Vol. 29 no. 5 2013, pages 658–660
doi:10.1093/bioinformatics/btt008

Advance Access publication January 10, 2013

CePa: an R package for finding significant pathways weighted by multiple network centralities

Zuguang Gu[1,2] and Jin Wang[1,*]

[1]The State Key Laboratory of Pharmaceutical Biotechnology and Jiangsu Engineering Research Center for MicroRNA Biology and Biotechnology, School of Life Science and [2]Key Laboratory of Advanced Photonic and Electronic Materials, Department of Physics, Nanjing University, Nanjing 210093, China

Associate Editor: Trey Ideker

ABSTRACT

Summary: CePa is an R package aiming to find significant pathways through network topology information. The package has several advantages compared with current pathway enrichment tools. First, pathway node instead of single gene is taken as the basic unit when analysing networks to meet the fact that genes must be constructed into complexes to hold normal functions. Second, multiple network centralities are applied simultaneously to measure importance of nodes from different aspects to make a full view on the biological system. CePa extends standard pathway enrichment methods, which include both over-representation analysis procedure and gene-set analysis procedure. CePa has been evaluated with high performance on real-world data, and it can provide more information directly related to current biological problems.

Availability: CePa is available at the Comprehensive R Archive Network (CRAN): http://cran.r-project.org/web/packages/CePa/

Contact: jwang@nju.edu.cn

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

Received on September 18, 2012; revised on January 3, 2013; accepted on January 4, 2013

1 INTRODUCTION

Biological pathways are basic integrated circuits to actualize specific biological process in regular biological system. Identifying significant pathways in which genes are perturbed is important for researchers to focus on the most relevant sets of genes. Finding significant pathways from expression data is a fundamental task in bioinformatics analysis, which is always named pathway enrichment analysis or gene-set enrichment analysis.

Currently, there are two categories of methodologies on this subject: over-representation analysis (ORA) and gene-set analysis (GSA). The former takes a list of differentially expressed genes and tries to find which pathways are enriched in the gene list compared with background genes by a 2×2 contingency table (Khatri and Draghici, 2005). One representative tool is DAVID (Huang *et al.*, 2009). The GSA procedure uses the whole-gene expression matrix from microarray experiments, by first calculating gene-level statistics in a pathway and then integrating into a pathway-level score (Ackermann and Strimmer, 2009). ORA procedure is convenient for microarray data with few replications, and GSA procedure can generate more reliable conclusions on the analysis.

However, current standard enrichment methods are limited for revealing significant pathways because genes are treated identical in these procedures. It should be noted that pathways are represented as networks, thus importance of genes varies from different views of network structure. Currently, there are only a few works considering pathway topology. The first category is to use topological information through a global network, which is always constructed from the entire genome (Glaab *et al.*, 2010). The drawback is that such global network, e.g. constructed from public protein–protein interaction database or predicted interactions, would contain a lot of noise and is also not specific to biological conditions. The second category is to design a pathway score to cover certain aspect of genes (e.g. genes are called important when they locate in the upstream of pathways, or when they directly regulate many other genes) (Draghici *et al.*, 2007). The limitation is that because of the complexity of biological pathways, such single measurement cannot fully capture the characteristics of different genes.

Here, we proposed the R package CePa that extends standard gene-set enrichment methods (both of ORA and GSA procedure) with pathway topology information. In our previous work, we have proved ORA extension shows better performance than the standard one (Gu *et al.*, 2012). Our method has two advantages. First, we take nodes rather than genes as the basic units when analysing pathways. It is to the fact that genes must be assembled into complexes to achieve normal biological functions. Also, one single gene may locate in different nodes in real biological systems. According to our previous study on an existing pathway database (Gu *et al.*, 2012), there are ~50% genes located in multiple nodes, and 50% nodes contain more than one gene. Second, we use network centralities in the pathway as the weight of nodes and do not set a fixed measurement for the importance of genes, but we allow multiple choices tested simultaneously. This feature enables users to look at the system from multiple aspects, thus users can make more complete conclusions on current biological problems. The package is designed to be flexible so that it can implement many current gene-set enrichment methods and various centrality measurements. Finally, CePa generates informative and interactive report for the analysis and also supports parallel computing.

*To whom correspondence should be addressed.

2 APPROACH

2.1 ORA extension

We first mapped differentially expressed genes to differentially affected nodes. That is, if any member gene in a node is differential, the node is called differentially affected. Such a rule is to the fact that a complex loses its normal function as long as any component gene is abnormal. The pathway score is defined as the summation of weights of differentially affected nodes in the pathway

$$s = \sum_{i=1}^{n} w_i d_i \tag{1}$$

where s is the pathway score, w_i is the centrality of the ith node and d_i is a binary variable to identify whether the ith node is differentially affected.

