BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 6 2012, pages 893–894
doi:10.1093/bioinformatics/bts041

Systems biology

Advance Access publication February 15, 2012

CytoSaddleSum: a functional enrichment analysis plugin for Cytoscape based on sum-of-weights scores

Aleksandar Stojmirović, Alexander Bliskovsky and Yi-Kuo Yu*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Associate Editor: John Quackenbush

ABSTRACT

Summary: *CytoSaddleSum* provides Cytoscape users with access to the functionality of *SaddleSum*, a functional enrichment tool based on sum-of-weight scores. It operates by querying *SaddleSum* locally (using the standalone version) or remotely (through an HTTP request to a web server). The functional enrichment results are shown as a term relationship network, where nodes represent terms and edges show term relationships. Furthermore, query results are written as Cytoscape attributes allowing easy saving, retrieval and integration into network-based data analysis workflows.

Availability: www.ncbi.nlm.nih.gov/CBBresearch/Yu/downloads.html The source code is placed in Public Domain.

Contact: yyu@ncbi.nlm.nih.gov

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

Received on October 24, 2011; revised on December 16, 2011; accepted on January 18, 2012

1 INTRODUCTION

CytoSaddleSum is a Cytoscape (Smoot *et al.*, 2011) plugin to access the functionality of *SaddleSum*, an enrichment analysis tool based on sum-of-weights-score (Stojmirović and Yu, 2010). Unlike most other enrichment tools, *SaddleSum* does not require users to directly select significant genes or perform extensive simulations to compute statistics. Instead, it uses weights derived from measurements, such as log expression ratios, to produce a score for each database term. It then estimates, depending on the number of genes involved, the *P*-value for that score by using the saddlepoint approximation (Lugannani and Rice, 1980) to the empirical distribution function derived from all weights. This approach was shown (Stojmirović and Yu, 2010) to yield accurate *P*-values and internally consistent retrievals.

As a popular and flexible platform for visualization, integration and analysis of network data, Cytoscape allows gene expression data import and hosts numerous plugins for functional enrichment analysis. However, none of these plugins are based on the 'gene set analysis approach' that takes into account gene weights. Therefore, to fill this gap, we have developed *CytoSaddleSum*, a Cytoscape interface to *SaddleSum*. To enable several desirable features of *CytoSaddleSum*, however, we had to significantly extend the original *SaddleSum* code (see descriptions below).

*To whom correspondence should be addressed.

2 IMPLEMENTATION

While *CytoSaddleSum* is implemented in Java using Cytoscape API, it functions by running either locally or remotely a separate instance of *SaddleSum*, written in C. In either mode, *CytoSaddleSum* takes the user input through a graphical user interface, validates it, and passes a query to *SaddleSum*. Upon receiving the entire query results, *CytoSaddleSum* stores them as the node and network attributes of the newly created term relationship graph. Consequently, the query output can be edited or manipulated within Cytoscape. Furthermore, saving term graph through Cytoscape also preserves the results for later use.

The most important extension to *SaddleSum* involved construction of extended term databases (ETDs). Each ETD contains the mappings of genes to Gene Ontology (Gene Ontology Consortium, 2010) terms and KEGG (Kanehisa *et al.*, 2008) pathways, as well as an abbreviated version of the NCBI Gene (Maglott *et al.*, 2011) database for all genes mapped to terms. Thanks to the latter, when using an ETD, *SaddleSum* is able to interpret the provided gene labels as NCBI Gene IDs, as gene symbols and as gene aliases. Each ETD also contains relations among terms that are used by *SaddleSum* for term graph construction.

3 USAGE

CytoSaddleSum operates on the currently selected Cytoscape network whose nodes represent genes or gene products. The queries are submitted through the query form embedded as a tab into the Cytoscape Control Panel, on the left of the screen. The selected network must contain at least one node mapped to a floating-point Cytoscape attribute, which would provide node weights. *CytoSaddleSum* considers only the selected nodes within the network. The user can select the weight attribute through a dropdown box on the query form. Any selected node without specified weight is assumed to have weight 0. The user-settable *cannonicalName* attribute, automatically created by Cytoscape for each network node, serves as the gene label.

