Package ‘ggmsa’

March 6, 2024

Title  Plot Multiple Sequence Alignment using 'ggplot2'

Version  1.8.0

Description  A visual exploration tool for multiple sequence alignment and associated data. Supports MSA of DNA, RNA, and protein sequences using 'ggplot2'. Multiple sequence alignment can easily be combined with other 'ggplot2' plots, such as phylogenetic tree Visualized by 'ggtree', boxplot, genome map and so on. More features: visualization of sequence logos, sequence bundles, RNA secondary structures and detection of sequence recombinations.

Depends  R (>= 4.1.0)

Imports  Biostrings, ggplot2, magrittr, tidyr, utils, stats, aplot, RColorBrewer, ggalt, ggforce, dplyr, R4RNA, grDevices, seqmagick, grid, methods, statebins, ggtree (>= 1.17.1)

Suggests  ggtreeExtra, ape, cowplot, knitr, BiocStyle, rmarkdown, readxl, ggnewscale, kableExtra, gggenes, testthat (>= 3.0.0)

License  Artistic-2.0

Encoding  UTF-8

URL  https://doi.org/10.1093/bib/bbac222(paper).

https://www.amazon.com/Integration-Manipulation-Visualization-Phylogenetic-Computational-ebook/dp/B0BSNLZR1Z/

(book)

BugReports  https://github.com/YuLab-SMU/ggmsa/issues

biocViews  Software, Visualization, Alignment, Annotation, MultipleSequenceAlignment

RoxygenNote  7.1.2

VignetteBuilder  knitr

Config/testthat/edition  3

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

git_branch  RELEASE_3_18
R topics documented:

adjust_ally ............................................. 3
assign_dms ............................................. 4
available_colors ...................................... 4
available_fonts ........................................ 5
available_msa .......................................... 5
extract_seq ............................................ 6
facet_msa .............................................. 6
geom_GC .................................................. 7
geom_helix ............................................... 8
geom_msa ................................................. 9
gem_msaBar ............................................. 11
gem_seed ................................................ 11
gem_seqlogo ........................................... 12
gghelix ................................................... 13
ggmaf ..................................................... 14
ggmsa ..................................................... 15
ggSeqBundle ........................................... 17
Gram-negative_AKL.fasta ............................... 18
Gram-positive_AKL.fasta ............................... 18
GVariation ............................................. 19
LeaderRepeat_All.fa ................................... 19
merge_seq ............................................... 20
plot ...................................................... 20
readSSfile ............................................. 21
read_maf ............................................... 22
reset_pos ............................................... 22
Rfam ....................................................... 23
sample.fasta ........................................... 23
seedSample.fa ......................................... 24
seqdiff .................................................. 24
seqlogo .................................................. 25
sequence-link-tree.fasta ............................... 26
show ...................................................... 26
simplify_hdata ......................................... 27
adjust_ally

Description

adjust the tree branch position after assigning ancestor node

Usage

adjust_ally(tree, node, sub = FALSE, seq_colname = "mol_seq")

Arguments

tree  ggtree object
node  internal node in tree
sub  logical value.
seq_colname  the colname of MSA on tree$data

Value

tree

Author(s)

Lang Zhou
assign_dms

<table>
<thead>
<tr>
<th>Description</th>
<th>assign dms value to alignments.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td><code>assign_dms(x, dms)</code></td>
</tr>
<tr>
<td>Arguments</td>
<td>x: data frame from tidy_msa()</td>
</tr>
<tr>
<td></td>
<td>dms: dms data frame</td>
</tr>
<tr>
<td>Value</td>
<td>tree</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Lang Zhou</td>
</tr>
</tbody>
</table>

available_colors

<table>
<thead>
<tr>
<th>Description</th>
<th>List Color Schemes currently available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td><code>available_colors()</code></td>
</tr>
<tr>
<td>Value</td>
<td>A character vector of available color schemes</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Lang Zhou</td>
</tr>
<tr>
<td>Examples</td>
<td><code>available_colors()</code></td>
</tr>
</tbody>
</table>
### available_fonts

**List Font Families currently available**

**Description**

This function lists font families currently available that can be used by 'ggmsa'.

