Package 'DART'

March 26, 2013

Type Package

Title Denoising Algorithm based on Relevance network Topology

Version 1.4.0

Date 2012-05-25

Author Yan Jiao, Katherine Lawler, Andrew E Teschendorff

Depends R (>= 2.10.0), igraph0

Suggests breastCancerVDX, breastCancerMAINZ, Biobase

biocViews GeneExpression, DifferentialExpression, GraphsAndNetworks,Pathways, Bioinformatics

Maintainer Katherine Lawler <katherine.lawler@kcl.ac.uk>

Description Denoising Algorithm based on Relevance network Topology (DART) is an algorithm designed to evaluate the consistency of prior information molecular signatures (e.g in-vitro perturbation expression signatures) in independent molecular data (e.g gene expression data sets). If consistent, a pruning network strategy is then used to infer the activation status of the molecular signature in individual samples.

License GPL-2

LazyLoad yes

R topics documented:

DART-package *Denoising Algorithm based on Relevance network Topology*

Description

Denoising Algorithm based on Relevance network Topology (DART) is an unsupervised algorithm which evaluates the consistency of model pathway/molecular signatures in independent molecular samples before estimating the activation status of the signature in these independent samples. DART was devised for application to cancer genomics problems. For instance, one may wish to infer the activation status of an in-vitro derived oncogenic perturbation gene expression signature in a primary tumour for which genome-wide expression data is available. Before estimating pathway activity in the tumour, DART will evaluate if the "model" in-vitro signature pattern of up and down regulation is consistent with the expression variation seen across an expression panel of primary tumours. If the consistency score is statistically significant, this justfies using the perturbation signature to infer the activation status of the oncogenic pathway in the independent tumour samples. However, in this case, DART will also prune/denoise the perturbation signature using a relevance network topology strategy. This denoising step implemented in DART has been shown to improve estimates of pathway activity in clinical tumour specimens. Other examples of model pathway signatures could be a pathway model of signal transduction, a curated list of genes predicted to be up or down regulated in response to pathway activation/inhibition, or predicted upregulated targets of a transcription factor from say ChIP-Chip/Seq experiments. Three internal functions implement the steps in DART and are provided as explicit functions to allow user flexibility. DoDART is the main user function which will automatically and sequentially run through the following internal functions:

(1) BuildRN: This function builds a relevance correlation network of the model pathway signature in the data set in which the pathway activity estimate is desired. Note that this step is totally unsupervised and does not use and phenotypic information of the samples.

(2) EvalConsNet: This function evaluates the consistency of the inferred network with the prior information of the model pathway signature. The up/down regulatory pattern given by the model signature implies predictions about the directionality of the gene-gene correlations in the independent data set. For instance, if gene "A" is upregulated and gene "B" is downregulated, then assuming that the model signature has any relevance in the independent data set, we would expect genes "A" and "B" to be anti-correlated. Thus, a consistency score can be computed. Only if the consistency score is higher than the score expected by random chance is it recommended that the model signature be used to infer pathway activity.

(3) PruneNet: This function obtains the pruned, i.e consistent, network, in which any edge represents a significant correlation in gene expression whose directionality agrees with that predicted by the prior information. This is the denoising step of the algorithm. The function returns the whole pruned network and its maximally connected component.

(4) PredActScore: Given the adjacency matrix of the maximally connected consistent subnetwork and given the regulatory weights of the corresponding model pathway signature, this function estimates a pathway activation score in each sample. This function can also be used to infer pathway activity in another independent data set using the inferred subnetwork.

Author(s)

Andrew E Teschendorff, Yan Jiao

References

Jiao Y, Lawler K, Patel GS, Purushotham A, Jones AF, Grigoriadis A, Ng T, Teschendorff AE. Denoising algorithm based on relevance network topology improves molecular pathway activity inference. Submitted.

Teschendorff AE, Gomez S, Arenas A, El-Ashry D, Schmidt M, et al. (2010) Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. BMC Cancer 10:604.