After selecting the network and the nodes within it, the user needs to select a term database and set the statistical and weight processing parameters. The latter enable users to transform the supplied weights within *SaddleSum*. This includes changing the sign of the weights, as well as applying a cutoff, by weight or by rank. All weights below the cutoff are set to 0. The statistical parameters are *E*-value cutoff, minimum term size, effective database size and statistical method. We define the effective database size as the number of terms in the term database that map to at least k genes among the selected nodes, where k is the minimum term size. Apart from the default

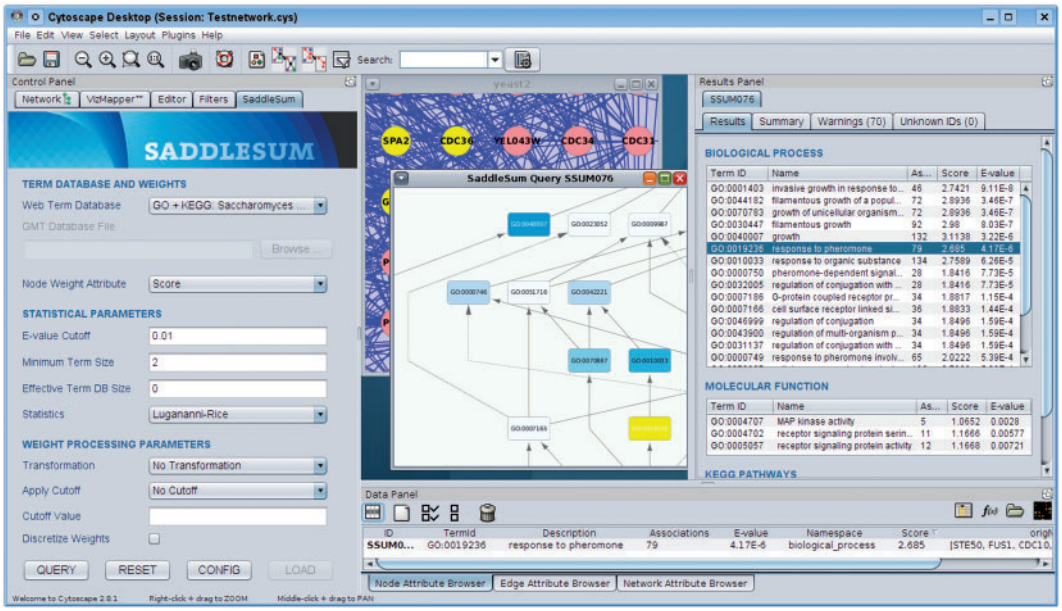

Fig. 1. *CytoSaddleSum* user interface consists of the query form (left), the results panel (right) and the term relationship network (center), which here partially covers the original network. The results stored as attributes of the term network can be edited through Cytoscape Data Panel.

'Lugannani-Rice' statistics, it is also possible to select 'One-sided Fisher's Exact test' statistics, which are based on the hypergeometric distribution. In that case, the user must select a cutoff under the weight processing parameters.

To run local queries, a user needs the command-line version of *SaddleSum* and the term databases, both available for download from our website, and install them on the same machine that runs Cytoscape. The advantages of running local queries include speed, independence of Internet connection and support of queries to custom databases in the GMT file format used by the GSEA tool (Subramanian *et al.*, 2005). Furthermore, the stand-alone program can be used outside of Cytoscape for large sets of queries. On the other hand, running remote queries require no installation of additional software, since queries are passed to the *SaddleSum* server over an HTTP connection. The disadvantage of running remote queries is that it can take much longer to run and that the choice of term databases is restricted to ETDs available only for some model organisms.

CytoSaddleSum also displays warning or error messages reported by *SaddleSum*. For example, when a provided gene label is ambiguous, depending on whether the ambiguity could be resolved, *CytoSaddleSum* will relay a warning or an error message reported by *SaddleSum*. *CytoSaddleSum* presents query results as a term relationship network (Fig. 1), consisting of significant terms or their ancestors linked by hierarchical relations available in the term database. The statistical significance of each term is indicated by the color of its corresponding node. To facilitate browsing of the results, *CytoSaddleSum* generates a set of summary tables, which contain the lists of significant terms and various details about the query. These summary tables are embedded into Cytoscape Results Panel, on the right of the screen. Clicking on a significant term in a summary table will select that term in the term relationship network and select all nodes mapping to it in the original network. The results can be exported as text or tab-delimited files and can be restored from tab-delimited files through the Export and Import menus of Cytoscape. Detailed instructions, explanations and examples can be found in *SaddleSum* manual (Supplementary Material).

Funding: Intramural Research Program of the National Library of Medicine at National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

Gene Ontology Consortium (2010) The gene ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Lugannani,R. and Rice,S. (1980) Saddle point approximation for the distribution of the sum of independent random variables. *Adv. Appl. Probab.*, **12**, 475–490.

Maglott,D. *et al.* (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.

Smoot,M.E. *et al.* (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.

Stojmirović,A. and Yu,Y.-K. (2010) Robust and accurate data enrichment statistics via distribution function of sum of weights. *Bioinformatics*, **26**, 2752–2759.

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.