# Package 'rGADEM'

# September 19, 2024

Type Package

Title de novo motif discovery		
<b>Version</b> 2.53.0		
<b>Date</b> 2014-04-01		
Author Arnaud Droit, Raphael Gottardo, Gordon Robertson and Leiping Li		
Maintainer Arnaud Droit <arnaud.droit@crchuq.ulaval.ca></arnaud.droit@crchuq.ulaval.ca>		
<b>Depends</b> R (>= 2.11.0), Biostrings, IRanges, BSgenome, methods, seqLogo		
Imports Biostrings, GenomicRanges, methods, graphics, seqLogo		
Suggests BSgenome. Hsapiens. UCSC.hg19, rtracklayer		
<b>Description</b> rGADEM is an efficient de novo motif discovery tool for large-scale genomic sequence data. It is an open-source R package, which is based on the GADEM software.		
License Artistic-2.0		
biocViews Microarray, ChIPchip, Sequencing, ChIPSeq, MotifDiscovery		
git_url https://git.bioconductor.org/packages/rGADEM		
git_branch devel		
git_last_commit 192d081		
git_last_commit_date 2024-04-30		
Repository Bioconductor 3.20		
Date/Publication 2024-09-18		
Contents		
align-class		
GADEM		
motif-class		
parameters-class		
readPWMfile		
Index		

2 align-class

align-class

Class "align"

# Description

This object contains the individual motifs identified but also the location (seqID and position) of the sites in the original sequence data. It also included the spaced dyad from which the motifs is derived, PWM score p-value cuttoff for the run.

# Objects from the Class

Objects can be created by calls of the form new("align", ...).

#### **Slots**

```
seq : Motif identified .
```

**chr** :Chromosome identified.

start : Sequence start.

end :Sequence end.

strand :Strand position.

seqID :Sequence identification.

pos :Position identification.

**pval** :p-Value for each identification.

fastaHeader :Fasta accession.

# Author(s)

Arnaud Droit <arnaud.droit@crchuq.ulaval.ca>

# See Also

```
gadem , motif, parameters
```

```
showClass("align")
```

GADEM 3

GADEM	Motif Analysis with rGADEM	

#### **Description**

It is an R implementation of GADEM, a powerful computational tools for de novo motif discovery.

# Usage

```
GADEM(Sequences, seed=1, genome=NULL, verbose=FALSE, numWordGroup=3, numTop3mer=20, numTop4mer=40, numTop5mer=60, numGeneration=5, populationSize=100, pValue=0.0002, eValue=0.0, extTrim=1, minSpaceWidth=0, maxSpaceWidth=10, useChIPscore=0, numEM=40, fEM=0.5, widthWt=80, fullScan=0, slideWinPWM=6, stopCriterion=1, numBackgSets=10, weightType=0, bFileName="NULL", Spwm="NULL", minSites =-1, maskR=0, nmotifs=25)
```

#### **Arguments**

Sequences Sequences from BED or FASTA file are converted into XString object view

seed When a seed is specified, the run results are deterministic

genome Specify the genome

verbose Print immediate results on screen [TRUE-yes (default), FALSE-no]. These re-

sults include the motif consensus sequence, number of sites (in sequences sub-

jected to EM optimization, see -fEM, above), and ln(E-value).

numWordGroup number of non-zero k-mer groups

numTop3mer Number of top-ranked trimers for spaced dyads (default: 20).

NumTop4mer Number of top-ranked tetramers for spaced dyads (default: 40).

NumTop5mer Number of top-ranked pentamers for spaced dyads (default: 60).

numGeneration Number of genetic algorithm (GA) generations (default: 5).

populationSize GA population size (default: 100). Both default settings should work well for

most datasets (ChIP-chip and ChIP-seq). The above two arguments are ignored in a seeded analysis, because spaced dyads and GA are no longer needed (num-Generation is set to 1 and populationSize is set to 10 internally, corresponding

to the 10 maxp choices).

pValue P-value cutoff for declaring BINDING SITES (default: 0.0002). Depending on

data size and the motif, you might want to assess more than one value. For ChIP-seq data (e.g., 10 thousand +/-200-bp max-center peak cores), p=0.0002 often seems appropriate. However, short motifs may require a less stringent setting.

eValue ln(E-value) cutoff for selecting MOTIFS (default: 0.0). If a seeded analysis fails

to identify the expected motif, run GADEM with -verbose 1 to show motif ln(E-value)s on screen, then rerun with a larger ln(E-value) cutoff. This can help in identifying short and/or low abundance motifs, for which the default E-value

threshold may be too low.

extTrim Base extension and trimming (1 -yes, 0 -no) (default: 1).

minSpaceWidth Minimal number of unspecified nucleotides in spaced dyads (default: 0).