**Usage**

```r
available_fonts()
```

**Value**

A character vector of available font family names

**Author(s)**

Lang Zhou

**Examples**

```r
available_fonts()
```

### available_msa

**List MSA objects currently available**

**Description**

This function lists MSA objects currently available that can be used by 'ggmsa'.

**Usage**

```r
available_msa()
```

**Value**

A character vector of available objects

**Author(s)**

Lang Zhou

**Examples**

```r
available_msa()
```
### extract_seq

**Description**

extract ancestor sequence from tree data

**Usage**

```r
extract_seq(tree_adjust, seq_colname = "mol_seq")
```

**Arguments**

- `tree_adjust`: ggtree object
- `seq_colname`: the colname of MSA on tree$data

**Value**

character

**Author(s)**

Lang Zhou

### facet_msa

**Description**

The MSA would be plot in a field that you set.

**Usage**

```r
facet_msa(field)
```

**Arguments**

- `field`: a numeric vector of the field size.

**Value**

ggplot layers

**Author(s)**

Lang Zhou
**Examples**

```r
library(ggplot2)
f <- system.file("extdata/sample.fasta", package="ggmsa")
# 2 fields
ggmsa(f, end = 120, font = NULL, color="Chemistry_AA") +
  facet_msa(field = 60)
# 3 fields
ggmsa(f, end = 120, font = NULL, color="Chemistry_AA") +
  facet_msa(field = 40)
```

**Description**

Multiple sequence alignment layer for ggplot2. It plot points of GC content.

**Usage**

```r
geom_GC(show.legend = FALSE)
```

**Arguments**

- `show.legend` logical. Should this layer be included in the legends?

**Value**

a ggplot layer

**Author(s)**

Lang Zhou

**Examples**

```r
#plot GC content
f <- system.file("extdata/LeaderRepeat_All.fa", package="ggmsa")
ggmsa(f, font = NULL, color="Chemistry_NT") + geom_GC()
```
Description

The layer of helix plot

Usage

```
geom_helix(helix_data, color_by = "length", overlap = FALSE, ...)
```

Arguments

- **helix_data**: a data frame. The file of nucleotide secondary structure and then read by readSSfile().
- **color_by**: generate colors for helices by various rules, including integer counts and value ranges one of "length" and "value"
- **overlap**: Logicals. If TRUE, two structures data called predict and known must be given (e.g., `helix_data = list(known = data1, predicted = data2)`), plots the predicted helices that are known on top, predicted helices that are not known on the bottom, and finally plots unpredicted helices on top in black.
- **...**: additional parameter

Value

`ggplot2` layers

Author(s)

Lang Zhou

Examples

```
RF03120 <- system.file("extdata/Rfam/RF03120_SS.txt", package="ggmsa")
RF03120_fas <- system.file("extdata/Rfam/RF03120.fasta", package="ggmsa")
SS <- readSSfile(RF03120, type = "Vienna")
ggmsa(RF03120_fas, font = NULL, border = NA,
      color = "Chemistry_NT", seq_name = FALSE) +
     geom_helix(SS)
```
Description

Multiple sequence alignment layer for ggplot2. It creates background tiles with/without sequence characters.

Usage

```r
geom_msa(
  data,
  font = "helvetical",
  mapping = NULL,
  color = "Chemistry_AA",
  custom_color = NULL,
  char_width = 0.9,
  none_bg = FALSE,
  by_conservation = FALSE,
  position_highlight = NULL,
  seq_name = NULL,
  border = NULL,
  consensus_views = FALSE,
  use_dot = FALSE,
  disagreement = TRUE,
  ignore_gaps = FALSE,
  ref = NULL,
  position = "identity",
  show.legend = FALSE,
  dms = FALSE,
  position_color = FALSE,
  ...
)
```

