Package ‘signal.hsmm’

March 8, 2015

Type Package
Title Predict Presence of Signal Peptides
Version 1.2.1
LazyData true
Date 2015-03-08
Description Predicts presence of signal peptides in eukaryotic protein using
hidden semi-Markov models. The implemented algorithm can be accessed both
from command line and GUI.
License GPL-3
URL https://github.com/michbur/signal.hsmm
Depends R (>= 3.0.0)
Imports biogram, seqinr, shiny
Suggests knitr, markdown, shinyAce
NeedsCompilation no
Repository CRAN
Author Michal Burdukiewicz [cre, aut],
Piotr Sobczyk [aut]
Maintainer Michal Burdukiewicz <michalburdukiewicz@gmail.com>
Date/Publication 2015-03-08 08:57:59

R topics documented:

  aaaggregation ......................................................... 2
  benchmark_dat ....................................................... 2
  duration_viterbi ..................................................... 3
  find_nhc .............................................................. 4
  gui_signal.hsmm ...................................................... 4
  hmm_pred ............................................................. 5
  hmm_pred_list ......................................................... 5
  is_protein ............................................................ 6
aaaggregation

Scheme for amino acid aggregation

Description

Amino acids are grouped together in larger sets based on their physicochemical properties important in the recognition of signal peptide.

Usage

aaaggregation

Format

a list of length four. Each element contains a character vector of amino acid names (one-letter abbreviations).

benchmark_dat

Benchmark data set for signal.hsmm

Description

Lists eukaryotic proteins added to UniProt database release 2014_07 between 2011 and 2014 (140 proteins with signal peptide and 280 randomly sampled proteins without signal peptide). All proteins were used in the benchmark test to compare the performance of signal.hsmm and other signal peptide predictors.

Usage

benchmark_dat

Format

a list of SeqFastaAA objects. Slot sig contains the range of signal peptide (if any).
duration_viterbi

Source

UniProt

Examples

summary(benchmark_dat)

duration_viterbi  Compute most probable path with extended Viterbi algorithm

Description

Viterbi algorithm for Hidden Markov Model with duration Valid only for special case on data

Usage

duration_viterbi(aa_sample, pipar, tpmpar, od, params)

Arguments

aa_sample character vector representing single aminoacid sequence.

pipar Probabilities of initial state in Markov Model.

tpmpar Matrix with transition probabilities between states.

od Matrix of response probabilities. Eg. od[1,2] is a probability of signal 2 in state 1.

params Matrix of probability distribution for duration. Eg. params[10,2] is probability of duration of time 10 in state 2.

Value

A list of length four:

• path a vector of most probable path
• viterbi values of probability in all intermediate points,
• psi matrix that gives for every signal and state the previous state in viterbi path,
• duration matrix that gives for every signal and state gives the duration in that state on viterbi path.

Note

Currently has very restricted application to specific input All computations are on logarithms of probabilities
find_nhc

Localize n-, h- and c-region in signal peptide

Description

Finds borders between distinct regions constituting signal peptides using a heuristic algorithm.

Usage

find_nhc(protein, signal = NULL)

Arguments

  protein   a vector of amino acids or object of class SeqFastaAA.
  signal    range of signal peptide. If NULL, the attribute sig of protein will be used.

Value

  a vector of length 4 containing positions of:
    1. start of n-region,
    2. start of h-region,
    3. start of c-region,
    4. cleavage site.

References


checksum 54157

---

gui_signal.hsmm

GUI for signal.hsmm

Description

A graphical user interface for predicting presence of signal peptides.

Usage

  gui_signal.hsmm()

Value

  null.
**hsmm_pred**

**Note**
Any ad-blocking software may be cause of malfunctions.

**See Also**

- run_signal.hsmm

---

### hsmm_pred

**hsmm_pred class**

**Description**
A single prediction of signal.hsmm.

**Details**
Always a named list of five elements

1. **sp_probability** is a probability of signal peptide presence.
2. **sp_start** is a start of potential signal peptide (naively 1 aminoacid).
3. **sp_end** is a position of last amino acid of signal peptide.
4. **struc** is numeric vector representing predicted structure of input protein.
5. **prot** is character vector containing input sequence of amino acids.
6. **str_approx** has value bigger than 0 if the predicted signal peptide structure was approximated (usually in case of sequences that have no signal peptides).

**See Also**

- summary.hsmm_pred
- plot.hsmm_pred

---

### hsmm_pred_list

**hsmm_pred_list class**

**Description**
A list of prediction(s) generated by run_signal.hsmm function.

**Details**
A named list. Each element belongs to the **hsmm_pred** class.

**See Also**

- summary.hsmm_pred_list
- pred2df
is_protein  

**Protein test**

**Description**

Checks if an object is a protein (contains letters from one-letter amino acid code).

**Usage**

```r
is_protein(object)
```

**Arguments**

- `object` character vector where each element represents one amino acid.

**Value**

TRUE or FALSE.

---

plot.hsmm_pred  

**Plot single signal.hsmm prediction**

**Description**

Plots objects of class `hsmm_pred`.

