Package 'DynamicCancerDriver'

December 9, 2021

Type Package

Title Dynamic cancer drivers A causal approach for cancer driver discovery based on bio-pathological trajectories

Version 1.4.1

Maintainer Andres M. Cifuentes-Bernal <andres.cifuentes_bernal@mymail.unisa.edu.au>

Description We propose a novel approach for causal inference of genes driving one or more core processes during cancer development (i.e. dynamic cancer driver). We use the concept of pseudotime for inferring the latent progression of samples along a biological transition during cancer and identifies a critical event when such a process is significantly deviated from normal to carcinogenic. We infer driver genes by assessing the causal effect they have on the process after such a critical event.

License GPL-3.0

Encoding UTF-8

LazyData true

Depends R (>= 3.5.0)

Imports CausalImpact(>= 1.2.7), tidyverse(>= 1.3.1)

RoxygenNote 7.1.2

Author Andres Mauricio Cifuentes_Bernal [aut, cre] (<https://orcid.org/0000-0002-1880-3624>), Vu VH Pham [aut] (<https://orcid.org/0000-0002-8884-3584>), Xiaomei Li [aut] (<https://orcid.org/0000-0002-8870-3186>), Lin Liu [aut], Jiuyong Li [aut], Thuc Le [aut] (<https://orcid.org/0000-0002-9732-4313>)

R topics documented:

CreatePPIRank *CreatePPIRank*

Description

CreatePPIRank takes each gene in geneIDs and determines the number of times such a gene appears as input node (n.in) and as output node (n.out) in the PPI network described by PPImatrix. If no PPImatrix is provided, the PPI network described at Vinayagam et al. (2011) is used as reference.
The rank of a gene corresponds to the total count of the times that such a gene is either an input node or an output node in the PPI network.

Usage

```
function(geneIDs = NULL, PPImatrix = NULL)
```

Arguments

geneIDs	A character vector containing genes IDs. Genes IDs can be Ensembl.ID (recommended), HGNC.ID, NCBI.ID or HGNC.symbol
PPImatrix	A 2 column dataframe (preferred) or matrix containing the input nodes (column 1), and the output nodes (column 2) of a PPI network. If NULL (default), the PPI network decribed in Vinayagam et al. (2011) is used as reference.

Value

A dataframe with the following variables:

1. IDs: Ensembl.ID, HGNC.ID, NCBI.ID or HGNC.symbol of the genes in the PPI network
2. n.in: Count of the gene as input node
3. n.out: Count of the gene as output node
4. total: Count of the total times the gene appears in the PPI network

Author(s)

Andres Mauricio Cifuentes_Bernal, Vu VH Pham, Xiaomei Li, Lin Liu, JiuyongLi and Thuc Duy Le

See Also

findDCD

Examples

```
## Not run:
   data("GSE75688_TPM_tumor", package = "DynamicCancerDriver")
   CreatePPIRank(colnames(GSE75688_TPM_tumor)[1:100])

## End(Not run)
```

findCovariate *findCovariate*

Description

For each feature in FS, findCovariate finds the non-PPI gene with the largest Pearson correlation with the feature. For a proper functioning, the data of all features in FS must be included as columns in GeneExpression.

Usage

```
function(GeneExpression, FS, PPIrank=NULL)
```

Arguments

GeneExpression	A dataframe (preferred) or a matrix containing mRNA gene expression. Columns represent mRNAs and rows represent samples. Column names should correspond to gene IDs. Gene IDs can be from any of the following nomenclatures: *Ensembl.ID* (recommended), *HGNC.ID, NCBI.ID* or *HGNC.symbol*.
FS	A character vector containing the names of the features for which a covariate is to be found.
PPIrank	(optional) A dataframe obtained from CreatePPIRank

Value

A dataframe with the following two variables:

1. *Feature:* The vector FS of features.
   The name of this variable can be 1 of *Ensembl.ID, HGNC.ID, NCBI.ID* or *HGNC.symbol*.
2. scontrol: For each *Feature*, the name of the non-PPI gene with the largest Pearson correlation

Author(s)

Andres Mauricio Cifuentes_Bernal, Vu VH Pham, Xiaomei Li, Lin Liu, JiuyongLi and Thuc Duy Le

See Also

findDCD

Examples

```
## Not run:
   data("GSE75688_TPM_tumor", package = "DynamicCancerDriver")
   FS <- colnames(GSE75688_TPM_tumor)[1:100]

   sControl <- findCovariate(GeneExpression = GSE75688_TPM_tumor[,1:500]
   , FS = FS)

## End(Not run)
```

findDCD *findDCD*

Description

findDCD identifies genes driving driving one (or more) significant biological processes along cancer progression based on the hypothesis that the causal relationship between a cancer driver gene and cancer development induces a significant deviation (also referred as causal impact) of a core process from normal to carcinogenic.

Usage

```
function(GeneExpression, z=NULL, pathCovariate =NULL
, findEvent = T, Step=1, chunk_size= 100
, PPItop = 0.3, alpha=0.05, CIniter=200
, returnModel=F, elbo_tol=1e-3, project = NULL)
```

Arguments

GeneExpression	A dataframe (preferred) or a matrix containing mRNA gene expression. Columns represent mRNAs and rows represent samples. Column names should correspond to gene IDs. Gene IDs can be from any of the following nomenclatures: *Ensembl.ID* (recommended), *HGNC.ID, NCBI.ID* or *HGNC.symbol*.
z	A numeric vector with pseudotime score for each sample. If NULL (default), pseudotime score is calculated by using phenopath package and the pathCovariate.
findEvent	If TRUE (default) samples are ordered in pseudotime order and deviations from normal to cancerogenic are assessed by using the CausalImpact function from the package CausalImpact. The sample with the largest (significant) CausalImpact is labeled as the "event". If FALSE, the sample where the change of sign (from negative to positive) occurs is labeled as the "event".
Step	An integer indicating the distance between samples to be assessed when findEvent = TRUE. Step = 1 (default) means that all samples are considered during the findEvent process.
chunk_size	An integer defining the number of genes to be analysed at a time. chunk_size = 100 (default) indicates that groups of 100 genes will be analised at a time.
PPItop	A numeric value between 0 and 1 indicating the percentage of PPI genes in the dataset to be selected as putative drivers. PPI genes with the most interactions are selected. PPItop = 0.3 (default) means that the 30 genes with the most interactions in the PPI network are selected.
alpha	Significance level for the statistical test. alpha=0.05 by default.
CIniter	number of iterations (200 by default) for CausalImpact modeling.
returnModel	If TRUE, the complete CausalImpact model is included in the outcome of findDCD. If FALSE (default), only the most relevant parameters of the CausalImpact model are included in the outcome of findDCD.
elbo_tol	A numeric value (elbo_tol = 1e-3 by default). The relative pct change in the evidence lower bound (ELBO) below which phenopath calculation of the pseudotime score is considered converged.

`project` (optional) A TCGA project name (e.g. BRCA). If provided, a dummy rank for the inferred dynamic cancer driver is calculated based on the frequency of mutations of those genes in the TCGA project dataset.

Value

A `list` consisting of the following elements:

`res` A `list` with the results of the *DynamicCancerDriver* inference process. Results are listed as follows:

1. `FS:` A `vector` containing the names of the putative cancer drivers
2. `CausalImpact:` Causal impact models of the putative drivers
3. `CDinfer:` Inferred Dynamic Cancer Drivers
4. `summary:` A table with a summary of the results

`eventAt` A `integer` containing the index (after pseudotime ordering) of the sample labeled as the "event".

`z` Pseudotime score

Author(s)

Andres Mauricio Cifuentes_Bernal, Vu VH Pham, Xiaomei Li, Lin Liu, JiuyongLi and Thuc Duy Le

See Also

findCovariate, parallelCI

Examples

```
## Not run:
   data("GSE75688_TPM_tumor", package = "DynamicCancerDriver")

----- Find Dynamic Cancer Drivers, PPI top 30% -----
DCD.HER2time_SC <- findDCD(GeneExpression = GSE75688_TPM_tumor
                           , pathCovariate = "HER2"
                           , PPItop = 0.3
                           , findEvent = TRUE)

## End(Not run)
```

`findZ` *findZ*

Description

`findZ` calculates a pseudotime score for each sample by using a `pathCovariate` that reasonably encodes the progression of one core biological process during cancer progression (in the sense described by Campbell 2018).
Pseudotime score calculation relies in the procedures implemented in the phenopath package.