Arguments

data: sequence alignment with data frame, generated by tidy_msa().
font: font families, possible values are 'helvetical', 'mono', and 'DroidSansMono', 'TimesNewRoman'. Defaults is 'helvetical'.
mapping: aes mapping If font = NULL, only plot the background tile.
custom_color: A data frame with two column called "names" and "color". Customize the color scheme.
char_width a numeric vector. Specifying the character width in the range of 0 to 1. Defaults is 0.9.

none_bg a logical value indicating whether background should be displayed. Defaults is FALSE.

by_conservation a logical value. The most conserved regions have the brightest colors.

position_highlight A numeric vector of the position that need to be highlighted.

seq_name a logical value indicating whether sequence names should be displayed. Defaults is 'NULL' which indicates that the sequence name is displayed when 'font = null', but 'font = char' will not be displayed. If 'seq_name = TRUE' the sequence name will be displayed in any case. If 'seq_name = FALSE' the sequence name will not be displayed under any circumstances.

border a character string. The border color.

consensus_views a logical value that opening consensus views.

use_dot a logical value. Displays characters as dots instead of fading their color in the consensus view.

disagreement a logical value. Displays characters that disagreement to consensus(excludes ambiguous disagreements).

ignore_gaps a logical value. When selected TRUE, gaps in column are treated as if that row didn’t exist.

ref a character string. Specifying the reference sequence which should be one of input sequences when ‘consensus_views’ is TRUE.

position Position adjustment, either as a string, or the result of a call to a position adjustment function, default is ‘identity’ meaning ‘position_identity()’.

show.legend logical. Should this layer be included in the legends?

dms logical.

position_color logical.

... additional parameter

Value

A list

Author(s)

Guangchuang Yu, Lang Zhou seq_name’ work position_highlight’ work border’ work none_bg’ work

Examples

library(ggplot2)
aln <- system.file("extdata", "sample.fasta", package = "ggmsa")
tidy_aln <- tidy_msa(aln, start = 150, end = 170)
ggplot() + geom_msa(data = tidy_aln, font = NULL) + coord_fixed()
**geom_msaBar**

**Description**

Multiple sequence alignment layer for ggplot2. It plot sequence conservation bar.

**Usage**

```r
geom_msaBar()
```

**Value**

A list

**Author(s)**

Lang Zhou

**Examples**

```r
#plot multiple sequence alignment and conservation bar.
f <- system.file("extdata/sample.fasta", package="ggmsa")
ggmsa(f, 221, 280, font = NULL, seq_name = TRUE) + geom_msaBar()
```

---

**geom_seed**

**Description**

Highlighting the seed in miRNA sequences

**Usage**

```r
geom_seed(seed, star = FALSE)
```

**Arguments**

- `seed` : a character string. Specifying the miRNA seed sequence like 'GAGGUAG'.
- `star` : a logical value indicating whether asterisks should be displayed.

**Value**

a ggplot layer
Author(s)

Lang Zhou

Examples

```r
miRNA_sequences <- system.file("extdata/seedSample.fa", package="ggmsa")
ggmsa(miRNA_sequences, font = 'DroidSansMono',
      color = "Chemistry_NT", none_bg = TRUE) +
geom_seed(seed = "GAGGUAG", star = FALSE)
ggmsa(miRNA_sequences, font = 'DroidSansMono',
      color = "Chemistry_NT") +
geom_seed(seed = "GAGGUAG", star = TRUE)
```

Description

Multiple sequence alignment layer for ggplot2. It plot sequence motifs.

Usage

```r
geom_seqlogo(
  font = "DroidSansMono",
  color = "Chemistry_AA",
  adaptive = TRUE,
  top = TRUE,
  custom_color = NULL,
  show.legend = FALSE,
  ...
)
```

Arguments

- `font`: font families, possible values are 'helvetica', 'mono', and 'DroidSansMono', 'TimesNewRoman'. Defaults is 'DroidSansMono'.
- `adaptive`: A logical value indicating whether the overall height of seqlogo corresponds to the number of sequences. If is FALSE, seqlogo overall height = 4, fixedly.
- `top`: A logical value. If TRUE, seqlogo is aligned to the top of MSA.
- `custom_color`: A data frame with two columns called "names" and "color". Customize the color scheme.
- `show.legend`: logical. Should this layer be included in the legends?
- `...`: additional parameter
Plots nucleotide secondary structure as helices in arc diagram

