Package 'LowMACA'

October 12, 2016

Type Package

Title LowMACA - Low frequency Mutation Analysis via Consensus Alignment

Version 1.4.2

Date 2015-04-29

Author Stefano de Pretis , Giorgio Melloni

Maintainer
Stefano de Pretis <ste.depo@gmail.com>, Giorgio Melloni <melloni.giorgio@gmail.com>

Description The LowMACA package is a simple suite of tools to investigate and analyze the mutation profile of several proteins or pfam domains via consensus alignment. You can conduct an hypothesis driven exploratory analysis using our package simply providing a set of genes or pfam domains of your interest.

License GPL-3

Depends R (>= 2.10)

Imports cgdsr, parallel, stringr, reshape2, data.table, RColorBrewer, methods, LowMACAAnnotation, BiocParallel, motifStack, Biostrings

Suggests BiocStyle, knitr, rmarkdown

VignetteBuilder knitr

biocViews SomaticMutation, SequenceMatching, WholeGenome, Sequencing, Alignment, DataImport, MultipleSequenceAlignment

SystemRequirements clustalo, gs, perl

NeedsCompilation no

R topics documented:

`LowMACA-package` *LowMACA : Low frequency Mutations Analysis via Consensus Alignment*

Description

The LowMACA package is a simple suite of tools to investigate and analyze the mutation profile of several proteins or pfam domains via consensus alignment. You can conduct an hypothesis driven exploratory analysis using our package simply providing a set of genes or pfam domains of your interest.

Details

LowMACA allows to collect, align, analyze and visualize mutations from different proteins or pfam domains.

1. *newLowMACA:* construct a LowMACA object with your proteins or pfam
2. *setup:* align sequences, get mutations and map mutations on the consensus sequence
3. *entropy:* calculate entropy score and pvalues for every position
4. *lfm:* retrieve significant position
5. *lmPlot:* visualize mutations on the consensus sequence, conservation and significant clusters

Author(s)

Stefano de Pretis , Giorgio Melloni

Maintainer: <ste.depo@gmail.com> <melloni.giorgio@gmail.com>

References

Melloni GEM, de Pretis S, Riva L, et al. LowMACA: exploiting protein family analysis for the identification of rare driver mutations in cancer. BMC Bioinformatics. 2016;17:80. doi:10.1186/s12859-016-0935-7

See Also

LowMACA project website

Examples

```
#Create an object of class LowMACA for RAS domain family
lm <- newLowMACA(pfam="PF00071" , genes=c("KRAS" , "NRAS" , "HRAS"))
#Select melanoma, breast cancer and colorectal cancer
lmParams(lm)$tumor_type <- c("skcm" , "brca" , "coadread")
#Align sequences, get mutation data and map them on consensus
lm <- setup(lm)
#Calculate statistics
lm <- entropy(lm)
#Retrieve original mutations
lfm(lm)
#Plot
bpAll(lm)
lmPlot(lm)
protter(lm)
```

---

alignSequences *Align sequences via clustalo*

Description

Align sequences for an object of class LowMACA

Usage

```
alignSequences(object, clustalo_filename=NULL , mail=NULL ,
perlCommand="perl", use_hmm=FALSE, datum=FALSE)
```

Arguments

`object`	an object of class LowMACA containing at least 2 sequences.
`clustalo_filename`	a character string that contains the file name where clustal omega alignment file will be stored. In case it's NULL no file will be written. Default=NULL
`mail`	a character string indicating the email address where error report should be sent in web mode
`perlCommand`	a character string containing the path to Perl executable. if missing, "perl" will be used as default
`use_hmm`	When analysing Pfam sequences, it is possible to use the Hidden Markov Model (HMM) of the specific Pfam to align the sequences. Default is FALSE.
`datum`	When analysing Pfam sequences, use all the genes that belong to the Pfam to generate the alignment. This creates a unique mapping between individual residues and consensus sequence, disregarding the set of sequences that are selected for the analysis. Default is FALSE.

Details

This method launches a system call to clustalo aligner and optionally creates a fasta file in clustal format. A warning is returned if at least one sequence has a pairwise similarity below 20% to any other sequence. If only one sequence is passed to `alignSequences`, the alignment will be skipped, but no warning will be raised. If mail is not NULL, a local installation of clustal omega is no longer required and the alignment is performed using clustal omega EBI web service. A limit of 2000 sequences is set in this case and perl must be installed in the system

Value

The method returns an object of class LowMACA updating the slot `alignment`. See `lmAlignment`

Warning

When a sequence has a similarity below 20%, a warning is raised. In order to produce strong results in terms of conservation of multiple mutations, consider to remove that sequence from the analysis. The alignment will obviously change.

Author(s)

Stefano de Pretis, Giorgio Melloni

References

Trident Score Clustal Omega Clustal Omega Web Service

See Also

`getMutations` , `mapMutations` , `setup`

Examples

```
#Create an object of class LowMACA for RAS domain family
lm <- newLowMACA(pfam="PF00071" , genes=c("KRAS" , "NRAS" , "HRAS"))
#Align sequences using local installation of clustalo
lm <- alignSequences(lm)
#Web service clustalomega aligner
lm <- alignSequences(lm , mail="lowmaca@gmail.com")
#Use HMM to align
lm <- alignSequences(lm , use_hmm=TRUE)
#Use "datum"
lm <- alignSequences(lm , datum=TRUE)
```

`allPfamAnalysis` *Global analysis of a repository of mutations*

Description

Given a repository of mutations, the method `allPfamAnalysis` launches the analysis of all the Pfams and single sequences which are involved with at least one mutation.

