# Package ‘odseq’

January 31, 2017

**Type** Package  
**Title** Outlier detection in multiple sequence alignments  
**Version** 1.2.0  
**Date** 2015-12-20  
**Author** José Jiménez  
**Maintainer** José Jiménez <jose@jimenezluna.com>  
**Description** Performs outlier detection of sequences in a multiple sequence alignment using bootstrap of predefined distance metrics. Outlier sequences can make downstream analyses unreliable or make the alignments less accurate while they are being constructed. This package implements the OD-seq algorithm proposed by Jehl et al (doi 10.1186/s12859-015-0702-1) for aligned sequences and a variant using string kernels for unaligned sequences.  
**License** MIT + file LICENSE  
**LazyData** True  
**Encoding** UTF-8  
**biocViews** Alignment, MultipleSequenceAlignment  
**VignetteBuilder** knitr  
**Suggests** knitr(>= 1.11)  
**Depends** R (>= 3.2.3)  
**Imports** msa (>= 1.2.1), kebabs (>= 1.4.1), mclust (>= 5.1)  
**NeedsCompilation** no

### R topics documented:

- odseq-package .................................................. 2  
- odmix .............................................................. 3  
- odseq .............................................................. 4  
- odseq_unaligned .................................................. 5  
- seqs ............................................................... 6

**Index**  

1
**odseq-package**  
*Outlier detection in multiple sequence alignments*

**Description**

Performs outlier detection of sequences in a multiple sequence alignment using bootstrap of predefined distance metrics. Outlier sequences can make downstream analyses unreliable or make the alignments less accurate while they are being constructed. This package implements the OD-seq algorithm proposed by Jehl et al (doi 10.1186/s12859-015-0702-1) for aligned sequences and a variant using string kernels for unaligned sequences.

**Details**

The DESCRIPTION file:

Package: odseq
Type: Package
Title: Outlier detection in multiple sequence alignments
Version: 1.2.0
Date: 2015-12-20
Author: José Jiménez
Maintainer: José Jiménez <jose@jimenezluna.com>
Description: Performs outlier detection of sequences in a multiple sequence alignment using bootstrap of predefined distance metrics.
License: MIT + file LICENSE
LazyData: True
Encoding: UTF-8
biocViews: Alignment, MultipleSequenceAlignment
VignetteBuilder: knitr
Suggests: knitr(>= 1.11)
Depends: R (>= 3.2.3)
Imports: msa (>= 1.2.1), kebabs (>= 1.4.1), mclust (>= 5.1)
NeedsCompilation: no

Index of help topics:

odmix Gaussian mixture modelling of distances in a multiple sequence alignment.
odseq Outlier detection in a multiple sequence alignment
odseq-package Outlier detection in multiple sequence alignments
odseq_unaligned Outlier detection provided a distance/similarity matrix of sequences.
seqs PFAM plus random data.

**Author(s)**

José Jiménez

Maintainer: José Jiménez <jose@jimenezluna.com>
odmix

References


See Also

odseq odseq_unaligned

Examples

```r
library(msa)
data(seqs)
al <- msa(seqs)
odseq(al, distance_metric = "affine", B = 1000, threshold = 0.025)
```

---

Gaussian mixture modelling of distances in a multiple sequence alignment.

Description

This function performs clustering of biological sequences via fitting a Gaussian mixture model of the distances defined by the odseq algorithm.

Usage

```r
odmix(msa_object, distance_metric, groups)
```

Arguments

- `msa_object`: An object of formal class MsaAMultipleAlignment, as provided by the msa package.
- `distance_metric`: A string indicating the type of distance metric to be computed. Either 'linear' and 'affine' is supported at the moment.
- `groups`: Number of groups to fit in the mixture model. If a numeric vector of size n, n models will be fitted and a list of BIC values will be given to choose a single model.

Value

A list containing the following items:

- `prob`: A numeric matrix of size n x groups where the probability of belonging to a group is provided for each sequence.
- `class`: The class assigned according to `prob`. Returns a numeric vector.
- `BIC`: BIC values for the models proposed in `groups`

Author(s)

José Jiménez <jose@jimenezluna.com>
See Also

odseq_unaligned odseq

Examples

library(msa)
data(seqs)
al <- msa(seqs)
odmix(al, distance_metric = "affine", groups = 2)

Description

This function will first compute a distance metric among every sequence in the multiple alignment. Then it will bootstrap an average score of these distance to provide information on the distribution of scores, which is used to distinguish outlier sequences with a certain threshold

Usage

odseq(msa_object, distance_metric = "linear", B = 100, threshold = 0.025)

Arguments

msa_object An object of formal class MsaAMultipleAlignment, as provided by the msa package.
distance_metric A string indicating the type of distance metric to be computed. Either 'linear' and 'affine' is supported at the moment.
B Integer indicating the number of bootstrap replicates to be run. The higher the more robust the detection should be.
threshold Float indicating the probability to be left at the right of the bootstrap scores distribution when computing outliers. This parameter may need some tuning depending on each specific problem

Value

Returns a logical vector, where TRUE indicates an outlier.

Author(s)

José Jiménez <jose@jimenezluna.com>

References


See Also

odseq_unaligned
Examples

```r
library(msa)
data(seqs)
al <- msa(seqs)
odseq(al, distance_metric = "affine", B = 1000, threshold = 0.025)
```

Description

Outlier detection provided a distance/similarity matrix of sequences.

Usage

```r
odseq_unaligned(distance_matrix, B = 100, threshold = 0.025, type = "similarity")
```

Arguments

- **distance_matrix**: A numeric matrix representing either similarity or distance among unaligned sequences. Package `kebabs` may be useful for this task.
- **B**: Integer indicating the number of bootstrap replicates to be run. The higher the more robust the detection should be.
- **threshold**: Float indicating the probability to be left at the right of the bootstrap scores distribution when computing outliers. This parameter may need some tuning depending on each specific problem.
- **type**: A string indicating the type of distance metric used. Either 'similarity' or 'distance'.

Value

Returns a logical vector, where TRUE indicates an outlier.

Author(s)

José Jiménez <jose@jimenezluna.com>

References


See Also

`odseq`
Examples

```r
library(kebabs)
data(seqs)
sp <- spectrumKernel(k = 3)
mat <- getKernelMatrix(sp, seqs)
odseq_unaligned(mat, B = 1000, threshold = 0.025, type = "similarity")
```

Description

Sequences from a certain PFAM family plus 100 random sequences.

Usage

```r
data("seqs")
```

Value

An object of class AAStringSet.

Examples

```r
data(seqs)
```
Index

odmix, 3
odseq, 3, 4, 4, 5
odseq-package, 2
odseq_unaligned, 3, 4, 5

seqs, 6