Instruction of IEApackage

This package includes Matlab scripts and several datasets for demo of IEA approach:

(a) IEA.m is a Matlab function for the routine of experimental analysis.

(b) IEArun.m is the main script to call IEA by supplying following parameters:

(1) dataPath: the directory locating the dataset.

(2) control: the label to indicate samples of control, and other samples are cases.

(3) data: the file name of dataset.

(4) genecard: the file name of disease genes.

(5) ppi: the file name of PPI.

(6) geneset: the file name of gene sets. E.g. KEGG pathways.

(c) _Funcs directory includes Matlab scripts for each step of IEA analysis, and called in IEA.m

(d) The input datasets include:

(1) **c2.cp.kegg.v4.0.symbols.xls**: the dataset of gene-sets like KEGG Pathways, pointed by parameter **geneset**. Its format is: each row represents a pathway; the first column is the name of pathway; the second column is the link of pathway in KEGG; the following columns list the member genes in this pathway.

(2) **String_v9.1_hs_900.csv:** the dataset of gene network like STRING, pointed by parameter **ppi**. Its format is: each row represents one interaction between two genes.

(3) 10-T2D-pan: custom directory to locate data and results, pointed by parameter **dataPath**. It is organized as: it saves the final output results (e.g. R1.txt, R2_GSP.txt, R3_1_PsC.txt and R3_2_PSC.txt); its sub-directory "data" is used to locate the input gene expression data pointed by parameter **data** (e.g. **GSE41762.xlsx**) and disease-gene data pointed by parameter **genecard** (e.g. **TD genecards.xlsx**); its another sub-directory "results" is used to save the intermediate files, which can speed up the replicate calculations. Besides, the format of gene expression data is an excel file, which have three sheets: the first sheet named data stores the raw gene expression, each column represents a sample, the first row is the labels of samples, and other rows represent genes; the second sheet named geneName stores the names of genes; the third sheet named disease stores the clinical information of samples, each column represent a sample, and each row represents a clinic, especially, the first row represents the labels of control samples or case samples (which is set by parameter **control**). Meanwhile, the format of disease-gene data is an excel file including gene list downloaded from GeneCards database.

(e) The analysis results are saved in directory pointed by parameter **dataPath** (e.g. 10-T2D-pan):

(1) R1.txt: it outputs the genes as DEGs or DEVGs.

(2) R2_GSP.txt: it outputs the enrichment as P-value for each gene set, HT1 is conventional method, HT2 is IEA method.

(3) R3_1_PsC.txt: it outputs the enrichment as P-value for each pair of pathways.

(4) R3_2_PSC.txt: it outputs the selected pathway crosstalks.

(f) As a demo, users can directly run IEArun.m in Matlab. This package has been tested in different computer environments as: Window 7 or above; Matlab 2010 or above.

(g) When users analyzed yourself new data, please:

(1) Prepare input datasets as introduced in (d).

(2) Clear the previous results in directory pointed by parameter dataPath.

(3) Clear intermediate files in sub-directory "results" of directory pointed by parameter dataPath. Notice, if you analyze a same dataset again, there is no need to this clearance, because the intermediate files can speed up the calculation on the same data.

(4) Set parameters in IEArun.m as introduced in (b).

(5) Run IEArun.m.

(h) In addition, the R function HT2 supplies the calculation of new enrichment score, whose scripts are in the directory "R". And the R package Rcpp is needed, and the working directory should be set by "setwd" in script.

Contact: Tao Zeng, zengtao@sibs.ac.cn