4 GADEM

maxSpaceWidth Maximal number of unspecified nucleotides in spaced dyads (default: 10). minSpaceWidth

and maxSpaceWidth control the lengths of spaced dyads, and, with exTrim, control motif lengths. Longer motifs can be discovered by setting maxSpaceWidth

to larger values (e.g. 50).

useChIPscore Use top-scoring sequences for deriving PWMs. Sequence (quality) scores are

stored in sequence header (see documentation). 0 - no (default, randomly select

sequences), 1 - yes.

numEM Number of EM steps (default: 40). One might want to set it to a larger value

(e.g. 80) in a seeded run, because such runs are fast.

fEM Fraction of sequences used in EM to obtain PWMs in an unseeded analysis

(default: 0.5). For unseeded motif discovery in a large dataset (e.g. >10 million nt), one might want to set -fEM to a smaller value (e.g., 0.3 or 0.4) to reduce run

time.

widthWt For -posWt 1 or 3, width of central sequence region with large EM weights for

PWM optimization (default: 50). This argument is ignored when weightType is

0 (uniform prior) or 2 (Gaussian prior).

fullScan GADEM keeps two copies of the input sequences internally: one (D) for discov-

ering PWMs and one (S) for scanning for binding sites using the PWMs Once a motif is identified, its instances in set D are always masked by Ns. However, masking motif instances in set S is optional, and scanning unmasked sequences

allows sites of discovered motifs to overlap.

slideWinPWM sliding window for comparing pwm similarity (default : 6).

stopCriterion Number of generations without new motifs before stopping analysis.

numBackgSets Number of sets of background sequences (default: 10). The background se-

quences are simulated using the [a,c,g,t] frequencies in the input sequences, with length matched between the two sets. The background sequences are used as the

random sequences for assessing motif enrichment in the input data.

weightType Weight profile for positions on the sequence. 0 - no weight (uniform spatial

prior, default), 1 (gaussian prior) and 2 (triangle prior) - small or zero weights for the ends and large weights for the center (e.g. the center 50 bp). Consider using 1 or 2 if you expect strong central enrichment (as in ChIP-seq) and your

sequences are long(e.g. >200 bp).

bFileName Reading user-specified background models.

Spwm File name for the seed PWM, when a seeded approach is used. can be used as

the starting PWM for the EM algorithm. This will help find an expected motif and is much faster than unseeded de novo discovery. Also, when a seed PWM is specified, the run results are deterministic, so only a single run is needed (repeat runs with the same settings will give identical results). In contrast, unseeded runs are stochastic, and we recommend comparing results from several repeat

runs.

minSites Minimal number of sites required for a motif to be reported (default: num-

Seq/20)

maskR Mask low-complexity sequences or repeats; 'aaaaaaaaa', 'ttttttt', 'cacacaca',

'tgtgtgtg', 'tatatatat', 'ggaggaggagga', 'gaggaggaggag', 'agaagaagaaga', 'ctcctcctcctc', 'tcctcctcctc', 'tc

or 'cagcagcagcag' (default: 0-no masking,1-masking)

nmotifs Number of motifs sought (default: 25)

#### Author(s)

Arnaud Droit <arnaud.droit@crchuq.ulaval.ca>

gadem-class 5

#### **Examples**

```
library(BSgenome.Hsapiens.UCSC.hg19)
library(rtracklayer)
BedFile<- system.file("extdata","Test_100.bed",package="rGADEM")
Sequences<-import(BedFile)
gadem<-GADEM(Sequences,verbose=1,genome=Hsapiens)</pre>
```

gadem-class

Class "gadem"

# **Description**

This object contains all gadem output information.

# **Objects from the Class**

Objects can be created by calls of the form new("gadem", ...).

# **Slots**

```
motifList List of input PWM. parameters List of rGADEM parameters.
```

#### Methods

```
[ signature(x = "gadem"): subset gadem object.
[[ signature(x = "gadem"): subset gadem object.
nMotifs signature(x = "gadem"): Number of motifs identified
names signature(x = "gadem"): Assign motifs names.
dim signature(x = "gadem"): Number of sequences identified for each motifs.
consensus signature(x = "gadem"):Sequence of consensus motifs.
nOccurrences signature(x = "gadem"):View of PWMs.
plot, gadem-method signature(x = "gadem"):Plot.
startPos signature(x = "gadem"):Start position for each sequences.
endPos signature(x = "gadem"):End position for each sequences.
getPWM signature(x = "gadem"):End position for each sequences.
```

# Author(s)

Arnaud Droit <arnaud.droit@crchuq.ulaval.ca>

#### See Also

```
motif, align, parameters
```

```
showClass("gadem")
```

6 parameters-class

motif-class

Class "motif"

# Description

This object contains PWM, motif consensus, motif length and all aligned sequences for a specific motif

# **Objects from the Class**

Objects can be created by calls of the form new("motif\_gadem", ...).