Usage

```
allPfamAnalysis(repos
   , allLowMACAObjects=NULL
, mutation_type=c("missense", "all", "truncating" , "silent")
, NoSilent=TRUE
, mail=NULL
, perlCommand="perl"
, verbose=FALSE
, conservation=0.1
, use_hmm=FALSE
, datum=FALSE
, clustal_cmd="clustalo"
, BPPARAM=bpparam("SerialParam"))
```

Arguments

`repos`	either a data.frame or a filename containing the data to analyze
`allLowMACAObjects`	filename of a RData file to save all the LowMACA object `allPfamsLM` produced by the function. It can be usefull for plotting a specific Pfam after the analysis, but it can be a pretty large object. Default NULL
`mutation_type`	type of mutation to be considered for the analysis. Default to missense.
`NoSilent`	logical indicating if Silent mutations should be deleted or not. Default TRUE
`mail`	if not NULL, it must be a valid email address to use EBI clustalo web service. Default is to use a local clustalo installation

perlCommand	a character string containing the path to Perl executable. if missing, "perl" will be used as default. Only used if mail is set
verbose	logical. verbose output or not
conservation	a number between 0 and 1. Represents the minimum level of conservation to test a mutation
use_hmm	When analysing Pfam sequences, it is possible to use the Hidden Markov Model (HMM) of the specific Pfam to align the sequences. Default is FALSE.
datum	When analysing Pfam sequences, use all the genes that belong to the Pfam to generate the alignment. This creates a unique mapping between individual residues and consensus sequence, disregarding the set of sequences that are selected for the analysis. Default is FALSE.
clustal_cmd	path to clustalomega executable. default is to check "clustalo" in the PATH
BPPARAM	An object of class BiocParallelParam specifiying parameters related to the parallel execution of some of the tasks and calculations within this function. See function bpparam() from the BiocParallel package.

Details

This function takes a data.frame or a tab delimited text file in LowMACA format (see LowMACA_AML) and perform a full analysis of the dataset. It basically divide the mutations into their Pfam and launch many LowMACA analysis as many Pfam are hit by mutations up to the lfm function. Every significant position after lfm is tested at gene level. A binomial test is performed to see if the ratio between the number of mutations in the significant position over the total number of mutations is higher than expected by chance at gene level. The significant mutations of all the lfm functions are aggregated in one single data.frame.

Value

A list of two dataframes named 'AlignedSequence' and 'SingleSequence'

The first dataframe is the result of the alignment based analysis. Every gene is aggregated by its corresponding Pfam domain.

Gene_Symbol	gene symbols of the analyzed genes
Multiple_Aln_pos	positions in the consensus relatively to the sequence analyzed.
Pfam_ID	Pfam name analyzed
binomialPvalue	pvalue of the single gene test, See details
Amino_Acid_Position	amino acidic positions relative to original protein
Amino_Acid_Change	amino acid changes in hgvs format
Sample	Sample barcode where the mutation was found
Tumor_Type	Tumor type of the Sample
Envelope_Start	start of the pfam domain in the protein
Envelope_End	end of the pfam domain in the protein

metric	qvalue of the position in the multiple alignment of Pfam domains
Entrez	entrez ids of the mutations
Entry	Uniprot entry of the protein
UNIPROT	other protein names for Uniprot
Chromosome	cytobands of the genes
Protein.name	extended protein names

The second dataframe represent the result of LowMACA on every couple gene-domain when it is not aligned with any other member of the same Pfam ID.

Gene_Symbol	gene symbols of the analyzed genes
Amino_Acid_Position	amino acidic positions relative to original protein
Amino_Acid_Change	amino acid changes in hgvs format
Sample	Sample barcode where the mutation was found
Tumor_Type	Tumor type of the Sample
Envelope_Start	start of the pfam domain in the protein
Envelope_End	end of the pfam domain in the protein
Multiple_Aln_pos	positions in the consensus relatively to the sequence analyzed. See warnings section
Entrez	entrez ids of the mutations
Entry	Uniprot entry of the protein
UNIPROT	other protein names for Uniprot
Chromosome	cytobands of the genes
Protein.name	extended protein names

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

lfm, LowMACA_AML

Examples

```
#Load Homeobox example
data(lmObj)
#Extract the data inside the object as a toy example
myData <- lmMutations(lmObj)$data
#Run allPfamAnalysis on every mutations
significant_muts <- allPfamAnalysis(repos=myData)
#Show the result of alignment based analysis
head(significant_muts$AlignedSequence)
```

```
#Show all the genes that harbor significant mutations
unique(significant_muts$AlignedSequence$Gene_Symbol)
#Show the result of the Single Gene based analysis
head(significant_muts$SingleSequence)
#Show all the genes that harbor significant mutations
unique(significant_muts$SingleSequence$Gene_Symbol)
```

BLOSUM62 *BLOSUM62 matrix*

Description

A substitution matrix used for sequence alignment of proteins. In LowMACA, it is used to calculate the trident conservation score.

Usage

```
data("BLOSUM62")
```

Format

A squared numeric matrix with aminoacids as rownames and colnames

Source

BLOSUM62 from NCBI

Examples

```
#Load BLOSUM62 and show its structure
data(BLOSUM62)
str(BLOSUM62)
```

---

bpAll *Draw a mutation barplot*

Description

bpAll draws a stacked barplot of the mutations mapped on the consensus sequence

Usage

```
bpAll(object)
```

Arguments

object	an object of class LowMACA

Details

Returns a barplot in which mutations are stacked per position on the consensus sequence. Every color represent mutations taht map on the same input sequence (either a protein or a pfam) The LowMACA object must pass through the methods `alignSequences` , `getMutations` , `mapMutations`

Value

NULL

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

`lmPlot`

Examples

```
#Load homeobox example and draw plot
data(lmObj)
lmObj <- entropy(lmObj)
bpAll(lmObj)
```

`entropy` *Calculate LowMACA statistics*

Description

`entropy` is a method for objects of class LowMACA. It calculates global entropy score of the mutation profile of the alignment and a test for every position in the consensus comparing the number of observed mutations against a weigthed random uniform distribution.