The detailed evaluation of ORA extension can be found in our previous work (Gu *et al.*, 2012).

2.2 GSA extension

In standard GSA procedure, there are always gene-level statistics to combine into a pathway-level score for a single pathway. Here, we first extend gene-level statistic to node-level statistic. If a node in the pathway only contains one gene, the gene's expression vector is set as the node's expression vector. And if node in pathways comprises multiple genes, first take the largest principle component according to the member genes' expression value, and the vector for the component is set as the node's expression vector. The node-level statistic d is then calculated from this vector. For a certain pathway, the score for GSA extension is calculated as

$$s = f(\mathbf{w} \cdot \mathbf{d}) \tag{2}$$

where s is the pathway score, $\mathbf{w}$ is the centrality vector and $\mathbf{d}$ is the node-level statistic vector for the pathway nodes. The transformation function f acts on the vectorized production of $\mathbf{w}$ and $\mathbf{d}$. In Equation (2), $\mathbf{d}$ is frequently taken as t-values of the node expression vectors. CePa provides options for three node-level statistics: t-value, absolute t-value and square of t-value, in which the latter two options are recommended because they can capture the both up- and downregulation in the pathway. Nevertheless, user-defined $\mathbf{d}$ can also be applied in the package, such as signal-to-noise ratio or other robust methods. The transformation function f provided in CePa includes max, min, median, mean, sum and rank statistic. Similarly, self-defined transformation of f is also allowed in CePa. With such flexible design for the package, many current gene-set enrichment analysis methods, such as SAM-GS (Dinu *et al.*, 2007) and maxmean (Efron and Tibshirani, 2007) can be extended.

The GSA extension has been evaluated on a p53 dataset. Compared with standard GSA, CePa can find more significant pathways and reveal more details about the key regulations in the pathways (see Supplementary Data).

3 IMPLEMENTATION

3.1 Pathway database

CePa has integrated pathways from Pathway Interaction Database (PID) (Schaefer *et al.*, 2009). PID provides four pathway catalogues: NCI-Nature, BioCarta, Reactome and KEGG (see PID's website and FTP site). The pathway data are parsed from extensible mark-up language (XML) format file and can be used directly by `data(PID.db)` in the package. Each catalogue contains at least three components: a pathway list in which pathways are represented by a list of interactions; an interaction list in which each interaction is composed of an input node and an output node; and a mapping list that provides mappings from nodes to genes. We provide a parsing script available at http://mcube.nju.edu.cn/jwang/lab/soft/cepa/ for users who want to parse other versions of PID data. The detailed parsing procedure can be found in the Supplementary Material and the package vignette.

3.2 Functions

The usage for CePa is designed to be simple. All the calculation can be achieved by the function `cepa.all()`, which means applying the CePa algorithm on a list of pathway and multiple centralities. The function is wrapper of both ORA extension and GSA extension. It chooses corresponding procedure according to the arguments specified. The example codes are as follows.

```
res.ora <- cepa.all(dif = dif, bk = bk,
                    pc = PID.db$NCI)
res.gsa <- cepa.all(mat = expr, label = label,
                    pc = PID.db$NCI)
```

In the example, if the arguments contain gene lists (e.g. `dif` for the differential gene list and `bk` for the background gene list), the calculation is sent to functions doing ORA extension. Although if the arguments contain an expression matrix (`mat`) and a phenotype label (`label`), the GSA extension will be evoked.

The `plot()` function provides rich visualizations for the result. It generates different figures according to the arguments specified. First, it draws heatmap of *P*-values/FDRs of all pathways or those only significant under several centrality measurements (Fig. 1A). Second, it draws the network graph of a certain pathway (Fig. 1B). Third, it draws null distributions of node-level statistics and pathway score for a single pathway under specific centrality measurement (Fig. 1C and D).

As calculations on a list of pathways are independent in the analysis, CePa also implements the algorithm parallel through function `cepa.all.parallel()`with dependence of snow package.

Finally, CePa can generate an informative report in HTML format through `report()` function. The report includes tables and figures that fully describe the results of the analysis. Additionally, pathways can be visualized interactively by Cytoscape Web (Lopes *et al.*, 2010) in the report. An example of the report can be found at http://mcube.nju.edu.cn/jwang/lab/soft/cepa/p53test/.

