Package 'FunCluster'

February 19, 2015

Version 1.09

Date 2008-06-19

Title Functional Profiling of Microarray Expression Data

Author Corneliu Henegar <corneliu@henegar.info>

Maintainer Corneliu Henegar <corneliu@henegar.info>

Depends R (>= 2.10), Hmisc, cluster

Description FunCluster performs a functional analysis of microarray
expression data based on Gene Ontology & KEGG functional
annotations. From expression data and functional annotations
FunCluster builds classes of putatively co-regulated biological
processes through a specially designed clustering procedure.

License GPL (>= 2)

URL http://corneliu.henegar.info/FunCluster.htm

Repository CRAN

Date/Publication 2012-07-23 10:35:30

NeedsCompilation no

R topics documented:

1

FunCluster.R-package   *Functional Profiling of Microarray Expression Data*

Description

FunCluster performs a functional analysis of microarray expression data based on Gene Ontology & KEGG annotations. FunCluster is designed to build functional classes of putatively co-regulated biological processes through a specially designed clustering procedure relying on expression data and functional annotations.

Details

Together with the FunCluster algorithm this package provide also:
1. GO and KEGG annotations (as of June 2008) automatically extracted from their respective web resources

2. The routine for the automated extraction and update of the functional annotations from their respective web resources. The use of this routine is simple: `annotations(date.annot = "")`. Under common circumstances these routine will provide up-to-date annotations, stored into environmental variables, directly formatted for FunCluster's use. Some errors may be seen when using this routine related to a lack of availability of the GO annotations for the current month. In case of extraction errors, explained most usually by a delay in updating GO web servers, the release date can be expressly indicated (see `annotations`).

3. The two test data sets used for the JBCB paper (see examples below). The first data set is related to the dichotomous functional analysis of the genes specifically expressed within adipocytes and stroma vascular fraction (SVF) cells, extracted from adipose tissue of morbidly obese subjects (see submitted paper and cited reference for further details). Two lists of transcripts significantly expressed within adipocytes and SVF cells respectively are provided together with the list of all initial transcripts available for the analysis (necessary for the accurate computation of transcript enrichment during automated annotation of transcript expression data performed by FunCluster). The second data set is structured in a similar way and is containing the hyperinsulinemic muscle clamp expression data.

The format of the data files should be respected in order to perform a successful analysis. All the files are tab separated text files which can be easily obtained from Excel data. The only transcript identification system acceptable for FunCluster analysis is EntrezGene GeneID's. Please see more details on this choice in the JBCB paper. The transcript expression data within the tab separated text files is organized within rows, one for each transcript, and columns with the first one containing the transcript identifiers for each transcript and the rest of them containing the expression level of that transcript in each of the available microarray samples. See test data and JBCB paper for more details.

The results of the FunCluster analysis of transcript expression data are stored as tab separated text files in the "Results" subfolder of the working folder. For each type of available biological annotations and for each list of transcript expression data to be analyzed (one or two), FunCluster

provides a ranked list with the significant functional clusters observed, stored within a separate text file. Detailed findings on the terminological composition and transcript enrichment significance of the resulting functional clusters are provided. In order to correctly access results files the best approach is to use Microsoft Excel XP or later (English version) as these files were specifically formatted for Excel use. Other tabular data processing software can also be used to read these files, although accessibility will be less optimal. Some difficulties in correctly accessing results files may be observed with older versions of Microsoft Excel (prior to XP version), as well as with Excel versions in other languages than English.

Author(s)

Corneliu Henegar `<corneliu@henegar.info>`

References

1. Henegar C, Cancello R, Rome S, Vidal H, Clement K, Zucker JD. Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes. J Bioinform Comput Biol. 2006 Aug;4(4) :833-52.

2. Cancello R, Henegar C, Viguerie N, Taleb S, Poitou C, Rouault C, Coupaye M, Pelloux V, Hugol D, Bouillot JL, Bouloumie A, Barbatelli G, Cinti S, Svensson PA, Barsh GS, Zucker JD, Basdevant A, Langin D, Clement K. Reduction of macrophage infiltration and chemoattractant gene expression changes in white adipose tissue of morbidly obese subjects after surgery-induced weight loss. Diabetes 2005; 54(8):2277-86.