Usage

```
entropy(object, bw = NULL , conservation=0.1)
```

Arguments

`object`	an object of class LowMACA
`bw`	a character string or a numeric positive value representing the desired bandwith to launch the function density for the uniform distribution. 0 will not launch density (every position is not aggregated to the surrounded ones) , 'auto' will let the simulation decide according to the Silverman's rule of thumb and every other number is a user defined bandwidth passed to the function `density`.
`conservation`	a number between 0 and 1. Represents the minimum level of conservation to test a mutation

Details

The parameter bw overwrites the bandwidth set with lmParams. Therefore, if bw is set to NULL, the method entropy uses the predefined bandwidth of the LowMACA object.

Value

entropy returns an object of class LowMACA updating the slot entropy and the slot alignment. The slot entropy becomes a list of 6 elements:

- bw the bandwidth used to calculate the null profile
- uniform a function to calculate the null profile
- absval absolute value of entrpy calculated
- log10pval p value of the entropy test in log 10
- pvalue p value of the entropy test
- conservation_thr the minimum conservation level accepted

The slot alignment is updated in the df element by adding 6 new columns

- mean a numeric vector representing the mean value of the empirical uniform function at every position in the consensus
- lTsh a numeric vector representing the limit inferior of the 95% confidence interval of the empirical uniform function at every position in the consensus
- uTsh a numeric vector representing the limit superior of the 95% confidence interval of the empirical uniform function at every position in the consensus
- profile a numeric vector representing the density of mutations at every position in the sample normalized by the number of position. In case of bandwidth 0, this vector is equal to the number of mutations divided by the total number of mutations
- pvalue a numeric vector representing the pvalue of the number of mutations found at every position against the weigthed random uniform distribution of mutations
- qvalue a numeric vector representing the corrected pvalues using FDR method. Only positions with a conservation score >= 10% are considered

Author(s)

Stefano de Pretis , Giorgio Melloni

References

doi:10.1186/gm563 923 Melloni et al.: *DOTS-Finder: a comprehensive tool for assessing driver genes in cancer genomes.* Genome Medicine 2014 6:44

Silverman, B. W. (1986) *Density Estimation*. London: Chapman and Hall.

See Also

alignSequences lmParams lmEntropy

Examples

```
#Load homeobox example and run entropy
data(lmObj)
lmObj <- entropy(lmObj)
lmEntropy(lmObj)
```

---

getMutations *Retrieve mutation data for a LowMACA object*

Description

Exploting the capabilities of the cgdsr package, this method downloads and parse the mutation data of the specified genes in the selected tumor types. It also aggregates and show the frequencies of mutations of every gene in the different tumor types.

Usage

```
getMutations(object, repos = NULL)
```

Arguments

object	a LowMACA class object
repos	a data.frame containing mutations for the specified genes in the LowMACA object in case of custom mutation data. Default NULL

Details

With repos=NULL, the method is a wrapper around getMutationData method from package cgdsr-package. The output of the method is moduled by the parameters in lmParams("LowMACA_object"). See lmParams for further information.

Value

An object of class LowMACA is returned with an update in the slot mutations. See lmMutations method.

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

lmParams getMutationData lmMutations

Examples

```
#Create an object of class LowMACA
lm <- newLowMACA(pfam="PF12906")
#Change some paramters
#By default, LowMACA retrieve only missense mutations.
#We want all mutations
lmParams(lm)[['mutation_type']] <- 'all'
#By default, LowMACA takes mutations from all the kinds of tumor
#We want just prostate cancer samples
lmParams(lm)[['tumor_type']] <- 'prad'
lm <- getMutations(lm)
```

`lfm` *Show significant clusters of mutations*

Description

The method `lfm` (low frequency mutations) retrieve the original mutations that created the significant clusters calculated with `entropy` on the consensus

Usage

```
lfm(object , metric='qvalue', threshold=.05, conservation=NULL)
```

Arguments

`object`	a LowMACA class object
`metric`	a character that defines whether to use 'pvalue' or 'qvalue' to select significant positions. Default: 'qvalue'
`threshold`	a numeric defining the threshold of significance for the defined metric. Default: 0.05
`conservation`	a numeric value in the range of 0-1 that defines the threshold of trident conservation score to include the specified position. The default value is inherited from the slot entropy, whose default is 0.1

Details

After the alignment, we lose every information about the original sequences used as input. The consensus sequence is in fact an alignment that could not represent the reality of human proteins. `lfm` allows to go back on the original dataset and retrieve the proteins and the real positions of the mutations that we consider 'conserved'.

Value

A data.frame with 13 columns corresponding to the mutations retrieved:

1. Gene_Symbol gene symbols of the mutations
2. Amino_Acid_Position amino acidic positions relative to original protein
3. Amino_Acid_Change amino acid changes in hgvs format
4. Sample Sample barcode where the mutation was found
5. Tumor_Type Tumor type of the Sample
6. Envelope_Start start of the pfam domain in the protein
7. Envelope_End end of the pfam domain in the protein
8. Multiple_Aln_pos positions in the consensus
9. Entrez entrez ids of the mutations
10. Entry Uniprot entry of the protein
11. UNIPROT other protein names for Uniprot
12. Chromosome cytobands of the genes
13. Protein.name extended protein names