Usage

\[ \text{gghelix}(\text{helix\_data}, \text{color\_by} = "length", \text{overlap} = \text{FALSE}) \]

Arguments

\begin{itemize}
  \item \text{helix\_data} \hspace{1cm} \text{a data frame. The file of nucleotide secondary structure and then read by \text{readSSfile}().}
  \item \text{color\_by} \hspace{1cm} \text{generate colors for helices by various rules, including integer counts and value ranges one of "length" and "value"}
  \item \text{overlap} \hspace{1cm} \text{Logicals. If \text{TRUE}, two structures data called predict and known must be given(e.g:\text{helix\_data} = \text{list(known = data1, predicted = data2)}, plots the predicted helices that are known on top, predicted helices that are not known on the bottom, and finally plots unpredicted helices on top in black.}
\end{itemize}

Value

\text{ggplot object}

Author(s)

Lang Zhou

Examples

\begin{verbatim}
RF03120 <- system.file("extdata/Rfam/RF03120_SS.txt", package="ggmsa")
helix_data <- readSSfile(RF03120, type = "Vienna")
gghelix(helix_data)
\end{verbatim}
ggmaf

Description

plot MAF

Usage

ggmaf(
    data,
    ref,
    block_start = NULL,
    block_end = NULL,
    facet_field = NULL,
    heights = c(0.4, 0.6),
    facet_heights = NULL
)

Arguments

data a tidy MAF data frame. You can get it by tidy_maf_df()
ref character, the name of reference genome. e.g: "hg38.chr1_KI270707v1_random"
block_start a numeric vector(>0). The start block to plot.
block_end a numeric vector(< max block). The end block to plot.
facet_field a numeric vector. The field in a facet panel.
heights two numeric vector. The plot proportion between "Genomic location" panel(upon) and "Alignment" panel(down). Default: c(0.4, 0.6)
facet_heights Numeric vectors. The facet proportion.

Value

ggplot object

Author(s)

Lang Zhou
Description

Plot multiple sequence alignment using ggplot2 with multiple color schemes supported.

Usage

```r
ggmsa(
  msa,
  start = NULL,
  end = NULL,
  font = "helvetical",
  color = "Chemistry_AA",
  custom_color = NULL,
  char_width = 0.9,
  none_bg = FALSE,
  by_conservation = FALSE,
  position_highlight = NULL,
  seq_name = NULL,
  border = NULL,
  consensus_views = FALSE,
  use_dot = FALSE,
  disagreement = TRUE,
  ignore_gaps = FALSE,
  ref = NULL,
  show.legend = FALSE
)
```

Arguments

- `msa`: Multiple aligned sequence files or objects representing either nucleotide sequences or AA sequences.
- `start`: a numeric vector. Start position to plot.
- `end`: a numeric vector. End position to plot.
- `font`: font families, possible values are 'helvetical', 'mono', and 'DroidSansMono', 'TimesNewRoman'. Defaults is 'helvetical'. If font = NULL, only plot the background tile.
- `custom_color`: A data frame with two column called "names" and "color". Customize the color scheme.
- `char_width`: a numeric vector. Specifying the character width in the range of 0 to 1. Defaults is 0.9.
none_bg  a logical value indicating whether background should be displayed. Defaults is FALSE.

by_conservation  a logical value. The most conserved regions have the brightest colors.

position_highlight  A numeric vector of the position that need to be highlighted.

seq_name  a logical value indicating whether sequence names should be displayed. Defaults is 'NULL' which indicates that the sequence name is displayed when 'font = null', but 'font = char' will not be displayed. If 'seq_name = TRUE' the sequence name will be displayed in any case. If 'seq_name = FALSE' the sequence name will not be displayed under any circumstances.

border  a character string. The border color.

consensus_views  a logical value that opening consensus views.

use_dot  a logical value. Displays characters as dots instead of fading their color in the consensus view.

disagreement  a logical value. Displays characters that disagreement to consensus(excludes ambiguous disagreements).

ignore_gaps  a logical value. When selected TRUE, gaps in column are treated as if that row didn’t exist.

ref  a character string. Specifying the reference sequence which should be one of input sequences when 'consensus_views' is TRUE.

show.legend  logical. Should this layer be included in the legends?

