Package 'ggmsa'

January 6, 2021

Title Plot Multiple Sequence Alignment using 'ggplot2'

Version 0.0.5

Description Supports visualizing multiple sequence alignment of DNA and protein sequences using 'ggplot2'. It supports a number of colour schemes, including Chemistry, Clustal, Shapely, Taylor and Zappo. Multiple sequence alignment can easily be combined with other 'ggplot2' plots, such as aligning a phylogenetic tree produced by 'ggtree' with multiple sequence alignment.

Depends R (>= 3.5.0)

Imports Biostrings, ggplot2, magrittr, tidyr, utils, stats, stringr, aplot, RColorBrewer

Suggests ape, cowplot, ggtree (>= 1.17.1), knitr, methods, prettydoc, rmarkdown, seqmagick,

License Artistic-2.0

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

VignetteBuilder knitr

NeedsCompilation no

Author Guangchuang Yu [aut, cre] (<https://orcid.org/0000-0002-6485-8781>), Lang Zhou [aut], Huina Huang [ctb]

Maintainer Guangchuang Yu <guangchuangyu@gmail.com>

Repository CRAN

Date/Publication 2021-01-06 15:50:11 UTC

R topics documented:

`available_colors` *List Color Schemes currently available*

Description

This function lists color schemes currently available that can be used by 'ggmsa'

Usage

```
available_colors()
```

Value

A character vector of available color schemes

Author(s)

Lang Zhou

Examples

```
available_colors()
```

available_fonts *List Font Families currently available*

Description

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

Usage

```
available_fonts()
```

Value

A character vector of available font family names

Author(s)

Lang Zhou

Examples

```
available_fonts()
```

available_msa *List MSA objects currently available*

Description

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

Usage

```
available_msa()
```

Value

A character vector of available objects

Author(s)

Lang Zhou

Examples

```
available_msa()
```

`color_Clustal` *A color scheme of Culstal. The algorithm to assign colors for Multiple Sequence.*

Description

A color scheme of Culstal. The algorithm to assign colors for Multiple Sequence.

Usage

```
color_Clustal(y)
```

Arguments

`y`	sequence alignment with data frame, generated by tidy_msa().

`color_scheme` *Assigning colors to sequence alignment.*

Description

Assigning colors to sequence alignment.

Usage

```
color_scheme(y, color)
```

Arguments

`y`	sequence alignment with data frame, generated by tidy_msa().
`color`	a Color scheme. One of 'Clustal', 'Chemistry_AA', 'Shapely_AA', 'Zappo_AA', 'Taylor_AA', 'LETTER', 'CN6', 'Chemistry_NT', 'Shapely_NT', 'Zappo_NT', 'Taylor_NT'. Defaults is 'Chemistry_AA'.

facet_msa *segment MSA*

Description

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

Usage

```
facet_msa(field)
```

Arguments

field	a numeric vector of the field size.

Author(s)

Lang Zhou

Examples

```
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)
```

geom_GC *geom_GC*

Description

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

Usage

```
geom_GC()
```

Author(s)

Lang Zhou

Examples

```
library(ggplot2)
#plot GC content
f <- system.file("extdata/LeaderRepeat_All.fa", package="ggmsa")
ggmsa(f, font = NULL, color="Chemistry_NT") + geom_GC()
```

geom_msa *geom_msa*

Description

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

Usage

```
geom_msa(
  data,
  font = "helvetical",
  mapping = NULL,
  color = "Chemistry_AA",
  char_width = 0.9,
  none_bg = FALSE,
  by_conservation = FALSE,
  posHighligthed = NULL,
  seq_name = NULL,
  border = NULL,
  consensus_views = FALSE,
  use_dot = FALSE,
  disagreement = TRUE,
  ignore_gaps = FALSE,
  ref = NULL,
  ...
)
```

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.
color	A Color scheme. One of 'Clustal', 'Chemistry_AA', 'Shapely_AA', 'Zappo_AA', 'Taylor_AA', 'LETTER','CN6',, 'Chemistry_NT', 'Shapely_NT', 'Zappo_NT', 'Taylor_NT'. Defaults is 'Chemistry_AA'.
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 disaplayed. Defaults is FALSE.
by_conservation	a logical value. The most conserved regions have the brightest colors.
posHighligthed	A numeric vector of the position that need to be highlighted.

seq_name	a logical value indicating whether seqence 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 opeaning 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 disagreememt 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.
...	additional parameter

Value

A list

Author(s)

Guangchuang Yu

`geom_msaBar` *geom_msaBar*

Description

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

Usage

```
geom_msaBar()
```

Value

A list

Author(s)

Lang Zhou

Examples

```
#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 *geom_seed*

Description

Highlighting the seed in miRNA sequences

Usage

```
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 disaplayed.