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

entropy

Examples

```
#Load homeobox example and launch entropy method
data(lmObj)
lmObj <- entropy(lmObj)
significant_muts <- lfm(lmObj)
#Display original mutations that formed significant clusters (column Multiple_Aln_pos)
head(significant_muts)
#Position 4 has a qvalue<0.05
#What are the genes mutated in position 4 in the consensus?
cluster_4_genes <- significant_muts[ significant_muts[['Multiple_Aln_pos']]==4 , 'Gene_Symbol']
#Display the genes and their number of mutation in consensus position 4
sort(table(cluster_4_genes))
```

`lfmSingleSequence` *Show significant clusters of mutations of every gene in a LowMACA object without alignment*

Description

The method `lfmSingleSequence` (low frequency mutations in Single Sequence) launch `lfm` method on every gene or domain inside a LowMACA object without aligning the sequences

Usage

```
lfmSingleSequence(object , metric='qvalue', threshold=.05
, conservation=0.1 , BPPARAM=bpparam("SerialParam") , mail=NULL , perlCommand="perl"
,verbose=FALSE)
```

Arguments

`object`	a LowMACA class object
`metric`	a character that defines whether to use 'pvalue' or 'qvalue' to select significant positions. Default: 'qvalue'
`threshold`	a numeric element between 0 and 1 defining the threshold of significance for the defined metric. Default: 0.05
`conservation`	a numeric value in the range of 0-1 that defines the threshold of trident conservation score to include the specified position. Default: 0.1
`BPPARAM`	An object of class `BiocParallelParam` specifiying parameters related to the parallel execution of some of the tasks and calculations within this function. See function `bpparam()` from the `BiocParallel` package.
`mail`	if not NULL, it must be a valid email address to use EBI clustalo web service. Default is to use a local clustalo installation
`perlCommand`	a character string containing the path to Perl executable. if missing, "perl" will be used as default. Only used in web mode
`verbose`	logical. verbose output or not

Details

This function completes a LowMACA analysis by analyzing every gene or domain in the LowMACA object as a 'single sequence' analysis was started in the first place. The result is a dataframe showing all the significant positions of every gene. If you have a LowMACA object composed by 100 genes, it will launch 100 LowMACA single gene analyses and aggregates the results of every `lfm` launched on these 100 objects. The output looks very similar to `lfm`, but in this case the column Multiple_Aln_pos has a different meaning. While in `lfm` it shows where the mutation falls in the consensus sequence, in this case it must be intended the consensus within the gene. If the original LowMACA object had mode equal to 'gene', the column Multiple_Aln_pos will be always equal to Amino_Acid_Position. If mode is 'pfam', it is the same unless a gene harbors more than one domain of the same type within its sequence. In that case, an internal alignment of every domain inside the protein is performed.

Value

A data.frame with 10 columns corresponding to the mutations retrieved:

1. Gene_Symbol gene symbols of the analyzed genes
2. Amino_Acid_Position amino acidic positions relative to original protein
3. Amino_Acid_Change amino acid changes in hgvs format
4. Sample Sample barcode where the mutation was found
5. Tumor_Type Tumor type of the Sample
6. Envelope_Start start of the pfam domain in the protein
7. Envelope_End end of the pfam domain in the protein
8. Multiple_Aln_pos positions in the consensus relatively to the sequence analyzed. See warnings section
9. Entrez entrez ids of the mutations
10. Entry Uniprot entry of the protein
11. UNIPROT other protein names for Uniprot
12. Chromosome cytobands of the genes
13. Protein.name extended protein names

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

lfm

Examples

```
#Load homeobox example
data(lmObj)
#Run lfmSingleSequence
significant_muts <- lfmSingleSequence(lmObj)
#Show the result
head(significant_muts)
#Show all the genes that harbor significant mutations without the alignment
unique(significant_muts$Gene_Symbol)
```

`lmAlignment` *Show Alignment Results from a LowMACA object*

Description

Method for objects of class LowMACA. It can show the results of the alignment procedure that has been performed on the LowMACA object

Usage

```
lmAlignment(object)
```

Arguments

`object`	object of class LowMaca

Value

A list containing the following elements:

- ALIGNMENT an object of class data.frame containing the mapping of the position of the original amino acids to the consensus sequence
- SCORE a list of two objects
  - DIST_MAT a matrix of the pairwise similarities between sequences as resulted after the multiple alignment (from 0% to 100%)
  - SUMMARY_SCORE a data.frame containing summary descriptives of the distance matrix
  - CLUSTAL an object of class "AAMultipleAlignment" as provided by Biostrings R package
  - df a dataframe containing the predicted consesus sequence and the trident conservation score at every position

Author(s)

Stefano de Pretis, Giorgio Melloni

See Also

`alignSequences`

Examples

```
data('lmObj')
str(lmAlignment(lmObj))
```

`lmEntropy` *Show Entropy Information Contained in a LowMACA object*

Description

Method for objects of class LowMACA. It can show the results of entropy analysis performed on the LowMACA object by the function `entropy`

Usage

```
lmEntropy(object)
```

Arguments

`object`	object of class LowMaca

Value

A list containing the following elements:

- bw a numeric value that represents the bandwidth used to calculate the Shannon entropy score
- uniform an object of class function that was used to calculate the score
- absval a numeric value representing the Shannon entropy of the sample data
- log10pval a numeric value representing the pvalue of the Shannon entropy score against a gamma distribution with same mean and variance as the empirical uniform distribution in -log10 scale
- pvalue a numeric value representing the pvalue of the Shannon entropy score against a gamma distribution with same mean and variance as the empirical uniform distribution

Author(s)

Stefano de Pretis, Giorgio Melloni

See Also

`entropy`

Examples

```
data('lmObj')
lmObj <- entropy(lmObj)
lmEntropy(lmObj)
```

`lmMutations` *Show Mutation Data Contained in a LowMACA object*

Description

Method for objects of class LowMACA. It can show the mutation data contained within the LowMACA object that has been retrieved from `getMutations` method.