Examples

```
### Example
### load in example data:
data(dataDART);
### dataDART$data: mRNA expression data of 67 ER negative breast cancer samples.
### dataDART$pheno: 51 basals and 16 HER2+ (ERBB2+).
### dataDART$sign: perturbation signature of ERBB2 activation.

### Build Relevance Network
rn.o <- BuildRN(dataDART$data,dataDART$sign,fdr=0.05);
### Evaluate Consistency
evalNet.o <- EvalConsNet(rn.o);
print(evalNet.o$cons)
### The consistency score, i.e fraction of consistent edges is 0.81
### P-value is significant, so proceed:
### Prune i.e denoise the network
prNet.o <- PruneNet(evalNet.o);
### print dimension of the maximally connected pruned network
print(dim(prNet.o$pradjMC));
### infer signature activation in the original data set
pred.o <- PredActScore(prNet.o$pradjMC,prNet.o$signMC,dataDART$data)
### check that activation is higher in HER2+ compared to basals
boxplot(pred.o$score ~ dataDART$pheno);
pv <- wilcox.test(pred.o$score ~ dataDART$pheno)$p.value;
text(x=1.5,y=10,labels=paste("P=",pv,sep=""));
```

BuildRN *Builds the relevance correlation network*

Description

This function builds the relevance correlation network for the genes in the model pathway signature in the given data set.

Usage

```
BuildRN(data.m, sign.v, fdr)
```

Arguments

data.m	Data matrix (numeric). Rows label features, columns label samples. It is assumed that number of features is much larger than number of samples. Rownames must be a valid gene or probe identifier.
sign.v	Model pathway signature vector (numeric). Elements correspond to the regulatory weights, i.e the sign indicates if up or downregulated. Names of sign.v must be a gene name (probe) identifier which must match the gene (probe) identifier used for the rows of data.m.
fdr	Desired false discovery rate (numeric) of significant edges in relevance correlation network.

Value

A list with following entries:

adj	Adjacency matrix of inferred relevance network
s	Model signature vector in data
sd	Gene signature data matrix
c	Correlations between signature genes
d	Data matrix
rep.idx	Indices of the genes in signature which could be found in data matrix

Author(s)

Andrew E Teschendorff, Yan Jiao

References

Jiao Y, Lawler K, Patel GS, Purushotham A, Jones AF, Grigoriadis A, Ng T, Teschendorff AE. Denoising algorithm based on relevance network topology improves molecular pathway activity inference. Submitted.

Teschendorff AE, Gomez S, Arenas A, El-Ashry D, Schmidt M, et al. (2010) Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. BMC Cancer 10:604.

Examples

```
data(dataDART)
rn.o <- BuildRN(dataDART$data, dataDART$sign, fdr=0.05)
## See ?DoDART and vignette('DART') for further examples.
```

dataDART *Example data for DART package*

Description

This data object consists of a gene expression data matrix over 67 oestrogen receptor negative breast cancers (Wang, Y. et al. Lancet 365, 671-9 (2005)), of which 51 are basals and 16 are HER2+/ERBB2+. This classification is based on the intrinsic subtype classifier (see Hu Z, et al., BMC Genomics. 2006 Apr 27;7:96), and while not equivalent to IHC or the amplification status at the ERBB2 locus, it broadly matches these other classifications. The model pathway signature is an in-vitro derived perturbation signature reflecting ERBB2 activation (see Creighton CJ, et al., Cancer Res 2006 Apr 1;66(7):3903-11). Both data and signature are annotated with Entrez Gene IDs.

Usage

```
dataDART
```

Format

A list with 4 elements: (i) data: is the gene expression matrix described above, (ii) sign: is the ERBB2 perturbation signature, (iii) pheno: basal/ERBB2 status of samples, (iv) phenoMAINZ: basal/ERBB2 status of a further set of samples.

Examples

```
data(dataDART)
names(dataDART)
```

DoDART *Main function of DART*

Description

This is the main function implementing DART. Given a data matrix and a model (pathway) signature it will construct the relevance correlation network for the genes in the signature over the data, evaluate the consistency of the correlative patterns with those predicted by the model signature, filter out the noise and finally obtain estimates of pathway activity for each individual sample. Specifically, it will call and run the following functions:

(1) BuildRN: This function builds a relevance correlation network of the model pathway signature in the data set in which the pathway activity estimate is desired. We point that this step is totally unsupervised and does not use and phenotypic information of the samples.

