Package 'PWMEnrich'

October 9, 2013

Imports seqLogo, gdata, evd

Maintainer Robert Stojnic <robert.stojnic@gmail.com>

License GPL-3

Title PWM enrichment analysis

Type Package

LazyLoad yes

Author Robert Stojnic, with contributions from Diego Diez

Description Asses the enrichment of already known PWMs (from JASPAR and MotifDb) in DNA sequences. The package implements multiple algorithms, including fixed-threshold (Z-score) and threshold-free (Lognormal normalization and Clover) methods. These can be applied to a single sequence (e.g. enhancer of interest) or a group of sequences (e.g. a set of ChIP-chip/seq peaks). The output is a ranked list of PWMs according to their level of enrichment compared to genomic background. Custom sets of PWMs and genomic background are also supported.

Version 2.2.0

biocViews Bioinformatics, SequenceMatching, GenomicSequence, Software

Date 2013-03-08

Depends methods, grid, BiocGenerics, Biostrings,

Suggests
MotifDb, BSgenome.Dmelanogaster.UCSC.dm3,PWMEnrich.Dmelanogaster.background, testthat, gtools, parallel

Collate
'AllDataClasses.R' 'AllGenerics.R' 'background.R' 'clover.R''diff.R' 'misc.R' 'MotifEnrichmentResults-methods.R''options.R' 'plot.R' 'PWMBackground-methods.R' 'PWM-methods.R''pwm.R' 'readData.R' 'seqLogoSupp.R' 'similarity.R'

R topics documented:

`.inputParamMotifs` *Normalizes the motifs input argument for multiple functions...*

Description

Normalizes the motifs input argument for multiple functions

Usage

```
.inputParamMotifs(motifs)
```

Arguments

motifs	a list of motifs either as frequency matrices (PFM) or as PWM objects. If PFMs are specified they are converted to PWMs using uniform background.

.inputParamSequences *Normalize the sequences input argument...*

Description

Normalize the sequences input argument

Usage

```
.inputParamSequences(sequences)
```

Arguments

sequences	a set of sequences to be scanned, a list of DNAString or other scannable objects

---

.inputPFMfromMatrixOrPWM *Check the frequency matrix input parameter for motifSimilarity...*

Description

Check the frequency matrix input parameter for motifSimilarity

Usage

```
.inputPFMfromMatrixOrPWM(m)
```

Arguments

m	either a PWM object or a matrix

Value

corresponding PFM

.normalize.bg.seq *check consistency of bg...*

Description

check consistency of bg.seq input parameter

Usage

```
.normalize.bg.seq(bg.seq)
```

Arguments

bg.seq	a set of background sequences, either a list of DNAString object or DNAStringSet object

.normargPfm *Input parameter normalization for PWMUnscaled...*

Description

Input parameter normalization for PWMUnscaled

Usage

```
.normargPfm(x)
```

Arguments

x	a frequency matrix

Details

This function is from Biostrings package. A Position Frequency Matrix (PFM) is also represented as an ordinary matrix. Unlike a PWM, it must be of type integer (it will typically be the result of consensusMatrix()).

.normargPriorParams *Input parameter normalization function for PWMUnscaled...*

Description

Input parameter normalization function for PWMUnscaled

Usage

```
.normargPriorParams(prior.params)
```

Arguments

prior.params	Typical 'prior.params' vector: c(A=0.25, C=0.25, G=0.25, T=0.25)

Details

This function is from Biostrings package

affinitySequenceSet *Calculate total affinity over a set of sequences...*

Description

Calculate total affinity over a set of sequences

Usage

```
affinitySequenceSet(scores, seq.len, pwm.len)
```

Arguments

scores	affinity scores for individual sequences
seq.len	lengths of sequences
pwm.len	lengths of PWMs

`cloverPvalue1seq` *Calculate the Clover P-value as described in the Clover paper...*

Description

Calculate the Clover P-value as described in the Clover paper

Usage

```
cloverPvalue1seq(scores, seq.len, pwm.len, bg.fwd, bg.rev, B=1000, verbose=TRUE, clover)
```

Arguments

`scores`	the affinity scores for individual sequences
`seq.len`	lengths of sequences
`pwm.len`	lengths of PWMs
`bg.fwd`	the raw score of forward strand
`bg.rev`	the raw scores of reverse strand
`B`	the number of random replicates
`verbose`	if to give verbose progress reports
`clover`	the clover scores if already calculated

Details

This function only take one background sequence as input, it also just calculates the P-value so it is more efficient.

Value

P-value

`cloverScore` *Calculate the Clover score using the recursive formula from Frith et al...*

Description

Calculate the Clover score using the recursive formula from Frith et al

Usage

```
cloverScore(scores, lr3=FALSE, verbose=FALSE)
```

Arguments

scores	a matrix of average odds scores, where columns are motifs, and rows sequences
lr3	if to return a matrix of LR3 scores, where columns correpond to motifs, and rows to subset sizes
verbose	if to produce verbose output of progress

Value

the LR4 score, which is the mean of LR3 scores over subset sizes

---

colMedians *Calculate medians of columns...*

Description

Calculate medians of columns

Usage

```
colMedians(x)
```

Arguments

x	a matrix

---

colSds *Calculate standard deviations of columns...*

Description

Calculate standard deviations of columns

Usage

```
colSds(x)
```

Arguments

x	a matrix

concatenateSequences *Concatenata DNA sequences into a single character object...*

Description

Concatenata DNA sequences into a single character object

Usage

```
concatenateSequences(sequences)
```

Arguments

sequences	either a list of DNAString objects, or a DNAStringSet

Value

a single character string

cutoffZscore *Z-score calculation for cutoff hits...*

Description

Z-score calculation for cutoff hits

Usage

```
cutoffZscore(scores, seq.len, pwm.len, bg.P)
```

Arguments

scores	the hit counts for the sequences
seq.len	the length distribution of sequences
pwm.len	the length distribution of the PWMs
bg.P	background probabilities of observing a motif hit at nucleotide resolution (scaled to sequence length, not 2 * length)

Details

The Z-score is calculated separately for each sequence

Value

Z-score

`cutoffZscoreSequenceSet` *Z-score calculation for cutoff hits for group of sequences...*

Description

Z-score calculation for cutoff hits for group of sequences

Usage

```
cutoffZscoreSequenceSet(scores, seq.len, pwm.len, bg.P)
```

Arguments

`scores`	the hit counts for the sequences
`seq.len`	the length distribution of sequences
`pwm.len`	the length distribution of the PWMs
`bg.P`	background probabilities of observing a motif hit at nucleotide resolution

Details

The Z-score is calculated as if the sequence came for one very long sequence

Value

Z-score

`divideRows` *Divide each row of a matrix with a vector...*

Description

Divide each row of a matrix with a vector

Usage

```
divideRows(m, v)
```

Arguments

`m`	matrix to be divided
`v`	the vector to use for division

DNAStringSetToList *Convert DNAStringSet to list of DNAString objects...*

Description

Convert DNAStringSet to list of DNAString objects

Usage

```
DNAStringSetToList(x)
```

Arguments

x	an object of class DNAStringSet

Details

as.list doesn't seem to always work for DNAStringSets, so implementing this ourselves.

empiricalPvalue *Calculate the empirical P-value by affinity of cutoff.*

Description

Calculate the empirical P-value by affinity of cutoff.

Usage

```
empiricalPvalue(scores, seq.len, pwm.len, bg.fwd, bg.rev, cutoff, B=10000,
    verbose=FALSE, exact.length=FALSE)
```

Arguments

scores	the scores obtained for the sequence
seq.len	the length of the sequence, if a single value will take a single sequence of given length. If a vector of values, will take sequences of given lengths and joint them together
pwm.len	the lengths of PWMs
bg.fwd	raw odds scores for the forward strand of background
bg.rev	raw odds scores for the reverse strand of background
cutoff	if not NULL, will use hit count above this cutoff. The cutoff should be specified in log2.
B	the number of random replicates
verbose	if to give verbose progress reports

exact.length	if to take into consideration that the actual sequence lengths differ for different PWMs. For very long sequences (i.e. seq.len » pwm.len) this make very little difference, however the run time with exact.length is much longer.

Details

This is the new backend function for empirical P-values for either affinity or cutoff. The function only works on single sequences.

`empiricalPvalueSequenceSet`

Empirical P-value for a set of sequences...