Usage

```
lmMutations(object)
```

Arguments

`object`	object of class LowMaca

Value

A list containing the following elements:

- data a data.frame describing the mutations on every genes and their effect the amino acids they belong to
- freq a data.frame containing the absolute number of mutated patients by gene and selected tumor types (this is useful to explore the mutational landscape of your genes in the different tumor types)
- aligned a matrix where rows represent proteins/pfam, and columns report the number of mutations on every position of the consensus

Author(s)

Stefano de Pretis, Giorgio Melloni

See Also

`getMutations`

Examples

```
data('lmObj')
str(lmMutations(lmObj))
```

lmObj *Example of a LowMACA object*

Description

An object of class LowMACA of the alignment and mapping of the homeobox domain. It is the example used in the vignette.

Usage

```
data("lmObj")
```

Format

An object of class LowMACA

Source

Created by LowMACA package

Examples

```
#Load lmObj and show its structure
data(lmObj)
str(lmObj)
```

lmParams *Show and set parameters*

Description

Method for objects of class LowMACA. It can show the most important user-definable parameters for a LowMACA analysis and allows to change them.

Usage

```
lmParams(object)
lmParams(object) <- value
```

Arguments

`object` an object of class LowMaca

`value` a named list containing:

1. `mutation_type` a character string among: 'missense' , 'truncating' , 'silent' ,'all'. Default 'missense'
2. `tumor_type` a character vector or string containing the tumor type barcode of the data in cBioPortal. Default 'all'.
3. `min_mutation_number` an integer value describing the minimum number of mutations accepted for a sequence. If a sequence does not harbor a sufficient number of mutations is discarded from the analysis. Default is 1
4. `density_bw` either a numeric value or 'auto'. A numeric value is passed directly to the function density while putting 0 will not launch density at all (every position is not aggregated to the surrounded ones). 'auto' will let the simulation decide according to the Silverman's rule of thumb the correct bandwidth. Default is 0.
5. `clustal_cmd` path to clustalo executable
6. `use_hmm` When analysing Pfam sequences, it is possible to use the Hidden Markov Model (HMM) of the specific Pfam to align the sequences. Default is FALSE.
7. `datum` When analysing Pfam sequences, use all the genes that belong to the Pfam to generate the alignment. This creates a unique mapping between individual residues and consensus sequence, disregarding the set of sequences that are selected for the analysis. Default is FALSE.

Details

LowMACA is a suite of tool that analyze conserved mutations, so it looks for clusters of gain of function alterations. With 'missense' mutation_type we intend all those mutations that change the original DNA but do not create stop codon nor alter the reading frame (these mutations are collectively defined as 'truncating' mutations). In addition we let the possibility to also choose 'silent' mutations even though they are currently not supported by the cBioPortal. To see all the available tumor types to run a LowMACA analysis, simply run `showTumorType`. The parameter `density_bw` has a strong effect on the statistical analysis of LowMACA. With the default bandwidth (0), the Shannon entropy calculation becomes descrete, while the continuos version is used in all the other cases.

Value

If `lmParams` is used as a show method it returns a named list of 5 elements: `mutation_type='missense'` , `tumor_type='all'` , `min_mutation_number=1` , `density_bw=0` , `clustal_cmd='clustalo'`

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

`showTumorType` `getMutations` `entropy` `density`

Examples

```
#Construct a LowMACA object
lm <- newLowMACA(pfam="PF12906")
#Show default parameters
lmParams(lm)
#Change all parameters
lmParams(lm) <- list(mutation_type='all'
                     , tumor_type=c('skcm','brca')
                     , min_mutation_number=0
                     , density_bw=0
                     , clustal_cmd='clustalo'
                     , use_hmm=FALSE
                     , datum=FALSE)
#Change just one parameter
lmParams(lm)[['tumor_type']] <- 'prad'
```

---

lmPlot *Draw a comprehensive LowMACA plot*

Description

LowMACA comprehensive plot is a four layers plot that summarize the entire LowMACA output

Usage

```
lmPlot(object , conservation=NULL, splitLen=NULL)
```

Arguments

object	a LowMACA class object
conservation	a numeric value in the range of 0-1 that defines the threshold of trident conservation score to include the specified position. The default value is inherited from the slot entropy, whose default is 0.1
splitLen	An integer, defines after how many amino acids the plot should be split By default this parameter is set to NULL, that mean that the plot is not split.

Details

The method returns a plot, which is divided into four layers. The LowMACA object must have been passed through the methods `alignSequences` , `getMutations` , `mapMutations` and `entropy`. The four layers of the plot are:

1. The bar plot visualized by `bpAll`
2. The distribution of mutations against the 95% confidence interval superior limit of the null hypothesis (dotted line) with orange bars representing a position with a pvalue <0.05 and a red star for qvalue<0.05
3. The Trident score distribution

4. The logo plot representing the consensus sequence

If this plot is used on a LowMACA object with a single protein, the result is formed by three layers only:

1. The bar plot visualized by `bpAll`
2. The Pfam domains structure inside the protein
3. The distribution of mutations against the 95% confidence interval superior limit of the null hypothesis (dotted line) with orange bars representing a position with a pvalue <0.05 and a red star for qvalue<0.05

Value

NULL

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

`alignSequences` `getMutations` `mapMutations` `entropy` `bpAll`

Examples

```
#Load homeobox example and draw the plot
data(lmObj)
#Calculate statistics for nullProfile
lmObj <- entropy(lmObj)
lmPlot(lmObj)
```

