Package 'SpectralTAD'

May 20, 2024

Title SpectralTAD: Hierarchical TAD detection using spectral clustering

Version 1.20.0

Description SpectralTAD is an R package designed to identify Topologically Associated Domains (TADs) from Hi-C contact matrices. It uses a modified version of spectral clustering that uses a sliding window to quickly detect TADs. The function works on a range of different formats of contact matrices and returns a bed file of TAD coordinates. The method does not require users to adjust any parameters to work and gives them control over the number of hierarchical levels to be returned.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.2.3

Imports dplyr, PRIMME, cluster, Matrix, parallel, BiocParallel, magrittr, HiCcompare, GenomicRanges, utils

Suggests BiocCheck, BiocManager, BiocStyle, knitr, rmarkdown, microbenchmark, testthat, covr

Depends R (>= 3.6)

VignetteBuilder knitr

biocViews Software, HiC, Sequencing, FeatureExtraction, Clustering

BugReports https://github.com/dozmorovlab/SpectralTAD/issues

URL https://github.com/dozmorovlab/SpectralTAD

git_url https://git.bioconductor.org/packages/SpectralTAD

git_branch RELEASE_3_19

git_last_commit a132d8c

git_last_commit_date 2024-04-30

Repository Bioconductor 3.19

Date/Publication 2024-05-19

Author Mikhail Dozmorov [aut, cre] (<https://orcid.org/0000-0002-0086-8358>),
Kellen Cresswell [aut],
John Stansfield [aut]

Maintainer Mikhail Dozmorov <mikhail.dozmorov@gmail.com>

Contents

`rao_chr20_25_rep` *Contact matrix from Rao 2014, chromosome 20, 25kb resolution*

Description

A sparse 3-column contact matrix

Usage

```
data(rao_chr20_25_rep)
```

Format

A data.frame with 3 columns and 2125980 rows:

V1 The genomic loci corresponding to a given row of the contact matrix

V2 The genomic loci corresponding to a given column of the contact matrix

V3 Number of contacts between Loci1 and Loci2

Value

A data.frame

Source

Data from Rao SS, Huntley MH, Durand NC, Stamenova EK et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014 Dec 18;159(7):1665-80. PMID: 25497547. Available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525

SpectralTAD *Hierarchical Spectral Clustering of TADs*

Description

Hierarchical Spectral Clustering of TADs

Usage

```
SpectralTAD(
  cont_mat,
  chr,
  levels = 1,
  qual_filter = FALSE,
  z_clust = FALSE,
  eigenvalues = 2,
  min_size = 5,
  window_size = 25,
  resolution = "auto",
  gap_threshold = 1,
  grange = FALSE,
  out_format = "none",
  out_path = chr
)
```

Arguments

cont_mat	Contact matrix in either sparse 3 column, n x n or n x (n+3) form where the first three columns are coordinates in BED format. If an x n matrix is used, the column names must correspond to the start point of the corresponding bin. If large mode is selected, then this matrix must be a tab-seperated n x n or n x (n+3) and it should be the path to a contact matrix. Required.
chr	The chromosome of the contact matrix being analyzed. Required.
levels	The number of levels of the TAD hierarchy to be calculated. The default setting is 1.
qual_filter	Option to turn on quality filtering which removes TADs with negative silhouette scores (poorly organized TADs). Default is FALSE.
z_clust	Option to filter sub-TADs based on the z-score of their eigenvector gaps. Default is TRUE.
eigenvalues	The number of eigenvectors to be calculated. The default and suggested setting is 2.
min_size	The minimum allowable TAD size measured in bins. Default is 5.
window_size	The size of the sliding window for calculating TADs. Smaller window sizes correspond to less noise from long-range contacts but limit the possible size of TADs

resolution	The resolution of the contact matrix. If none selected, the resolution is estimated by taking the most common distance between bins. For n x (n+3) contact matrices, this value is automatically calculated from the first three columns.
gap_threshold	Corresponds to the percentage of zeros allowed before a column/row is removed from the analysis. 1=100%, .7 = 70%, etc. Default is 1.
grange	Parameter to determine whether the result should be a GRangeList object. Defaults to FALSE
out_format	Specifies the format of the file which SpectralTAD outputs. If "none, no file is output. "juicebox" or "bedpe" returns a bedpe file compatible with juicebox. "hicexplorer" or "bed" returns a bed file compatible with hicexplorer. Default is none
out_path	Path of output file. Default is the chromosome