Description

Empirical P-value for a set of sequences

Usage

```
empiricalPvalueSequenceSet(scores, seq.len, pwm.len, bg.fwd, bg.rev, cutoff, B=10000,
    verbose=FALSE)
```

Arguments

scores	a matrix of scores, rows for sequences, columns for PWMs
seq.len	the lengths of sequences
pwm.len	the lengths of PWMs
bg.fwd	raw odds scores for the forward strand of background
bg.rev	raw odds scores for the reverse strand of background
cutoff	if not NULL, will use hit count above this cutoff. The cutoff should be specified in log2.
B	the number of random replicates
verbose	if to give verbose progress reports

Details

Calculate empirical P-value for a set of sequences, using either affinity or cutoff. When cutoff is used, the score is a number of motif hits above a certain log-odds cutoff.

getBackgroundFrequencies *Get the four nucleotides background frequencies...*

Description

Get the four nucleotides background frequencies

Usage

```
getBackgroundFrequencies(organism="dm3", pseudo.count=1, quick=FALSE)
```

Arguments

organism	either a name of the organisms for which the background should be compiled (currently only supported name is "dm3" for Drosophila Melanogaster), or a BSgenome object (see BSgenome package).
pseudo.count	the number to which the frequencies sum up to, by default 1
quick	if to preform fitting on a reduced set of 100 promoters. This will not give as good results but is much quicker than fitting to all the promoters (~10k). Usage of this parameter is recommended only for testing and rough estimates.

Details

Estimate the background frequencies of A,C,G,T on a set of promoters from an organism

Author(s)

Robert Stojnic, Diego Diez

Examples

```
## Not run:
getBackgroundFrequencies("dm3")

## End(Not run)
```

gevPerSequence *Apply GEV background normalization per every sequence...*

Description

Apply GEV background normalization per every sequence

Usage

```
gevPerSequence(scores, seq.len, pwm.len, bg.loc, bg.scale, bg.shape)
```

Arguments

scores	affinity scores for the PWMs, can contain scores for more than one sequence (as rows), P-values are extracted separately
seq.len	the length distribution of the sequences
pwm.len	the lengths of PWMs
bg.loc	list of linear regression for location parameter
bg.scale	list of linear regression for scale parameter
bg.shape	list of linear regression for shape parameter

keepFinite *Replace all infinite values by 0...*

Description

Replace all infinite values by 0

Usage

```
keepFinite(x)
```

Arguments

x	a vector of values

`logNormPval` *Calculate the P-value from lognormal distribution with background of equal length...*

Description

Calculate the P-value from lognormal distribution with background of equal length

Usage

```
logNormPval(scores, seq.len, pwm.len, bg.mean, bg.sd, bg.len)
```

Arguments

`scores`	affinity scores for the PWMs, can contain scores for more than one sequence (as rows), P-values are extracted separately
`seq.len`	the length distribution of the sequences
`pwm.len`	the leggths of PWMs
`bg.mean`	the mean values from the background for PWMs
`bg.sd`	the sd values from the background
`bg.len`	the length distribution of the background (we currently support only constant length)

`logNormPvalSequenceSet` *Lognormal P-value for a set of sequences...*

Description

Lognormal P-value for a set of sequences

Usage

```
logNormPvalSequenceSet(scores, seq.len, pwm.len, bg.mean, bg.sd, bg.len)
```

Arguments

`scores`	a matrix of per-sequence affinity scores
`seq.len`	lengths of sequences
`pwm.len`	lengths of pwms
`bg.mean`	mean background at length of bg.len
`bg.sd`	standard deviation of background at length of bg.len
`bg.len`	the length for which mean and sd are calculated

Value

P-value

---

makeBackground *Make a background for a set of position frequency matrices...*

Description

Make a background for a set of position frequency matrices

Usage

```
makeBackground(motifs, organism="dm3", type="logn", quick=FALSE, ...)
```

Arguments

motifs	a list of position frequency matrices (4xL matrices)
organism	either a name of the organisms for which the background should be compiled (currently only supported name is "dm3" for Drosophila Melanogaster), or a BSgenome object (see BSgenome package).
type	the type of background to be compiled. Possible types are: • "logn" - estimate a lognormal background • "cutoff" - estimate a Z-score background with fixed log-odds cutoff (in log2) • "pval" - estimate a Z-score background with a fixed P-value cutoff. Note that this may require a lot of memory since the P-value of motif hits is first estimated from the empirical distribution. • "empirical" - create an empirical P-value background. Note that this may require a lot of memory (up to 10GB in default "slow" mode (quick=FALSE) for 126 JASPAR motifs and 1000 D. melanogaster promoters). • "GEV" - estimate a generalized extreme value (GEV) distribution background by fitting linear regression to distribution parameters in log space
quick	if to preform fitting on a reduced set of 100 promoters. This will not give as good results but is much quicker than fitting to all the promoters (~10k). Usage of this parameter is recommended only for testing and rough estimates.
...	other named parameters that backend function makePWM***Background functions take.

Details

This is a convenience front-end function to compile new backgrounds for a set of PFMs. Currently only supports D. melanogaster, but in the future should support other common organisms as well.

Author(s)

Robert Stojnic, Diego Diez

Examples

```
# load in the two example de-novo motifs
motifs = readMotifs(system.file(package="PWMEnrich", dir="extdata", file="example.transfac"), remove.acc=TRUE)

## Not run:
# construct lognormal background
bg.logn = makeBackground(motifs, organism="dm3", type="logn")

# alternatively, any BSgenome object can also be used
if(require("BSgenome.Dmelanogaster.UCSC.dm3"))
bg.logn = makeBackground(motifs, organism=Dmelanogaster, type="logn")

# construct a Z-score of hits with P-value background
bg.pval = makeBackground(motifs, organism="dm3", type="pval", p.value=1e-3)

# now we can use them to scan for enrichment in sequences (in this case there is a consensus Tin binding site)
motifEnrichment(DNAString("TGCATCAAGTGTGTAGTG"), bg.logn)
motifEnrichment(DNAString("TGCATCAAGTGTGTAGTG"), bg.pval)

## End(Not run)
```

`makePriors` *Make priors from background sequences...*

Description

Make priors from background sequences

Usage

```
makePriors(bg.seq, bg.pseudo.count)
```

Arguments

`bg.seq`	a set of background sequences
`bg.pseudo.count`	the total pseudocount shared between nucleotides

Details

These priors serve both as background nucleotide frequencies and pseudo-counts for PWMs.

Examples

```
# some example sequences
sequences = list(DNAString("AAAGAGAGTGACCGATGAC"), DNAString("ACGATGAGGATGAC"))
# make priors with pseudo-count of 1 shared between them
makePriors(sequences, 1)
```

makePWMCutoffBackground *Make a cutoff background...*

Description

Make a cutoff background

Usage

```
makePWMCutoffBackground(bg.seq, motifs, cutoff=log2(exp(4)), bg.pseudo.count=1, bg.source="",
    verbose=TRUE)
```

Arguments

bg.seq	a set of background sequences, either a list of DNAString object or DNAStringSet object
motifs	a set of motifs, either a list of frequency matrices, or a list of PWM objects. If frequency matrices are given, the background distribution fitted from bg.seq. Same ratios are used for pseudo counts that sum up to bg.pseudo.count for the 4 nucleotides.
cutoff	the cutoff at which the background should be made, i.e. at which a motif hit is called significant
bg.pseudo.count	the pseudo count which is shared between nucleotides when frequency matrices are given
bg.source	a free-form textual description of how the background was generated
verbose	if to produce verbose output

Details

Make a background based on number of motifs hits above a certain threshold.

Examples

```
## Not run:
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel.PFM)

# make background for MotifDb motifs using 2kb promoters of all D. melanogaster transcripts using cutoff of 5
if(require("BSgenome.Dmelanogaster.UCSC.dm3"))
makePWMCutoffBackground(Dmelanogaster$upstream2000, MotifDb.Dmel.PFM, cutoff=log2(exp(5)))
}

## End(Not run)
```

makePWMEmpiricalBackground *Make an empirical P-value background...*

Description

Make an empirical P-value background

Usage

```
makePWMEmpiricalBackground(bg.seq, motifs, bg.pseudo.count=1, bg.source="", verbose=TRUE, ...)
```

Arguments

bg.seq	a set of background sequences, either a list of DNAString object or DNAStringSet object
motifs	a set of motifs, either a list of frequency matrices, or a list of PWM objects. If frequency matrices are given, the background distribution fitted from bg.seq. Same ratios are used for pseudo counts that sum up to bg.pseudo.count for the 4 nucleotides.
bg.pseudo.count	the pseudo count which is shared between nucleotides when frequency matrices are given
bg.source	a free-form textual description of how the background was generated
verbose	if to produce verbose output
...	currently unused (this is for convenience for makeBackground function)

Details

Make a background appropriate for empirical P-value calculation. The provided set of background sequences is contcatenated into a single long sequence which is then scanned with the motifs and raw scores are saved. This object can be very large.