(2) EvalConsNet: This function evaluates the consistency of the inferred network with the prior information of the model pathway signature. The up/down regulatory pattern given by the model signature implies predictions about the directionality of the gene-gene correlations in the independent data set. For instance, if gene "A" is upregulated and gene "B" is downregulated, then assuming that the model signature has any relevance in the independent data set, we would expect genes "A"

and "B" to be anti-correlated. Thus, a consistency score can be computed. Only if the consistency score is higher than the score expected by random chance is it recommended that the model signature be used to infer pathway activity.

(3) PruneNet: This function obtains the pruned, i.e consistent, network, in which any edge represents a significant correlation in gene expression whose directionality agrees with that predicted by the prior information. This is the denoising step of the algorithm. The function returns the whole pruned network and its maximally connected component.

(4) PredActScore: Given the adjacency matrix of the maximally connected consistent subnetwork and given the regulatory weights of the corresponding model pathway signature, this function estimates a pathway activation score in each sample. This function can also be used to infer pathway activity in another independent data set using the inferred subnetwork.

Usage

DoDART(data.m, sign.v, fdr)

Arguments

data.m	Data matrix (numeric). Rows label features, columns label samples. It is assumed that number of features is much larger than number of samples. Rownames of data.m must be valid unique gene (probe) identifiers.
sign.v	Model pathway signature vector (numeric). Elements represent the regulatory weights (i.e if up or downregulated). Names of sign.v are the gene identifiers, which must match the gene (probe) identifiers of the rows of data.m
fdr	Desired false discovery rate (numeric) which determines allowed false positive in the relevance network.

Value

netcons	A vector summarising the properties of the correlation network and the consistency with the model pathway signature: nG is the number of genes in the signature, nE is the number of edges of the relevance network generated by function BuildRN, fE is the ratio of the number of edges in the relevance network to the maximum possible number, fconsE is the fraction of edges whose sign (i.e sign of correlation) is the same as the directionality predicted by the model signature, Pval(consist) is a p-value reflecting the significance of fconsE, and is estimated as the fraction of randomisations that yielded an average connectivity larger than the observed one.
adj	Adjacency matrix of maximally connected consistent relevance network.
sign	Model pathway signature vector of genes found in data set and in maximally connected component.
score	Predicted activation scores of the model signature in the samples of data set data.m.
degree	Degrees/connectivities of the genes in the DART network.

Author(s)

Andrew E Teschendorff, Yan Jiao

References

Jiao Y, Lawler K, Patel GS, Purushotham A, Jones AF, Grigoriadis A, Ng T, Teschendorff AE. Denoising algorithm based on relevance network topology improves molecular pathway activity inference. Submitted.

Teschendorff AE, Gomez S, Arenas A, El-Ashry D, Schmidt M, et al. (2010) Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. BMC Cancer 10:604.

Examples

```
### Example
### load in example data:
data(dataDART);
### dataDART$data: mRNA expression data of 67 ER negative breast cancer samples.
### dataDART$pheno: 51 basals and 16 HER2+ (ERBB2+).
### dataDART$phenoMAINZ: 24 basals and 8 HER2+ (ERBB2+).
### dataDART$sign: perturbation signature of ERBB2 activation.


### Using DoDART
dart.o <- DoDART(dataDART$data,dataDART$sign,fdr=0.05);
### check that activation is higher in HER2+ compared to basals
boxplot(dart.o$score ~ dataDART$pheno);
pv <- wilcox.test(dart.o$score ~ dataDART$pheno)$p.value;
text(x=1.5,y=10,labels=paste("P=",pv,sep=""));
```

EvalConsNet — *Evaluates the consistency of the inferred relevance correlation network with the correlations predicted by the model pathway signature.*

Description

This function evaluates statistical consistency of the inferred relevance network with the correlations predicted by the model pathway signature. Only if the consistency score is higher than the score expected by random chance, is it recommended to use the signature to infer pathway activity.

