTION

The analysis of biological networks has become essential to study functional genomic data. For this, the pathway activity index, initially proposed to identify altered pathways (Tomfohr *et al.*, 2005), is a method that has been successfully applied recently. For example, it has also been used to detect pathways related to circadian rhythm (Ovacik *et al.*, 2010), to predict cancer classification (Su *et al.*, 2009) and survival (Chen *et al.*, 2010) and to identify chemical features of drug toxicity (Antczak *et al.*, 2010). Nevertheless, there is no current software that facilitates pathway activity estimations. One of the wide uses of the activity indexes is the detection of altered gene sets (Emmert-Streib and Glazko, 2011). For this, other methods that depend on a list or rank of interesting or differential expressed genes (DEG) such as GSEA, David, Babelomics and Webgestalt (Medina *et al.*, 2010; Sherman *et al.*, 2007; Subramanian *et al.*, 2005; Zhang *et al.*, 2005) attempt to identify those gene sets that are statistically over-represented or under-represented in such list. Because of the list dependency, these tools may limit the number of associations found and could miss other subtle relations (Allison *et al.*, 2006; Emmert-Streib and Glazko, 2011; Subramanian *et al.*, 2005). Therefore, approaches that consider all genes in a pathway should be theoretically more powerful (Jiang and Gentleman, 2007), and such analysis can be complementary because small, but coordinated, changes of several genes in a pathway would be also biologically interesting. PLAGE (Tomfohr *et al.*, 2005) is one of the approaches that consider all genes in a gene set using dimensionality reduction applying singular value decomposition (SVD) to detect differential expressed pathways (Jiang and Gentleman, 2007). However, there is no tool currently available to perform this type of analyses.

Since the transformation performed by SVD can be seen as decomposition by principal component analysis (PCA), we hypothesized that other linear and non-linear decompositions that have also been used to analyse gene expression may also detect coordinated differential expressed gene sets, for example, independent component analysis (ICA) (Frigyesi *et al.*, 2006), non-linear Isomaps (Dawson *et al.*, 2005) and non-negative matrix factorization (NMF) (Schachtner *et al.*, 2008). Nevertheless, such decompositions may be biased by DEG. For example, PCA designates components by the magnitude of the explained variance; consequently, top components may reflect the variance of few DEG instead of coordinated alterations of several genes. Results shown in Supplementary Material demonstrate that some gene sets that would be called significant by PLAGE and that contain non-significant numbers of DEG are no longer significant after its removal, suggesting that those significant gene set calls were false positives. Moreover, we observed that several gene sets can be detected as altered even when DEG have been removed from the analysis. Therefore, we developed Compadre (COMponent Pathway Analysis of Differential

*To whom correspondence should be addressed.

expressed genes REmoved), an R package for bioinformatics-minded users, that implements gene set decomposition to generate activity indexes. In addition, we implemented a web application for biologists to detect differential expressed gene sets.

2 IMPLEMENTATION

The main features of our implementation are as follows: (i) Four decomposition methods (PCA, ICA, Isomap and NMF); (ii) Five statistical tests (*f*-test, SAM, Kruskal–Wallis, Welch and Brown–Forsythe) to estimate DEG (Chen *et al.*, 2005) and significant gene sets; (iii) Several ready to use gene sets (KEGG, Biocarta, GO, Reactome and those from MSigDB); (iv) Users may provide its own decomposition R function for (i), *P*-values or R function for (ii), or gene sets for (iii); and (v) Results are delivered before and after removing DEG, including the activity indexes data set, easy to interpret gene set statistics and summarizing heatmap plots. Two uses of Compadre were developed, a web interface and an R package. The web uses the R package, and the results are stored in a file server at least one month, whose link is sent via e-mail when the process has finished. The Compadre algorithm involves three procedures (Fig. 1). The first procedure is ran for each gene set decomposing the corresponding gene expression sub-matrix into components representing activity indexes, generating an activity-based data set, which is used to test the difference between groups of samples. In the second procedure, DEG are detected and removed from the input data set and processed as in the first procedure. In the third procedure, the detected DEG are used to perform an over-representation test. By performing these three procedures, Compadre is allowed to recognize whether a gene set may be altered because it contains: (i) a significant number of DEG (third procedure, Evidence = Strong), (ii) several subtle altered genes (second procedure, Evidence = High) or (iii) few DEG (first procedure, Evidence = Plage, see Supplementary Material).