For reliable P-value calculation the size of the background set needs to be at least seq.len / min.P.value. For instance, to get P-values at a resolution of 0.001 for a single sequence of 500bp, we would need a background of at least 500/0.001 = 50kb. This ensures that we can make 1000 independent 500bp samples from this background to properly estimate the P-value. For a group of sequences, we would take seq.len to be the total length of all sequences in a group.

Examples

```
## Not run:
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel.PFM)

# make empirical background by saving raw scores for each bp in the sequence - this can be very large in memory!
if(require("BSgenome.Dmelanogaster.UCSC.dm3"))
```

```
makePWMEmpiricalBackground(Dmelanogaster$upstream2000[1:100], MotifDb.Dmel.PFM)
}

## End(Not run)
```

`makePWMGEVBackground` *Make a GEV background distribution...*

Description

Make a GEV background distribution

Usage

```
makePWMGEVBackground(bg.seq, motifs, bg.pseudo.count=1, bg.len=seq(200, 2000, 200),
    bg.source="", verbose=TRUE, fit.log=TRUE)
```

Arguments

`bg.seq`	a set of background sequences, either a list of DNAString object or DNAStringSet object
`motifs`	a set of motifs, either a list of frequency matrices, or a list of PWM objects. If frequency matrices are given, the background distribution fitted from bg.seq. Same ratios are used for pseudo counts that sum up to bg.pseudo.count for the 4 nucleotides.
`bg.pseudo.count`	the pseudo count which is shared between nucleotides when frequency matrices are given
`bg.len`	the length range of background chunks
`bg.source`	a free-form textual description of how the background was generated
`verbose`	if to produce verbose output
`fit.log`	if to fit log odds (instead of odds)

Details

Construct a lognormal background distribution for a set of sequences. Sequences concatenated are binned in 'bg.len' chunks and lognormal distribution fitted to them.

Examples

```
## Not run:
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel.PFM)

# make background for MotifDb motifs using 2kb promoters of all D. melanogaster transcripts
if(require("BSgenome.Dmelanogaster.UCSC.dm3"))
makePWMGEVBackground(Dmelanogaster$upstream2000, MotifDb.Dmel.PFM)
```

```
}

## End(Not run)
```

---

makePWMLognBackground *Make a lognormal background distribution...*

Description

Make a lognormal background distribution

Usage

```
makePWMLognBackground(bg.seq, motifs, bg.pseudo.count=1, bg.len=1000, bg.source="",
    verbose=TRUE)
```

Arguments

bg.seq	a set of background sequences, either a list of DNAString object or DNAStringSet object
motifs	a set of motifs, either a list of frequency matrices, or a list of PWM objects. If frequency matrices are given, the background distribution fitted from bg.seq. Same ratios are used for pseudo counts that sum up to bg.pseudo.count for the 4 nucleotides.
bg.pseudo.count	the pseudo count which is shared between nucleotides when frequency matrices are given
bg.len	the length of background chunks
bg.source	a free-form textual description of how the background was generated
verbose	if to produce verbose output

Details

Construct a lognormal background distribution for a set of sequences. Sequences concatenated are binned in 'bg.len' chunks and lognormal distribution fitted to them.

Examples

```
## Not run:
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel.PFM)

# make background for MotifDb motifs using 2kb promoters of all D. melanogaster transcripts
if(require("BSgenome.Dmelanogaster.UCSC.dm3"))
makePWMLognBackground(Dmelanogaster$upstream2000, MotifDb.Dmel.PFM)
}

## End(Not run)
```

makePWMPvalCutoffBackground *Construct a cutoff background from empirical background...*

Description

Construct a cutoff background from empirical background

Usage

```
makePWMPvalCutoffBackground(bg.p, p.value=0.001, bg.source="")
```

Arguments

bg.p	an object of class PWMEmpiricalBackground
p.value	the P-value used to find cuttoffs for each of the motifs
bg.source	textual description of background source

Details

This function takes already calculated empirical background distribution and chooses cutoff for each motif based on P-value cutoff for individual sites.

Value

an object of type PWMCutoffBackground

Examples

```
## Not run:
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel.PFM)

# make empirical background - here we use only 100 sequences for illustrative purposes
if(require("BSgenome.Dmelanogaster.UCSC.dm3"))
bg.p = makePWMEmpiricalBackground(Dmelanogaster$upstream2000[1:100], MotifDb.Dmel.PFM)

# use the empirical background to pick a threshold and make cutoff background
makePWMPvalCutoffBackground(bg.p, 0.001)
}

## End(Not run)
```

makeStartEndPos *Divide total...*

Description

Divide total.len into fragments of length len by providing start,end positions

Usage

```
makeStartEndPos(total.len, len)
```

Arguments

total.len	total available length to be subdivided
len	size of the individual chunk

Value

a data.frame containing paired up start,end positions

---

matrixShuffleZscorePerSequence
Obtain z-score for motif column shuffling...

Description

Obtain z-score for motif column shuffling

Usage

```
matrixShuffleZscorePerSequence(scores, sequences, pwms, cutoff, B=30)
```

Arguments

scores	a set of already calculated scores
sequences	either one sequence or a list/set of sequences (objects of type DNAString or DNAStringSet)
pwms	a list of PWMs
cutoff	if NULL, will use affinity, otherwise will use number of hits over this log2 odds cutoff
B	number of replicates, i.e. PWM column shuffles

Details

All PWMs are shuffled at the same time. This function would be too slow to produce empirical P-values, thus we return a z-score from a small number of shuffles.

The z-scores are calculated for each sequence individually.

maxAligned *Returned the aligned motif parts...*

Description

Returned the aligned motif parts

Usage

```
maxAligned(m1, m2, offset)
```

Arguments

m1	frequency matrix of first motif
m2	frequency matrix of second motif
offset	a number of nucleotides by which the first motif is offsetted compared to the second

Details

This function takes the offset of first motif relative to second and chops off the end of both motifs that are not aligned. It returns a list containing only the columns that align.

Value

a list of column-trimmed motifs m1, m2

motifDiffEnrichment *Differential motif enrichment*

Description

Test for differential enrichment between two groups of sequences

Usage

```
motifDiffEnrichment(sequences1, sequences2, pwms, score="autodetect", bg="autodetect",
    cutoff=log2(exp(4)), verbose=TRUE, res1, res2)
```

Arguments

sequences1	First set of sequences. Can be either a single sequence (an object of class DNAString), or a list of DNAString objects, or a DNAStringSet object.
sequences2	Second set of sequences. Can be either a single sequence (an object of class DNAString), or a list of DNAString objects, or a DNAStringSet object.
pwms	this parameter can take multiple values depending on the scoring scheme and background correction used. When the method parameter is set to "autodetect", the following default algorithms are going to be used: • if pwms is a list containing either frequency matrices or a list of PWM objects then the "affinity" algorithm is selected. If frequency matrices are given, they are converted to PWMs using uniform background. For best performance, convert frequency matrices to PWMs before calling this function using realistic genomic background. • Otherwise, appropriate scoring scheme and background correction are selected based on the class of the object (see below).
score	this parameter determines which scoring scheme to use. Following scheme as available: • "autodetect" - default value. Scoring method is determined based on the type of pwms parameter. • "affinity" - use threshold-free affinity scores without a background. The pwms parameter can either be a list of frequency matrices, PWM objects, or a PWMLognBackground object. • "cutoff" - use number of motif hits above a score cutoff as a measure of enrichment. No background correction is performed. The pwms parameter can either be a list of frequency matrices, PWM objects, or a PWMCutoffBackground object.
bg	this parameter determines which background correction to use, if any. • "autodetect" - default value. Background correction is determined based on the type of the pwms parameter. • "logn" - use a lognormal distribution background pre-computed for a set of PWMs. This requires pwms to be of class PWMLognBackground. • "z" - use a z-score for the number of significant motif hits compared to background number of hits. This requires pwms to be of class PWMCutoffBackground. • "none" - no background correction
cutoff	the score cutoff for a significant motif hit if scoring scheme "cutoff" is selected.
res1	the output of motifEnrichment if already calculated for sequences1
res2	the output of motifEnrichment if already calculated for sequences2
verbose	if to produce verbose output

Details

This function calls motifEnrichment on two groups of sequences and calculates the difference statistics when possible.