3. FunCluster website: <http://corneliu.henegar.info/FunCluster.htm>

See Also

cluster.

Examples

```
## Not run:
## most common use
FunCluster(go.direct = FALSE, alpha = 0.05, clusterm = "cc",
        org = "HS", location = FALSE, compare =
        "common.correl.genes", corr.th = 0.85,
        corr.met = "greedy", two.lists = TRUE,
        restrict = TRUE)

## when only GO direct annotations are to be used and detailed
findings are needed
FunCluster(go.direct = TRUE, alpha = 0.05, clusterm = "cc",
        org = "HS", location = FALSE, compare =
        "common.correl.genes", corr.th = 0.85,
        corr.met = "greedy", two.lists = TRUE,
        restrict = TRUE, details = TRUE)

## hierarchical agglomerative clustering and Silhouette computations
can be used for the preliminary step of building clusters of
co-expressed transcripts
```

```
        FunCluster(go.direct = TRUE, alpha = 0.05, clusterm = "cc",
                org = "HS", location = FALSE, compare =
                "common.correl.genes", corr.th = 0.85,
                corr.met = "hierarchical", two.lists = TRUE,
                restrict = TRUE)

        ## use only common annotated transcripts for the annotation clustering
        FunCluster(go.direct = FALSE, alpha = 0.05, clusterm = "cc",
                org = "HS", location = FALSE, compare =
                "common. genes",
                two.lists = TRUE, restrict = TRUE)

        ## the following example forces the use of a previous GO release
        (e.g. January 2006) for updating annotations
        annotations(date.annot = "200601")

  ## End(Not run)
```

---

adipose   *Gene Expression in Human White Adipose Tissue*

Description

This data set resulted from microarray experiments performed on human white adipose tissue after the separation of its two cellular components: mature adipocytes and stroma vascular fraction (SVF) cells, the non adipose component of the tissue. The purpose of this experimental model was to distinguish the two cellular fractions from a functional perspective and, in the meantime, to help establish the contribution of adipose and non adipose cells in the expression of inflammatory molecules in morbid obesity.

Usage

```
data(adipose)
```

Format

Three data frames containing significant genes specifically expressed in human mature adipocytes or in stroma vascular fraction cells of human white adipose tissue and a reference list with all the genes tested for differential expression during this experiment.

Source

http://corneliu.henegar.info/FunCluster.htm

References

Cancello R, Henegar C, Viguerie N, Taleb S, Poitou C, Rouault C, Coupaye M, Pelloux V, Hugol D, Bouillot JL, Bouloumie A, Barbatelli G, Cinti S, Svensson PA, Barsh GS, Zucker JD, Basdevant A, Langin D, Clement K. Reduction of macrophage infiltration and chemoattractant gene expression changes in white adipose tissue of morbidly obese subjects after surgery-induced weight loss. Diabetes 2005; 54(8):2277-86.

See Also

FunCluster, annotations.

Annotations *Functional Profiling of Microarray RNA Expression Data*

Description

This routine belongs to the package FunCluster and performs an automated extraction and update of the Gene Ontology & KEGG annotations which are needed for FunCluster analysis.

Usage

```
annotations(date.annot = "")
```

Arguments

date.annot   allows to specify the GO release to be used for annotations update. It has no effect on KEGG annotations.

Details

For details concerning FunCluster please see FunCluster help or man page help(FunCluster). The "Annotations" routine is allowing the automated extraction and update of the functional annotations from their respective web resources. Under common circumstances these routine will provide up-to-date annotations, stored into environmental variables and directly formatted for FunCluster use. Some errors may be seen when using this routine, related to the availability of GO annotations for the current month. In case of extraction errors, explained most usually by a delay in updating GO web servers, the date of the GO release to be used can be expressly indicated through the parameter annot.date (see example below). The transcript identification system used for FunCluster analysis is EntrezGene GeneID's. Please see more details on this choice within the JBCB paper.