Value

a ggplot layer

Author(s)

Lang Zhou

Examples

```
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)
```

geom_seqlogo *geom_seqlogo*

Description

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

Usage

```
geom_seqlogo(
  font = "DroidSansMono",
  color = "Chemistry_NT",
  adaptive = TRUE,
  top = TRUE,
  ...
)
```

Arguments

font	font families, possible values are 'helvetical', 'mono', and 'DroidSansMono', 'TimesNewRoman'. Defaults is 'DroidSansMono'.
color	A Color scheme. One of 'Clustal', 'Chemistry_AA', 'Shapely_AA', 'Zappo_AA', 'Taylor_AA', 'LETTER', 'CN6', 'Chemistry_NT', 'Shapely_NT', 'Zappo_NT', 'Taylor_NT'. Defaults is 'Chemistry_AA'.
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.
...	additional parameter

Value

A list

Author(s)

Lang Zhou

Examples

```
#plot multiple sequence alignment and sequence motifs
f <- system.file("extdata/LeaderRepeat_All.fa", package="ggmsa")
ggmsa(f,font = NULL,color = "Chemistry_NT") + geom_seqlogo()
```

get_consensus *calling the consensus sequence.*

Description

calling the consensus sequence.

Usage

```
get_consensus(tidy, ignore_gaps = FALSE, ref = NULL)
```

Arguments

tidy	sequence alignment with data frame, generated by tidy_msa().
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.

ggmsa *ggmsa*

Description

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

Usage

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

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.
color	a Color scheme. One of 'Clustal', 'Chemistry_AA', 'Shapely_AA', 'Zappo_AA', 'Taylor_AA', 'LETTER', 'CN6', 'Chemistry_NT', 'Shapely_NT', 'Zappo_NT', 'Taylor_NT'. Defaults is 'Chemistry_AA'.
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 disaplayed. Defaults is FALSE.
by_conservation	a logical value. The most conserved regions have the brightest colors.
posHighligthed	A numeric vector of the position that need to be highlighted.
seq_name	a logical value indicating whether seqence 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 opeaning 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 disagreememt 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.

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")

#XMultipleAlignment objects can be used as input in the 'ggmsa'
#AAMultipleAlignment <- Biostrings::readAAMultipleAlignment(fasta)
#ggmsa(AAMultipleAlignment, 164, 213, color="Chemistry_AA")

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

#Xbin objects from 'seqmagick' can be used as input in the 'ggmsa'
#AAbin <- seqmagick::fa_read(fasta)
#ggmsa(AAbin, 164, 213, color="Chemistry_AA")
```

---

msa2tidy *This function converts the msa_data to the tidy data.*

Description

This function converts the msa_data to the tidy data.

Usage

```
msa2tidy(msaData)
```

Arguments

msaData	sequence alignment data generated by msa_data().

---

seqlogo *seqlogo*

Description

plot sequence logo for MSA based 'ggolot2'

Usage

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

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.
color	A Color scheme. One of 'Clustal', 'Chemistry_AA', 'Shapely_AA', 'Zappo_AA', 'Taylor_AA', 'LETTER', 'CN6','Chemistry_NT', 'Shapely_NT', 'Zappo_NT', 'Taylor_NT'. Defaults is 'Chemistry_AA'.
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.

Value

ggplot object

Author(s)

Lang Zhou

Examples

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

tidy_color *cleaning the needless sequeces' color according to the consensus sequcence (only used in the consensus views).*

Description

cleaning the needless sequeces' color according to the consensus sequcence (only used in the consensus views).

Usage

```
tidy_color(y, consensus, disagreement, ref)
```

Arguments

y	a data frame, sequence alignment with specified color.
consensus	the consensus sequence which can be called by get_consensus().
disagreement	a logical value. Displays characters that disagreememt to consensus(excludes ambiguous disagreements).
ref	a character string. Specifying the reference sequence which should be one of input sequences when 'consensus_views' is TRUE.

tidy_msa *tidy_msa*

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 alignemnt

Value

tibble data frame

Author(s)

Guangchuang Yu

Index