Package 'CoGAPS'

September 24, 2012

Version 1.6.0

Date 2011-09-02

Title Coordinated Gene Activity in Pattern Sets

Author Elana J. Fertig

Description Coordinated Gene Activity in Pattern Sets (CoGAPS) infers biological processes which are active in individual gene sets from corresponding microarray measurements. CoGAPS achieves this inference by combining a MCMC matrix decomposition algorithm (GAPS) with a novel statistic inferring activity on gene sets.

Maintainer Elana J. Fertig <ejfertig@jhmi.edu>, Michael F. Ochs <mfo@jhu.edu>

SystemRequirements GAPS-JAGS (==1.0.2)

Depends R (>= 2.9.0), R.utils (>= 1.2.4)

Imports graphics, grDevices, methods, stats, utils

License GPL (== 2)

URL http://www.cancerbiostats.onc.jhmi.edu/CoGAPS.cfm

biocViews GeneExpression, Microarray, Bioinformatics

R topics documented:

AGS — *Simulated amplitude matrix with gene set activity.*

Description

Simulated amplitude matrix specifying activity in two gene sets (gs).

Usage

```
AGS
```

Format

Matrix of 30 rows by 3 columns with simulated amplitude matrix.

calcCoGAPSStat — *CoGAPS gene set statistic*

Description

Computes the p-value for the association of underlying patterns from microarray data to activity in gene sets.

Usage

```
calcCoGAPSStat(Amean, Asd, GStoGenes, numPerm=500)
```

Arguments

Amean	Sampled mean value of the amplitude matrix $\mathbf{A}$. `row.names(Amean)` must correspond to the gene names contained in GStoGenes.
Asd	Sampled standard deviation of the amplitude matrix $\mathbf{A}$.
GStoGenes	List or data frame containing the genes in each gene set. If a list, gene set names are the list names and corresponding elements are the names of genes contained in each set. If a data frame, gene set names are in the first column and corresponding gene names are listed in rows beneath each gene set name.
numPerm	Number of permuations used for the null distribution in the gene set statistic. (optional; default=500)

Details

This script links the patterns identified in the columns of $\mathbf{P}$ to activity in each of the gene sets specified in GStoGenes using a novel z-score based statistic developed in Ochs et al. (2009). Specifically, the z-score for pattern p and gene set G_i containing \$G\$ total genes is given by

$$Z_{i,p} = \frac{1}{G} \sum_{g in \mathcal{G}_i} \frac{\mathbf{A_{gp}}}{\sigma_{gp}},$$

where g indexes the genes in the set and σ_{gp} is the standard deviation of $\mathbf{A}_{gp}$ obtained from MCMC sampling. CoGAPS then uses the specified `numPerm` random sample tests to compute a consistent p value estimate from that z score.

Value

A list containing:

GSUpreg	p-values for upregulation of each gene set in each pattern.
GSDownreg	p-values for downregulation of each gene set in each pattern.
GSActEst	p-values for activity of each gene set in each pattern.

Author(s)

Elana J. Fertig <ejfertig@jhmi.edu>

References

M.F. Ochs, L. Rink, C. Tarn, S. Mburu, T. Taguchi, B. Eisenberg, and A.K. Godwin. (2009) Detection and treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. Cancer Research, 69:9125-9132.

See Also

CoGAPS, GAPS

CoGAPS *CoGAPS driver script*

Description

Runs the CoGAPS algorithm to infer underlying patterns in microarray data and their association to activity in gene sets.

