Iterative network guided cMapping and validation

Supplementary Material and Methods - Supplementary Code: CONNECTION_SCORES

This document describes functions, scripts and data objects used in the software enclosed to the paper entitled *A semi-supervised approach for refining transcriptional signatures of drug response and repositioning predictions*, by Francesco Iorio et al, submitted as research paper to PLoS ONE.

Author: Francesco Iorio (iorio@ebi.ac.uk)

Paper website: http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission

April 30, 2014

`CS` *Connection scores to multiple ranked lists and statistical significance*

Description

This function computes connections scores of a signature generated with one among the functions

`DeriveSingleSignature`,
`DeriveConsistentSignature`,
`DeriveInconsistentSignature`,
`DeriveMSTSignature`

(all contained in `ITERATIVE_CMAPPING_library.R`) to multiple ranked lists of genes (sorted according to their differential expression, in decreasing order), by computing also statistical significance.
Empirical p-values are computed by simulating a null model through permutation of the ranked lists, by using the `est_emp_Cs` function.

Usage

```
CS(signature, RANKED_LISTS, show_progress = TRUE)
```

Arguments

`signature`	A signature of genes generated as described above
`RANKED_LISTS`	A data frame where each column contains a genome-wide ranked lists of genes or probe-sets compatible with the input signature. This data frame should have more than one column.
`show_progress`	A boolean parameter specifying whether a progress bar should be visualised or not (default = TRUE)

Value

A list of numerical vectors containing for all the columns of `RANKED_LISTS` (i.e. for each inputted ranked list):

`CS`	The obtained connection score

`Pval` The p-value of the obtained connection score

`adjP`

The p-value of the obtained connection score after correction for multiple hypothesys testing

`NCS` The normalised connection score, computed as described in [1]

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

Paper website: http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission

Refer ences

[1] Lamb,J. et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313, 1929. [2] Iorio,F. et al. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. Proceedings of the National Academy of Sciences, 107, 14621.

See Also

est_emp_Cs

Examples

```
## loading functions and data objects needed to perform iterative connectivity mapping
source('CODE/ITERATIVE_CMAPPING_library.R')

## generating the optimal signature of digoxin (a cardiac glycoside), as described in [2]
digoxinSig<-DeriveSingleSignature(seed='digoxin')

## querying the prototype ranked lists of digoxin and digoxigenin, digitoxigenin,
## and ouabain (other cardiac glycosides) with the optimal signature of digoxin
CS(digoxinSig,DRUG_PRLs[,c('digoxin','digoxigenin','digitoxigenin','ouabain')])
```

---

cMap_CS *Connection scores computation*

Description

This function computes connections scores of a genome-wide ranked lists of genes (sorted according to their differential expression, in decreasing order) and a signature composed by two sets of genes (up-regulated and down-regulated respoctively), as described in [1,2], by means of unweighted GSEA [3]

Usage

```
cMap_CS(ranked_list, opsig1, returnRS = FALSE)
```

Arguments

ranked_list	A string vector containing a genome-wide ranked list of genes sorting according to their differential expression, in decreasing order
opsig1	A list composed by two string vectors (UP and DOWN) containing the up-regulated (resp. down-regulated) genes of the signature
returnRS	A boolean parameter specifying if the individual enrichment scores (for the two parts of the signatures), together with the two corresponding obtained running sums should be returned or not (default = FALSE)

Value

The obtained connection score or (if returnRS == TRUE) a structure containing the following objects:

TES	The obtained connection score
ESUP	The enrichment score of the up regulated part of the input signature (i.e. opsig1$UP)
ESDOWN	The enrichment score of the up-regulated part of the input signature (i.e. opsig1$DOWN)
RSUP	A numerical vector with the obtained running sum for the up-regulated part of the input signature (i.e. opsig1$UP)
RDOWN	A numerical vector with the obtained running sum for the up-regulated part of the input signature (i.e. opsig1$DOWN)

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

Paper website: http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission

Refer ences

[1] Lamb,J. (2007) The Connectivity Map: a new tool for biomedical research. Nature Reviews Cancer, 7, 54-60.

[2] Lamb,J. et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313, 1929.

[3] Subramanian,A. 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, 15545.