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
# load the background file for drosophila and lognormal correction
data(PWMLogn.dm3.MotifDb.Dmel)

# get the differential enrichment
diff = motifDiffEnrichment(DNAString("TGCATCAAGTGTGTAGTGTGAGATTAGT"), DNAString("TGAACGAGTAGGACGATGAGAGATTGATG"

# motifs differentially enriched in the first sequence (with lognormal background correction)
head(sort(diff$group.bg, decreasing=TRUE))

# motifs differentially enriched in the second sequence (with lognormal background correction)
head(sort(diff$group.bg))
}
```

`motifEnrichment` *Motif enrichment*

Description

Calculate motif enrichment using one of available scoring algorithms and background corrections.

Usage

```
motifEnrichment(sequences, pwms, score="autodetect", bg="autodetect", cutoff,
    verbose=TRUE, motif.shuffles=30, B=1000, group.only=FALSE)
```

Arguments

`sequences`	the sequences to be scanned for enrichment. Can be either a single sequence (an object of class DNAString), or a list of DNAString objects, or a DNAStringSet object.
`pwms`	this parameter can take multiple values depending on the scoring scheme and background correction used. When the `method` parameter is set to "autodetect", the following default algorithms are going to be used: • if `pwms` is a list containing either frequency matrices or a list of PWM objects then the "affinity" algorithm is selected. If frequency matrices are given, they are converted to PWMs using uniform background. For best performance, convert frequency matrices to PWMs before calling this function using realistic genomic background. • Otherwise, appropriate scoring scheme and background correction are selected based on the class of the object (see below).
`score`	this parameter determines which scoring scheme to use. Following scheme as available: • "autodetect" - default value. Scoring method is determined based on the type of `pwms` parameter.

- "affinity" - use threshold-free affinity scores without a background. The pwms parameter can either be a list of frequency matrices, PWM objects, or a PWMLognBackground object.
- "cutoff" - use number of motif hits above a score cutoff as a measure of enrichment. No background correction is performed. The pwms parameter can either be a list of frequency matrices, PWM objects, or a PWMCutoffBackground object.
- "clover" - use the Clover algorithm (Frith et al, 2004). The Clover score of a single sequence is identical to the affinity score, while for a group of sequences is an average of products of affinities over all sequence subsets.

bg — this parameter determines which background correction to use, if any.

- "autodetect" - default value. Background correction is determined based on the type of the pwms parameter.
- "logn" - use a lognormal distribution background pre-computed for a set of PWMs. This requires pwms to be of class PWMLognBackground.
- "z" - use a z-score for the number of significant motif hits compared to background number of hits. This requires pwms to be of class PWMCutoffBackground.
- "pval" - use empirical P-value based on a set of background sequences. This requires pwms to be of class PWMEmpiricalBackground. Note that PWMEmpiricalBackground objects tend to be very large so that the empirical P-value can be calculated in reasonable time.
- "ms" - shuffle columns of motif matrices and use that as basis for P-value calculation. Note that since the sequences need to rescanned with all of the new shuffled motifs this can be very slow. Also, this also works only no *individual* sequences, not groups.
- "none" - no background correction

cutoff — the score cutoff for a significant motif hit if scoring scheme "cutoff" is selected.

verbose — if to print verbose output

motif.shuffles — number of times to shuffle motifs if using "ms" background correction

B — number of replicates when calculating empirical P-value

group.only — if to produce statistical only for the group of sequences, not individual sequences. This is useful when one wants to calculate the empirical P-value for the whole group, but not individual sequences (which might take quite a long time).

Details

This function provides and interface to all algorithms available in PWMEnrich to find motif enrichment in a single or a group of sequences with/without background correction.

Since for all algorithms the first step involves calculating raw scores without background correction, the output always contains the scores without background correction together with (optional) background-corrected scores.

Unless otherwise specified the scores are returned both separately for each sequence (without/with background) and for the whole group of sequences (without/with background).

To use a background correction you need to supply a set of PWMs with precompiled background distribution parameters (see function makeBackground). When such an object is supplied as the pwm parameter, the scoring scheme and background correction are automatically determined.

There are additional packages with already pre-computed background (e.g. see package PWMEnrich.Dmelanogaster.backgr

Please refer to (Stojnic & Adryan, 2012) for more details on the algorithms.

Value

a MotifEnrichmentResults object containing a subset following elements:

- "score" - scoring scheme used
- "bg" - background correction used
- "params" - any additional parameters
- "sequences" - the set of sequences used
- "pwms" - the set of pwms used
- "sequence.nobg" - per-sequence scores without any background correction. For "affinity" and "clover" a matrix of mean affinity scores; for "cutoff" number of significant hits above a cutoff
- "sequence.bg" - per-sequence scores after background correction. For "logn" and "pval" the P-value (smaller is better); for "z" and "ms" background corrections the z-scores (bigger is better).
- "group.nobg" - aggregate scores for the whole group of sequences without background correction. For "affinity" and "clover" the mean affinity over all sequences in the set; for "cutoff" the total number of hits in all sequences.
- "group.bg" - aggregate scores for the whole group of sequences with background correction. For "logn" and "pval", the P-value for the whole group (smaller is better); for "z" and "ms" the z-score for the whole set (bigger is better).
- "sequence.norm" - (only for "logn") the length-normalized scores for each of the sequences. Currently only implemented for "logn", where it returns the values normalized from LogN(0,1) distribution
- "group.norm" - (only for "logn") similar to sequence.norm, but for the whole group of sequences

References

- R. Stojnic & B. Adryan: Identification of functional DNA motifs using a binding affinity lognormal background distribution, submitted.
- MC Frith et al: Detection of functional DNA motifs via statistical over-representation, Nucleid Acid Research (2004).

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
###
# load the pre-compiled lognormal background
data(PWMLogn.dm3.MotifDb.Dmel)
```

```
# scan two sequences for motif enrichment
sequences = list(DNAString("GAAGTATCAAGTGACCAGTAGATTGAAGTAGACCAGTC"), DNAString("AGGTAGATAGAACAGTAGGCAATGGGGGAA
res = motifEnrichment(sequences, PWMLogn.dm3.MotifDb.Dmel)

# most enriched in both sequences (lognormal background P-value)
head(motifRankingForGroup(res))

# most enriched in both sequences (raw affinity, no background)
head(motifRankingForGroup(res, bg=FALSE))

# most enriched in the first sequence (lognormal background P-value)
head(motifRankingForSequence(res, 1))

# most enriched in the first sequence (raw affinity, no background)
head(motifRankingForSequence(res, 1, bg=FALSE))

###
# Load the pre-compiled background for hit-based motif counts with cutoff of P-value = 0.001
data(PWMPvalueCutoff1e3.dm3.MotifDb.Dmel)

res.count = motifEnrichment(sequences, PWMPvalueCutoff1e3.dm3.MotifDb.Dmel)

# Enrichment in the whole group, z-score for the number of motif hits
head(motifRankingForGroup(res))

# First sequence, sorted by number of motif hits with P-value < 0.001
head(motifRankingForSequence(res, 1, bg=FALSE))

}
```

`MotifEnrichmentResults-class` *A wrapper class for results of motifEnrichment() that should make it easier to access the results.*

Description

A wrapper class for results of motifEnrichment() that should make it easier to access the results.

Details

Note that this is only a wrapper around a list which is the return value in PWMEnrich 1.3 and as such it provides the same interface as a list (for backward compatibility), with some additional methods.