Important note for Microsoft Windows users: in order to use this routine you will need additional software for handling TAR and GZIP archives. Such software is available for Windows under the GNU license.

For TAR packages you can go to:
http://gnuwin32.sourceforge.net/packages/tar.htm. Please do not forget to place the TAR

executable and its dependencies (DLL's) somewhere into the PATH (like "C:/Windows" for example). For GZIP you can go to:
http://gnuwin32.sourceforge.net/packages/gzip.htm. The same observation as for the TAR executables and dependencies applies also here.

Note

This package is related to a paper published in the Journal of Bioinformatics and Computational Biology: Henegar C, Cancello R, Rome S, Vidal H, Clement K, Zucker JD. Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes. J Bioinform Comput Biol. 2006 Aug;4(4) :833-52.

References

1. Henegar C, Cancello R, Rome S, Vidal H, Clement K, Zucker JD. Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes. J Bioinform Comput Biol. 2006 Aug;4(4) :833-52.

2. Cancello R, Henegar C, Viguerie N, Taleb S, Poitou C, Rouault C, Coupaye M, Pelloux V, Hugol D, Bouillot JL, Bouloumie A, Barbatelli G, Cinti S, Svensson PA, Barsh GS, Zucker JD, Basdevant A, Langin D, Clement K. Reduction of macrophage infiltration and chemoattractant gene expression changes in white adipose tissue of morbidly obese subjects after surgery-induced weight loss. Diabetes 2005; 54(8):2277-86.

3. FunCluster website: http://corneliu.henegar.info/FunCluster.htm

See Also

FunCluster.

Examples

```
## Not run:
## the following example forces the use of a previous GO release
(e.g. January 2006) for updating annotations;
KEGG annotations are not affected by this parameter.
annotations(date.annot = "200601")

## End(Not run)
```

FunCluster   *Functional Profiling of Microarray Expression Data*

Description

FunCluster performs a functional analysis of microarray expression data based on Gene Ontology & KEGG annotations. FunCluster is designed to build functional classes of putatively co-regulated biological processes through a specially designed clustering procedure relying on expression data and functional annotations.

Usage

```
FunCluster(wd = "", org = "HS", go.direct = FALSE, clusterm = "cc",
      compare = "common.correl.genes", corr.met = "greedy",
      corr.th = 0.85, two.lists = TRUE, restrict = FALSE,
      alpha = 0.05, location = FALSE, details = FALSE)
```

Arguments

wd	sets the working directory where the expression data files are to be found and where results are to be stored.
org	indicates the biological species to which analyzable transcript expression data is related; currently only three possibilities are available with FunCluster: "HS" for human expression data, "MM" for mouse (Mus Musculus) expression data and "SC" for yeast (Saccharomyces Cerevisiae) expression data. Default value is "HS".
two.lists	possible values are TRUE if a discriminatory functional analysis of two lists of transcripts is required (e.g. significantly up-regulated transcripts versus down-regulated transcripts) or FALSE if only one list of transcripts is to be analyzed. In the case of differential analysis of two lists of transcripts, FunCluster expects to find within the working folder two tab separated text files containing the transcript expression data named "up.txt" and "down.txt" respectively (names are mandatory). In the case of only one list of transcripts FunCluster expects to find within its working directory a single tab separated text file named "genes.txt". Please see the example dataset for the format of the data files. The default value of this parameter is TRUE.
restrict	possible values are TRUE if a reference list of transcripts is provided for the statistical significance calculation of the transcript enrichment of the biological annotations or FALSE if such a restriction is not imposed and the transcript enrichment significance is therefore estimated with regards of the whole genome. The purpose of this parameter was to correct the enrichment significance calculations for those situations in which expression data is not available for the whole genome but only for a fraction of it, either because of microarray processing errors which limits the number of transcripts available for analysis, or for the case of dedicated microarrays, designed to scan only a fraction of the genome. For the case in which such a restriction is needed a tab separated text file named "ref.txt" should be provided, containing the list of all the transcripts initially available for the analysis (after filtering for missing data). The transcripts should be identified only by their LocusLink ID number or by their EntrezGene ID number. The default value for this parameter is FALSE.
go.direct	if TRUE it restricts the transcript enrichment calculations for the GO (Gene Ontology) annotations only to directly annotated transcripts, without taking into account the ontological lattice and the subsuming relations inside Gene Ontology. Default value is FALSE, which means that, when calculating the transcript enrichment significance of a GO functional annotation, directly annotated transcripts are considered together with transcripts annotated by the directly subsumed terms within the ontological lattice.