Details

Given a sparse 3 column, an n x n contact matrix, or n x (n+3) contact matrix, SpectralTAD returns a list of TAD coordinates in BED format. SpectralTAD works by using a sliding window that moves along the diagonal of the contact matrix. By default, we use the biologically relevant maximum TAD size of 2Mb and minimum size of 5 bins to determine the size of this window. Within each window, we calculate a Laplacian matrix and determine the location of TAD boundaries based on gaps between eigenvectors calculated from this matrix. The number of TADs in a given window is calculated by finding the number that maximizes the silhouette score. A hierarchy of TADs is created by iteratively applying the function to sub-TADs. The number of levels in each hierarchy is determined by the user.

Value

A list where each entry is in BED format corresponding to the level of the hierarchy.

Examples

```
#Read in data
data("rao_chr20_25_rep")
#Find TADs
spec_table <- SpectralTAD(rao_chr20_25_rep, chr= 'chr20')
```

SpectralTAD_Par *Parallelized Hierarchical Spectral Clustering of TADs*

Description

Parallelized Hierarchical Spectral Clustering of TADs

Usage

```
SpectralTAD_Par(
  cont_list,
  chr,
  levels = 1,
  qual_filter = FALSE,
  z_clust = FALSE,
  eigenvalues = 2,
  min_size = 5,
  window_size = 25,
  resolution = "auto",
  grange = FALSE,
  gap_threshold = 1,
  cores = "auto",
  labels = NULL
)
```

Arguments

cont_list	List of contact matrices where each is in either sparse 3 column, n x n or n x (n+3) form, where the first 3 columns are chromosome, start and end coordinates of the regions. If an x n matrix is used, the column names must correspond to the start point of the corresponding bin. Required.
chr	Vector of chromosomes in the same order as their corresponding contact matrices. Must be same length as cont_list. Required.
levels	The number of levels of the TAD hierarchy to be calculated. The default setting is 1.
qual_filter	Option to turn on quality filtering which removes TADs with negative silhouette scores (poorly organized TADs). Default is FALSE.
z_clust	Option to filter sub-TADs based on the z-score of their eigenvector gaps. Default is TRUE.
eigenvalues	The number of eigenvectors to be calculated. The default and suggested setting is 2.
min_size	The minimum allowable TAD size measured in bins. Default is 5.
window_size	The size of the sliding window for calculating TADs. Smaller window sizes correspond to less noise from long-range contacts but limit the possible size of TADs
resolution	The resolution of the contact matrix. If none selected, the resolution is estimated by taking the most common distance between bins. For n x (n+3) contact matrices, this value is automatically calculated from the first 3 columns.
grange	Parameter to determine whether the result should be a GRangeList object. Defaults to FALSE
gap_threshold	Corresponds to the percentage of zeros allowed before a column/row is removed from analysis. 1=100%, .7 = 70%, etc. Default is 1.
cores	Number of cores to use. Defaults to total available cores minus one.

`labels` Vector of labels used to name each contact matrix. Must be same length as cont_list. Default is NULL.

Details

This is the parallelized version of the SpectralTAD() function. Given a sparse 3 column, an n x n contact matrix, or n x (n+3) contact matrix, SpectralTAD returns a list of TAD coordinates in BED format. SpectralTAD works by using a sliding window that moves along the diagonal of the contact matrix. By default we use the biologically relevant maximum TAD size of 2Mb and minimum size of 5 bins to determine the size of this window. Within each window, we calculate a Laplacian matrix and determine the location of TAD boundaries based on gaps between eigenvectors calculated from this matrix. The number of TADs in a given window is calculated by finding the number that maximize the silhouette score. A hierarchy of TADs is created by iteratively applying the function to sub-TADs. The number of levels in each hierarchy is determined by the user.

Value

List of lists where each entry is a list of data frames or GRanges in BED format corresponding to TADs seperated by hierarchies

Examples

```
#Read in data
data("rao_chr20_25_rep")
#Make a list of matrices
mat_list = list(rao_chr20_25_rep, rao_chr20_25_rep)
#Make a vector of chromosomes
chr = c("chr20", "chr20")
#Make a vector of labels
labels = c("run1", "run2")
spec_table <- SpectralTAD_Par(mat_list, chr= chr, labels = labels, cores = 2)
```

Index