Value

ggplot object

Author(s)

Guangchuang Yu

Examples

#plot multiple sequences by loading fasta format
fasta <- system.file("extdata", "sample.fasta", package = "ggmsa")
ggmsa(fasta, 164, 213, color="Chemistry_AA")

## Not run:
#XMultipleAlignment objects can be used as input in the 'ggmsa'
AAMultipleAlignment <- readAAMultipleAlignment(fasta)
ggmsa(AAMultipleAlignment, 164, 213, color="Chemistry-AA")

#XStringSet objects can be used as input in the 'ggmsa'
AAStringSet <- readAAStringSet(fasta)
ggmsa(AAStringSet, 164, 213, color="Chemistry-AA")

#Xbin objects from 'seqmagick' can be used as input in the 'ggmsa'
ggSeqBundle

AAbin <- fa_read(fasta)
ggmsa(AAbin, 164, 213, color="Chemistry_AA")

## End(Not run)

---

**Description**

plot Sequence Bundles for MSA based 'ggolot2'

**Usage**

ggSeqBundle(
  msa,  
  line_width = 0.3,  
  line_thickness = 0.3,  
  line_high = 0,  
  spline_shape = 0.3,  
  size = 0.5,  
  alpha = 0.2,  
  bundle_color = c("#2ba0f5","#424242"),  
  "Y", "N", "Q", "D", "E", "K", "R", "H")
)

**Arguments**

- **msa**: Multiple sequence alignment file(FASTA) or object for representing either nucleotide sequences or peptide sequences. Also receives multiple MSA files. eg: msa = c("Gram-negative_AKL.fasta", "Gram-positive_AKL.fasta").
- **line_width**: The width of bundles at each site, default is 0.3.
- **line_thickness**: The thickness of bundles at each site, default is 0.3.
- **line_high**: The high of bundles at each site, default is 0.
- **spline_shape**: A numeric vector of values between -1 and 1, which control the shape of the spline relative to the control points. From geom_xspline() in ggalt package.
- **size**: A numeric vector of values between 0 and 1, which control the size of each lines.
- **alpha**: A numeric vector of values between 0 and 1, which control the alpha of each lines.
- **bundle_color**: The colors of each sequence bundles. eg: bundle_color = c("#2ba0f5","#424242").
**Value**

ggplot object

**Author(s)**

Lang Zhou

**Examples**

```r
aln <- system.file("extdata", "Gram-negative_AKL.fasta", package = "ggmsa")
ggSeqBundle(aln)
```

---

**Description**

Amino acids in the adenylate kinase lid (AKL) domain from Gram-negative bacteria.

**Format**

A MSA fasta with 100 sequences and 36 positions.

**Source**

http://biovis.net/year/2013/info/redesign-contest

---

**Description**

Amino acids in the adenylate kinase lid (AKL) domain from Gram-positive bacteria.

**Format**

A MSA fasta with 100 sequences and 36 positions.

**Source**

http://biovis.net/year/2013/info/redesign-contest
**Description**

A folder containing 4 MAS files as a sample data set to identify the sequence recombination event.

**Format**

a folder

**Details**

- A.Mont.fas MSA with sequences of 'Mont' and 'CF_YL21'
- B.Oz.fas MSA with sequences of 'Oz' and 'CF_YL21'
- C.Wilga5.fas MSA with sequences of 'Wilga5' and 'CF_YL21'
- sample_alignment.fa MSA with sequences of 'Mont', 'CF_YL21', 'Oz', and 'Wilga5'

**Source**


---

**LeaderRepeat_All.fa**  

**A sample DNA alignment sequences**

**Description**

DNA alignment sequences with 24 sequences and 56 positions.