compare	refers to the approach used for clustering highly co-expressed transcripts from available data needed in order to identify, compare and group functional annotations sharing a significant number of highly co-expressed transcripts. The default value is "common.correl.genes" which implies that a detailed analysis in search for shared co-expressed transcripts is performed (very expensive computationally and requiring enough microarray samples for correlation calculations on transcript expression profiles). If "common.genes" is selected this means that when comparing two functional annotations only the transcripts commonly annotated by the two terms are considered, without taking into account transcripts expression. If "correl.mean.exp" is selected the comparison of two functional annotations is based only on the correlations of their "mean expression profile" computed for each annotation as a vector of mean expression levels of annotated transcripts for each available microarray sample.
corr.th	it allows varying the correlation threshold used to search for and build clusters of highly co-expressed transcripts with the greedy approach. Default value based on currently available literature data is 0.85 corresponding to a Spearman correlation coefficient Rs > = 0.85.
corr.met	indicates the procedure to be used to build transcript expression clusters. It counts only if "compare" is set to "common.correl.genes". Two values are possible: "hierarchical" will use a hierarchical agglomerative procedure combined with Silhouette computing; "greedy" will use an original greedy clustering procedure conditioned by a correlation threshold specified by "corr.th" to assure homogeneity of clusters.
clusterm	is related to the algorithm used to group terms (biological annotations), having significant transcript enrichment within the analyzed data, in order to build functional classes of putatively co-regulated biological functions. Default value is "cc" and it should not be modified.
alpha	signifies the threshold of p-values significance (alpha) resulting from statistical calculations concerning transcript enrichment of biological annotations. Default value is 0.05.
location	allows to perform an analysis of the transcript enrichment of genome locations based on available genome location data (chromosome and cytoband transcript locations). If TRUE is selected it provides two lists, one containing chromosome transcript enrichment data and the other cytoband transcript enrichment data, separately for each list of analyzed transcripts. Default value is FALSE.
details	specifies if intermediary results (detailed annotation data) has to be saved.

Details

FunCluster can be used with the currently available R distributions (tested with distributions posterior to 2.0.0), either with Microsoft Windows operating environments (tested with Windows XP) or, better, with a Linux operating environment (tested with Fedora Core 3 and 4 and Suse Linux 10.0). Please be aware that FunCluster analysis implies a lot of computations and therefore high processing power and good stability of the operating system are absolute requirements.

Together with the FunCluster algorithm this package provide also:
1. GO and KEGG annotations (as of June 2008) automatically extracted from their respective web resources

2. The routine for the automated extraction and update of the functional annotations from their respective web resources. The use of this routine is simple: `annotations(date.annot = "")`. Under common circumstances these routine will provide up-to-date annotations, stored into environmental variables, directly formatted for FunCluster's use. Some errors may be seen when using this routine related to a lack of availability of the GO annotations for the current month. In case of extraction errors, explained most usually by a delay in updating GO web servers, the release date can be expressly indicated (see `annotations`).

3. The two test data sets used for the JBCB paper (see examples below). The first data set is related to the dichotomous functional analysis of the genes specifically expressed within adipocytes and stroma vascular fraction (SVF) cells, extracted from adipose tissue of morbidly obese subjects (see submitted paper and cited reference for further details). Two lists of transcripts significantly expressed within adipocytes and SVF cells respectively are provided together with the list of all initial transcripts available for the analysis (necessary for the accurate computation of transcript enrichment during automated annotation of transcript expression data performed by FunCluster). The second data set is structured in a similar way and is containing the hyperinsulinemic muscle clamp expression data.

