Package ‘SpectralTAD’

November 6, 2019

Title SpectralTAD: Hierarchical TAD detection using spectral clustering

Version 1.2.0

Author Kellen Cresswell <cresswellkg@vcu.edu>,
John Stansfield <stansfieldjc@vcu.edu>,
Mikhail Dozmorov <mikhail.dozmorov@vcuhealth.org>

Maintainer Kellen Cresswell <cresswellkg@vcu.edu>

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 6.1.1

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

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

Depends R (>= 3.6)

VignetteBuilder knitr

biocViews Software, HiC, Sequencing, FeatureExtraction, Clustering

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

git_branch RELEASE_3_10

git_last_commit 9bde8c8

Date/Publication 2019-11-05

R topics documented:

rao_chr20_25_rep .......................................................... 2
SpectralTAD ............................................................... 2
SpectralTAD_Par ........................................................... 4
rao_chr20_25_rep  Contact matrix from Rao 2014, chromosome 20, 25kb resolution

Description
A sparse 3-column contact matrix

Usage
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

SpectralTAD  Hierarchical Spectral Clustering of TADs

Description
Hierarchical Spectral Clustering of TADs

Usage
SpectralTAD(cont_mat, chr, levels = 1, qual_filter = FALSE, z_clust = TRUE, eigenvalues = 2, min_size = 5, resolution = “auto”, gap_threshold = 1, grange = FALSE)
**SpectralTAD**

**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. 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.

- **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 GRanges object. Defaults to FALSE

**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**

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

Description
Parallelized Hierarchical Spectral Clustering of TADs

Usage
SpectralTAD_Par(cont_list, chr, levels = 1, qual_filter = FALSE, z_clust = TRUE, eigenvalues = 2, min_size = 5, 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 n 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.
- **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 separated by hierarchies

**Examples**

```r
# 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)
```
Index

*Topic datasets
  rao_chr20_25_rep, 2
  rao_chr20_25_rep, 2
  SpectralTAD, 2
  SpectralTAD_Par, 4