Package: bc3net (via r-universe)

August 31, 2024

Version 1.0.4

Date 2016-11-26

Title Gene Regulatory Network Inference with Bc3net

Author Ricardo de Matos Simoes [aut, cre], Frank Emmert-Streib [aut]

Maintainer Ricardo de Matos Simoes
<ricardo_dematossimoes@dfci.harvard.edu>

Depends R (>= 3.0.0), c3net, infotheo, igraph, Matrix, lattice

Description Implementation of the BC3NET algorithm for gene regulatory network inference (de Matos Simoes and Frank Emmert-Streib, Bagging Statistical Network Inference from Large-Scale Gene Expression Data, PLoS ONE 7(3): e33624, <doi:10.1371/journal.pone.0033624>).

License GPL (>= 2)

NeedsCompilation no

Date/Publication 2016-11-28 08:21:04

Repository https://bio-complexity.r-universe.dev

RemoteUrl https://github.com/cran/bc3net

RemoteRef HEAD

RemoteSha bc56bcae2bafe85f44ef2c9a46a9bff6dfe259bc

Contents

bc3net-package *BC3NET Gene Regulatory network Inference*

Description

The basic idea of BC3NET is to generate from one dataset D_s, consisting of s samples, an ensemble of B independent bootstrap datasets D_k by sampling from D(s) with replacement by using a non-parametric bootstrap (Efron 1993). Then, for each generated data set D_k in the ensemble, a network G^b_k is inferred by using C3NET (Altay 2010a). From the ensemble of networks G^b_k we construct one weighted network G^b_w which is used to determine the statistical significance of the connection between gene pairs. This results in the final binary, undirected network G.

A base component of BC3NET is the inference method C3NET introduced in Altay (2010a), which we present in the following in a modified form to obtain a more efficient implementation. Briefly, C3NET consists of three main steps. First, mutual information values among all gene pairs are estimated. Second, an extremal selection strategy is applied allowing each of the p genes in a given dataset to contribute at most one edge to the inferred network. That means we need to test only p different hypotheses and not p(p-1)/2. This potential edge corresponds to the hypothesis test that needs to be conducted for each of the p genes. Third, a multiple testing procedure is applied to control the type one error. In the above described context, this results in a network G^b_k.

Details

Package:	bc3net
Type:	Package
Version:	1.0.0
Date:	2012-01-12
License:	GPL (>=2)

bc3net.R c3mtc.R makenull.R mimwrap.R getpval.R mat2igraph.R

Author(s)

Ricardo de Matos Simoes <r.dematossimoes@qub.ac.uk> Frank Emmert-Streib <f.emmert-streib@qub.ac.uk>
Maintainer: Ricardo de Matos Simoes <r.dematossimoes@qub.ac.uk>

References

de Matos Simoes R, Emmert-Streib F., Bagging statistical network inference from large-scale gene expression data., PLoS One. 2012;7(3):e33624. Epub 2012 Mar 30.

See Also

C3NET, MINET, INFOTHEO

Examples

```
data(expmat)
bnet=bc3net(expmat)

data(expmat)
cnet=c3mtc(expmat)
```

`bc3net` *Bc3net gene regulatory network inference*

Description

The basic idea of BC3NET is to generate from one dataset D_s, consisting of s samples, an ensemble of B independent bootstrap datasets D_k by sampling from D(s) with replacement by using a non-parametric bootstrap (Efron 1993). Then, for each generated data set D_k in the ensemble, a network G^b_k is inferred by using C3NET (Altay 2010a). From the ensemble of networks G^b_k we construct one weighted network G^b_w which is used to determine the statistical significance of the connection between gene pairs. This results in the final binary, undirected network G.

A base component of BC3NET is the inference method C3NET introduced in Altay (2010a), which we present in the following in a modified form to obtain a more efficient implementation. Briefly, C3NET consists of three main steps. First, mutual information values among all gene pairs are estimated. Second, an extremal selection strategy is applied allowing each of the p genes in a given dataset to contribute at most one edge to the inferred network. That means we need to test only p different hypotheses and not p(p-1)/2. This potential edge corresponds to the hypothesis test that needs to be conducted for each of the p genes. Third, a multiple testing procedure is applied to control the type one error. In the above described context, this results in a network G^b_k.