Usage

```
CoGAPS(data, unc, GStoGenes, outputDir, outputBase="", sep="\t",
       isPercentError=FALSE, numPatterns, MaxAtomsA=2^32, alphaA=0.01,
       MaxAtomsP=2^32, alphaP=0.01, SAIter=1000000000, iter = 500000000,
thin=-1, nPerm=500, verbose=TRUE, plot=FALSE, keepChain=FALSE)
```

Arguments

data	The matrix of m genes by n arrays of expression data. The input can be either the data matrix itself or the file containing this data. If the latter, CoGAPS will read in the data using `read.table(data, sep=sep, header=T, row.names=1)`.
unc	The matrix of m genes by n arrays of uncertainty (standard deviation) for the expression data. The input can be either a file containing the uncertainty (using the format from data), a matrix containing the uncertainty, or a constant value. If unc is a constant value, it can represent either a constant uncertainty or a constant percentage of the values in data as determined by isPercentError.
GStoGenes	List or data frame containing the genes in each gene set. If a list, gene set names are the list names and corresponding elements are the names of genes contained in each set. If a data frame, gene set names are in the first column and corresponding gene names are listed in rows beneath each gene set name.

numPatterns	Number of patterns into which the data will be decomposed. Must be less than the number of genes and number of arrays in the data.
outputDir	Directory to which to output result and diagnostic files created by CoGAPS. (Use "" to output results to the current directory).
outputBase	Prefix for all result and diagnostic files created by CoGAPS (optional; default="")
sep	Text delimiter for tables in data and unc (if specified in file) and any output tables (optional; default="\t")
isPercentError	Boolean indicating whether constant value in unc is the value of the uncertainty or the percentage of the data that is the uncertainty.
MaxAtomsA	Maximum number of atoms in the atomic domain used for the prior of the amplitude matrix in the decomposition (see Sibisi and Skilling, 1997). The default value will typically be sufficient for most applications (optional; default=2^32).
alphaA	Sparsity parameter reflecting the expected number of atoms per element of the amplitude matrix in the decomposition. To enforce sparsity, this parameter should typically be less than one. (optional; default=0.01)
MaxAtomsP	Maximum number of atoms in the atomic domain used for the prior of the pattern matrix in the decomposition (see Sibisi and Skilling, 1997). The default value will typically be sufficient for most applications (optional; default=2^32).
alphaP	Sparsity parameter reflecting the expected number of atoms per element of the pattern matrix in the decomposition. To enforce sparsity, this parameter should typically be less than one. (optional; default=0.01)
SAIter	Number of burn-in iterations for the MCMC matrix decomposition (optional; default=1000000000)
iter	Number of iterations to represent the distribution of amplitude and pattern matrices with the MCMC matrix decomposition (optional; default=500000000)
thin	Double whose integer part represents the number of iterations at which the samples are kept and decimal part provides an identifier for the output files from this implementation of CoGAPS. If thin is an integer or not specified, this decimal file identifier is assigned randomly. (optional; default=-1; code assigns number of iterations kept to be iter/1000 and file identifier to be runif(1))
nPerm	Number of permuations used for the null distribution in the gene set statistic. (optional; default=500)
verbose	Boolean which specifies the amount of output to the user about the progress of the program. (optional; default=TRUE)
plot	Boolean which specifies whether plots representing the resulting amplitude and pattern matrices should be made. (optional; default=FALSE)
keepChain	Boolean which specifies if chain values of $\mathbf{A}$ and $\mathbf{P}$ are saved in outputDir (optional; default=FALSE).

Details

CoGAPS first decomposes the data matrix using GAPS, $\mathbf{D}$, into a basis of underlying patterns and then determines the gene set activity in each of these patterns.

The GAPS decomposition is achieved by finding amplitude and pattern matrices ($\mathbf{A}$ and $\mathbf{P}$, respectively) for which

$$\mathbf{D} = \mathbf{AP} + \Sigma,$$

where Σ is the matrix of uncertainties given by unc. The matrices $\mathbf{A}$ and $\mathbf{P}$ are assumed to have the atomic prior described in Sibisi and Skilling (1997) and are found with MCMC sampling implemented within JAGS.