**Format**

A MSA fasta
merge_seq

merge_seq

Description
merge two MSA

Usage
merge_seq(previous_seq, gap, subsequent_seq, adjust_name = TRUE)

Arguments
- previous_seq: previous MSA
- gap: gap length
- subsequent_seq: subsequent MSA
- adjust_name: logical value. merge seq name or not

Value
tidy MSA data frame

Author(s)
Lang Zhou

plot
plot method for SeqDiff object

Description
plot method for SeqDiff object

Usage
## S4 method for signature 'SeqDiff,ANY'
plot(
x,
width = 50,
title = "auto",
lab = "Nucleotide Position",
by = "bar",
fill = "firebrick",
colors = c(A = "#ff6d6d", C = "#769dcc", G = "#f2be3c", T = "#74ce98"),
xlim = NULL
)
Arguments

- `x`: SeqDiff object
- `width`: bin width
- `title`: plot title
- `xlab`: xlab
- `by`: one of 'bar' and 'area'
- `fill`: fill color of upper part of the plot
- `colors`: color of lower part of the plot
- `xlim`: limits of x-axis

Value

plot

Author(s)

guangchuang yu

Examples

```
fas <- list.files(system.file("extdata", "GVariation", package="ggmsa"),
                 pattern="fas", full.names=TRUE)
x1 <- seqdiff(fas[1], reference=1)
plot(x1)
```

Description

Read secondary structure file

Usage

```
readSSfile(file, type = NULL)
```

Arguments

- `file`: A text file in connect format
- `type`: file type. one of "Helix", "Connect", "Vienna" and "Bpseq"

Value

data frame
Author(s)

Lang Zhou

Examples

RF03120 <- system.file("extdata/Rfam/RF03120_SS.txt", package="ggmsa")
helix_data <- readSSfile(RF03120, type = "Vienna")

Description

read 'multiple alignment format'(MAF) file

Usage

read_maf(multiple_alignment_format)

Arguments

multiple_alignment_format

a multiple alignment format(MAF) file

Value

data frame

Author(s)

Lang Zhou

reset_pos

reset MSA position

Usage

reset_pos(seq_df)

Arguments

seq_df MSA data
Value
data frame

Author(s)
Lang Zhou

Description
A folder containing seed alignment sequences and corresponding consensus RNA secondary structure.

Format
a folder

Details
- RF00458.fasta seed alignment sequences of Cripavirus internal ribosome entry site (IRES)
- RF03120.fasta seed alignment sequences of Sarbecovirus 5’UTR
- RF03120_SS.txt consensus RNA secondary structure of Sarbecovirus 5’UTR

Source
https://rfam.xfam.org/

Description
A dataset containing the alignment sequences of the phenylalanine hydroxylase protein (PH4H) within nine species

Format
A MSA fasta with 9 sequences and 456 positions.
seedSample.fa  microRNA data used in ggmsa

Description
Fasta format sequences of mature miRNA sequences from miRBase

Format
A MSA fasta with 6 sequences and 22 positions.

Source
https://www.mirbase.org/ftp.shtml

seqdiff

Description
calculate difference of two aligned sequences

Usage
seqdiff(fasta, reference = 1)

Arguments
fasta  fasta file
reference  which sequence serve as reference, 1 or 2

Value
SeqDiff object

Author(s)
guangchuang yu

Examples
fas <- list.files(system.file("extdata", "GVariation", package="ggmsa"),
                 pattern="fas", full.names=TRUE)
seqdiff(fas[1], reference=1)
Description

plot sequence logo for MSA based 'ggolot2'

Usage

seqlogo(
  msa,
  start = NULL,
  end = NULL,
  font = "DroidSansMono",
  color = "Chemistry_AA",
  adaptive = FALSE,
  top = FALSE,
  custom_color = NULL
)

Arguments

msa  Multiple sequence alignment file or object for representing either nucleotide sequences or peptide sequences.
start  Start position to plot.
end  End position to plot.
font  font families, possible values are 'helvetical', 'mono', and 'DroidSansMono', 'TimesNewRoman'. Defaults is 'DroidSansMono'. If font=NULL, only the background tiles is drawn.
adaptive  A logical value indicating whether the overall height of seqlogo corresponds to the number of sequences. If FALSE, seqlogo overall height = 4, fixedly.
top  A logical value. If TRUE, seqlogo is aligned to the top of MSA.
custom_color  A data frame with two cloumn called "names" and "color". Customize the color scheme.