Slots

`res`: (`list`) a list of old results with elements such as: sequence.bg, sequence.nobg, group.bg, group.nobg

Methods

names signature(x = "MotifEnrichmentResults"): Name of different pieces of information associated with MotifEnrichmentResults

$ signature(x = "MotifEnrichmentResults"): Access a property by name

show signature(object = "MotifEnrichmentResults"): show method for MotifEnrichmentResults

motifRankingForGroup signature(obj = "MotifEnrichmentResults"): Get a ranking of motifs by their enrichment in the whole set of sequences

motifRankingForSequence signature(obj = "MotifEnrichmentResults"): Get a ranking of motifs by their enrichment in one specific sequence

plotTopMotifsGroup signature(obj = "MotifEnrichmentResults"): Plot the top N enrichment motifs in a group of sequences

plotTopMotifsSequence signature(obj = "MotifEnrichmentResults"): Plot the top N enrichment motifs in a single sequence

---

motifIC *Information content for a PWM or PFM...*

Description

Information content for a PWM or PFM

Usage

```
motifIC(motif, prior.params=c(A = 0.25, C = 0.25, G = 0.25, T = 0.25),
    bycol=FALSE)
```

Arguments

motif	a matrix of frequencies, or a PWM object
prior.params	the prior parameters to use when a matrix is given (ignored if motif is already a PWM)
bycol	if to return values separately for each column

Value

information content in bits (i.e. log2)

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel)
data(MotifDb.Dmel.PFM)

motifIC(MotifDb.Dmel$ttk) # the nucleotide distribution is taken from the PWM (in this case genomic background)
motifIC(MotifDb.Dmel.PFM$ttk) # information content with default uniform background because the input is a matrix,
}
```

`motifPrAUC` *Calculate PR-AUC for motifs ranked according to some scoring scheme...*

Description

Calculate PR-AUC for motifs ranked according to some scoring scheme

Usage

```
motifPrAUC(seq.res)
```

Arguments

`seq.res`	a matrix where each column represents a PWM and each row a result for a different sequence.

Details

Note that this function asssumes that smaller values are better!

`motifRankingForGroup,MotifEnrichmentResults-method` *Get a ranking of motifs by their enrichment in the whole set of sequences...*

Description

Get a ranking of motifs by their enrichment in the whole set of sequences

Usage

```
## S4 method for signature 'MotifEnrichmentResults'
motifRankingForGroup(obj, bg=TRUE, id=FALSE, order=FALSE, rank=FALSE, unique=FALSE, ...)
```

Arguments

`obj`	a MotifEnrichmentResults object
`bg`	if to use background P-values to do the ranking (if available)
`id`	if to show PWM IDs instead of target TF names
`order`	if to output the ordering of PWMs instead of actual P-values or raw values
`rank`	if the output should be rank of a PWM instead of actual P-values or raw values
`unique`	if TRUE, only the best rank is taken for each TF (only when id = FALSE, order = FALSE)
`...`	currently unused

Value

a vector of P-values or raw enrichments sorted such that the first motif is most enriched

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
###
# load the pre-compiled lognormal background
data(PWMLogn.dm3.MotifDb.Dmel)

# scan two sequences for motif enrichment
sequences = list(DNAString("GAAGTATCAAGTGACCAGTAAGTCCCAGATGA"), DNAString("AGGTAGATAGAACAGTAGGCAATGAAGCCGATG"))
res = motifEnrichment(sequences, PWMLogn.dm3.MotifDb.Dmel)

# most enriched in both sequences (sorted by lognormal background P-value)
head(motifRankingForGroup(res))

# Return a non-redundant set of TFs
head(motifRankingForGroup(res, unique=TRUE))

# sorted by raw affinity instead of P-value
head(motifRankingForGroup(res, bg=FALSE))

# show IDs instead of target TF names
head(motifRankingForGroup(res, id=TRUE))

# output the rank instead of P-value
head(motifRankingForGroup(res, rank=TRUE))
}
```

---

motifRankingForSequence,MotifEnrichmentResults-method

Get a ranking of motifs by their enrichment in one specific sequence...

Description

Get a ranking of motifs by their enrichment in one specific sequence

Usage

```
## S4 method for signature 'MotifEnrichmentResults'
motifRankingForSequence(obj, seq.id, bg=TRUE, id=FALSE, order=FALSE, rank=FALSE, unique=FALSE,
    ...)
```

Arguments

obj	a MotifEnrichmentResults object
seq.id	either the sequence number or sequence name

bg	if to use background P-values to do the ranking (if available)
id	if to show PWM IDs instead of target TF names
order	if to output the ordering of PWMs instead of actual P-values or raw values
rank	if the output should be rank of a PWM instead of actual P-values or raw values
unique	if TRUE, only the best rank is taken for each TF (only when id = FALSE, order = FALSE)
...	currently unused

Value

a vector of P-values or raw enrichments sorted such that the first motif is most enriched

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
###
# load the pre-compiled lognormal background
data(PWMLogn.dm3.MotifDb.Dmel)

# scan two sequences for motif enrichment
sequences = list(DNAString("GAAGTATCAAGTGACCAGTAAGTCCCAGATGA"), DNAString("AGGTAGATAGAACAGTAGGCAATGAAGCCGATG"))
res = motifEnrichment(sequences, PWMLogn.dm3.MotifDb.Dmel)

# most enriched in the second sequences (sorted by lognormal background P-value)
head(motifRankingForSequence(res, 2))

# return unique TFs enriched in sequence 2
head(motifRankingForSequence(res, 2, unique=TRUE))

# sorted by raw affinity instead of P-value
head(motifRankingForSequence(res, 2, bg=FALSE))

# show IDs instead of target TF names
head(motifRankingForSequence(res, 2, id=TRUE))

# output the rank instead of P-value
head(motifRankingForSequence(res, 2, rank=TRUE))
}
```

`motifRecoveryAUC` *Calculate Recovery-AUC for motifs ranked according to some scoring scheme...*

Description

Calculate Recovery-AUC for motifs ranked according to some scoring scheme

Usage

```
motifRecoveryAUC(seq.res)
```

Arguments

seq.res	a matrix where each column represents a PWM and each row a result for a different sequence.

Details

Note that this function asssumes that smaller values are better!

---

motifScores *Motif affinity of number of hits over a threshold...*

Description

Motif affinity of number of hits over a threshold

Usage

```
motifScores(sequences, motifs, raw.scores=FALSE, verbose=TRUE, cutoff)
```

Arguments

sequences	a set of sequences to be scanned, a list of DNAString or other scannable objects
motifs	a list of motifs either as frequency matrices (PFM) or as PWM objects. If PFMs are specified they are converted to PWMs using uniform background.
raw.scores	if to return raw scores (odds) for each position in the sequence. Note that scores for forward and reverse strand are concatenated into a single long vector of scores (twice the length of the sequence)
verbose	if to print verbose output
cutoff	if not NULL, will count number of matches with score above value specified (instead of returning the average affinity). Can either be one value, or a vector of values for each of the motifs.

Details

Scan a number of sequences either to find overall affinity, or a number of hits over a score threshold.

Value

if raw.scores=FALSE, returns a matrix of mean scores (after cutoff if any), where columns are motifs. The returned values are either mean odd scores (not log-odd), or number of hits above a threshold; otherwise if raw.scores=TRUE, returns a list of raw score values (before cutoff)

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel)

affinity = motifScores(DNAString("CGTAGGATAAAGTAACTAGTTGATGATGAAAG"), MotifDb.Dmel) # affinity scores
counts = motifScores(DNAString("CGTAGGATAAAGTAACTAGTTGATGATGAAAG"), MotifDb.Dmel, cutoff=log2(exp(4))) # motif hi
print(affinity)
print(counts)

# scanning multiple sequences
sequences = list(DNAString("CGTAGGATAAAGTAACTAGTTGATGATGAAAG"), DNAString("TGAGACGAAGGGGATGAGATGCGGAAGAGTGAAA")
affinity2 = motifScores(sequences, MotifDb.Dmel)
print(affinity2)
}
```

`motifScoresBigMemory` *This is a memory intensive version of motifScore() which is abount 2 times faster...*

Description

This is a memory intensive version of motifScore() which is abount 2 times faster

Usage

```
motifScoresBigMemory(sequences, motifs, raw.scores=FALSE, verbose=TRUE, cutoff)
```

Arguments

`sequences`	set of input sequences
`motifs`	set of input PWMs or PFMs
`raw.scores`	if to return scores for each base-pair
`verbose`	if to produce verbose output
`cutoff`	the cutoff for calling binding sites (in base 2 log).

Details

The parameters and functionality are the same as `motifScores`. Please refer to documentation of this function for detailed explanation of functionality.

This function is not meant to be called directly, but is indirectly called by motifScores() once a global parameters useBigMemory is set.