Then, the patterns identified in the columns of $\mathbf{P}$ are linked to activity in each of the gene sets specified in GStoGenes using a novel z-score based statistic developed in Ochs et al. (2009). Specifically, the z-score for pattern p and gene set G_i containing G total genes is given by

$$Z_{i,p} = \frac{1}{G} \sum_{g in \mathcal{G}_i} \frac{\mathbf{A_{gp}}}{Asd_{gp}},$$

where g indexes the genes in the set and Asd_{gp} is the standard deviation of $\mathbf{A}_{gp}$ obtained from MCMC sampling. CoGAPS then uses the specified `nPerm` random sample tests to compute a consistent p value estimate from that z score. Note that the data from Ochs et al. (2009) are provided with this package in GIST_TS_20084.RData and TFGSList.RData are also provided with this package for further validation with nIter=5e+07.

Value

A list containing:

`D`	Microarray data matrix.
`Sigma`	Data matrix with uncertainty of D.
`Amean`	Sampled mean value of the amplitude matrix $\mathbf{A}$.
`Asd`	Sampled standard deviation of the amplitude matrix $\mathbf{A}$.
`Pmean`	Sampled mean value of the pattern matrix $\mathbf{P}$.
`Psd`	Sampled standard deviation of the pattern matrix $\mathbf{P}$.
`meanMock`	Mock data obtained from matrix decomposition for sampled mean values (= Amean %*% Pmean).
`meanChi2`	χ^2 value for the sampled mean values (Amean and Pmean) of the matrix decomposition.
`GSUpreg`	p-values for upregulation of each gene set in each pattern.
`GSDownreg`	p-values for downregulation of each gene set in each pattern.
`GSActEst`	p-values for activity of each gene set in each pattern.

Note

Running GAPS will create the folder ouptutDir, create diagnostic files with χ^2 and number of atoms, files with the mean and standard deviation of $\mathbf{A}$ and $\mathbf{P}$, files with p-values for upregulation/downregulation/activity of each gene set, and optionally values of $\mathbf{A}$ and $\mathbf{P}$ from the MCMC chain.

Author(s)

Elana J. Fertig <ejfertig@jhmi.edu>

References

M.F. Ochs, L. Rink, C. Tarn, S. Mburu, T. Taguchi, B. Eisenberg, and A.K. Godwin. (2009) Detection and treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. Cancer Research, 69:9125-9132.

M. Plummer. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, and A. Zeileis, editors, Proceedings of the Third International Workshop on Distributed Statistical Computing, Vienna, Austria.

S. Sibisi and J. Skilling. (1997) Prior distributions on measure space. Journal of the Royal Statistical Society, B, 59:217-235.

See Also

GAPS, calcCoGAPSStat

Examples

```
## Not run:
## Load data
data(EasySimGS)

## Run the CoGAPS matrix decomposition
nIter <- 5e+05
results <- CoGAPS(data=DGS, unc=0.01, isPercentError=FALSE,
                  GStoGenes=gs,
                  numPatterns=3,
                  SAIter = 2*nIter, iter = nIter,
                  outputDir='GSResults', plot=TRUE)

## End(Not run)
```

DGS *Simulated gene expresssion data.*

Description

Gene expression data simulated from 3 known true patterns (PGS) with activity in two gene sets (gs) specified in the simulated amplitude (AGS).

Usage

```
DGS
```

Format

Matrix of 30 rows by 25 columns of simulated expression measurements.

GAPS *GAPS matrix decomposition script*

Description

Decomposes microarray data into underlying patterns and corresponding amplitude.