Value

ggplot object

Author(s)

Lang Zhou
Examples

```r
# plot sequence motif independently
nt_sequence <- system.file("extdata", "LeaderRepeat_All.fa", 
  package = "ggmsa")
seqlogo(nt_sequence, color = "Chemistry_NT")
```

sequence-link-tree.fasta

```
sequence-link-tree
```

Description

Alignment sequences used to demonstrate circular MSA layout

Format

A MSA fasta with 28 sequences and 480 positions.

show

```
show
```

show method

Description

show method

Usage

```r
show(object)
```

Arguments

```r
object SeqDiff object
```

Value

message

Examples

```r
fas <- list.files(system.file("extdata", "GVariation", package="ggmsa"), 
  pattern="fas", full.names=TRUE)
x1 <- seqdiff(fas[1], reference=1)
x1
```
### simplify_hdata

**Description**
reset hdata data position

**Usage**
simplify_hdata(hdata, sim_msa)

**Arguments**
- **hdata**: data from tidy_hdata()
- **sim_msa**: MSA data frame

**Value**
data frame

**Author(s)**
Lang Zhou

### simplot

**Description**
Sequence similarity plot

**Usage**
simplot(
  file,
  query,
  window = 200,
  step = 20,
  group = FALSE,
  id,
  sep,
  sd = FALSE,
  smooth = FALSE,
  smooth_params = list(method = "loess", se = FALSE)
)
theme_msa

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>file</td>
<td>alignment fast file</td>
</tr>
<tr>
<td>query</td>
<td>query sequence</td>
</tr>
<tr>
<td>window</td>
<td>sliding window size (bp)</td>
</tr>
<tr>
<td>step</td>
<td>step size to slide the window (bp)</td>
</tr>
<tr>
<td>group</td>
<td>whether grouping sequence (eg. For &quot;A-seq1,A-seq2,B-seq1 and B-seq2&quot;, using sep = &quot;,&quot; and id = 1 to divide sequences into groups A and B)</td>
</tr>
<tr>
<td>id</td>
<td>position to extract id for grouping; only works if group = TRUE</td>
</tr>
<tr>
<td>sep</td>
<td>separator to split sequence name; only works if group = TRUE</td>
</tr>
<tr>
<td>sd</td>
<td>whether display standard deviation of similarity among each group; only works if group=TRUE</td>
</tr>
<tr>
<td>smooth</td>
<td>FALSE(default) or TRUE; whether display smoothed spline.</td>
</tr>
<tr>
<td>smooth_params</td>
<td>a list that add params for geom_smooth, (default: smooth_params = list(method = &quot;loess&quot;, se = FALSE))</td>
</tr>
</tbody>
</table>

Value

ggplot object

Author(s)

guangchuang yu

Examples

fas <- system.file("extdata/GVariation/sample_alignment.fa", package="ggmsa")
simplot(fas, 'CF_YL21')

theme_msa

Description

Theme for ggmsa.

Usage

theme_msa()

Author(s)

Lang Zhou
**tidy_hdata**

**Description**

tidy protein-protein interactive position data

**Usage**

tidy_hdata(gap, inter, previous_seq, subsequent_seq)

**Arguments**

- **gap**
  - gap length
- **inter**
  - protein-protein interactive position data
- **previous_seq**
  - previous MSA
- **subsequent_seq**
  - subsequent MSA

**Value**

helix data

**Author(s)**

Lang Zhou

---

**tidy_maf_df**

**Description**

tidy MAF data frame

**Usage**

tidy_maf_df(maf_df, ref)

**Arguments**

- **maf_df**
  - a MAF data frame. You can get it by read_maf()
- **ref**
  - character, the name of reference genome. eg: "hg38.chr1_KI270707v1_random"

**Value**

data frame
**Author(s)**

Lang Zhou

**Description**

Convert msa file/object to tidy data frame.