Usage

```
bc3net(dataset, boot=100, estimator="pearson", disc="equalwidth", mtc1=TRUE,
alpha1=0.05, nullit=NA, null=c(), adj1="bonferroni", mtc2=TRUE,
alpha2=0.05, adj2="bonferroni",
weighted=TRUE, igraph=TRUE, verbose=FALSE)
```

Arguments

`dataset`	gene expression dataset where rows define genes and columns samples
`boot`	default 100 bootstrap datasets are generated to infer an ensemble of c3net gene regulatory networks
`estimator`	estimators for continuous variables "pearson", "spearman", "kendall", "spearman" estimators for discrete variables "emp", "mm","sg","shrink"
`disc`	required for discrete estimators, method for discretize function (see infotheo package) "equalwidth" (default), "equalfreq", "globalequalwidth"

`nullit` nullit defines the size of the generated null distribution vector used for hypothesis testing of significant edges inferred by c3net. The null distribution of mutual information is generated from sample and gene label randomization.

number of iterations, where the default is defined by

nullit=ceiling(10^5/(((genes*genes)/2)-genes))

genes: number of genes

`null` assign alternatively an external null distribution vector

`mtc1` consider multiple hypothesis testing for edges inferred by c3net

`alpha1` significance level for mtc1

`adj1` if mtc1==TRUE default multiple hypothesis testing procedure for c3net inferred edges using "bonferroni" (default)

alternatively use "holm", "hochberg", "hommel", "bonferroni", "BH", "BY","fdr", "none" (see ?p.adjust())

`mtc2` Consider multiple hypothesis testing for edges inferred by bc3net. A binomial test is performed for each gene pair with an ensemble consensus rate >0 consider multiple hypothesis testing for edges inferred by bc3net

`alpha2` significance level for mtc2

`adj2` Consider multiple hypothesis testing for edges inferred by bc3net. if mtc2==TRUE "bonferroni" is used as multiple hypothesis testing procedure. alternatively use "holm", "hochberg", "hommel", "bonferroni", "BH", "BY","fdr", "none"

`weighted` A weighted network is returned, where the weights denote the ensemble consensus rate of bc3net.

`igraph` A bc3net igraph object is returned.

`verbose` Return processing information of running procedures.

Details

BC3NET Gene regulatory network inference

Value

'bc3net' returns a gene regulatory network formated as adjacency matrix, as weighted matrix where the edge weights are defined by the corresponding mutual information values or as undirected weighted or unweighted igraph object.

Author(s)

de Matos Simoes R, Emmert-Streib F.

References

Altay G, Emmert-Streib F. Inferring the conservative causal core of gene regulatory networks. BMC Syst Biol. 2010 Sep 28;4:132.

de Matos Simoes R, Emmert-Streib F. Bagging statistical network inference from large-scale gene expression data. PLoS One. 2012;7(3):e33624, Epub 2012 Mar 30, <doi:10.1371/journal.pone.0033624>.

de Matos Simoes R, Emmert-Streib F. Influence of statistical estimators of mutual information and data heterogeneity on the inference of gene regulatory networks. PLoS One. 2011;6(12):e29279. Epub 2011 Dec 29.

See Also

C3NET c3mtc

Examples

```
data(expmat)
bnet=bc3net(expmat)
```

---

c3mtc *'c3mtc' gene regulatory network inference using c3net with multiple testing correction procedure*

Description

We present in the following the inference method C3NET introduced in Altay (2010a) in a modified form to obtain a more efficient implementation. Briefly, C3NET consists of three main steps. First, mutual information values among all gene pairs are estimated. Second, an extremal selection strategy is applied allowing each of the p genes in a given dataset to contribute at most one edge to the inferred network. That means we need to test only p different hypotheses and not p(p-1)/2. This potential edge corresponds to the hypothesis test that needs to be conducted for each of the p genes. Third, a multiple testing procedure is applied to control the type one error.

In order to determine the statistical significance of the mutual information values between genes we test for each pair of genes the following null hypothesis.

H_0^I: The mutual information between gene i and j is zero.

Because we are using a nonparametric test we need to obtain the corresponding null distribution for H_0^I from a randomization of the data.

The formulated null hypothesis is performed by permuting the sample and gene labels for all genes of the entire expression matrix at once. The vector of the mutual information null distribution is obtained from repeated randomizations for a given number of iterations.