---

lmPlotSingleSequence *Draw a LowMACA comprehensive plot of a specified gene within a LowMACA object*

Description

LowMACA comprehensive plot is a four layers plot that summarize the entire LowMACA output

Usage

```
lmPlotSingleSequence(object , gene , mail=NULL , perlCommand="perl")
```

Arguments

object	a LowMACA class object
gene	a Gene Symbol that identifies one of the gene analyzed in the LowMACA object
mail	if not NULL, it must be a valid email address to use EBI clustalo web service. Default is to use a local clustalo installation
perlCommand	a character string containing the path to Perl executable. if missing, "perl" will be used as default. Only used in web mode

Details

If the specified gene has more than one domain of the same type and mode is pfam, the plot is composed by four layers:

1. The bar plot visualized by bpAll
2. The distribution of mutations against the 95% confidence interval superior limit of the null hypothesis (dotted line) with orange bars representing a position with a pvalue <0.05 and a red star for qvalue<0.05
3. The Trident score distribution
4. The logo plot representing the consensus sequence

If the specified gene has only one domain of the same type and mode is pfam, the plot is composed by two layers:

1. The bar plot visualized by bpAll
2. The distribution of mutations against the 95% confidence interval superior limit of the null hypothesis (dotted line) with orange bars representing a position with a pvalue <0.05 and a red star for qvalue<0.05

If mode is gene, the plot is composed by three layers:

1. The bar plot visualized by bpAll
2. The Pfam domains structure inside the protein
3. The distribution of mutations against the 95% confidence interval superior limit of the null hypothesis (dotted line) with orange bars representing a position with a pvalue <0.05 and a red star for qvalue<0.05

Value

NULL

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

lmPlot bpAll

Examples

```
#Load homeobox example and draw the plot
data(lmObj)
#DUXA has a significant cluster of mutation
#Plot Mutations on DUXA gene in the
#original sequences of its domains PF00046
lmPlotSingleSequence(lmObj , gene="DUXA")
```

LowMACA-class *Class* "LowMACA"

Description

LowMACA class object describing the properties of mutations mapped on pfam domains or proteins

Objects from the Class

Objects can be created by calls of the form `newLowMACA(genes, pfam)`.

Constructor

newLowMACA(genes=character_vector , pfam=character_vector)

Slots

arguments Object of class `"list"` with 6 elements:

- genes : vector of selected genes for the analysis in Hugo names format. NULL if `mode="pfam"`.
- pfam : vector of selected domains for the analysis in pfam ids format. NULL if `mode="genes"`.
- input : data.frame describing the input data as gene symbols, pfam ids, entrez ids, envelope start and end of the domain relative to the protein, name of the canonical protein in uniprot format, amino acidic sequence.
- mode : character. automatically set by the constructor as either `"pfam"` or `"genes"`. If `pfam=NULL` then `mode="genes"`, `"pfam"` otherwise.
- params : named list of starting parameters for the LowMaca analysis. Call `lmParams(object)` to show default. See `lmParams` for further details.
- parallelize : named list of logicals. `getMutations`=FALSE is the default for the `getMutations` method and `makeAlignment`=TRUE is the default for the `alignSequences` method. See `parallelize` for further details.

alignment Object of class `"list"` with 4 elements:

- ALIGNMENT : data.frame of the result of the alignment. Every row represents a position of a sequence and the relative mapping to the consensus sequence.
- SCORE : list of two elements. `DIST_MAT` is a matrix of pairwise similarities between sequences as described by clustalo. `SUMMARY_SCORE` is a dataframe of summary descriptive statistics of the `DIST_MAT` matrix

- CLUSTAL : an object of class AAMultipleAlignment from package Biostrings
- df : a data.frame describing the consensus sequence, its per-position degree of conservation and its mutations null profile density. See entropy and lmPlot for further details

mutations Object of class "list" with 3 elements:

- data : data.frame derived from the query to the cBioPortal query, getMutationData Every row represents a mutation stratified by position, gene and tumor type.
- freq : data.frame of absolute frequency of mutation stratified by gene and tumor type.
- aligned : matrix representing the number of mutations at every position in the consensus sequence (columns) and in each original sequence (rows)

entropy Object of class "list" with 5 elements:

- bw : numeric value. user defined bandwidth for the function entropy
- uniform : function that generate the uniform null profile
- absval : numeric value. Shannon entropy of the mutation data profile according to the defined bandwidth
- log10pval : numeric value. pvalue of the entropy test in -log10 scale
- pvalue : numeric value. pvalue of the entropy test

Methods

alignSequences alignSequences(object = "LowMACA"): ...

bpAll bpAll(object = "LowMACA"): ...

entropy entropy(object = "LowMACA"): ...

getMutations getMutations(object = "LowMACA"): ...

lfm lfm(object = "LowMACA"): ...

lmPlot lmPlot(object = "LowMACA"): ...

mapMutations mapMutations(object = "LowMACA"): ...

nullProfile signature(object = "LowMACA"): ...

parallelize parallelize(object = "LowMACA"): ...

parallelize<- signature(object = "LowMACA"): ...

lmParams params(x = "LowMACA"): ...

lmParams<- signature(object = "LowMACA"): ...

protter protter(object = "LowMACA"): ...

setup setup(object = "LowMACA"): ...

show show(object = "LowMACA"): ...

lfmSingleSequence lfmSingleSequence(object = "LowMACA"): ...

lmPlotSingleSequence lmPlotSingleSequence(object = "LowMACA"): ...