3.3 Centralities

Centralities are calculated by `igraph` package. By default, the `cepa.all()` function uses six centralities: equal weight (defined as 1 for all nodes), in-degree, out-degree, betweenness, in-reach and out-reach. The latter two centralities measure whether the node locates in the upstream or downstream of

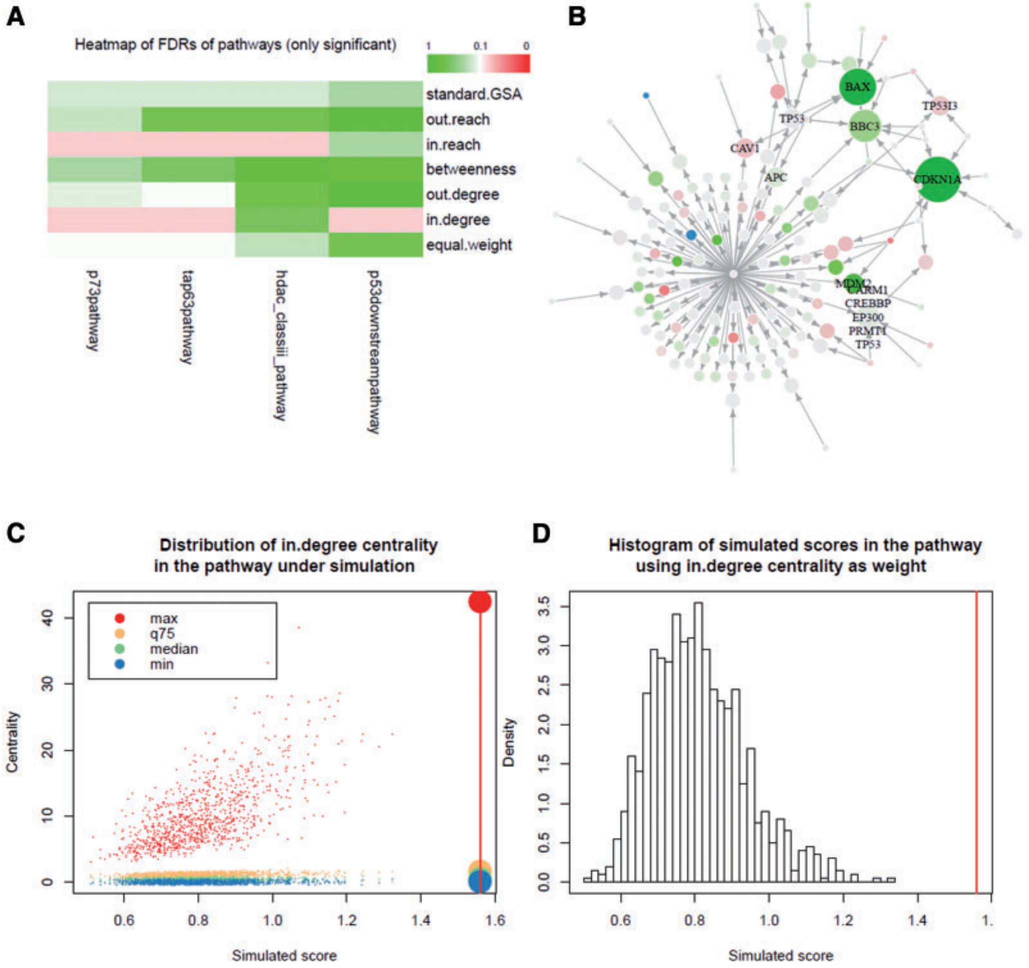

Fig. 1. Figures generated by the CePa package. (**A**) Heatmap of FDRs in significant pathways under several centrality measures; (**B**) graph view of a certain pathway in which different colours of nodes correspond to differential expression, and sizes of nodes correspond to centrality values; (**C**) distribution of node-level statistics in a pathway under simulation; and (**D**) histogram of simulated pathway scores

the pathway. Choice for the default centralities is that they have more clear biological meanings. Also, users can set their own options for centralities through a lot of other methods.

Funding: National Natural Science Foundation of China (30890044, 60990311 and 31171273); Major State Basic Research Project (2011CB301900); Nature Science Foundation of Jiangsu Province (BK2011010).

Conflict of Interest: none declared.

REFERENCES

Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.

Dinu,I. *et al.* (2007) Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, **8**, 242.

Draghici,S. *et al.* (2007) A systems biology approach for pathway level analysis. *Genome Res.*, **17**, 1537–1545.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Glaab,E. *et al.* (2010) TopoGSA: network topological gene set analysis. *Bioinformatics*, **26**, 1271–1272.

Gu,Z. *et al.* (2012) Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. *BMC Syst. Biol.*, **6**, 56.

Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Lopes,C.T. *et al.* (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.

Schaefer,C.F. *et al.* (2009) PID: the pathway interaction database. *Nucleic Acids Res.*, **37**, D674–D679.