**Usage**

```
tidy_msa(msa, start = NULL, end = NULL)
```

**Arguments**

- `msa` multiple sequence alignment file or sequence object in DNAStringSet, RNAStringSet, AAStringSet, BStringSet, DNAMultipleAlignment, RNAMultipleAlignment, AAMultipleAlignment, DNAbin or AAbin
- `start` start position to extract subset of alignment
- `end` end position to extract subset of alignment

**Value**

tibble data frame

**Author(s)**

Guangchuang Yu

**Examples**

```r
fasta <- system.file("extdata", "sample.fasta", package = "ggmsa")
aln <- tidy_msa(msa = fasta, start = 10, end = 100)
```

**Description**

Alignment sequences of used to show graphical combination

**Format**

A MSA fasta with 5 sequences and 404 positions.
TP53_genes.xlsx

Description
The local genome map shows the 30000 sites around the TP53 gene.

Format
xlsx

treeMSA_plot

Description
plot Tree-MSA plot

Usage
treeMSA_plot(
p_tree,
tidymsa_df,
ancestral_node = "none",
sub = FALSE,
panel = "MSA",
font = NULL,
color = "Chemistry_AA",
seq_colname = NULL,
...
)

Arguments
p_tree tree view
tidymsa_df tidy MSA data
ancestral_node vector, internal node in tree. Assigning a internal node to display "ancestral sequences". If ancestral_node = "none" hides all ancestral sequences, if ancestral_node = "all" shows all ancestral sequences.
sub logical value. Displaying a subset of ancestral sequences or not.
panel panel name for plot of MSA data
font font families, possible values are 'helvetical', 'mono', and 'DroidSansMono', 'TimesNewRoman'. Defaults is 'helvetical'. If font = NULL, only plot the background tile.
seq_colname  the colname of MSA on tree$data
...  additional parameters for 'geom_msa'

Details
'treeMSA_plot()' automatically re-arranges the MSA data according to the tree structure,

Value
ggplot object

Author(s)
Lang Zhou
Index

* datasets
  Gram-negative_AKL.fasta, 18
  Gram-positive_AKL.fasta, 18
  GVariation, 19
  LeaderRepeat_All.fa, 19
  Rfam, 23
  sample.fasta, 23
  seedSample.fa, 24
  sequence-link-tree.fasta, 26
  tp53.fa, 30
  TP53_genes.xlsx, 31

adjust_ally, 3
assign_dms, 4
available_colors, 4
available_fonts, 5
available_msa, 5

extract_seq, 6

facet_msa, 6

geom_GC, 7
geom_helix, 8
geom_msa, 9
geom_msaBar, 11
geom_seed, 11
geom_seqlogo, 12
gghelix, 13
ggmaf, 14
ggmsa, 15
ggSeqBundle, 17
Gram-negative_AKL.fasta, 18
Gram-positive_AKL.fasta, 18
GVariation, 19

LeaderRepeat_All.fa, 19

merge_seq, 20

plot, 20

plot, SeqDiff, ANY-method (plot), 20

read_maf, 22
readSSfile, 21
reset_pos, 22
Rfam, 23

sample.fasta, 23
seedSample.fa, 24
seqdiff, 24
SeqDiff-class (show), 26
seqlogo, 25
sequence-link-tree.fasta, 26
show, 26
show, SeqDiff-method (show), 26
simplify_hdata, 27
simplot, 27

theme_msa, 28
tidy_hdata, 29
tidy_maf_df, 29
tidy_msa, 30
tp53.fa, 30
TP53_genes.xlsx, 31
treeMSA_plot, 31