Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

diffEnrich: An R Package to Compare Functional Enrichment Between Two Experimentally-derived Groups of Genes by Connecting to the KEGG REST API

Harry A. Smith
Department of Biostatistics and Informatics
Colorado School of Public Health
Skaggs School of Pharmacy
and Pharmaceutical Sciences

Laura Saba
Skaggs School of Pharmacy
and Pharmaceutical Sciences

Abstract

Motivation: To aid in the biological interpretation of a list of candidate genes and proteins generated as part of omics studies, researchers quantitate the enrichment of known pathways or biological functions among the genes of interest. With the advent of new technologies and new experimental designs, it is often of interest to compare enrichment of a particular pathway between two gene lists (i.e., differential enrichment). **Results:** This package provides a number of functions that are intended to be used in a pipeline. Briefly, a function within the package will map species-specific ENTREZ gene IDs to their respective Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways by accessing the KEGG REST API. The KEGG API is used to guarantee the most up-to-date pathway data from KEGG. Next, another function will identify significantly enriched pathways in two gene sets independently. The user can then identify pathways that are differentially enriched between the two gene sets using a third function. This package also provides a plotting function. **Availability and implementation:** diffEnrich is freely available on the Comprehensive R Archive Network (CRAN). Issues and bug reports can be submitted to the GitHub page `https://github.com/SabaLab/diffEnrich/issues`. **Supplementary information:** A step-by-step tutorial is provided on the diffEnrich GitHub page `https://github.com/SabaLab/diffEnrich`, and example data are included in the package.

Keywords: differential enrichment, KEGG REST API, R.

1. Introduction

Often high throughput omics studies include a functional enrichment analysis to glean biological insight from a list of candidate genes, proteins, metabolites, etc. Functional enrichment examines whether the number of genes in the list associated with a biological function or particular pathway is more than would be expected by chance. As an example, enrichment of a particular pathway among a list of genes that are differentially expressed after an experimental manipulation may indicate that the pathway has been altered by that manipulation. This analysis is rather straight forward and many solutions have been offered (e.g., Huang *et al.* (2009); Kuleshov *et al.* (2016); Liao *et al.* (2019); Subramanian *et al.* (2005)). A wide variety of databases have also been used to define these pathways (e.g., Kanehisa and Goto (2000)) and ontologies (e.g., Ashburner *et al.* (2000)).

One key component of a statistically rigorous functional enrichment analysis is the definition of a background data set that can be used to estimate the number of candidate genes that are "expected" to be associated with the pathway by chance, e.g., if 5% of genes in the background data set are associated with a pathway then 5% of candidate gene are expected to be associated with the pathway by chance. For many study designs, the background data set is relatively simple to define (e.g., RNA-Seq analyses where the background data set includes genes expressed above background).

However, for some newer omics technologies, the background data set is hard to define. For example, LC-MS analysis can be used to identify carbonylated proteins (Petersen *et al.* (2018); Shearn *et al.* (2019); Shearn *et al.* (2018)). With this study design, carbonylated proteins are isolated using a BH-derivation and then LC-MS is used to identify peptides in this isolated sample. The most appropriate background data set would be proteins present in that tissue, but this would require a separate analytical analysis. Furthermore, most functional enrichment analyses involve a single gene list. However, in protein modification studies, the typical experimental design compares the presence or absence of particular modified proteins between multiple groups.

When there are two or more gene lists to compare and the background gene list is not clearly defined, as is often the case in protein modification experiments, we propose a differential enrichment analysis. In this analysis, we compare the proportion of genes/proteins from one gene list associated with a particular pathway to the proportion of genes/proteins from a second gene list that are associated with that pathway. To easily execute this analysis, we have designed an R package that uses the KEGG REST API to obtain the most recent version of the KEGG PATHWAY (Kanehisa and Goto (2000)) database to initially identify functional enrichment within a gene list using the entire KEGG transcriptome as the background data set and then to identify differentially enriched pathways between two gene lists. This R package includes a function to generate a "differential enrichment" graphic.