Examples

```
## loading the prototype ranked lists for all the drugs in the connectivity map[1,2] dataset
load('DATA/DRUG_PRLs.ro')

## selecting the PRL of metformin
rankedList<-DRUG_PRLs[,'metformin']

## generating a random signature
signature<-list(UP=DRUG_PRLs[sample(1:5000,250),1],DOWN=DRUG_PRLs[sample(17000:22000,250),1])
```

```
## computing the connection score of the ranked list to the signature
cMap_CS(rankedList,signature)
```

`combine_2CS` *Combining connection score sets obtained with two different signatures*

Description

This function combines the connection score sets obtained by using two different signatures by qurying with the a set of genome-wide ranked lists of genes, through the function `CS`

Usage

```
combine_2CS(CS1, CS2, printToFile = FALSE, fn = "")
```

Arguments

`CS1`	A list of numerical vectors outputted by the `CS` function when using the first signature as input
`CS2`	A list of numerical vectors outputted by the `CS` function when using the second signature as input
`printToFile`	A boolean parameter specifying if the output of this function should be stored in a tab delimited txt file (default = FALSE). If TRUE then a file, whose name is speficied in the `fn` parameter is created in the `~/OUTPUT` directory (where ~ is the working directory)
`fn`	A string containing the file storing the results. This parameter is ignored if `printToFile = FALSE`

Details

For usage examples see the pipeline described at
http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission/iterativeCmappingPL/IterativeCmappingPipeline.html

Value

A data frame containing a row for each queryied ranked list of genes (corresponding to column names). With the following columns:

`cons S CS`	Connection scores obtained with the first signature
`cons S pvalue`	Empirical p-values of connection scores obtained with the first signature
`cons S fdr`	False discovery rate for connection scores obtained with the first signature
`incons S NCS`	Normalised connection scores obtained with the first signature
`incons S CS`	Connection scores obtained with the second signature
`incons S pvalue`	Empirical p-values of connection scores obtained with the second signature
`incons S fdr`	False discovery rate for connection scores obtained with the second signature
`incons S NCS`	Normalised connection scores obtained with the second signature
`avg NCS`	Normlised connection scores averaged across the two signatures

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

Paper website: http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission

---

`combine_3CS` *Combining connection score sets obtained with three different signatures on a user defined sub-sets of ranke lists*

Description

This function combines the connection score sets obtained by using two different signatures by qurying with the a set of genome-wide ranked lists of genes, through the function `CS`

Usage

```
combine_3CS(CS1, CS2, CS3, previousNeighBr = "", printToFile = FALSE, fn = "")
```

Arguments

`CS1`	A list of numerical vectors outputted by the `CS` function when using the first signature as input
`CS2`	A list of numerical vectors outputted by the `CS` function when using the second signature as input
`CS3`	A list of numerical vectors outputted by the `CS` function when using the third signature as input
`previousNeighBr`	A string list containing the names of the ranked lists the analysis should focus on
`printToFile`	A boolean parameter specifying if the output of this function should be stored in a tab delimited txt file (default = FALSE). If TRUE then a file, whose name is speficied in the `fn` parameter is created in the `~/OUTPUT` directory (where ~ is the working directory)
`fn`	A string containing the file storing the results. This parameter is ignored if `printToFile = FALSE`

Details

For usage examples see the pipeline described at
http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission/iterativeCmappingPL/IterativeCmappingPipeline.html

Value

A data frame containing a row for each queryied ranked list of genes (corresponding to column names). With the following columns:

`P/PI cons S NCS`
: Normalised connection scores obtained with the first signature

`P/PI incons S NCS`
: Normalised connection scores obtained with the second signature

`MST S NCS`
: Normalised connection scores obtained with the third signature

`avg NCS`
: Normlised connection scores averaged across the three signatures

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

Paper website: http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission

---

`est_emp_Cs` *Connection score null model simulation by ranked list permutation*

Description

This function estimates an empirical null distribution of connection scores for a signature of a given size and a set of genome-wide ranked lists of genes. Given the tri-modal nature of the modeled distribution [1], this function returns a 3-gaussian mixture distribution that can be used to estimate connection scores p-values

Usage

```
est_emp_Cs(signature, nt, RANKED_LISTS, show_progress = TRUE)
```

Arguments

`signature`
: A list composed by two string vectors (`UP` and `DOWN`) containing the up-regulated (resp. down-regulated) genes of the signature

`nt`
: An integer specifying the number of permutations of the ranked lists to be performed

`RANKED_LISTS`
: A data frame where each column contains a genome-wide ranked lists of genes or probe-sets compatible with the input signature. This data frame should have more than one column.

`show_progress`
: A boolean parameter specifying whether a progress bar should be visualised or not (default = TRUE)

Value

A list of class mixEM

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

Paper website: http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission

Refer ences

[1] Lamb,J. et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313, 1929.

[2] Iorio,F. et al. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. Proceedings of the National Academy of Sciences, 107, 14621.