Usage

```
c3mtc(dataset, null=NULL, mtc=TRUE, adj="bonferroni", alpha=0.05, nullit=NA,
estimator="pearson", disc="none", adjacency=FALSE, igraph=TRUE)
```

Arguments

`dataset` gene expression dataset where rows define genes and columns samples

`nullit` nullit defines the size of the generated null distribution vector used for hypothesis testing of significant edges inferred by c3net. The null distribution of mutual information is generated from sample and gene label randomization.

default number of iterations: nullit=ceiling(10^5/(((genes*genes)/2)-genes)) genes: number of genes

`estimator` minet package (continuous estimators) "pearson", "spearman", "kendall", "spearman"

minet package (discrete estimators) "mi.empirical", "mi.mm","mi.sg","mi.shrink"

c3net gaussian estimator (pearson) "gaussian"

bspline requires installation of "mis_calc" "bspline"

`disc` only required for discrete estimators (minet package) "equalfreq", "equalwidth"

`mtc` consider multiple hypothesis testing for edges inferred by c3net

`adj` if mtc==TRUE default multiple hypothesis testing procedure for c3net inferred edges using "bonferroni" (default)

alternatively use "holm", "hochberg", "hommel", "bonferroni", "BH", "BY","fdr", "none" (see ?p.adjust())

`alpha` significance level for mtc after multiple hypothesis testing correction

`adjacency` return an adjacency matrix

`igraph` return igraph object

`null` If NULL a null distribution vector is generated from a sample label and gene label permutation of the gene expression matrix. For the ensemble inference of one dataset an external null distribution vector is suggested for decreasing running time.

Value

'c3mtc' returns a gene regulatory network formated as adjacency matrix, as weighted matrix where the edge weights are defined by the corresponding mutual information values or as undirected weighted or unweighted igraph object.

Author(s)

de Matos Simoes R, Emmert-Streib F.

References

Altay G, Emmert-Streib F. Inferring the conservative causal core of gene regulatory networks. BMC Syst Biol. 2010 Sep 28;4:132. PubMed PMID: 20920161; PubMed Central PMCID: PMC2955605.

de Matos Simoes R, Emmert-Streib F. Bagging statistical network inference from large-scale gene expression data. PLoS One. 2012;7(3):e33624. Epub 2012 Mar 30. PubMed PMID: 22479422; PubMed Central PMCID: PMC3316596.

de Matos Simoes R, Emmert-Streib F. Influence of statistical estimators of mutual information and data heterogeneity on the inference of gene regulatory networks. PLoS One. 2011;6(12):e29279. Epub 2011 Dec 29. PubMed PMID: 22242113; PubMed Central PMCID: PMC3248437.

See Also

`c3`

Examples

```
data(expmat)
net=c3mtc(expmat)
```

enrichment *Function that performs a functional enrichment analysis based on a one-sided Fisher's exact teset (hypergeometric test).*

Description

For a given set of candidate genes, reference genes and a list object of gene sets (e.g. Gene Ontology terms or gene sets from pathways) a one-sided Fisher's exact test ("greater") is performed for each gene set in the collection.

Usage

```
enrichment(genes, reference, genesets, adj = "fdr", verbose = FALSE)
```

Arguments

`genes` A character vector of gene identifiers used as a candidate gene list that is assessed by the functional enrichment analysis. The candidate gene list is a subset of the reference gene list.

`reference` A character vector of gene identifiers used as the reference gene list. Note all candidate genes are included in the reference gene list.

`genesets` A named list object of a collection of gene sets. The identifiers used for the candidate and reference genes need to match the identifier types used for the gene sets. An example list of gene sets is given in data(exgensets) showing an example list object of gene sets from pathways with gene symbols.

$‘Reactome:REACT_115566:Cell Cycle‘ [1] "APITD1" "TAOK1" "CDC23" [4] "RB1" "PRKCA" "HIST1H4J" [7] "MCM10" "PPP1CC" "NUP153" ...

$‘Reactome:REACT_152:Cell Cycle, Mitotic‘ [1] "APITD1" "TAOK1" "CDKN2C" [4] "RB1" "PRKCA" "MCM10" [7] "HIST1H2BH" "NUP153" "TUBGCP3" [10] "APEX1" "RPA2" "PRKACA" ...

`adj` The default value is <fdr> (False discovery rate using the Benjamini-Hochberg approach). Multiple testing correction based on the function stats::p.adjust() with available options for "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr" and "none"

`verbose` The default value is <FALSE>. If this option is set <TRUE> the number and name of the gene sets during their processing is reported.

Details

The enrichment analysis is based on a one-sided Fisher's exact test.

Value

The function returns a data.frame object with the columns

"TermID" the given name of a gene set i from the named gene set collection list object S. "genes" the number of candidate genes present in the gene set i "all" the number of all genes present in the gene set i "pval" the nominal p-value from a one-sided fisher's exact test "padj" the adjusted p-value to consider for multiple testing

Author(s)

Ricardo de Matos Simoes

References

de Matos Simoes R, Tripathi S, Emmert-Streib F. Organizational structure and the periphery of the gene regulatory network in B-cell lymphoma. BMC Syst Biol. 2012; 6:38.

Inference and Analysis of Gene Regulatory Networks in R: Applications in Biology, Medicine, and Chemistry, DOI: 10.1002/9783527694365.ch10 In book: Computational Network Analysis with R, 2016, pp.289-306

Examples

```
# In the following candidate genes are defined from a
# giant connected component of an example igraph network
# where we use the remaining genes of a given network
# as a reference list.

data(exanet)
data(exgensets)

candidate=V(getgcc(exanet))$name
reference=V(exanet)$name

# hypergeometric test is performed for
# each defined set of genes in the
# list object exgensets
tab.hypg=enrichment(candidate, reference, exgensets, verbose=TRUE)
```

exanet *Example gene regulatory network for testing purposes only*

Description

The network consists of 814 genes and 1,103 interactions.

Usage

```
data(exanet)
```

Format

An undirected weighted igraph object with 814 vertices and 1,103 egdges.

References

de Matos Simoes R, Emmert-Streib F., "Bagging statistical network inference from large-scale gene expression data" PLoS One. 2012;7(3):e33624. Epub 2012 Mar 30. PubMed PMID: 22479422; PubMed Central PMCID: PMC3316596.

Examples

```
data(expmat)
```

exgensets *Example gene sets defined from the CPDB database*

Description

This list data object contains 25 gene sets defined from the CPDB database and is used for testing purposes.

Usage

```
data(exgensets)
```

Format

A list object with 25 character vectors.

Details

An list object of 25 pathway gene sets from CPDB with gene symbols.

Source

http://consensuspathdb.org/

References

A. Kamburov, U. Stelzl, H. Lehrach and R. Herwig, The ConsensusPathDB interaction database: 2013 update, Nucleic Acids Research, Volume 41, Issue D1, Pp. D793-D800

Examples

```
data(exgensets)

# The gene set collection from CPDB can be formatted by:
# "CPDB_pathways_genes.tab" is available in the download section in http://consensuspathdb.org/
# cpdb = readLines("CPDB_pathways_genes.tab", warn = FALSE)
# cpdb = lapply(cpdb, function(x) strsplit(x, "\t")[[1]])
# names(cpdb) = sapply(cpdb, function(x) paste(x[3], x[2], x[1], sep=":") )
# cpdb = lapply(cpdb, function(x) x[-c(1:3)])
# cpdb = cpdb[-length(cpdb)]
# cpdb = lapply(cpdb, function(x) strsplit(x, ",")[[1]])
```

expmat *Test gene expression dataset*

Description

The dataset is test dataset of a gene expression matrix with 100 genes and 100 samples

Usage

```
data(expmat)
```

Format

A matrix with 100 observations and 100 variables.

References

de Matos Simoes R, Emmert-Streib F., "Bagging statistical network inference from large-scale gene expression data" PLoS One. 2012;7(3):e33624. Epub 2012 Mar 30. PubMed PMID: 22479422; PubMed Central PMCID: PMC3316596.

Examples

```
data(expmat)
```

getgcc *Extracting the giant connected component from an igraph object*

Description

The function uses igraph::clusters and igraph::induced.subgraph to extract the giant component of an igraph object.

Usage

```
getgcc(net)
```

Arguments

`net`	An igraph object.

Details

In a connected component of an undirected graph all pairs of vertices u and v are reachable by a path. The giant connected component describes the largest connected component of a graph object.

Value

`getgcc` returns an igraph object with the giant connected component of an igraph object.

References

G. Csardi and T. Nepusz, The igraph software package for complex network research, InterJournal, Complex Systems,1695, 2006, http://igraph.org

Examples

```
# The example network "exanet" contains 814 genes with 1103 edges
data(exanet)

# The giant connected component of the example
# network contains 382 genes with 989 edges
getgcc(exanet)
```

gpea *Gene pair enrichment analysis (GPEA)*

Description

When a network G contains n interactions, of which k interactions are among genes from the given gene set S, then a p-value for the enrichment of gene pairs of this gene set S can be calculated based on a e.g., one-sided Fisher's exact test. For p genes there is a total of N=p(p-1)/2 different gene pairs (clique graph) with the assumption that all genes within a gene set are associated to each other. If there are pS genes for a particular gene set (S) then the total number of gene pairs for this gene set is mS=pS(pS-1)/2.

Usage

```
gpea(gnet, genesets, verbose = TRUE, cmax = 1000, cmin = 3,
                 adj = "bonferroni")
```

Arguments

`gnet` igraph object (e.g., inferred from bc3net) of a given network where the gene identifiers [V(net)$names] correspond to the provided gene identifiers in the reference gene sets.

`genesets` A named list object of a collection of gene sets. The identifiers used for the candidate and reference genes need to match the identifier types used for the gene sets. An example list of gene sets is given in data(exgensets) showing an example list object of gene sets from pathways with gene symbols.

$'Reactome:REACT_115566:Cell Cycle' [1] "APITD1" "TAOK1" "CDC23" [4] "RB1" "PRKCA" "HIST1H4J" [7] "MCM10" "PPP1CC" "NUP153" ...

$'Reactome:REACT_152:Cell Cycle, Mitotic' [1] "APITD1" "TAOK1" "CDKN2C" [4] "RB1" "PRKCA" "MCM10" [7] "HIST1H2BH" "NUP153" "TUBGCP3" [10] "APEX1" "RPA2" "PRKACA" ...

`verbose` The default value is <FALSE>. If this option is set <TRUE> the number and name of the gene sets during their processing is reported.

`cmax` All provided genesets with more than cmax genes will be excluded from the analysis (default cmax=1000).

`cmin` All provided genesets with less than cmin genes will be excluded from the analysis (default cmin>=3).

`adj` The default value is <fdr> (False discovery rate using the Benjamini-Hochberg approach). Multiple testing correction based on the function stats::p.adjust() with available options for "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr" and "none"

Details

The enrichment analysis is based on a one-sided Fisher's exact test.

Value

The function returns a data.frame object with the columns

"TermID" the given name of a gene set i from the named gene set collection list object S. "edges" the number of connected gene pairs present a given geneset "genes" the number of candidate genes present in the gene set i "all" the number of all genes present in the gene set i "pval" the nominal p-value from a one-sided fisher's exact test "padj" the adjusted p-value to consider for multiple testing

Author(s)

Ricardo de Matos Simoes

References

Inference and Analysis of Gene Regulatory Networks in R: Applications in Biology, Medicine, and Chemistry, DOI: 10.1002/9783527694365.ch10 In book: Computational Network Analysis with R, 2016, pp.289-306

Urothelial cancer gene regulatory networks inferred from large-scale RNAseq, Bead and Oligo gene expression data, BMC Syst Biol. 2015; 9: 21.

See Also

See Also as `enrichment`

Examples

```
data(exanet)
data(exgensets) ## example gene sets from the CPDB database (http://www.consensuspathdb.org)

res = gpea(exanet, exgensets, cmax=1000, cmin=2)
```

`mimwrap` *Wrapper function for mutual information matrix estimators*

Description

Mutual information matrix estimation wrapper function for various mutual information estimators. Depends on infotheo package for mutual information estimators on discrete variables.

Usage

```
mimwrap(dataset, estimator="pearson", disc="equalwidth", bins = sqrt(ncol(dataset)))
```

Arguments

- `dataset` Data gene expression matrix where rows denote genes (features) and columns samples.
- `estimator` estimators for continuous variables "pearson" (default), "spearman", "kendall", "spearman"

  estimators for discrete variables (infotheo package) "emp", "mm","sg","shrink"
- `disc` only required for discrete estimators (see infotheo package) "equalwidth" (default), "globalequalwidth" , "equalfreq"
- `bins` number of bins for the descretize function (infotheo), default sqrt(ncol(dataset))

Details

A mutual information matrix is estimated from a gene expression data set

Value

`mimwrap` returns a symmetric mutual information matrix for various mutual information estimators.

References

Patrick E Meyer, Frederic Lafitte and Gianluca Bontempi, minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information, BMC Bioinformatics 2008, 9:461

Carsten O. Daub, Ralf Steuer, Joachim Selbig, and Sebastian Kloska, Estimating mutual information using B-spline functions - an improved similarity measure for analysing gene expression data, BMC Bioinformatics. 2004; 5: 118

de Matos Simoes R, Emmert-Streib F., Bagging statistical network inference from large-scale gene expression data., PLoS One. 2012;7(3):e33624. Epub 2012 Mar 30.

Examples

```
data(expmat)
mim <- mimwrap(expmat)
```

Index