Package 'SIMLR'

October 18, 2017

Version 1.2.3

Date 2017-09-29

Title SIMLR: Single-cell Interpretation via Multi-kernel LeaRning

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 Single-cell RNA-seq technologies enable high throughput gene expression measurement of individual cells, and allow the discovery of heterogeneity within cell populations. Measurement of cell-to-cell gene expression similarity is critical to identification, visualization and analysis of cell populations. However, single-cell data introduce challenges to conventional measures of gene expression similarity because of the high level of noise, outliers and dropouts. We develop a novel similarity-learning framework, SIMLR (Single-cell Interpretation via Multi-kernel LeaRning), which learns an appropriate distance metric from the data for dimension reduction, clustering and visualization. SIMLR is capable of separating known subpopulations more accurately in single-cell data sets than do existing dimension reduction methods. Additionally, SIMLR demonstrates high sensitivity and accuracy on high-throughput peripheral blood mononuclear cells (PBMC) data sets generated by the GemCode single-cell technology from 10x Genomics.

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 Bo Wang [aut],
Daniele Ramazzotti [aut, cre],
Luca De Sano [aut],
Junjie Zhu [ctb],
Emma Pierson [ctb],
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.

`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_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

```
SIMLR_Large_Scale(X = ZeiselAmit$in_X, c = ZeiselAmit$n_clust, k = 5, kk = 5)
```

ZeiselAmit *test dataset for SIMLR large scale*

Description

example dataset to test SIMLR large scale, reduced version 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