Package 'scTEP'

October 14, 2022

Type Package

Title Single-Cell Trajectory Inference using Ensemble Pseudotimes

Version 0.1.0

Maintainer Yifan Zhang <yfzhang@nevada.unr.edu>

Description A single-cell trajectory inference method using 'Autoencoder' and Minimum Spanning Tree (MST) from high dimensional 'scRNA-Seq' data. The software run clustering methods six times using 'scDHA' with the number of clusters set from 6 to 10. Then, the 'scTEP' calculates pseudotime based on multiple clustering results. Lastly, the 'scTEP' generates trajectory using MST algorithm and fine-tunes it according to the pseudotime of clusters.

License LGPL

Encoding UTF-8

LazyData true

LazyDataCompression xz

Depends R (>= 3.6.0)

Imports stats, dplyr, igraph, scDHA, foreach, BiocGenerics, Matrix, SummarizedExperiment, doParallel, ggsci, psych, tibble, rlang, SingleCellExperiment

RoxygenNote 7.2.1

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Author Yifan Zhang [aut, cre],
Duc Tran [aut],
Tin Nguyen [fnd],
Sergiu M. Dascalu [fnd],
Frederick C. Harris, Jr. [fnd]

Repository CRAN

Date/Publication 2022-09-26 13:10:02 UTC

R topics documented:

`clustering` *scTEP*

Description

The 'clustering' function conducts multiple clustering to the sc-RNA seq data using the 'scDHA' function from the 'scDHA' package. The 'scDHA' allows the user to set a specific cluster number for the clustering process. We set the cluster number from 6 to 10 and run the 'scDHA' function six times. The 'trajectoryinference' function will using those total six clustering results to generate pseudotimes.

Usage

```
clustering(data, ncores = 10L, seed = NULL)
```

Arguments

`data`	A list consists of gene expression matrix.
`ncores`	Number of processor cores to use. This values is set to `seed = 10L` by default.
`seed`	A parameter to set a seed for reproducibility.

Value

List with the following keys:

- allCluster - A list consists of clutering results using scDHA with `k = 5:10`.

References

1. Duc Tran, Hung Nguyen, Bang Tran, Carlo La Vecchia, Hung N. Luu, Tin Nguyen (2021). Fast and precise single-cell data analysis using a hierarchical autoencoder. Nature Communications, 12, 1029. doi: 10.1038/s41467-021-21312-2

Examples

```
# Load the package and the example data (goolam data set)
library(scTEP)
#Load pathway genesets
data('genesets')
#Load example data (SCE dataset)
data("goolam")
#Get data matrix and label
expr <- as.matrix(t(SummarizedExperiment::assay(goolam)))[, 1:100]

#Get data matrix and label
data = preprocessing(expr)

#Get clustering results using scDHA with k from 6 to 10.
allCluster = scTEP::clustering(data, ncores = 2)
```

`genesets` *genesets*

Description

A list consists of pathway genesets of Homo sapiens (human) and Mus musculus (house mouse).

Usage

```
genesets
```

Format

An object of class `list` of length 2.

`goolam` *goolam*

Description

goolam datas set save as a SCE object.

Usage

```
goolam
```

Format

An object of class `SingleCellExperiment` with 41336 rows and 124 columns.

preprocessing *preprocessing*

Description

Conduct preprocessing, including remove all zero columns and scale gene expression smaller than 100 by log transformation with 2 as base.

Usage

```
preprocessing(expr)
```

Arguments

expr	The gene expression matrix.

Value

List with the following keys:

- expr - Gene expression matrix, with rows represent samples and columns represent genes.

Examples

```
#Load the package
library(scTEP)
#Load example data
data("goolam")
#Get data matrix
expr <- as.matrix(t(SummarizedExperiment::assay(goolam)))

data = preprocessing(expr)
```

scTEP.fa *scTEP.fa*

Description

The 'scTEP.fa' function first selects the corresponding pathway gene sets of the data set from KEGG, then intersect the genes in the expression matrix with each pathway to have an intersect gene expression matrix for all pathways.