See Also

`motifScores`

`motifSimilarity` *Calculates similarity between two PFMs.*

Description

Calculates similarity between two PFMs.

Usage

```
motifSimilarity(m1, m2, trim=0.4, self.sim=FALSE)
```

Arguments

`m1`	matrix with four rows representing the frequency matrix of first motif
`m2`	matrix with four rows representing the frequency matrix of second motif
`trim`	bases with information content smaller than this value will be trimmed off both motif ends
`self.sim`	if to calculate self similarity (i.e. without including offset=0 in alignment)

Details

This function calculates the normalized motif correlation as a measure of motif frequency matrix similarity.

This score is essentially a normalized version of the sum of column correlations as proposed by Pietrokovski (1996). The sum is normalized by the average motif length of m1 and m2, i.e. (ncol(m1)+ncol(m2))/2. Thus, for two idential motifs this score is going to be 1. For unrelated motifs the score is going to be typically around 0.

Motifs need to aligned for this score to be calculated. The current implementation tries all possible ungapped alignment with a minimal of two basepair matching, and the maximal score over all alignments is returned.

Motif 1 is aligned both to Motif 2 and its reverse complement. Thus, the motif similarities are the same if the reverse complement of any of the two motifs is given.

References

Pietrokovski S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucleic Acids Res 1996;24:3836-3845.

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel.PFM)

# calculate the similarity of tin and vnd motifs (which are almost identical)
motifSimilarity(MotifDb.Dmel.PFM$tin, MotifDb.Dmel.PFM$vnd)
```

```
# similarity of two unrelated motifs
motifSimilarity(MotifDb.Dmel.PFM$tin, MotifDb.Dmel.PFM$ttk)
}
```

---

operators-MotifEnrichmentResults *Names of variables*

Description

Name of different pieces of information associated with MotifEnrichmentResults

Usage

```
## S4 method for signature 'MotifEnrichmentResults'
names(x)
## S4 method for signature 'MotifEnrichmentResults'
x$name
```

Arguments

x	the MotifEnrichmentResults object
name	the variable name

Value

names,MotifEnrichmentResults-method: the names of the variables

---

operators-PWM *Names of variables*

Description

Name of different pieces of information associated with PWM

Usage

```
## S4 method for signature 'PWM'
names(x)
## S4 method for signature 'PWM'
x$name
## S4 method for signature 'PWM'
length(x)
```

Arguments

x	the PWM object
name	the variable name

Details

length,PWM-method: Returns the motif length, i.e. the number of columns in the PWM.

Value

names,PWM-method: the names of the variables

---

operators-PWMCutoffBackground
Names of variables

Description

Name of different pieces of information associated with PWMCutoffBackground

Usage

```
## S4 method for signature 'PWMCutoffBackground'
names(x)
## S4 method for signature 'PWMCutoffBackground'
x$name
```

Arguments

x	the PWMCutoffBackground object
name	the variable name

Value

names,PWMCutoffBackground-method: the names of the variables

operators-PWMEmpiricalBackground
Names of variables

Description

Name of different pieces of information associated with PWMEmpiricalBackground

Usage

```
## S4 method for signature 'PWMEmpiricalBackground'
names(x)
## S4 method for signature 'PWMEmpiricalBackground'
x$name
```

Arguments

x	the PWMEmpiricalBackground object
name	the variable name

Value

names,PWMEmpiricalBackground-method: the names of the variables

operators-PWMGEVBackground
Names of variables

Description

Name of different pieces of information associated with PWMGEVBackground

Usage

```
## S4 method for signature 'PWMGEVBackground'
names(x)
## S4 method for signature 'PWMGEVBackground'
x$name
```

Arguments

x	the PWMGEVBackground object
name	the variable name

Value

names,PWMGEVBackground-method: the names of the variables

operators-PWMLognBackground
Names of variables

Description

Name of different pieces of information associated with PWMLognBackground

Usage

```
## S4 method for signature 'PWMLognBackground'
names(x)
## S4 method for signature 'PWMLognBackground'
x$name
```

Arguments

x	the PWMLognBackground object
name	the variable name

Value

names,PWMLognBackground-method: the names of the variables

PFMtoPWM *Convert frequencies into motifs using PWMUnscaled...*

Description

Convert frequencies into motifs using PWMUnscaled

Usage

```
PFMtoPWM(motifs, id=names(motifs), name=names(motifs), seq.count, ...)
```

Arguments

motifs	a list of motifs represented as matrices of frequencies (PFM)
id	the set of IDs for the motifs (defaults to names of the 'motifs' list)
name	the set of names for the motifs (defaults to names of the 'motifs' list)
seq.count	if frequencies in the motifs are normalized to 1, provides a vector of sequence counts (e.g. for MotifDb motifs)
...	other parameters to PWMUnscaled

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel.PFM)

PFMtoPWM(MotifDb.Dmel.PFM) # convert to PWM with uniform background

prior = getBackgroundFrequencies("dm3", quick=TRUE) # get background for drosophila (quick mode on a reduced datase
PFMtoPWM(MotifDb.Dmel.PFM, prior.params=prior) # convert with genomic background
}
```

---

pickGenome *A helper function to pick a genome for an organism...*

Description

A helper function to pick a genome for an organism

Usage

```
pickGenome(organism)
```

Arguments

organism	either organism name (such as "dm3") or a BSgenome object

Value

a BSgenome object

---

plot,PWM,missing-method *Plotting for the PWM class...*

Description

Plotting for the PWM class

Usage

```
## S4 method for signature 'PWM,missing'
plot(x, y, ...)
```

Arguments

x	the PWM object
y	unused
...	other parameters to pass to seqLogo's plot function

Details

This function produces a sequence logo (via package seqLogo).

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel)

# plot the tinman motif from MotifDb
plot(MotifDb.Dmel$tin)
}
```

---

plotMultipleMotifs *Plot mulitple motifs in a single plot...*

Description

Plot mulitple motifs in a single plot

Usage

```
plotMultipleMotifs(pwms, titles=names(pwms), rows=ceiling(sqrt(length(pwms))),
    cols=ceiling(sqrt(length(pwms))), xmargin.scale=1, ymargin.scale=1,
    ...)
```

Arguments

pwms	a list of PWM objects or frequency matrices
titles	a characater vector of titles for each of the plots
rows	number of rows in the grid
cols	number or cols in the grid
xmargin.scale	the scaling parameter for the X-axis margin. Useful when plotting more than one logo on a page
ymargin.scale	the scaling parameter for the Y-axis margin. Useful when plotting more than one logo on a page
...	other parameters passed to seqLogoGrid()

Details

Individual motif logos are plotted on a rows x cols grid. This function is a convenience interface for the seqLogoGrid function that deals with viewpoint placement in a matrix-like grid layout.

By default will try to make a square grid plot that would fit all the motifs and use list names as captions.

plotPFM *Plot a PFM (not PWM) using seqLogo...*

Description

Plot a PFM (not PWM) using seqLogo

Usage

```
plotPFM(pfm, ...)
```

Arguments

pfm	a matrix where rows are the four nucleotides
...	additional parameters for plot()

plotTopMotifsGroup,MotifEnrichmentResults-method
Plot the top N enrichment motifs in a group of sequences...

Description

Plot the top N enrichment motifs in a group of sequences

Usage

```
## S4 method for signature 'MotifEnrichmentResults'
plotTopMotifsGroup(obj, n, bg=TRUE, id=FALSE, ...)
```

Arguments

obj	a MotifEnrichmentResults object
n	the number of top ranked motifs to plot
bg	if to use background P-values to do the ranking (if available)
id	if to show PWM IDs instead of target TF names
...	other parameters passed to plotMultipleMotifs()

Examples

```
## Not run:
if(require("PWMEnrich.Dmelanogaster.background")){
###
# load the pre-compiled lognormal background
data(PWMLogn.dm3.MotifDb.Dmel)

# scan two sequences for motif enrichment
sequences = list(DNAString("GAAGTATCAAGTGACCAGTAAGTCCCAGATGA"), DNAString("AGGTAGATAGAACAGTAGGCAATGAAGCCGATG"))

res = motifEnrichment(sequences, PWMLogn.dm3.MotifDb.Dmel)

# plot the top 4 motifs in a 2x2 grid
plotTopMotifsGroup(res, 4)

# plot top 3 motifs in a single row
plotTopMotifsGroup(res, 3, row=1, cols=3)
}

## End(Not run)
```

`plotTopMotifsSequence,MotifEnrichmentResults-method`
Plot the top N enrichment motifs in a single sequence...

Description

Plot the top N enrichment motifs in a single sequence

Usage

