Package 'SVAlignR'

July 20, 2022

Version 0.2.2

Date 2022-03-01

Title Recovering Structure of Long Molecules from Structural Variation Data

Author Kevin R. Coombes

Maintainer Kevin R. Coombes <krc@silicovore.com>

Description Implements a method to combine multiple levels of multiple sequence alignment to uncover the structure of complex DNA rearrangements.

Depends R (>= 3.5.0)

Imports methods, graphics, grDevices, oompaBase, msa, Biostrings, NameNeedle, dendextend, ape, stringr, igraph, Polychrome, colorspace

Suggests viridisLite, knitr, rmarkdown

VignetteBuilder knitr

License Apache License (== 2.0)

URL http://oompa.r-forge.r-project.org/

NeedsCompilation no

R topics documented:

AlignedCluster-class 2
Cipher-class 3
SequenceCluster-class 4
StringGraph-class 5
SVAkignR-data 7

Index **8**

AlignedCluster-class *Class* "AlignedCluster"

Description

The AlignedCluster class is used to a;ign a set of clustered seqeunces. The alignClusters function returns a new object of the AlignedCluster class. The alignAllClusters function takes a SequenceCluster object and returns a list of AlignedCluster objects.

Usage

```
alignCluster(sequences, mysub = NULL, gapO = 10, gapE = 0.2)
alignAllClusters(sc, mysub = NULL, gapO = 10, gapE = 0.2)
makeSubsMatrix(match = 5, mismatch = -2)
## S4 method for signature 'AlignedCluster'
image(x, col = "black", cex = 1, main = "", ...)
```

Arguments

sequences	A character vector that contains all sequences to be aligned.
mysub	A square (usually symmetric) substituion matrix.
gapO	A numeric value defining the penalty for opening a gap.
gapE	A numeric value defining the penalty for extending a gap.
sc	An object of the SequenceCluster class.
match	A numeric value defining the reward for matching symbols from two sequences.
mismatch	A numeric value defining the penalty for mismatching symbols from two sequences.
x	An object of the AlignedCluster class.
col	A character setting the color of annotations in the image.
main	Character; the plot title.
cex	Numeric; size of teh text inside the image of the alignment matrix.
...	Extra arguments for generic or plotting routines.

Value

The alignCluster function returns a new object of the AlignedCluster class. The alignAlllClusters function returns a list of AlignedCluster objects. The makeSubMatrix function returns a symmetric substitution matrix.

Objects from the Class

Objects should be defined using the alignCluster or alignAllCluster functions. You typically pass in a character vector of sequences that have already been found to forma cluster.

Slots

alignment: .

weights A numeric vactor; the numbof times each unique raw sequence occurs.

consensus: .

Author(s)

Kevin R. Coombes <krc@silicovore.com>

Examples

```
data(longreads)
seqs <- longreads$connection[1:15]
pad <- c(rep("0", 9), rep("", 6))
names(seqs) <- paste("LR", pad, 1:length(seqs), sep = "")
seqs <- seqs[!duplicated(seqs)]
mysub <- makeSubsMatrix(match = 2, mismatch = -6)
ab <- alignCluster(seqs, mysub)
image(ab)
```

Cipher-class *Class* "Cipher"

Description

The Cipher class is used to change between different alphabets (and so behaves as a simple substituion cipher). The Cipher function returns a new object of the Cipher class.

Usage

```
Cipher(sampleText, split = "-", extras = c("-" = ":", "?" = "?"))
encode(cipher, text)
decode(cipher, text)
```

Arguments

sampleText	A character vector that contains all symbols you want to be able to transliterate. Duplicate symbols are automatically removed.
split	A single character used to split words into symbols. Defaults to a hyphen for our applications.
extras	Additional characters to be added for reverse tranlsiteration, since they may appear as the results of alignments in consensus sequences.
cipher	An object of the Cipher class.
text	A character vector of words to be transliterated.

Value

The Cipher function returns a new object of the Cipher class. The encode and decode functions return character vectors that are the same size as their input text parameters.