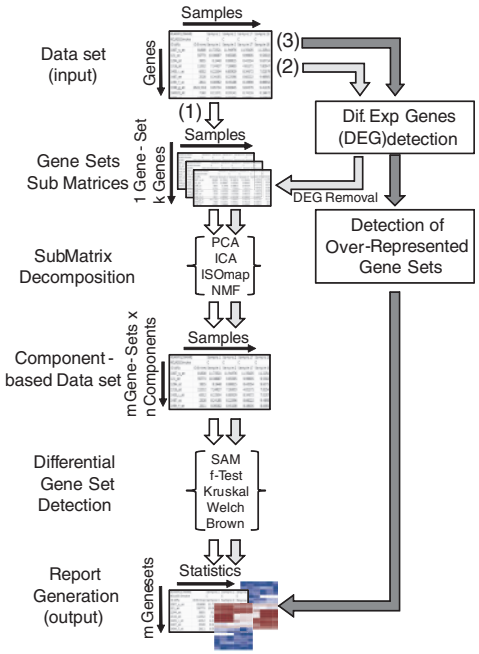

Fig. 1. Compadre involves three procedures (shaded arrows)

3 CONCLUSION

We provide an R package and web tool to estimate gene sets activity indexes and to detect altered gene sets. The simplicity of use allows easy incorporation to bioinformatics pipelines. Real data examples shown in supplementary examples for breast cancer, circadian rhythm and class prediction illustrate the versatility and usability of Compadre. Supplementary simulation results suggest that Compadre detects more pathways than over-representation tools like David, Babelomics and Webgestalt and less false positives than PLAGE. Results also imply that our novel analysis before and after DEG removal is critical to decrease gene sets false positives. In terms of true and false positives, PCA was the best; thus, it is recommended for a first approach. If results are scarce, ICA, NMF or Isomaps may be useful. We conclude that Compadre is a valuable tool to estimate activity indexes and to detect differentially expressed pathways from functional genomics data.

Funding: ITESM Cátedra de Bioinformática, Cátedra de Terapia Celular and CONACyT grant 83929.

Conflict of Interest: none declared.

REFERENCES

Allison,D.B. *et al.* (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.

Antczak,P. *et al.* (2010) Mapping drug physico-chemical features to pathway activity reveals molecular networks linked to toxicity outcome. *PLoS One*, **5**, e12385.

Chen,D. *et al.* (2005) Selecting genes by test statistics. *J. Biomed. Biotechnol.*, **2005**, 132–138.

Chen,X., Wang,L. and Ishwaran,H. (2010) An integrative pathway-based clinical-genomic model for cancer survival prediction. *Stat Probab Lett*, **80**, 1313–1319.

Dawson,K. *et al.* (2005) Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, a nonlinear algorithm. *BMC Bioinformatics*, **6**, 195.

Emmert-Streib,F. and Glazko,G.V. (2011) Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS Comput Biol*, **7**, e1002053.

Frigyesi,A. *et al.* (2006) Independent component analysis reveals new and biologically significant structures in micro array data. *BMC Bioinformatics*, **7**, 290.

Jiang,Z. and Gentleman,R. (2007) Extensions to gene set enrichment. *Bioinformatics*, **23**, 306–313.

Medina,I. *et al.* (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res.*, **38**, W210–W213.

Ovacik,M.A. *et al.* (2010) Circadian signatures in rat liver: from gene expression to pathways. *BMC Bioinformatics*, **11**, 540.

Schachtner,R. *et al.* (2008) Knowledge-based gene expression classification via matrix factorization. *Bioinformatics*, **24**, 1688–1697.

Sherman,B.T. *et al.* (2007) DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics*, **8**, 426.

Su,J., Yoon,B.J. and Dougherty,E.R. (2009) Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PLoS One*, **4**, e8161.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Tomfohr,J. *et al.* (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.

Zhang,B. *et al.* (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, **33**, W741–W748.