KEGG is a database resource for understanding high-level functions of a biological system, such as a cell, an organism and an ecosystem, from genomic and molecular-level information `https://www.kegg.jp/kegg/kegg1a.html`. KEGG is an integrated database resource consisting of eighteen databases that are clustered into 4 main categories: 1) systems information (e.g. hierarchies and maps), 2) genomic information (e.g. genes and proteins), 3) chemical information (e.g. biochemical reactions), and 4) health information (e.g. human disease and drugs) `https://www.kegg.jp/kegg/kegg1a.html`.

In 2012 KEGG released its first application programming interface (API), and has been adding

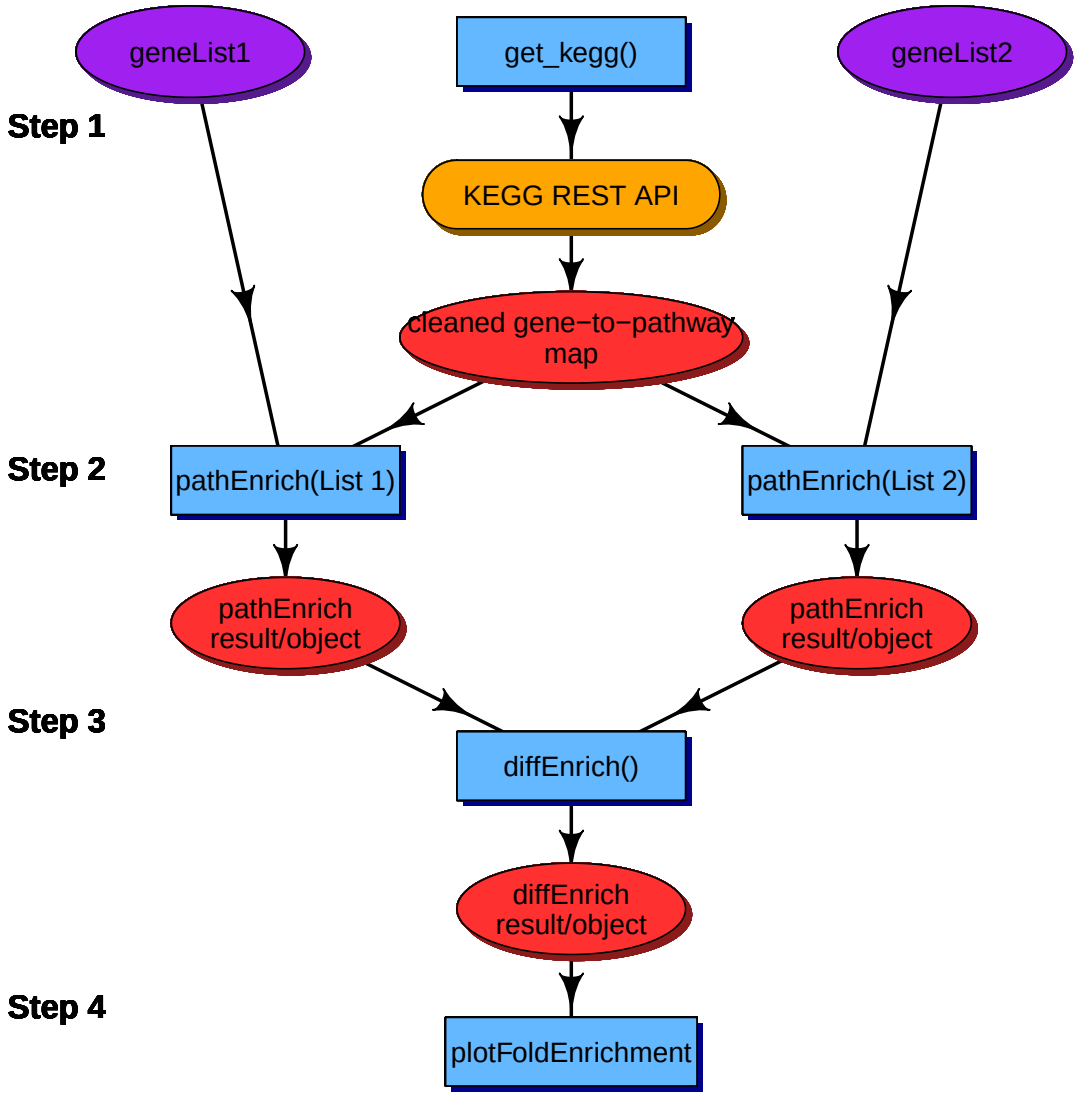

Figure 1: **diffEnrich Analysis pipeline.** Functions within the diffEnrich package are represented by blue rectangles. The data that must be provided by the user is represented by the purple ovals. Data objects generated by a function in diffEnrich are represented by red ovals. The external call of the `get_kegg` function to the KEGG REST API is represented in yellow.