Objects from the Class

Objects should be defined using the Cipher constructor. You typically pass in a character vector of "words" that contain all the symbols that are contained in the text to translated (i.e., encoded and decoded) between languages. A standard tqarget alphabet is created along with forward and reverse transliteration rules.

Slots

`forward`: A named character vector.

`reverse`: A named character vector.

Note

Attempting to manipulate a `Cipher` object using text containing NAs, missing values, or previously unknown symbols will result in an error.

Author(s)

Kevin R. Coombes <krc@silicovore.com>

Examples

```
motif <- "0-50-74-0-50-74-25-26-35"
alfa <- Cipher(motif)
alfa
en <-encode(alfa, motif)
en
de <- decode(alfa, en)
de
```

SequenceCluster-class *Class* "SequenceCluster"

Description

The `SequenceCluster` class is used to cluster seqeunces of "words" from an arbitrarily long alphabet. The `SequenceCluster` function returns a new object of the `SequenceCluster` class.

Usage

```
SequenceCluster(rawseq, method = c("needelman", "levenshtein"), NC = 5)
## S4 method for signature 'SequenceCluster,missing'
plot(x, type = "rooted", main = "Colored Clusters", ...)
updateClusters(sc, NC)
heat(x, ...)
```

Arguments

`rawseq`	A character vector that contains all words or "sequences" to be clustered.
`method`	The algorithm to use to compute distances between sequences. The choices are "levenshstein", which uses the Levenshtein edit distance, or "needelman", which uses the Needelman-Wunsch global alignment algorithm.
`x`	An object of the `SequenceCluster` class.
`sc`	An object of the `SequenceCluster` class.
`NC`	An integer; the number of clusters to cut from the dendrogram.
`type`	A character strnig; the type of lpot to make. Val;id types are "rooted", "clippe", or "unrooted".
`main`	Character; the plot title.
`...`	extra arguments for generic or plotting routines

Value

The SequenceCluster function returns a new object of the SequenceCluster class.

Objects from the Class

Objects should be defined using the SequenceCluster constructor. You typically pass in a character vector of "words" to be clustered.

Slots

method: A character vector describing which algorithm was used.

rawSequences A character vector that contains the input words or "sequences" tthat were clustered.

weights A numeric vactor; the numbof times each unique raw sequence occurs.

distance: A dist object.

hc: An hclust object.

NC: An integer; the number of to cut from the dendrogram.

clusters: An integer vector containng cluster assignments.

Author(s)

Kevin R. Coombes <krc@silicovore.com>

Examples

```
data(longreads)
sequences <- longreads$connection[1:30]        # named character vector
sequences <- sequences[!duplicated(sequences)] # dedup
sc <- SequenceCluster(sequences)               # cluster
plot(sc)                                     # visualize
sc <- updateClusters(sc, NC = 7)
plot(sc, type = "unrooted")
```

StringGraph-class *Class* "StringGraph"

Description

The StringGraph class is used to represent graphs relating strings that arise from strings reprsenting long-read breakpoint sequences. The basic examples are: (1) "Motif Graphs" where the edges are subtring relations, and (2) "Decomposition Graphs" where the edges are restricted subtring relations that decompose a long read.

Usage

```
MotifGraph(motifNodes, alfa, name = "motif")
DecompositionGraph(decomp, alfa, motifNodes, name = "decomp")
exportSG(sg)
## S4 method for signature 'StringGraph,ANY'
plot(x, y, ...)
```

Arguments

motifNodes	A list of node names and counts, separated by length. In particular, motifNodes[[L]] should contain the nodes of length L.
alfa	A Cipher object.
name	A character vector of length one.
decomp	A decomposito object; see details.
sg	An object of the StringGraph class.
x	An object of the StringGraph class.
y	Anything; it is ignored.
...	Extra graphical parameters.

Value

The MotifGraph and DecompositionGraph functions return a new object of the StringGraph class. The plot method and exportSG function nothing and are called for their side effects.

Objects from the Class

Objects should be defined using the MotifGraph or DecompositionGraph constructor. You typically pass in a "motifNodes" object, which is a list of sequence-strings separated by length, along with some auxiliary information.

Slots

name: A character vector of length one.

edgelist: A matrix representing a graph as a list of edges.

nodelist: A matrix representing the nodes of the graph, along with their properties.

graph: An igraph object.

layout: A matrix containng x-y locations for the nodes.

Note

Attempting to manipulate a StringGraph object using text containing NAs, missing values, or previously unknown symbols will result in an error.

Author(s)

Kevin R. Coombes <krc@silicovore.com>

SVAkignR-data *SVAlignR Sample Data*

Description

These data sets contain binary versions of data describing breakpoints and long read sequences from an HPV-positive head-and-neck cancer.

Usage

```
data("longreads")
```

Format

`longreads` A data frame with 197 rows and 5 columns. Each row represents a single Oxford Nanopore long read from a study of a cell line from an HPV-positive head-and-neck squamous cell tumor. The five columns contain (i) a uniqie identifier of each long read, (ii) the length of the read, in bytes, (iii) the ordered sequence of break points, represented as a hyphen separated list of numeric identifiers, (iv) manually estimated natural groups of reads, and (v) a manual curated indication of whether certain long reads should be omitted from the analysis.

`breakpoints` A data frame with 82 rows and 11 columns. Each row represents a single breakpoint from a study of a cell line from an HPV-positive head-and-neck squamous cell tumor. The columns contain (1) a unique identifier that is used in the long read connections, (2-4) a description of the chromosomal segment to the left of the breakpoint, (5-7) a description of the chromosomal segment to the right of the breakpoint, (8-9) the orientation of the two chromosomal segments, (10) a shorthand description of the breakpoint with the segment names separated by a vertical bar and negative strands contained in parentheses, and (11) a shorthand representation of the reverse orientation of the breakpoint.

Author(s)

Kevin R. Coombes <krc@silicovore.com>

Source

Long read (Oxford Nanopore) sequencing was performed on samples prepared at the laboratory of Maura Gillison and David Symer. Characterization of long reads as a sequence of well-defined break points was performed by Keiko Akagi.

Examples

```
data(longreads)
head(longreads)

alphabet <- Cipher(longreads$connection)
en <- encode(alphabet, "0-50-74-0-50-74-35")
en
decode(alphabet, en)
```

Index

∗ **cluster**
AlignedCluster-class, 2
SequenceCluster-class, 4
∗ **datasets**
SVAkignR-data, 7
∗ **math**
Cipher-class, 3
StringGraph-class, 5

alignAllClusters *(*AlignedCluster-class*)*, 2
alignCluster *(*AlignedCluster-class*)*, 2
AlignedCluster-class, 2

breakpoints *(*SVAkignR-data*)*, 7

Cipher *(*Cipher-class*)*, 3
Cipher-class, 3

decode *(*Cipher-class*)*, 3
DecompositionGraph *(*StringGraph-class*)*, 5

encode *(*Cipher-class*)*, 3
exportSG *(*StringGraph-class*)*, 5

heat *(*SequenceCluster-class*)*, 4

image,AlignedCluster-method *(*AlignedCluster-class*)*, 2

longreads *(*SVAkignR-data*)*, 7

makeSubsMatrix *(*AlignedCluster-class*)*, 2
MotifGraph *(*StringGraph-class*)*, 5

plot,SequenceCluster,missing-method *(*SequenceCluster-class*)*, 4
plot,StringGraph,ANY-method *(*StringGraph-class*)*, 5

SequenceCluster *(*SequenceCluster-class*)*, 4
SequenceCluster-class, 4
StringGraph *(*StringGraph-class*)*, 5
StringGraph-class, 5
SVAkignR-data, 7
SVAlignR-data *(*SVAkignR-data*)*, 7

updateClusters *(*SequenceCluster-class*)*, 4