The format of the data files should be respected in order to perform a successful analysis. All the files are tab separated text files which can be easily obtained from Excel data. The only transcript identification system acceptable for FunCluster analysis is EntrezGene GeneID's. Please see more details on this choice in the JBCB paper. The transcript expression data within the tab separated text files is organized within rows, one for each transcript, and columns with the first one containing the transcript identifiers for each transcript and the rest of them containing the expression level of that transcript in each of the available microarray samples. See test data and JBCB paper for more details.

The results of the FunCluster analysis of transcript expression data are stored as HTML files in the "Results" subfolder of the working folder. For each type of available biological annotations and for each list of transcript expression data to be analyzed (one or two), FunCluster provides a ranked list with the significant functional clusters observed, stored within a separate file. Detailed findings on the terminological composition and transcript enrichment significance of the resulting functional clusters are provided.

References

1. Henegar C, Cancello R, Rome S, Vidal H, Clement K, Zucker JD. Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes. J Bioinform Comput Biol. 2006 Aug;4(4) :833-52.

2. Cancello R, Henegar C, Viguerie N, Taleb S, Poitou C, Rouault C, Coupaye M, Pelloux V, Hugol D, Bouillot JL, Bouloumie A, Barbatelli G, Cinti S, Svensson PA, Barsh GS, Zucker JD, Basdevant A, Langin D, Clement K. Reduction of macrophage infiltration and chemoattractant gene expression changes in white adipose tissue of morbidly obese subjects after surgery-induced weight loss. Diabetes 2005; 54(8):2277-86.

3. FunCluster website: http://corneliu.henegar.info/FunCluster.htm

See Also

cluster, annotations.

Examples

```
## Not run:
## load adipose tissue data (see Diabetes and JBCB papers for details)
data(adipose)

## or load hyperinsulinemic muscle clamp data (see JBCB paper for details)
data(insulin)

## most common use
FunCluster(go.direct = FALSE, alpha = 0.05, clusterm = "cc",
        org = "HS", location = FALSE, compare =
        "common.correl.genes", corr.th = 0.85,
        corr.met = "greedy", two.lists = TRUE,
        restrict = TRUE)

## when only GO direct annotations are to be used and detailed
findings are needed
FunCluster(go.direct = TRUE, alpha = 0.05, clusterm = "cc",
        org = "HS", location = FALSE, compare =
        "common.correl.genes", corr.th = 0.85,
        corr.met = "greedy", two.lists = TRUE,
        restrict = TRUE, details = TRUE)

## hierarchical agglomerative clustering and Silhouette computations
can be used for the preliminary step of building clusters of
co-expressed transcripts
FunCluster(go.direct = TRUE, alpha = 0.05, clusterm = "cc",
        org = "HS", location = FALSE, compare =
        "common.correl.genes", corr.th = 0.85,
        corr.met = "hierarchical",
        two.lists = TRUE, restrict = TRUE)

## use only common annotated transcripts for the annotation clustering
FunCluster(go.direct = FALSE, alpha = 0.05, clusterm = "cc",
        org = "HS", location = FALSE, compare =
        "common.genes", two.lists = TRUE,
        restrict = TRUE)

## End(Not run)
```

---

insulin   *Insulin Coordinated Regulation of Gene Expression in Human Skeletal*
                            Muscle

Description

This data set resulted from a study investigating insulin coordinated regulation of gene expression in human skeletal muscle during a 3-hour hyperinsulinemic euglycemic clamp.

Usage

```
data(insulin)
```

Format

Three data frames containing significant up-regulated and down-regulated genes and a reference list with all the genes tested for differential expression during this experiment.

Source

http://corneliu.henegar.info/FunCluster.htm

References

S. Rome, K. Clement, R. Rabasa-Lhoret, E. Loizon, C. Poitou, G. S. Barsh, J. P. Riou, M. Laville and H. Vidal. Microarray profiling of human skeletal muscle reveals that insulin regulates approximately 800 genes during a hyperinsulinemic clamp, J Biol Chem 278(20), 18063-8 (2003).

See Also

FunCluster, annotations.

Index