features and functionality ever since. There are benefits to using an API. First, API's, like KEGG's, allow users to perform customized analyses with the most up-to-date versions of the data contained in the database. In addition, accessing the KEGG API is very easy using statistical programming tools like R or Python and integrating data retrieval into user's code makes the program reproducible. To further enforce reproducibilty diffEnrich adds a date and KEGG release tag to all data files it generates from accessing the API. For update histories and release notes for the KEGG REST API please visit https://www.kegg.jp/kegg/rest/.

2. Features

The goal of the diffEnrich package is to compare functional enrichment between two experimentally-derived groups of genes or proteins. This package provides four functions that are intended to be used in an ordered pipeline (Figure 1).

You can install the released version of diffEnrich from CRAN with:

```
install.packages("diffEnrich")
```

2.1. `get_kegg`: Download and prepare pathways from KEGG API

First, the `get_kegg` function is used to connect to the KEGG REST API and download the data sets required to perform downstream analysis. Currently, this function supports three species: Homo sapiens, Mus musculus, and Rattus norvegicus. For a given species, three data sets are generated: 1) ncbi_to_kegg: this data set maps NCBI/ENTREZ gene IDs to KEGG gene IDs, 2) kegg_to_pathway: this data set maps KEGG gene IDs to their respective KEGG pathway IDs, and 3) pathway_to_species: this data set maps KEGG pathway IDs to their respective pathway descriptions. This function allows the user to save versioned (based on KEGG release) and time-stamped text files of the three data sets described above. In addition to these flat files, `get_kegg` will also create a named list in R with the three data sets. `get_kegg` can also read in the saved text files from a past run and format them for downstream analyses. This function typically completes in a few seconds, but it is important to note that the finishing time is dependent on the time it takes to connect to the KEGG API.

References

Ashburner, *et al.* (2000). "Gene ontology: tool for the unification of biology." *Nature genetics*, **25**(1), 25–29. doi:10.1038/75556.

Huang D, *et al.* (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." *Nucleic acids research*, **37**(1), 1–13. doi:10.1093/nar/gkn923.

Kanehisa M, Goto S (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic acids research*, **28**(1), 27–30. doi:10.1093/nar/28.1.27.

Kuleshov M, *et al.* (2016). "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update." *Nucleic acids research*, **44**(1), 90–97. doi:10.1093/nar/gkw377.

Liao Y, *et al.* (2019). "Gene set analysis toolkit with revamped UIs and APIs." *Nucleic acids research*, **47**(1), 199–205. doi:10.1093/nar/gkz401.

Petersen D, *et al.* (2018). "Elevated Nrf-2 responses are insufficient to mitigate protein carbonylation in hepatospecific PTEN deletion mice." *PLoS one*, **13**(5). doi:10.1371/journal.pone.0198139.

Shearn C, *et al.* (2018). "Knockout of the Gsta4 Gene in Male Mice Leads to an Altered Pattern of Hepatic Protein Carbonylation and Enhanced Inflammation Following Chronic Consumption of an Ethanol Diet." *Alcoholism clinical and experimental research*, **42**(7), 1192–1205. doi:10.1111/acer.13766.

Shearn C, *et al.* (2019). "Cholestatic liver disease results increased production of reactive aldehydes and an atypical periportal hepatic antioxidant response." *Free radical biology and medicine*, **143**(1), 101–114. doi:10.1016/j.freeradbiomed.2019.07.036.

Subramanian T, *et al.* (2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43), 15545–15550. doi:10.1073/pnas.0506580102.

Affiliation:

Laura Saba
University of Colorado
Skaggs School of Pharmacy and Pharmaceutical Sciences
Mail Stop C238
12850 E. Montview Blvd. V20-2124
Aurora, CO 80045
E-mail: Laura.Saba@cuanschutz.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd