Package 'vDiveR'

January 9, 2024

Type Package

Title Visualization of Viral Protein Sequence Diversity Dynamics

Version 1.2.1

Description To ease the visualization of outputs from Diversity Motif Analyser ('DiMA'; <https://github.com/BVU-BILSAB/DiMA>). 'vDiveR' allows visualization of the diversity motifs (index and its variants – major, minor and unique) for elucidation of the underlying inherent dynamics. Please refer <https://vdiver-manual.readthedocs.io/en/latest/> for more information.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports DT, dplyr, gghalves, ggplot2, ggpubr, grid, gridExtra, ggtext, magrittr, plyr, tidyr, stringr, rlang, rentrez, readr, scales, maps

RoxygenNote 7.2.3

Depends R (>= 2.10)

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

NeedsCompilation no

Author Pendy Tok [aut, cre], Li Chuin Chong [aut], Evgenia Chikina [aut], Yin Cheng Chen [aut], Mohammad Asif Khan [aut]

Maintainer Pendy Tok <pendytok0518@gmail.com>

Repository CRAN

Date/Publication 2024-01-09 20:20:02 UTC

R topics documented:

`concat_conserved_kmer` *k-mer sequences concatenation*

Description

This function concatenates completely (index incidence = 100 index incidence < 100 k-mer position or are adjacent to each other and generate the CCS/HCS sequence in either CSv or FASTA format

Usage

```
concat_conserved_kmer(
  data,
  conservation_level = "HCS",
  kmer = 9,
  threshold_pct = NULL
)
```

Arguments

`data`	DiMA JSON converted csv file data
`conservation_level`	CCS (completely conserved) / HCS (highly conserved)
`kmer`	size of the k-mer window
`threshold_pct`	manually set threshold of index.incidence for HCS

Value

A list wit csv and fasta dataframes

Examples

```
csv<-concat_conserved_kmer(proteins_1host)$csv
csv_2hosts<-concat_conserved_kmer(protein_2hosts, conservation_level = "CCS")$csv
fasta <- concat_conserved_kmer(protein_2hosts, conservation_level = "HCS")$fasta
```

---

extract_from_GISAID *Extract metadata via fasta file from GISAID*

Description

This function get the metadata from each header of GISAID fasta file

Usage

```
extract_from_GISAID(file_path)
```

Arguments

file_path	path of fasta file

---

extract_from_NCBI *Extract metadata via fasta file from ncbi*

Description

This function get the metadata from each head of fasta file

Usage

```
extract_from_NCBI(file_path)
```

Arguments

file_path	path of fasta file

json2csv *JSON2CSV*

Description

This function converts DiMA (v4.1.1) JSON output file to a dataframe with 17 predefined columns which further acts as the input for other functions provided in this vDiveR package.

Usage

```
json2csv(
  json_data,
  host_name = "unknown host",
  protein_name = "unknown protein"
)
```

Arguments

json_data	DiMA JSON output dataframe
host_name	name of the host species
protein_name	name of the protein

Value

A dataframe which acts as input for the other functions in vDiveR package

Examples

```
inputdf<-json2csv(JSON_sample)
```

JSON_sample *DiMA (v4.1.1) JSON Output File*

Description

A sample DiMA JSON Output File which acts as the input for JSON2CSV()

Usage

```
JSON_sample
```

Format

A Diversity Motif Analyzer (DiMA) tool JSON file

metadata *Metadata Input Sample*

Description

A dummy dataset that acts as an input for plot_worldmap() and plot_time()

Usage

```
metadata
```

Format

A data frame with 1000 rows and 3 variables:

ID unique identifier of the sequence

country country of the sequence collection

date collection date of the sequence

metadata_extraction *Metadata Extraction from NCBI/GISAID EpiCoV FASTA file*

Description

This function retrieves metadata (ID, country, date) from the NCBI/GISAID EpiCoV FASTA file (default FASTA header expected).

Usage

```
metadata_extraction(file_path, source)
```

Arguments

file_path	path of fasta file
source	the source of fasta file, either "ncbi" or "GISAID"

Value

A dataframe that has three columns consisting ID, collected country and collected date

Examples

```
filepath <- system.file('extdata','GISAID_EpiCoV.faa', package = 'vDiveR')
meta_gisaid <- metadata_extraction(filepath, 'GISAID')
```

plot_conservationLevel *Conservation Levels Distribution Plot*

Description

This function plots conservation levels distribution of k-mer positions, which consists of completely conserved (black) (index incidence = 100%), highly conserved (blue) (90% <= index incidence < 100%), mixed variable (green) (20% < index incidence <= 90%), highly diverse (purple) (10% < index incidence <= 20%) and extremely diverse (pink) (index incidence <= 10%).

Usage

```
plot_conservationLevel(
  df,
  protein_order = "",
  conservation_label = 1,
  host = 1,
  base_size = 11,
  line_dot_size = 2,
  label_size = 2.6,
  alpha = 0.6
)
```

Arguments

df	DiMA JSON converted csv file data
protein_order	order of proteins displayed in plot
conservation_label	0 (partial; show present conservation labels only) or 1 (full; show ALL conservation labels) in plot
host	number of host (1/2)
base_size	base font size in plot
line_dot_size	lines and dots size
label_size	conservation labels font size
alpha	any number from 0 (transparent) to 1 (opaque)

Value

A plot

Examples

```
plot_conservationLevel(proteins_1host, conservation_label = 1,alpha=0.8, base_size = 15)
plot_conservationLevel(protein_2hosts, conservation_label = 0, host=2)
```

plot_correlation *Entropy and total variant incidence correlation plot*

Description

This function plots the correlation between entropy and total variant incidence of all the provided protein(s).

Usage

```
plot_correlation(
  df,
  host = 1,
  alpha = 1/3,
  line_dot_size = 3,
  base_size = 11,
  ylabel = "k-mer entropy (bits)\n",
  xlabel = "\nTotal variants (%)",
  ymax = ceiling(max(df$entropy)),
  ybreak = 0.5
)
```

Arguments

df	DiMA JSON converted csv file data
host	number of host (1/2)
alpha	any number from 0 (transparent) to 1 (opaque)
line_dot_size	dot size in scatter plot
base_size	base font size in plot
ylabel	y-axis label
xlabel	x-axis label
ymax	maximum y-axis
ybreak	y-axis breaks

Value

A scatter plot

Examples

```
plot_correlation(proteins_1host)
plot_correlation(protein_2hosts, base_size = 2, ybreak=1, ymax=10, host = 2)
```

`plot_dynamics_protein` *Dynamics of Diversity Motifs (Protein) Plot*

Description

This function compactly display the dynamics of diversity motifs (index and its variants: major, minor and unique) in the form of dot plot(s) as well as violin plots for all the provided individual protein(s).

Usage

```
plot_dynamics_protein(
  df,
  host = 1,
  protein_order = "",
  base_size = 8,
  alpha = 1/3,
  line_dot_size = 3,
  bw = "nrd0",
  adjust = 1
)
```

Arguments

`df`	DiMA JSON converted csv file data
`host`	number of host (1/2)
`protein_order`	order of proteins displayed in plot
`base_size`	base font size in plot
`alpha`	any number from 0 (transparent) to 1 (opaque)
`line_dot_size`	dot size in scatter plot
`bw`	smoothing bandwidth of violin plot (default: nrd0)
`adjust`	adjust the width of violin plot (default: 1)

Value

A plot

Examples

```
plot_dynamics_protein(proteins_1host)
```

`plot_dynamics_proteome` *Dynamics of Diversity Motifs (Proteome) Plot*

Description

This function compactly display the dynamics of diversity motifs (index and its variants: major, minor and unique) in the form of dot plot as well as violin plot for all the provided proteins at proteome level.

Usage

```
plot_dynamics_proteome(
  df,
  host = 1,
  line_dot_size = 2,
  base_size = 15,
  alpha = 1/3,
  bw = "nrd0",
  adjust = 1
)
```

Arguments

`df`	DiMA JSON converted csv file data
`host`	number of host (1/2)
`line_dot_size`	size of dot in plot
`base_size`	word size in plot
`alpha`	any number from 0 (transparent) to 1 (opaque)
`bw`	smoothing bandwidth of violin plot (default: nrd0)
`adjust`	adjust the width of violin plot (default: 1)

Value

A plot

Examples

```
plot_dynamics_proteome(proteins_1host)
```

plot_entropy *Entropy plot*

Description

This function plot entropy (black) and total variant (red) incidence of each k-mer position across the studied proteins and highlight region(s) with zero entropy in yellow. k-mer position with low support is marked with a red triangle underneath the x-axis line.

Usage

```
plot_entropy(
  df,
  host = 1,
  protein_order = "",
  kmer_size = 9,
  ymax = 10,
  line_dot_size = 2,
  base_size = 8,
  all = TRUE,
  highlight_zero_entropy = TRUE
)
```

Arguments

df	DiMA JSON converted csv file data
host	number of host (1/2)
protein_order	order of proteins displayed in plot
kmer_size	size of the k-mer window
ymax	maximum y-axis
line_dot_size	size of the line and dot in plot
base_size	word size in plot
all	plot both the entropy and total variants (pass FALSE in to plot only the entropy)
highlight_zero_entropy	highlight region with zero entropy (default: TRUE)

Value

A plot

Examples

```
plot_entropy(proteins_1host)
plot_entropy(protein_2hosts, host = 2)
```

plot_time *Time Distribution of Sequences Plot*

Description

This function plots the time distribution of provided sequences in the form of bar plot with 'Month' as x-axis and 'Number of Sequences' as y-axis. Aside from the plot, this function also returns a dataframe with 2 columns: 'Date' and 'Number of sequences'. The input dataframe of this function is obtainable from metadata_extraction(), with NCBI Protein / GISAID EpiCoV FASTA file as input.

Usage

```
plot_time(
  metadata,
  date_format = "%Y-%m-%d",
  base_size = 8,
  date_break = "2 month",
  scale = "count",
  only_plot = F
)
```

Arguments

metadata	a dataframe with 3 columns, 'ID', 'country', and 'date'
date_format	date format of the input dataframe
base_size	word size in plot
date_break	date break for the scale_x_date
scale	plot counts or log scale the data
only_plot	logical, return only plot or dataframe info as well, default FALSE

Value

A single plot or a list with 2 elements (a plot followed by a dataframe, default)

Examples

```
time_plot <- plot_time(metadata)$plot
time_df <- plot_time(metadata)$df
```

plot_worldmap *Geographical Distribution of Sequences Plot*

Description

This function plots a worldmap and color the affected geographical region(s) from light (lower) to dark (higher), depends on the cumulative number of sequences. Aside from the plot, this function also returns a dataframe with 2 columns: 'Country' and 'Number of Sequences'. The input dataframe of this function is obtainable from metadata_extraction(), with NCBI Protein / GISAID EpiCoV FASTA file as input.

Usage

```
plot_worldmap(meta, base_size = 8)
```

Arguments

meta	a dataframe with 3 columns, 'ID', 'country', and 'date'
base_size	word size in plot

Value

A list with 2 elements (a plot followed by a dataframe)

Examples

```
geographical_plot <- plot_worldmap(metadata)$plot
geographical_df <- plot_worldmap(metadata)$df
```

proteins_1host *DiMA (v4.1.1) JSON converted-CSV Output Sample 1*

Description

A dummy dataset with two proteins (A and B) from one host, human

Usage

```
proteins_1host
```

Format

A data frame with 806 rows and 17 variables:

proteinName name of the protein

position starting position of the aligned, overlapping k-mer window

count number of k-mer sequences at the given position

lowSupport k-mer position with sequences lesser than the minimum support threshold (TRUE) are considered of low support, in terms of sample size

entropy level of variability at the k-mer position, with zero representing completely conserved

indexSequence the predominant sequence (index motif) at the given k-mer position

index.incidence the fraction (in percentage) of the index sequences at the k-mer position

major.incidence the fraction (in percentage) of the major sequence (the predominant variant to the index) at the k-mer position

minor.incidence the fraction (in percentage) of minor sequences (of frequency lesser than the major variant, but not singletons) at the k-mer position

unique.incidence the fraction (in percentage) of unique sequences (singletons, observed only once) at the k-mer position

totalVariants.incidence the fraction (in percentage) of sequences at the k-mer position that are variants to the index (includes: major, minor and unique variants)

distinctVariant.incidence incidence of the distinct k-mer peptides at the k-mer position

multiIndex presence of more than one index sequence of equal incidence

host species name of the organism host to the virus

highestEntropy.position k-mer position that has the highest entropy value

highestEntropy highest entropy values observed in the studied protein

averageEntropy average entropy values across all the k-mer positions

---

`protein_2hosts` *DiMA (v4.1.1) JSON converted-CSV Output Sample 2*

Description

A dummy dataset with 1 protein (Core) from two hosts, human and bat

Usage

```
protein_2hosts
```

Format

A data frame with 200 rows and 17 variables:

proteinName name of the protein

position starting position of the aligned, overlapping k-mer window

count number of k-mer sequences at the given position

lowSupport k-mer position with sequences lesser than the minimum support threshold (TRUE) are considered of low support, in terms of sample size

entropy level of variability at the k-mer position, with zero representing completely conserved

indexSequence the predominant sequence (index motif) at the given k-mer position

index.incidence the fraction (in percentage) of the index sequences at the k-mer position

major.incidence the fraction (in percentage) of the major sequence (the predominant variant to the index) at the k-mer position

minor.incidence the fraction (in percentage) of minor sequences (of frequency lesser than the major variant, but not singletons) at the k-mer position

unique.incidence the fraction (in percentage) of unique sequences (singletons, observed only once) at the k-mer position

totalVariants.incidence the fraction (in percentage) of sequences at the k-mer position that are variants to the index (includes: major, minor and unique variants)

distinctVariant.incidence incidence of the distinct k-mer peptides at the k-mer position

multiIndex presence of more than one index sequence of equal incidence

host species name of the organism host to the virus

highestEntropy.position k-mer position that has the highest entropy value

highestEntropy highest entropy values observed in the studied protein

averageEntropy average entropy values across all the k-mer positions

Index