Usage

```
function(GeneExpression, FS, pathCovariate, elbo_tol = 1e-3)\\{}
```

Arguments

GeneExpression	A dataframe (preferred) or a matrix containing mRNA gene expression. Columns represent mRNAs and rows represent samples. Column names should correspond to gene IDs. Gene IDs can be from any of the following nomenclatures: *Ensembl.ID* (recommended), *HGNC.ID, NCBI.ID* or *HGNC.symbol*.
FS	A character vector containing the names of the features to be used for the calculation of the pseudotime score.
pathCovariate	A named vector containing the data of a path covariate.
elbo_tol	A numeric value (elbo_tol = 1e-3 by default). The relative pct change in the evidence lower bound (ELBO) below which phenopath calculation of the pseudotime score is considered converged.

Value

A dataframe with the following two variables:

1. Feature: The vector FS of features
2. scontrol: For each Feature, the name of the non-PPI gene with the largest Pearson correlation.

Author(s)

Andres Mauricio Cifuentes_Bernal, Vu VH Pham, Xiaomei Li, Lin Liu, JiuyongLi and Thuc Duy Le

See Also

findDCD

Examples

```
## Not run:
   data("GSE75688_TPM_tumor", package = "DynamicCancerDriver")
   FS <- colnames(GSE75688_TPM_tumor)[1:100]
   GE <- GSE75688_TPM_tumor[,1:500]
   sControl <- findCovariate(GeneExpression = GSE75688_TPM_tumor[,1:500]
   , FS = FS)

   #toy example, using "VIM" as path covariate
   z <- findZ(GeneExpression = GE
              , FS, pathCovariate = GSE75688_TPM_tumor[,"ENSG00000026025"]
              , elbo_tol = 1e-3)

## End(Not run)
```

parallelCI *parallelCI*

Description

parallelCI uses parallel package for implementing a parallelised calculation of the CausalImpact on gene expression. It is assumed that the provided GeneExpression matrix contains pseudotime ordered gene expression. z is the pseudotime score used for ordering the gene expression, and eventAt indicates the sample at the most significant change (from normal to carcinogenic) occurs.

Usage

```
function(GeneExpression,sControl,z,eventAt,
chunk_size = 50, returnModel = F)\\{}
```

Arguments

GeneExpression	A dataframe (preferred) or a matrix containing mRNA gene expression. Columns represent mRNAs and rows represent samples. Column names should correspond to gene IDs. Gene IDs can be from any of the following nomenclatures: *Ensembl.ID* (recommended), *HGNC.ID, NCBI.ID* or *HGNC.symbol*.
sControl	A 2 column matrix containing gene IDs (1st column) and non-PPI gene (1 per gene ID) to be used as covariate for CausalImpact modelling. For a correct functioning, the pseudotime ordered data of all elements in sControl need to be included as columns in GeneExpression.
z	A numeric vector containing the pseudotime score used for ordering the samples in GeneExpression. For a correct functioning, z needs to follow ascending order and this order must agree with the order of the samples (rows) of the GeneExpression matrix.
eventAt	An integer with the index (in pseudotime order) of the sample where the most significant change is inferred to happen.
CIniter	number of iterations (200 by default) for CausalImpact modeling.
chunk_size	An integer indicating the number of genes to be passed to each worker during the parallel calculation. (50 by default)
returnModel	A boolean. If TRUE, the full causal impact model (as calculated by CausalImpact package) is returned. if FALSE (default), only the main parameters of the CausalImpact are returned,

Value

A list where each element is the full CausalImpact model (if returnModel = TRUE) or the simplified CausalImpact model (if returnModel = FALSE) of one gene in the 1st column of sControl.

Author(s)

Andres Mauricio Cifuentes_Bernal, Vu VH Pham, Xiaomei Li, Lin Liu, JiuyongLi and Thuc Duy Le

See Also

findCovariate, findDCD

Examples

```
## Not run:
   data("GSE75688_TPM_tumor", package = "DynamicCancerDriver")
   FS <- colnames(GSE75688_TPM_tumor)[1:100]
   GE <- GSE75688_TPM_tumor[,1:500]
   sControl <- findCovariate(GeneExpression = GSE75688_TPM_tumor[,1:500]
   , FS = FS)

   #toy example, using "VIM" as path covariate
   z <- findZ(GeneExpression = GE
              , FS, pathCovariate = GSE75688_TPM_tumor[,"ENSG00000026025"]
              , elbo_tol = 1e-3)
   GE <- GE[order(z),,drop=F]
   z <- z[order(z),1, drop=F]

   parCI <-parallelCI(GE,sControl,z,eventAt=7)


## End(Not run)
```

Index