Package 'SIMLR'

April 12, 2018

Version 1.4.1

Date 2018-03-09

Title Title: SIMLR and CIMLR Multi-kernel LeaRning methods

Maintainer Daniele Ramazzotti <daniele.ramazzotti@yahoo.com>

Depends R (>= 3.4),

Imports parallel, Matrix, stats, methods, Rcpp, pracma, RcppAnnoy, RSpectra

Suggests BiocGenerics, BiocStyle, testthat, knitr, igraph

Description In this package we provide implementations of both SIMLR and CIMLR. These methods were originally applied to single-cell and cancer genomic data, but they are in principle capable of effectively and efficiently learning similarities in all the contexts where diverse and heterogeneous statistical characteristics of the data make the problem harder for standard approaches.

Encoding UTF-8

LazyData TRUE

License file LICENSE

URL https://github.com/BatzoglouLabSU/SIMLR

BugReports https://github.com/BatzoglouLabSU/SIMLR

biocViews Clustering, GeneExpression, Sequencing, SingleCell

RoxygenNote 6.0.1

LinkingTo Rcpp

NeedsCompilation yes

VignetteBuilder knitr

Author Daniele Ramazzotti [aut, cre],
Bo Wang [aut],
Luca De Sano [aut],
Serafim Batzoglou [ctb]

R topics documented:

`BuettnerFlorian` *test dataset for SIMLR*

Description

example dataset to test SIMLR from the work by Buettner, Florian, et al.

Usage

```
data(BuettnerFlorian)
```

Format

gene expression measurements of individual cells

Value

list of 6: in_X = input dataset as an (m x n) gene expression measurements of individual cells, n_clust = number of clusters (number of distinct true labels), true_labs = ground true of cluster assignments for each of the n_clust clusters, seed = seed used to compute the results for the example, results = result by SIMLR for the inputs defined as described, nmi = normalized mutual information as a measure of the inferred clusters compared to the true labels

Source

Buettner, Florian, et al. "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells." Nature biotechnology 33.2 (2015): 155-160.

`CIMLR` *CIMLR*

Description

perform the CIMLR clustering algorithm

Usage

```
CIMLR(X, c, no.dim = NA, k = 10, cores.ratio = 1)
```

Arguments

X	a list of multi-omic data each of which is an (m x n) data matrix of measurements of cancer patients
c	number of clusters to be estimated over X
no.dim	number of dimensions
k	tuning parameter
cores.ratio	ratio of the number of cores to be used when computing the multi-kernel

Value

clusters the patients based on CIMLR and their similarities

list of 8 elements describing the clusters obtained by CIMLR, of which y are the resulting clusters: y = results of k-means clusterings, S = similarities computed by CIMLR, F = results from network diffiusion, ydata = data referring the the results by k-means, alphaK = clustering coefficients, execution.time = execution time of the present run, converge = iterative convergence values by T-SNE, LF = parameters of the clustering

Examples

```
CIMLR(X = GliomasReduced$in_X, c = 3, cores.ratio = 0)
```

CIMLR_Estimate_Number_of_Clusters
CIMLR Estimate Number of Clusters

Description

estimate the number of clusters by means of two huristics as discussed in the CIMLR paper

Usage

```
CIMLR_Estimate_Number_of_Clusters(all_data, NUMC = 2:5, cores.ratio = 1)
```

Arguments

all_data	is a list of multi-omic data each of which is an (m x n) data matrix of measurements of cancer patients
NUMC	vector of number of clusters to be considered
cores.ratio	ratio of the number of cores to be used when computing the multi-kernel

Value

a list of 2 elements: K1 and K2 with an estimation of the best clusters (the lower values the better) as discussed in the original paper of SIMLR

Examples

```
CIMLR_Estimate_Number_of_Clusters(GliomasReduced$in_X,
   NUMC = 2:5,
   cores.ratio = 0)
```

GliomasReduced *test dataset for CIMLR*

Description

example dataset to test CIMLR. This is a reduced version of the dataset from the work by The Cancer Genome Atlas Research Network.

Usage

```
data(GliomasReduced)
```

Format

multi-omic data of cancer patients

Value

list of 1 element: in_X = input dataset as a list of 4 (reduced) multi-omic data each of which is an (m x n) measurements of cancer patients

Source

Cancer Genome Atlas Research Network. "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas." New England Journal of Medicine 372.26 (2015): 2481-2498.