Author(s)

Stefano de Pretis, Giorgio Melloni

References

LowMACA website

See Also

newLowMACA

Examples

```
#ANALYSIS OF SOME OF THE PROTEINS THAT SHARE THE HOMEOBOX DOMAIN
#Genes to analyze
Genes <- c("ADNP","ALX1","ALX4","ARGFX","CDX4","CRX"
   ,"CUX1","CUX2","DBX2","DLX5","DMBX1","DRGX"
,"DUXA","ESX1","EVX2","HDX","HLX","HNF1A"
,"HOXA1","HOXA2","HOXA3","HOXA5","HOXB1","HOXB3"
,"HOXD3","ISL1","ISX","LHX8")
#Pfam to analyze
Pfam <- "PF00046"
#Construct a new LowMACA object
lm <- newLowMACA(genes=Genes , pfam=Pfam)
#Change some parameters
lmParams(lm)[['tumor_type']] <- c("skcm" , "stad" , "ucec" , "luad" , "lusc" , "coadread" , "brca")
lmParams(lm)[['min_mutation_number']] <- 1
lmParams(lm)[['density_bw']] <- 0
#Run if you have clustalo installed
lm <- setup(lm)
#Calculate staistics
lm <- entropy(lm)
#Retrieve original mutations
lfm(lm)
#Plot
bpAll(lm)
lmPlot(lm)
protter(lm)
```

LowMACA_AML *Example of a LowMACA object*

Description

A data frame containing TCGA AML data in the format accepted by LowMACA

Usage

```
data("LowMACA_AML")
```

Format

A data.frame of 8 columns:

1. Entrez gene ID number
2. Gene_Symbol HGNC official gene symbol
3. Amino_Acid_Letter original amino acid letter in the position of the mutation
4. Amino_Acid_Position position of the mutation relative to the protein
5. Amino_Acid_Change amino acid change in hgvs format, like G12V
6. Mutation_Type classification of mutation according to MAF format.
7. Sample name of the sample where the mutation was found
8. Tumor_Type type of tumor, if applicable

Source

Adapted from TCGA ftp repository

See Also

MAF format specification HGVS

Examples

```
#Load LowMACA_AML and show its structure
data(LowMACA_AML)
str(LowMACA_AML)
```

---

mapMutations *Map mutations on consensus sequence*

Description

`mapMutations` is a method for the class LowMACA that re-maps the mutations on a sequence to the relative position in a consensus sequence.

Usage

```
mapMutations(object)
```

Arguments

`object`	an object of class LowMACA

Details

Every position in the consensus alignement correspond to different positions in the single aligned sequences. The mutations are mapped according to this scheme that can be evinced from the slot `alignment`. `mapMutations` must be called after `alignSequences` and `getMutations`

Value

An object of class LowMACA with an update in the slot `mutations`. `mapMutations` add a object named `aligned` of class matrix in this slot that represents the absolute number of mutations in each sequence/position in the consensus as a matrix.

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

`getMutations` `alignSequences` `LowMACA-class`

Examples

```
#Create an object of class LowMACA
lm <- newLowMACA(pfam="PF12906")
#Align the sequences, requires clustalo
## Not run: lm <- alignSequences(lm)
#Get mutations from the corresponding genes
## Not run: lm <- getMutations(lm)
#Map mutations on the consensus sequence
## Not run: lm <- mapMutations(lm)
```

---

newLowMACA *Construct a LowMACA object*

Description

Constructor for the class LowMACA. It initializes a LowMACA object with default parameters

Usage

```
newLowMACA(genes = NULL, pfam = NULL)
```

Arguments

`genes`	a character vector of gene symbols in HGNC format or a integer vector of Entrez IDs. If `pfam` is defined, it can be set to NULL
`pfam`	a character vector of pfam IDs. If `genes` is defined, it can be set to NULL

Details

When a LowMACA object is initialized, the `arguments` slot is filled with the input data and default parameters and path to clustalomega aligner. See `lmParams` and `parallelize` to change them.

Value

An object of class "LowMACA". See LowMACA-class

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

lmParams parallelize

Examples

```
#Set Genes and pfam for the analysis
Genes <- c("ADNP","ALX1","ALX4","ARGFX","CDX4","CRX"
     ,"CUX1","CUX2","DBX2","DLX5","DMBX1","DRGX"
,"DUXA","ESX1","EVX2","HDX","HLX","HNF1A"
,"HOXA1","HOXA2","HOXA3","HOXA5","HOXB1","HOXB3"
,"HOXD3","ISL1","ISX","LHX8")
Pfam <- "PF00046"
#LowMACA object of pfam PF00046 filtered by Genes
lm <- newLowMACA(genes=Genes , pfam=Pfam)
#LowMACA object of the entire pfam PF00046
lm <- newLowMACA(pfam=Pfam)
#LowMACA object of entire canonical proteins associated to Genes
lm <- newLowMACA(genes=Genes)
```

---

nullProfile *Draw a mutational profile plot*

Description

nullProfile is a method for objects of class LowMACA that draw a barplot highlighting the significant clusters of mutations found by LowMACA statistics

Usage

```
nullProfile(object , conservation=NULL, windowlimits=NULL)
```

Arguments

object	an object of class LowMACA
conservation	a numeric value in the range of 0-1 that defines the threshold of trident conservation score to include the specified position. The default value is inherited from the slot entropy, whose default is 0.1
windowlimits	A vector indicating which amino acids residues will be plotted. The vector refers to the positions in the global alignment. By default this parameter is set to NULL, that means that all the amino acids will be displayed.

Details

This method draw the second layer of the `lmPlot` of LowMACA. The blue dotted line is a curve that pass through all the points of the upper limit of the 95% confidence interval for the single position test performed by `entropy` (one point per position in the consensus). The black bars represents the density of mutations in our sample. If a bar passes the blue line, it will be depicted in orange (significant pvalue). After the correction for multiple testing, red stars appears at the top of the orange bars if a cluster is below 0.05 for the qvalue and has a conservation trident score of at least 0.1.