```
## S4 method for signature 'MotifEnrichmentResults'
plotTopMotifsSequence(obj, seq.id, n, bg=TRUE, id=FALSE, ...)
```

Arguments

`obj`	a MotifEnrichmentResults object
`seq.id`	either the sequence number or sequence name
`n`	the number of top ranked motifs to plot
`bg`	if to use background P-values to do the ranking (if available)
`id`	if to show PWM IDs instead of target TF names
`...`	other parameters passed to `plotMultipleMotifs()`

Examples

```
## Not run:
if(require("PWMEnrich.Dmelanogaster.background")){
###
# load the pre-compiled lognormal background
data(PWMLogn.dm3.MotifDb.Dmel)

# scan two sequences for motif enrichment
sequences = list(DNAString("GAAGTATCAAGTGACCAGTAAGTCCCAGATGA"), DNAString("AGGTAGATAGAACAGTAGGCAATGAAGCCGATG"))

res = motifEnrichment(sequences, PWMLogn.dm3.MotifDb.Dmel)

# plot the top 4 motifs in a 2x2 grid
plotTopMotifsSequence(res, 1, 4)

# plot top 3 motifs in a single row
plotTopMotifsSequence(res, 1, 3, row=1, cols=3)
}

## End(Not run)
```

PWM-class *A class that represents a Position Weight Matrix (PWM)...*

Description

A class that represents a Position Weight Matrix (PWM)

Slots

id: (character) a systematic ID given to this PWM, could include the source, version, etc

name: (character) the name of the transcription factor (TF) to which the PWM corresponds to

pfm: (matrix) Position Frequency Matrix (PFM) from which the PWM is derived

prior.params: (vector) Defines prior frequencies of the four bases (A,C,G,T), a named vector. These will be added to individual values for the PFM and at the same time used as background probabilities

pwm: (matrix) Final Position Weight Matrix (PWM) constructed using prior.params with logarithm base 2

Methods

plot signature(x = "PWM", y = "missing"): Plotting for the PWM class

names signature(x = "PWM"): Name of different pieces of information associated with PWM

$ signature(x = "PWM"): Access a property by name

length signature(x = "PWM"): Length of the motif

reverseComplement signature(x = "PWM"): Reverse complement for the PWM object

show signature(object = "PWM"): show method for PWM

PWMCutoffBackground-class

Hit count background distribution for a set of PWMs...

Description

Hit count background distribution for a set of PWMs

Slots

bg.source: (character) textual description of where the background distribution is derived from

bg.cutoff: (numeric) the cutoff score used to find significant motif hits (in log2 odds), either a single value or a vector of values

bg.P: (numeric) the density of significant motif hits per nucleotide in background

pwms: (list) the pwms for which the background has been compiled

Methods

show signature(object = "PWMCutoffBackground"): show method for PWMCutoffBackground

names signature(x = "PWMCutoffBackground"): Name of different pieces of information associated with PWMCutoffBackground

$ signature(x = "PWMCutoffBackground"): Access a property by name

---

PWMEmpiricalBackground-class

Background for calculating empirical P-values...

Description

Background for calculating empirical P-values

Details

This object contains raw scores for one very long sequence, thus it can be very large.

Slots

bg.source: (character) textual description of where the background distribution is derived from

bg.fwd: (matrix) affinity scores (odds) for the forward strand. PWMs as columns.

bg.rev: (matrix) affinity scores (odds) for the reverse strand. PWMs as columns.

pwms: (list) the pwms for which the background has been compiled

Methods

show signature(object = "PWMEmpiricalBackground"): show method for PWMEmpiricalBackground

names signature(x = "PWMEmpiricalBackground"): Name of different pieces of information associated with PWMEmpiricalBackground

$ signature(x = "PWMEmpiricalBackground"): Access a property by name

---

PWMGEVBackground-class

Generalized Extreme Values (GEV) background for P-values...

Description

Generalized Extreme Values (GEV) background for P-values

Details

The three parameters of the GEV distribution are fitted by doing linear regression on log of sequence length.

Slots

bg.source: (character) textual description of where the background distribution is derived from

bg.loc: (list) linear regression model for estimating the location parameter based on log(L), list of lm objects of PWMs

bg.scale: (list) linear regression model for estimating the scale parameter based on log(L), list of lm objects of PWMs

bg.shape: (list) linear regression model for estimating the shape parameter based on log(L), list of lm objects of PWMs

pwms: (list) the pwms for which the background has been compiled

Methods

show signature(object = "PWMGEVBackground"): show method for PWMGEVBackground

names signature(x = "PWMGEVBackground"): Name of different pieces of information associated with PWMGEVBackground

$ signature(x = "PWMGEVBackground"): Access a property by name

PWMLognBackground-class *Lognormal background distribution for a set of PWMs...*

Description

Lognormal background distribution for a set of PWMs

Slots

`bg.source`: (`character`) textual description of where the background distribution is derived from

`bg.len`: (`numeric`) the length to which the background is normalized to. This is a vector of values, can have a different value for each motif.

`bg.mean`: (`numeric`) the mean value of the lognormal distribution at bg.len

`bg.sd`: (`numeric`) the standard deviation of the lognormal distribution at bg.len

`pwms`: (`list`) the pwms for which the background has been compiled

Methods

`show signature(object = "PWMLognBackground")`: show method for PWMLognBackground

`names signature(x = "PWMLognBackground")`: Name of different pieces of information associated with PWMLognBackground

`$ signature(x = "PWMLognBackground")`: Access a property by name

PWMUnscaled *Create a PWM from PFM*

Description

The PWM function from Biostrings without unit scaling

Usage

```
PWMUnscaled(x, id = "", name = "", type=c("log2probratio", "prob"), prior.params=c(A
    = 0.25, C = 0.25, G = 0.25, T = 0.25), pseudo.count=prior.params,
    unit.scale=FALSE, seq.count)
```

Arguments

x	the integer count matrix representing the motif, rows as nucleotides
id	a systematic ID given to this PWM, could include the source, version, etc
name	the name of the transcription factor (TF) to which the PWM corresponds to
type	the type of PWM calculation, either as log2-odds, or posterior probability (frequency matrix)
prior.params	the pseudocounts for each of the nucleotides
pseudo.count	the pseudo-count values if different from priors
unit.scale	if to unit.scale the pwm (default is no unit scaling)
seq.count	if x is a normalised PFM (i.e. with probabilities instead of sequence counts), then this sequence count will be used to convert x into a count matrix

Details

By default the Biostrings package scales the log-odds score so it is within 0 and 1. In this function we take a more traditional approach with no unit scaling and offer unit scaling as an additional parameter.

See ?PWM from Biostrings for more information on input arguments.

Value

a new PWM object representing the PWM

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel.PFM)

PWMUnscaled(MotifDb.Dmel.PFM$ttk, id="ttk-JASPAR", name="ttk") # make a PWM with uniform background
PWMUnscaled(MotifDb.Dmel.PFM$ttk, id="ttk-JASPAR", name="ttk", prior.params=c("A"=0.2, "C"=0.3, "G"=0.3, "T"=0.2

prior = getBackgroundFrequencies("dm3", quick=TRUE) # get background for drosophila (quick mode on a reduced datase
PWMUnscaled(MotifDb.Dmel.PFM$ttk, id="ttk-JASPAR", name="ttk", prior.params=prior) # convert using genomic backgr
}
```

---

rankingProcessAndReturn

A helper function for motifRankingForGroup and motifRankingForSequence with the common code...

Description

A helper function for motifRankingForGroup and motifRankingForSequence with the common code

Usage

```
rankingProcessAndReturn(res, r, id, order, rank, unique, decreasing)
```

Arguments

res	the list of results from MotifEnrichmentResults object
r	the vector of raw results that needs to be processed
id	if to return IDs instead of names
order	if to return the ordering of motifs
rank	if to return the rank of motifs
unique	if to remove duplicates
decreasing	specifies the sorting order

---

readJASPAR *Read motifs in JASPAR format...*

Description

Read motifs in JASPAR format

Usage

```
readJASPAR(file, remove.ids=FALSE)
```

Arguments

file	the filename
remove.ids	if to strip JASPAR ID's from motif names, e.g. "MA0211.1 bap" would become just "bap"

Value

a list of matrices representing motifs (with four nucleotides as rows)

readMotifs *Read in motifs in JASPAR or TRANSFAC format...*

Description

Read in motifs in JASPAR or TRANSFAC format

Usage

```
readMotifs(file, remove.acc=FALSE)
```

Arguments

file	the filename
remove.acc	if to remove accession numbers. If TRUE, the AC entry in TRANSFAC files is ignored, and the accession is stripped from JASPAR, e.g. motif with name "MA0211.1 bap" would become just "bap". If FALSE, botht he AC and ID are used to generate the TRANSFAC name and the original motif names are preserved in JASPAR files.