SIMLR *SIMLR*

Description

perform the SIMLR clustering algorithm

Usage

```
SIMLR(X, c, no.dim = NA, k = 10, if.impute = FALSE, normalize = FALSE,
  cores.ratio = 1)
```

Arguments

X	an (m x n) data matrix of gene expression measurements of individual cells or and object of class SCESet
c	number of clusters to be estimated over X
no.dim	number of dimensions
k	tuning parameter
if.impute	should I traspose the input data?
normalize	should I normalize the input data?
cores.ratio	ratio of the number of cores to be used when computing the multi-kernel

Value

clusters the cells based on SIMLR and their similarities

list of 8 elements describing the clusters obtained by SIMLR, of which y are the resulting clusters: y = results of k-means clusterings, S = similarities computed by SIMLR, F = results from network diffiusion, ydata = data referring the the results by k-means, alphaK = clustering coefficients, execution.time = execution time of the present run, converge = iterative convergence values by T-SNE, LF = parameters of the clustering

Examples

```
SIMLR(X = BuettnerFlorian$in_X, c = BuettnerFlorian$n_clust, cores.ratio = 0)
```

SIMLR_Estimate_Number_of_Clusters

SIMLR Estimate Number of Clusters

Description

estimate the number of clusters by means of two huristics as discussed in the SIMLR paper

Usage

```
SIMLR_Estimate_Number_of_Clusters(X, NUMC = 2:5, cores.ratio = 1)
```

Arguments

X	an (m x n) data matrix of gene expression measurements of individual cells
NUMC	vector of number of clusters to be considered
cores.ratio	ratio of the number of cores to be used when computing the multi-kernel

Value

a list of 2 elements: K1 and K2 with an estimation of the best clusters (the lower values the better) as discussed in the original paper of SIMLR

Examples

```
SIMLR_Estimate_Number_of_Clusters(BuettnerFlorian$in_X,
   NUMC = 2:5,
   cores.ratio = 0)
```

`SIMLR_Feature_Ranking` *SIMLR Feature Ranking*

Description

perform the SIMLR feature ranking algorithm. This takes as input the original input data and the corresponding similarity matrix computed by SIMLR

Usage

```
SIMLR_Feature_Ranking(A, X)
```

Arguments

A	an (n x n) similarity matrix by SIMLR
X	an (m x n) data matrix of gene expression measurements of individual cells

Value

a list of 2 elements: pvalues and ranking ordering over the n covariates as estimated by the method

Examples

```
SIMLR_Feature_Ranking(A = BuettnerFlorian$results$S, X = BuettnerFlorian$in_X)
```

`SIMLR_Large_Scale` *SIMLR Large Scale*

Description

perform the SIMLR clustering algorithm for large scale datasets

Usage

```
SIMLR_Large_Scale(X, c, k = 10, kk = 100, if.impute = FALSE,
  normalize = FALSE)
```

Arguments

X	an (m x n) data matrix of gene expression measurements of individual cells or and object of class SCESet
c	number of clusters to be estimated over X
k	tuning parameter
kk	number of principal components to be assessed in the PCA
if.impute	should I traspose the input data?
normalize	should I normalize the input data?

Value

clusters the cells based on SIMLR Large Scale and their similarities

list of 8 elements describing the clusters obtained by SIMLR, of which y are the resulting clusters: y = results of k-means clusterings, S0 = similarities computed by SIMLR, F = results from the large scale iterative procedure, ydata = data referring the the results by k-means, alphaK = clustering coefficients, val = distances from the k-nearest neighbour search, ind = indeces from the k-nearest neighbour search, execution.time = execution time of the present run

Examples

```
resized = ZeiselAmit$in_X[, 1:340]
## Not run:
SIMLR_Large_Scale(X = resized, c = ZeiselAmit$n_clust, k = 5, kk = 5)

## End(Not run)
```

ZeiselAmit *test dataset for SIMLR large scale*

Description

example dataset to test SIMLR large scale. This is a reduced version of the dataset from the work by Zeisel, Amit, et al.

Usage

```
data(ZeiselAmit)
```

Format

gene expression measurements of individual cells

Value

list of 6: in_X = input dataset as an (m x n) gene expression measurements of individual cells, n_clust = number of clusters (number of distinct true labels), true_labs = ground true of cluster assignments for each of the n_clust clusters, seed = seed used to compute the results for the example, results = result by SIMLR for the inputs defined as described, nmi = normalized mutual information as a measure of the inferred clusters compared to the true labels

Source

Zeisel, Amit, et al. "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq." Science 347.6226 (2015): 1138-1142.

Index