Value

NULL

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

`lmPlot` `entropy`

Examples

```
#Load homeobox example
data(lmObj)
#Calculate statistics
lmObj <- entropy(lmObj)
nullProfile(lmObj)
```

`parallelize` *Show and set parallelization options*

Description

Method for objects of class LowMACA. It can show parallelization parameters of an object of class LowMACA and switch off and on parallelization of `alignSequences` and `getMutations` method

Usage

```
parallelize(object)
parallelize(object) <- value
```

Arguments

`object`	object of class LowMaca
`value`	a named list containing logical values. Default list(getMutations=FALSE , makeAlignment=TRUE)

Details

With `getMutations=TRUE`, the `getMutations` method runs in parallel during the queries to the different tumor_types. This can result in an overload to the cBioPortal database and the function returns error. With `makeAlignment=TRUE`, clustalo should run in parallel. Nevertheless, clustalo can be parallelized only if the OpenMP C library is correctly functioning.

Value

If `parallelize` is used as a show method it returns a named list of two elements: `getMutations` and `makeAlignment`

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

`getMutations`

Examples

```
#Construct a LowMACA object
lm <- newLowMACA(pfam="PF12906")
#Show parallelize default
parallelize(lm)
#Change all parameters
parallelize(lm) <- list(getMutations=TRUE , makeAlignment=FALSE)
#Change just one parameter
parallelize(lm)[['getMutations']] <- TRUE
```

---

protter *Draw a Protter plot*

Description

This is a wrapper around Protter web service for LowMACA class objects that draw a protter style plot.

Usage

```
protter(object, filename = "protter.png", threshold = 0.05 , conservation=NULL)
```

Arguments

`object`	an object of class LowMACA
`filename`	a character string that identifies the file name where protter plot will be stored. Default "protter.png"
`threshold`	a numeric value in the interval (0 , 1] that identifies the significant mutations. Default 0.05
`conservation`	a numeric value in the range of 0-1 that defines the threshold of trident conservation score to include the specified position. The default value is inherited from the slot entropy, whose default is 0.1

Details

Using the information in the slot `alignment`, a request is send to Protter server. Protter will predict a possible sencondary structure for the consensus sequence (if possible) and highlights the significant clusters of mutations found by LowMACA (if any). A significant pvalue is colored in orange, a significant qvalue is colored in red.

Value

NULL

Author(s)

Stefano de Pretis , Giorgio Melloni

References

Protter website

See Also

`LowMACA-class` `entropy`

Examples

```
#Load homeobox example
data(lmObj)
#Calculate statistics
lmObj <- entropy(lmObj)
#Create protter.png
protter(lmObj)
```

setup *Setup of a LowMACA object*

Description

A wrapper around `alignSequences` , `getMutations` and `mapMutations` in order to execute all these methods at once.

Usage

```
setup(object, repos = NULL, clustalo_filename=NULL
, mail=NULL , perlCommand="perl", use_hmm=FALSE, datum=FALSE)
```

Arguments

`object`	an object of class LowMACA
`repos`	a data.frame containing mutations for the specified genes in the LowMACA object in case of custom mutation data. Default NULL
`clustalo_filename`	a character string that contains the file name where clustal omega alignment file will be stored. In case it's NULL no file will be written. Default=NULL
`mail`	a character string indicating the email address where error report should be sent in web mode. Default is NULL, to use a local clustalo installation
`perlCommand`	a character string containing the path to Perl executable. if missing, "perl" will be used as default
`use_hmm`	When analysing Pfam sequences, it is possible to use the Hidden Markov Model (HMM) of the specific Pfam to align the sequences. Default is FALSE.
`datum`	When analysing Pfam sequences, use all the genes that belong to the Pfam to generate the alignment. This creates a unique mapping between individual residues and consensus sequence, disregarding the set of sequences that are selected for the analysis. Default is FALSE.

Details

If mail is not NULL, a local installation of clustal omega is no longer required and the alignment is performed using clustal omega EBI web service. A limit of 2000 sequences is set in this case and perl is required with XML::Simple and LWP modules installed

Value

An object of class LowMACA with all the updates provided by `alignSequences` , `getMutations` and `mapMutations` methods.

Author(s)

Stefano de Pretis , Giorgio Melloni

References

Trident Score Clustal Omega Clustal Omega Web Service

See Also

alignSequences getMutations mapMutations

Examples

```
#Create an object of class LowMACA for RAS domain family
lm <- newLowMACA(pfam="PF00071" , genes=c("KRAS" , "NRAS" , "HRAS"))
#Select a few tumor types
lmParams(lm)$tumor_type <- c("skcm" , "brca" , "coadread")
#Align sequences, get mutation data and map them on consensus
lm <- setup(lm)
#Same as above, but using web service
lm <- setup(lm , mail="lowmaca@gmail.com")
#Use HMM to align
lm <- setup(lm , use_hmm=TRUE)
#Use "datum"
lm <- setup(lm , datum=TRUE)
```

---

showTumorType *List of tumor types*

Description

Show all the possible tumor types accepted by LowMACA

Usage

```
showTumorType()
```

Details

This method is a wrapper around getCancerStudies and show all the barcodes for the tumor types as used by cBioPortal.

Value

A named vector of all the tumor types available in cgdsr package that can be passed to the method lmParams. Every element is the aggregation of all the available sequenced data from all the studies involved in a particular tumor type.

Author(s)

Stefano de Pretis , Giorgio Melloni

See Also

lmParams getCancerStudies

Examples

```
data('lmObj')
out <- showTumorType()
chosenTumors <- out[1:3]
lmParams(lmObj)$tumor_type <- chosenTumors
```

Index