Usage

```
EvalConsNet(buildRN.o)
```

Arguments

buildRN.o	Output list object from function BuildRN

Value

A list with following entries:

netcons	A vector summarising the properties of the network and the consistency with the prior information: nG is the number of genes in the signature, nE is the number of edges of the relevance network generated by function BuildRN, fE is the ratio of the number of edges in the relevance network to the maximum possible number, fconsE is the fraction of edges whose sign (i.e sign of correlation) is the same as the directionality predicted by the model signature, Pval(consist) is a p-value reflecting the significance of fconsE, and is estimated as the fraction of randomisations that yielded an average connectivity larger than the observed one.
netsign	A matrix of dimension 2 times number of edges in network comparing directionality of prior info and that in the observed data
adj	Adjacency matrix of inferred relevance network
s	Model signature vector
c	Correlation matrix between model signature genes

Author(s)

Andrew E Teschendorff, Yan Jiao

References

Jiao Y, Lawler K, Patel GS, Purushotham A, Jones AF, Grigoriadis A, Ng T, Teschendorff AE. Denoising algorithm based on relevance network topology improves molecular pathway activity inference. Submitted.

Teschendorff AE, Gomez S, Arenas A, El-Ashry D, Schmidt M, et al. (2010) Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. BMC Cancer 10:604.

Examples

```
data(dataDART)
rn.o <- BuildRN(dataDART$data, dataDART$sign, fdr=0.05)
evalNet.o <- EvalConsNet(rn.o)
## See ?DoDART and vignette('DART') for further examples.
```

PredActScore — *Computes the DART activation score of the model signature in the samples of a data set.*

Description

This function computes the DART activation score of the model signature for the samples of a data matrix. This data matrix can be the same data matrix in which DART was applied or it could be a totally independent data set.

Usage

```
PredActScore(pradjMC.m, signMC.v, data.m)
```

Arguments

pradjMC.m	The adjacency matrix (numeric) of the maximally connected pruned network inferred using DART.
signMC.v	The corresponding model signature vector (numeric).
data.m	A data matrix (numeric) with rows labeling genes/probes and columns labeling samples in which the signature activity scores are desired.

Value

A list with following elements:

adj	Adjacency matrix of the DART network for the genes found in the data set given by data.m.
sign	The corresponding model signature vector for the genes found in the data set.
score	A vector of the signature/pathway activity scores for each sample in the data set.

Author(s)

Andrew E Teschendorff, Yan Jiao

References

Jiao Y, Lawler K, Patel GS, Purushotham A, Jones AF, Grigoriadis A, Ng T, Teschendorff AE. Denoising algorithm based on relevance network topology improves molecular pathway activity inference. Submitted.

Teschendorff AE, Gomez S, Arenas A, El-Ashry D, Schmidt M, et al. (2010) Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. BMC Cancer 10:604.

Examples

```
data(dataDART)
rn.o <- BuildRN(dataDART$data, dataDART$sign, fdr=0.05)
evalNet.o <- EvalConsNet(rn.o)
prNet.o <- PruneNet(evalNet.o)
pred.o <- PredActScore(prNet.o$pradjMC,prNet.o$signMC,dataDART$data)
## See ?DoDART and vignette('DART') for further examples.
```

PruneNet — *Prunes relevance network to allow only edges that are consistent with the predictions of the model signature*

Description

Prunes relevance network to allow only edges that are consistent with the predictions of the model signature, and extracts the maximally connected component. This is the denoising step in DART.

Usage

```
PruneNet(evalNet.o)
```

Arguments

evalNet.o	Output list object from EvalConsNet

Value

A list with following entries:

pradj	The adjacency matrix of the pruned i.e consistent network.
sign	The model signature vector of genes in pruned network.
score	The fraction of edges surviving the pruning/denoising.
netconst	Same output as for EvalConsNet.
pradjMC	The adjacency matrix of the maximally connected component of pruned network.
signMC	The model signature vector of the genes in the maximally connected component.

Author(s)

Andrew E Teschendorff, Yan Jiao

References

Jiao Y, Lawler K, Patel GS, Purushotham A, Jones AF, Grigoriadis A, Ng T, Teschendorff AE. Denoising algorithm based on relevance network topology improves molecular pathway activity inference. Submitted.

Teschendorff AE, Gomez S, Arenas A, El-Ashry D, Schmidt M, et al. (2010) Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. BMC Cancer 10:604.

Examples

```
data(dataDART)
rn.o <- BuildRN(dataDART$data, dataDART$sign, fdr=0.05)
evalNet.o <- EvalConsNet(rn.o)
prNet.o <- PruneNet(evalNet.o)
pred.o <- PredActScore(prNet.o$pradjMC,prNet.o$signMC,dataDART$data)
## See ?DoDART and vignette('DART') for further examples.
```

Index