Details

The format is autodetected based on file format. If the autodetection fail then the file cannot be read.

Value

a list of 4xL matrices representing motifs (four nucleotides as rows)

Examples

```
# read in example TRANSFAC motifs without accession codes (just IDs)
readMotifs(system.file(package="PWMEnrich", dir="extdata", file="example.transfac"), remove.acc=TRUE)

# read in the JASPAR insects motifs provided as example
readMotifs(system.file(package="PWMEnrich", dir="extdata", file="jaspar-insecta.jaspar"), remove.acc=TRUE)
```

readTRANSFAC *Read in motifs in TRANSFAC format...*

Description

Read in motifs in TRANSFAC format

Usage

```
readTRANSFAC(file, remove.acc=TRUE)
```

Arguments

file	the filename
remove.acc	if to ignore transfac accession numbers

Value

a list of matrices representing motifs (with four nucleotides as rows)

---

registerCoresPWMEnrich

Register than PWMEnrich can use parallel CPU cores...

Description

Register than PWMEnrich can use parallel CPU cores

Usage

```
registerCoresPWMEnrich(numCores=NA)
```

Arguments

numCores	number of cores to use (default to take all cores), or NULL if no parallel execution is to be used

Details

Certain functions (like motif scanning) can be parallelized in PWMEnrich. This function registers a number of parallel cores (via core package parallel) to be used in code that can be parallelized. After this function is called, all further PWMEnrich function calls will run in parallel if possible.

By default parallel execution is turned off. To turn it off after using it, call this function by passing NULL.

Examples

```
## Not run:
registerCoresPWMEnrich(4) # use 4 CPU cores in PWMEnrich
registerCoresPWMEnrich() # use maximal number of CPUs
registerCoresPWMEnrich(NULL) # do not use parallel execution

## End(Not run)
```

reverseComplement,PWM-method *Reverse complement for the PWM object...*

Description

Reverse complement for the PWM object

Usage

```
## S4 method for signature 'PWM'
reverseComplement(x, ...)
```

Arguments

x	an object of type PWM
...	unused

Details

Finds the reverse complement of the PWM

Value

an object of type PWM that is reverse complement of x

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel.PFM)

reverseComplement(MotifDb.Dmel.PFM$ttk) # reverse complement of the ttk PWM
}
```

scanWithPWM *Scan the whole sequence on both strands...*

Description

Scan the whole sequence on both strands

Usage

```
scanWithPWM(pwm, dna, pwm.rev, odds.score=FALSE, both.strands=FALSE,
    strand.fun="mean")
```

Arguments

pwm	PWM object
dna	a DNAString or other sequence from Biostrings
pwm.rev	the reverse complement for a pwm (if it is already pre-computed)
odds.score	if to return raw scores in odds (not logodds) space
both.strands	if to return results on both strands
strand.fun	which function to use to summarise values over two strands (default is "mean")

Details

The whole sequence is scanned with a PWM and scores returned beginning at each position. Partial motif matches are not done, thus the last #[length of motif]-1 scores are NA.

The function returns either an odds average (*not* log-odds average), maximal score on each strand, or scores on both strands.

The function by default returns the score in log2 following the package Biostrings.

Value

a vector representing scores starting at each position, or a matrix with score in the two strands

Examples

```
if(require("PWMEnrich.Dmelanogaster.background")){
data(MotifDb.Dmel)

scanWithPWM(MotifDb.Dmel$ttk, DNAString("CGTAGGATAAAGTAACT")) # odds average over the two strands expressed as log
scanWithPWM(MotifDb.Dmel$ttk, DNAString("CGTAGGATAAAGTAACT"), both.strands=TRUE) # log2-odds scores on both stran
}
```

---

seqLogoGrid *Draw a motif logo on an existing viewport...*

Description

Draw a motif logo on an existing viewport

Usage

```
seqLogoGrid(pwm, ic.scale=TRUE, xaxis=TRUE, yaxis=TRUE, xfontsize=10, yfontsize=10,
    xmargin.scale=1, ymargin.scale=1, title="", titlefontsize=15)
```

Arguments

pwm	numeric The 4xW position weight matrix.
ic.scale	logical If TRUE, the height of each column is proportional to its information content. Otherwise, all columns have the same height.
xaxis	logical If TRUE, an X-axis will be plotted.
yaxis	logical If TRUE, a Y-axis will be plotted.
xfontsize	numeric Font size to be used for the X-axis.
yfontsize	numeric Font size to be used for the Y-axis.
xmargin.scale	the scaling parameter for the X-axis margin. Useful when plotting more than one logo on a page
ymargin.scale	the scaling parameter for the Y-axis margin. Useful when plotting more than one logo on a page
title	to be shown on the top
titlefontsize	the fontsize of the title

Details

This function comes from the seqLogo package. It has been modified to remove some unneccessary code as suggested by W Huber (https://stat.ethz.ch/pipermail/bioconductor/2010-September/035267.html).

Use this function for more advanced plotting where the viewports are directly set up and maintained (see package grid).

---

show,MotifEnrichmentResults-method

show method for MotifEnrichmentResults...

Description

show method for MotifEnrichmentResults

Usage

```
## S4 method for signature 'MotifEnrichmentResults'
show(object)
```

Arguments

object	the MotifEnrichmentResults object

show,PWM-method *show method for PWM...*

Description

show method for PWM

Usage

```
## S4 method for signature 'PWM'
show(object)
```

Arguments

object	the PWM object

show,PWMCutoffBackground-method *show method for PWMCutoffBackground...*

Description

show method for PWMCutoffBackground

Usage

```
## S4 method for signature 'PWMCutoffBackground'
show(object)
```

Arguments

object	the PWMCutoffBackground object

show,PWMEmpiricalBackground-method
show method for PWMEmpiricalBackground...

Description

show method for PWMEmpiricalBackground

Usage

```
## S4 method for signature 'PWMEmpiricalBackground'
show(object)
```

Arguments

object	the PWMEmpiricalBackground object

show,PWMGEVBackground-method
show method for PWMGEVBackground...

Description

show method for PWMGEVBackground

Usage

```
## S4 method for signature 'PWMGEVBackground'
show(object)
```

Arguments

object	the PWMGEVBackground object

`show,PWMLognBackground-method` *show method for PWMLognBackground...*

Description

show method for PWMLognBackground

Usage

```
## S4 method for signature 'PWMLognBackground'
show(object)
```

Arguments

`object`	the PWMLognBackground object

`tryAllMotifAlignments` *Try all motif alignments and return max score...*

Description

Try all motif alignments and return max score

Usage

```
tryAllMotifAlignments(m1, m2, min.align=2, exclude.zero=FALSE)
```

Arguments

`m1`	frequency matrix of motif 1
`m2`	frequency matrix of motif 2
`min.align`	minimal number of basepairs that need to align
`exclude.zero`	if to exclude offset=0, useful for calculating self-similarity

Details

This function tries all offsets of motif1 compared to motif2 and returns the maximal (unnormalized) correlation score.

The correlation score is essentially the sum of correlations of individual aligned columns as described in Pietrokovski (1996).

Value

single maximal score

References

Pietrokovski S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucleic Acids Res 1996;24:3836-3845.

`useBigMemoryPWMEnrich` *If to use a faster implementation of motif scanning that requires abount 5 to 10 times more memory...*

Description

If to use a faster implementation of motif scanning that requires abount 5 to 10 times more memory

Usage

```
useBigMemoryPWMEnrich(useBigMemory=FALSE)
```

Arguments

`useBigMemory`	a boolean value denoting if to use big memory implementation

Examples

```
## Not run:
useBigMemoryPWMEnrich(TRUE) # switch to big memory implementation globally
useBigMemoryPWMEnrich(FALSE) # switch back to default implementation

## End(Not run)
```

Index