Usage

```
GAPS(data, unc, outputDir, outputBase="", sep="\t", isPercentError=FALSE,
     numPatterns, MaxAtomsA=2^32, alphaA=0.01, MaxAtomsP=2^32, alphaP=0.01,
     SAIter=1000000000, iter = 500000000, thin=-1, verbose=TRUE,
     keepChain=FALSE)
```

Arguments

data	The matrix of m genes by n arrays of expression data. The input can be either the data matrix itself or the file containing this data. If the latter, GAPS will read in the data using `read.table(data, sep=sep, header=T, row.names=1)`.
unc	The matrix of m genes by n arrays of uncertainty (standard deviation) for the expression data. The input can be either a file containing the uncertainty (using the format from data), a matrix containing the uncertainty, or a constant value. If unc is a constant value, it can represent either a constant uncertainty or a constant percentage of the values in data as determined by isPercentError.
numPatterns	Number of patterns into which the data will be decomposed. Must be less than the number of genes and number of arrays in the data.
outputDir	Directory to which to output result and diagnostic files created by GAPS. (Use "" to output results to the current directory).
outputBase	Prefix for all result and diagnostic files created by GAPS (optional; default="")
sep	Text delimiter for tables in data and unc (if specified in file) and any output tables (optional; default="\t")
isPercentError	Boolean indicating whether constant value in unc is the value of the uncertainty or the percentage of the data that is the uncertainty.
MaxAtomsA	Maximum number of atoms in the atomic domain used for the prior of the amplitude matrix in the decomposition (see Sibisi and Skilling, 1997). The default value will typically be sufficient for most applications (optional; default=2^32).
alphaA	Sparsity parameter reflecting the expected number of atoms per element of the amplitude matrix in the decomposition. To enforce sparsity, this parameter should typically be less than one. (optional; default=0.01)
MaxAtomsP	Maximum number of atoms in the atomic domain used for the prior of the pattern matrix in the decomposition (see Sibisi and Skilling, 1997). The default value will typically be sufficient for most applications (optional; default=2^32).
alphaP	Sparsity parameter reflecting the expected number of atoms per element of the pattern matrix in the decomposition. To enforce sparsity, this parameter should typically be less than one. (optional; default=0.01)
SAIter	Number of burn-in iterations for the MCMC matrix decomposition (optional; default=1000000000)

iter	Number of iterations to represent the distribution of amplitude and pattern matrices with the MCMC matrix decomposition (optional; default=500000000)
thin	Double whose integer part represents the number of iterations at which the samples are kept and decimal part provides an identifier for the output files from this implementation of GAPS. If thin is an integer or not specified, this decimal file identifier is assigned randomly. (optional; default=-1; code assigns number of iterations kept to be iter/10000 and file identifier to be runif(1))
verbose	Boolean which specifies the amount of output to the user about the progress of the program. (optional; default=TRUE)
keepChain	Boolean which specifies if chain values of $\mathbf{A}$ and $\mathbf{P}$ are saved in outputDir (optional; default=FALSE).

Details

The decomposition in GAPS is achieved by finding amplitude and pattern matrices ($\mathbf{A}$ and $\mathbf{P}$, respectively) for which

$$\mathbf{D} = \mathbf{AP} + \Sigma$$

, where Σ is the matrix of uncertainties given by unc. The matrices $\mathbf{A}$ and $\mathbf{P}$ are assumed to have the atomic prior described in Sibisi and Skilling (1997) and are found with MCMC sampling implemented within JAGS.

Value

A list containing:

D	Microarray data matrix.
Sigma	Data matrix with uncertainty of D.
Amean	Sampled mean value of the amplitude matrix $\mathbf{A}$.
Asd	Sampled standard deviation of the amplitude matrix $\mathbf{A}$.
Pmean	Sampled mean value of the pattern matrix $\mathbf{P}$.
Psd	Sampled standard deviation of the pattern matrix $\mathbf{P}$.
meanMock	Mock data obtained from matrix decomposition for sampled mean values (= Amean %*% Pmean).
meanChi2	χ^2 value for the sampled mean values (Amean and Pmean) of the matrix decomposition.

Note

Running GAPS will create the folder ouptutDir, create diagnostic files with χ^2 and number of atoms, files with the mean and standard deviation of $\mathbf{A}$ and $\mathbf{P}$, and optionally values of $\mathbf{A}$ and $\mathbf{P}$ from the MCMC chain.