Usage

```
scTEP.fa(data, genesets, data_org = "hsa", ncores = 10L, seed = NULL)
```

Arguments

data	A list consists of gene expression matrix.
genesets	A list consists of Homo sapiens and Mus musculus gene sets.
data_org	The organism of the data set, mmu or hsa.
ncores	Number of processor cores to use. This values is set to seed = 10L by default
seed	A parameter to set a seed for reproducibility.

Value

List with the following keys:

- faData - A large matrix consists of concatenated 2 dimensional factor analysis results of pathways.

Examples

```
# Load the package and the example data (goolam datas set)
library(scTEP)
#Load pathway genesets
data('genesets')
#Load example data (SCE dataset)
data("goolam")
#Get data matrix and label
expr <- as.matrix(t(SummarizedExperiment::assay(goolam)))[1:10, 1:100]

#Get data matrix and label
data = preprocessing(expr)

#Generate factor analysis results for all the intersections between data matrix and genesets
data_fa = scTEP.fa(data, genesets, ncores = 2, data_org = 'mmu', seed = 1)
```

---

trajectoryinference *trajectoryinference*

Description

This is the main function that performs sc-RNA seq data trajectory inference. The 'scTEP' first load the latent representation and clustering result of scDHA. Second, the 'trajectoryinference' function iterates through all the clustering results and calculates the distance between clusters as pseudotimes. Third, it calculates the average pseudotime for every cluster in the clustering result obtained from the first step. Fourth, it generates an MST and fine-tunes it by the pseudotime. Therefore, we have the trajectory for the data set.

Usage

```
trajectoryinference(
  data,
  start.idx,
  scDHA_res,
  allCluster,
  ncores = 10L,
  seed = NULL
)
```

Arguments

data	A list consists of gene expression matrix.
start.idx	The indexes of the start cells, given by user.
scDHA_res	The 'scDHA' results, consists of latent and clustering result.
allCluster	A list consists of clustering results using 'scDHA' with k = 5:10.
ncores	Number of processor cores to use. This values is set to seed = 10L by default.
seed	A parameter to set a seed for reproducibility.

Value

List with the following keys:

- pseudotime - The pseudotime of cells in the data set.
- cluster - The clustering results of data set.
- data_clus_cent - The center of all the clusters.
- milestone_network - The milestone network of the inferred trajectory.
- g - An 'igraph' object of the inferred trajectory

References

1. Duc Tran, Hung Nguyen, Bang Tran, Carlo La Vecchia, Hung N. Luu, Tin Nguyen (2021). Fast and precise single-cell data analysis using a hierarchical autoencoder. Nature Communications, 12, 1029. doi: 10.1038/s41467-021-21312-2

Examples

```
# Load the package and the example data (goolam data set)
library(scTEP)
#Load pathway genesets
data('genesets')
#Load example data (SCE dataset)
data("goolam")
#Get data matrix and label
expr <- as.matrix(t(SummarizedExperiment::assay(goolam)))
label <- as.character(goolam$label)
```

```
stages = goolam@metadata$cell.stages

#Get data matrix and label
data = preprocessing(expr)

#Generate factor analysis results for all the intersections between data matrix and genesets
data_fa = scTEP.fa(data, genesets, ncores = 2, data_org = 'mmu', seed = 1)

#Get clustering results using 'scDHA' with k from 6 to 10.
allCluster = scTEP::clustering(data, ncores = 2)

scDHA_res <- scDHA(data_fa, do.clus = T, gen_fil = T, ncores = 2, seed = 1)
#Conduct trajectory inference to the data matrix
idx = which(label == stages[1])
out = trajectoryinference(data, idx, scDHA_res, allCluster, ncores = 2, seed = 1)
```

Index