#### **Slots**

```
pwm: PWM results.consensus: Sequences consensus.alignList: List of sequences alignment.name: Name of sequences.
```

# Author(s)

Arnaud Droit <arnaud.droit@crchuq.ulaval.ca>

#### See Also

```
gadem, align, parameters
```

# **Examples**

```
showClass("gadem")
```

parameters-class

Class "parameters"

# Description

This object contains contains parameters of GADEM analysis

# **Objects from the Class**

Objects can be created by calls of the form new("motif\_gadem", ...).

parameters-class 7

#### **Slots**

numWordGroup: Number of non-zero k-mer groups. **numTop3mer**: Number of top-ranked trimers for spaced dyads (default: 20). **verbose**: Print immediate results on screen [1-yes (default), 0-no]. **numTop4mer**: Number of top-ranked tetramers for spaced dyads (default: 40). **numTop5mer**: Number of top-ranked pentamers for spaced dyads (default: 60). **numGeneration**: Number of genetic algorithm (GA) generations (default: 5). **populationSize**: GA population size (default: 100). **pValue**: P-value cutoff for declaring BINDING SITES (default: 0.0002). eValue :ln(E-value) cutoff for selecting MOTIFS (default: 0.0). **extTrim**: Base extension and trimming (1 -yes, 0 -no) (default: 1). minSpaceWidth: Minimal number of unspecified nucleotides in spaced dyads (default: 0). maxSpaceWidth: Maximal number of unspecified nucleotides in spaced dyads (default: 10). **useChIPscore**: Use top-scoring sequences for deriving PWMs. numEM: Number of EM steps (default: 40). **fEM**: Fraction of sequences used in EM to obtain PWMs in an unseeded analysis (default: 0.5). widthWt: For -posWt 1 or 3, width of central sequence region with large EM weights for PWM optimization (default: 50). **fullScan**: GADEM keeps two copies of the input sequences internally. slideWinPWM :Sliding window for comparing pwm similarity (default : 6). stopCriterion numBackgSets: Number of sets of background sequences (default: 10). weight Type: Weight profile for positions on the sequence. **bFileName**: Reading user-specified background models. Spwm: File name for the seed PWM, when a seeded approach is used. **nSequences**: Number of input sequences. maskR : Mask low-complexity sequences or repeats. nmotifs: Maximal number of motifs sought. Author(s) Arnaud Droit <arnaud.droit@crchug.ulaval.ca>

# See Also

```
gadem, align, motif
```

```
showClass("parameters")
```

8 readPWMfile

readPWMfile

Read Transfac File

# Description

This function is use to read standard Transfac type file.

# Usage

```
readPWMfile(file)
```

# **Arguments**

file

Transfac file's name.

#### **Details**

This function is designed to read standard Transfac type file. For more information about the format, please refere to http://mcast.sdsc.edu/doc/transfac-format.html

# Value

A list of matrix.

# Author(s)

Arnaud Droit <<arnaud.droit@ircm.qc.ca>>

```
#####Database and Scores#####
path <- system.file("extdata","jaspar2009.txt",package="rGADEM")
jaspar <- readPWMfile(path)</pre>
```

# **Index**

```
* GADEM
    GADEM, 3
* MOTIFS
    GADEM, 3
* classes
    align-class, 2
    gadem-class, 5
    motif-class, 6
    parameters-class, 6
* misc
    readPWMfile, 8
[,gadem,ANY,ANY,ANY-method
        (gadem-class), 5
[,gadem,ANY,ANY-method(gadem-class), 5
[,gadem-method(gadem-class), 5
[[, gadem, ANY, ANY-method (gadem-class), 5
[[, gadem-method (gadem-class), 5
align, 5-7
align (align-class), 2
align-class, 2
consensus (gadem-class), 5
consensus, gadem-method (gadem-class), 5
dim, gadem-method (gadem-class), 5
endPos (gadem-class), 5
endPos, gadem-method (gadem-class), 5
GADEM, 3
gadem, 2, 6, 7
gadem (gadem-class), 5
gadem-class, 5
getPWM (gadem-class), 5
getPWM, gadem-method (gadem-class), 5
getPWM, motif-method (gadem-class), 5
length, gadem-method (gadem-class), 5
motif, 2, 5, 7
motif (motif-class), 6
motif-class, 6
names, gadem-method (gadem-class), 5
```

```
names<-,gadem-method(gadem-class),5</pre>
nMotifs (motif-class), 6
nMotifs, gadem-method (gadem-class), 5
nOccurrences (gadem-class), 5
nOccurrences, gadem-method
        (gadem-class), 5
parameters, 2, 5, 6
parameters (parameters-class), 6
parameters, gadem-method (gadem-class), 5
parameters-class, 6
plot, gadem, ANY-method (gadem-class), 5
plot,gadem-method(gadem-class), 5
plot, motif, ANY-method (gadem-class), 5
readPWMfile, 8
startPos (gadem-class), 5
startPos, gadem-method (gadem-class), 5
summary, list-method (gadem-class), 5
```