Author(s)

Elana J. Fertig <ejfertig@jhmi.edu>

References

M. Plummer. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, and A. Zeileis, editors, Proceedings of the Third International Workshop on Distributed Statistical Computing, Vienna, Austria.

S. Sibisi and J. Skilling. (1997) Prior distributions on measure space. Journal of the Royal Statistical Society, B, 59:217-235.

See Also

CoGAPS

Examples

```
## Not run:
## Load data
data(ModSim)

## Run GAPS matrix decomposition
nIter <- 500000
results <- GAPS(data=ModSim.D, unc=0.01, isPercentError=FALSE,
                numPatterns=3, SAIter=2*nIter, iter = nIter,
                outputDir='ModSimResults')

## Plot the results
plotGAPS(results$Amean, results$Pmean)

## End(Not run)
```

GIST.D *Sample GIST gene expression data from Ochs et al. (2009).*

Description

Gene expression data from gastrointestinal stromal tumor cell lines treated with Gleevec.

Usage

```
GIST_TS_20084
```

Format

Matrix with 1363 genes by 9 samples of mean gene expression data.

References

Ochs, M., Rink, L., Tarn, C., Mburu, S., Taguchi, T., Eisenberg, B., and Godwin, A. (2009). Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. Cancer Res, 69(23), 9125-9132.

GIST.S *Sample GIST gene expression data from Ochs et al. (2009).*

Description

Standard deviation of gene expression data from gastrointestinal stromal tumor cell lines treated with Gleevec.

Usage

```
GIST_TS_20084
```

Format

Matrix with 1363 genes by 9 samples containing standard deviation (GIST.S) of the gene expression data.

References

Ochs, M., Rink, L., Tarn, C., Mburu, S., Taguchi, T., Eisenberg, B., and Godwin, A. (2009). Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. Cancer Res, 69(23), 9125-9132.

gs *Simulated gene sets.*

Description

Simulated gene sets.

Usage

```
gs
```

Format

List containing simulated genes regulated in "gs1" and "gs2".

ModSim.D — *Simulated gene expresssion data.*

Description

Gene expression data simulated from 3 known true patterns (ModSim.P.true).

Usage

```
ModSim
```

Format

Matrix of 25 rows by 20 columns of simulated expression measurements.

ModSim.P.true — *Simulated gene expresssion data.*

Description

Known true patterns used to simulate gene expression data (ModSim.D).

Usage

```
ModSim
```

Format

Matrix of 3 rows by 20 columns containing true patterns used to simulate gene expression data.

PGS — *Simulated pattern matrix.*

Description

Simulated true patterns for gene expression with activity in two gene sets (gs).

Usage

```
PGS
```

Format

Matrix of 3 rows by 25 columns containing simulated patterns.

plotGAPS *Plotter for GAPS decomposition results*

Description

Plots the A and P matrices obtained from the GAPS matrix decomposition.

Usage

```
plotGAPS(A, P, outputPDF="")
```

Arguments

A	The amplitude matrix $\mathbf{A}$ obtained from GAPS.
P	The pattern matrix $\mathbf{P}$ obtained from GAPS.
outputPDF	Name of an pdf file to which the results will be output. (Optional; default="" will output plots to screen).

Note

If the plot option is true in CoGAPS, this function will be called automatically to plot results to the screen.

Author(s)

Elana J. Fertig <efertig@jhmi.edu>

See Also

CoGAPS

tf2ugFC *Gene sets defined by transcription factors defined from TRANSFAC.*

Description

List of genes contained in gastrointestinal stromal tumor cell line measurements that are regulated by transcription factors in the TRANSFAC database. Used for the gene set analysis in Ochs et al. (2009).

Usage

```
TFGSList
```

Format

Data.frame containing genes (rows) regulated by each transcription factor (columns).

References

Ochs, M., Rink, L., Tarn, C., Mburu, S., Taguchi, T., Eisenberg, B., and Godwin, A. (2009). Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. Cancer Res, 69(23), 9125-9132.

Index