Examples

```
## loading prototype ranked lists for the connectivity map [1] drugs
load('DATA/DRUG_PRLs.ro')

## loading functions and data objects needed to perform iterative connectivity mapping
source('CODE/ITERATIVE_CMAPPING_library.R')

## generating optimal signature for tamoxifen [2]
tamoxifenSig<-DeriveSingleSignature(seed='tamoxifen')

## converting signature format
tamoxifenSig<-list(UP=as.character(tamoxifenSig$seedUPreg$ProbeSets),
                   DOWN=as.character(tamoxifenSig$seedDOWNreg$ProbeSets))

## estimating connection scores null distribution for the tamoxifen siganture
## by executing 10000 permutation of the drug prototype ranked lists
tamoxifenNull<-est_emp_Cs(tamoxifenSig,nt=10000,DRUG_PRLs)

## visualising an histogram with the simulated connection scores
hist(tamoxifenNull$x,100)

## visualising the parameters of the 3-guassian distributions in the mixture model
summary(tamoxifenNull)
```

`getDrugName` *Drug name from internal identifiers*

Description

This function returns the name of the drug whose internal identifier is given in input

Usage

```
getDrugName(id)
```

Arguments

id	A string specifying the internal identifier of the drug under consideration

Value

A string specifying the name of the drug

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

Paper website: http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission

---

getDrugTarget *Drug target from internal identifiers*

Description

This function returns the target of the drug whose internal identifier is given in input

Usage

```
getDrugTarget(id)
```

Arguments

id	A string specifying the internal identifier of the drug under consideration

Value

A string specifying the target of the drug

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

Paper website: http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission

pnormmix *Connection scores empirical p-value computation*

Description

This fuction computes the empirical p-value of a connection score, given an empirical null distribution described as a 3-gaussian mixture model (generated by `est_emp_Cs`)

Usage

```
pnormmix(x, mixture)
```

Arguments

`x`	The connection score whose significance should be evaluated
`mixture`	A list of class mixEM generated by `est_emp_Cs` by giving in input the same signature and ranked list used to generate `x`

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

Paper website: http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission

Refer ences

[1] Lamb,J. et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313, 1929.

[2] Iorio,F. et al. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. Proceedings of the National Academy of Sciences, 107, 14621.

See Also

est_emp_Cs

Examples

```
## loading prototype ranked lists for the connectivity map [1] drugs
load('DATA/DRUG_PRLs.ro')

## loading functions and data objects needed to perform iterative connectivity mapping
source('CODE/ITERATIVE_CMAPPING_library.R')

## generating optimal signature for valproic acid (a histone deacetylase inhibitor) [2]
vaSig<-DeriveSingleSignature(seed='valproic_acid')

## converting signature format
vaSig<-list(UP=as.character(vaSig$seedUPreg$ProbeSets),
            DOWN=as.character(vaSig$seedDOWNreg$ProbeSets))
```

```
## estimating connection scores null distribution for the valproic acid siganture
## by executing 10000 permutation of the drug prototype ranked lists
vaNull<-est_emp_Cs(vaSig,nt=10000,DRUG_PRLs)

## computing the connection score of the prototype ranked list of trichostatin A
## (another histone deacetylase inhibitor) to the valproic acid optimal signature
cs<-cMap_CS(DRUG_PRLs[,'trichostatin_A'],vaSig)

## computing empirical p-value of the obtained connection score
pnormmix(cs,vaNull)
```

---

qES *Quick Enrichment Score*

Description

This function performs unweighted gene set enrichment analysis (GSEA) [1] by querying a genome-wide ranked list of genes with an input gene signature. It also visualise the obtained running sum.

Usage

```
qES(RANKEDLIST, REGULON, display = TRUE, returnRS = FALSE)
```

Arguments

RANKEDLIST	A string vector containing a genome-wide ranked list of genes sorting according to their differential expression, in decreasing order
REGULON	A signature of genes (i.e. a subset of the genes contained in RANKEDLIST)
display	A boolean parameter specifying if the obtained running sum should be visualised or not (default = TRUE)
returnRS	A boolean parameter specifying if the obtained running sum should be returned as vector of doubles (default = FALSE)

Value

The obtained enrichment score or (if returnRS == TRUE) a structure containing the following objects:

ES	The obtained enrichment score
RS	A numerical vector with same length of RANKEDLIST containing the obtained running sum
POSITION	The index position of the genes in REGULON along the list contained in RANKEDLIST
PEAK	The index position at which the running sum in RS reaches the maximal divergence from zero

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

Paper website: http://www.ebi.ac.uk/~iorio/PLoS_ONE_Submission

Refer ences

[1] Subramanian,A. 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, 15545.

[2] Garnett,M.J. et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 483, 570-575.

Examples

```
## Loading the genome wide ranked lists of the GDSC [2] cell lines,
## where the genes are sorted according to their basal expression statistics
load('DATA/GDSC_basal_ELstats_rankedLists.ro')

## select a ranked list
rankedList<-gdsc_basal_ELstats_rankedLists[,1]

## selecting a random gene signature
signature<-gdsc_basal_ELstats_rankedLists[sample(1:5000,200),1]

## computing the enrichment score and visualising the obtained running sum
qES(rankedList,signature)
```

Index