**Usage**

```r
## S3 method for class 'hsmm_pred'
plot(x, add_legend = TRUE, ...)
```

**Arguments**

- `x` object of class `hsmm_pred`.
- `add_legend` logical, if TRUE, legend is added to the plot.
- `...` ignored.

**Value**

Nothing.
**pred2df**

*Convert list of signal.hsmm predictions*

**Description**

Converts objects of class `hsmm_pred_list` to data frame.

**Usage**

`pred2df(object)`

**Arguments**

- `object`: Of class `hsmm_pred_list`.

**Value**

Data frame which columns contain respectively the probability of signal peptide presence as well as the start and the end of predicted signal peptide.

---

**predict.sighsnn_model**  
*Predict sighsnn_model object*

**Description**

Predicts

**Usage**

```r
## S3 method for class 'sighsnn_model'
predict(object, newdata, ...)```

**Arguments**

- `object`: sighsnn_model object.
- `newdata`: unknown sequence of class character or character. Alternatively, a list of sequences in mentioned formats.
- `...`: further arguments passed to or from other methods.

**Examples**

```r
#remember to remove it
## Not run:
pos_train_ultrahard <- read_uniprot("pos_ultrahard_data.txt", euk = TRUE)
model1 <- train_hsmm(pos_train_ultrahard, aa_group = aaaggregation)
predict(model1, benchmark_dat[1L:5])

## End(Not run)```
read_txt

*Read sequences from .txt file*

**Description**
Read sequence data saved in text file.

**Usage**
```r
read_txt(connection)
```

**Arguments**
- `connection`: a connection to the text (.txt) file.

**Details**
The input file should contain one or more amino acid sequences separated by empty line(s).

**Value**
a list of sequences. Each element has class `SeqFastaAA`. If connection contains no characters, function prompts warning and returns NULL.

---

read_uniprot

*Read data from UniProt database*

**Description**
Read data saved in UniProt original flat text format.

**Usage**
```r
read_uniprot(connection, euk)
```

**Arguments**
- `connection`: a connection to UniProt data in text format.
- `euk`: logical value if data has an eukaryotic origin.

**Value**
a list of sequences. Each element has a class `SeqFastaAA`. Slot `sig` contains the range (start and end) of signal peptide. Sequence with more than one cleavage site or atypical aminoacids are removed without any notice.
run_signal.hsmm

Predict presence of signal peptide in protein

Description
Using the hidden semi-Markov model predict presence of signal peptide in eukaryotic proteins.

Usage
run_signal.hsmm(test_data)

Arguments
test_data single protein sequence (character vector) or list of sequences. It may be an object of class SeqFastaAA.

Details
Function signal.hsmm returns respectively probability of presence of signal peptide, start of signal peptide and the probable cleavage site localization. If input consists of more than one sequence, result is a data.frame where each column contains above values for different proteins.

Value
An object of class hsmm_pred_list.

Note
Currently start of signal peptide is naively set as 1 amino acid.

See Also
hsmm_pred_list hsmm_pred

Examples
#run signal.hsmm on one sequence
x1 <- run_signal.hsmm(benchmark_dat[[1]])

#run signal.hsmm on one sequence, but input is a character vector
#run signal.hsmm on list of sequences
x3 <- run_signal.hsmm(benchmark_dat[1:3])
#see summary of results
summary(x3)
#print results as data frame
pred2df(x3)
#summary one result
summary(x3[[1]])
plot(x3[[1]])

summaryNhsmm_pred

Summarize single signal.hsmm prediction

Description

Summarizes objects of class hsmm_pred.

Usage

## S3 method for class 'hsmm_pred'
summary(object, ...)
summary.hsmm_pred_list

Arguments

object of class hsmm_pred.

... ignored

Value

Nothing.

Description

Summarizes objects of class hsmm_pred_list.

Usage

## S3 method for class 'hsmm_pred_list'
summary(object, ...)

Arguments

object of class hsmm_pred_list.

... ignored

Value

nothing.

train_hsmm

Train sighsmm_model object

Description

Train sighsmm_model object

Usage

train_hsmm(train_data, aa_group, max_length = 32)

Arguments

train_data training data.

aa_group method of aggregating amino acids.

max_length maximum length of signal peptide.
Value

object of class sighsmm_model
## Index

<table>
<thead>
<tr>
<th>Topic</th>
<th>Functions/Variables</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>classify</strong></td>
<td>run_signal.hmm</td>
<td>9</td>
</tr>
<tr>
<td><strong>datasets</strong></td>
<td>aaaggregation</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>benchmark_dat</td>
<td>2</td>
</tr>
<tr>
<td><strong>hplot</strong></td>
<td>plot.hsmm_pred</td>
<td>6</td>
</tr>
<tr>
<td><strong>manip</strong></td>
<td>pred2df</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>read_txt</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>read_uniprot</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>summary.hsmm_pred_list</td>
<td>11</td>
</tr>
<tr>
<td><strong>methods</strong></td>
<td>plot.hsmm_pred</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>summary.hsmm_pred</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>summary.hsmm_pred_list</td>
<td>11</td>
</tr>
<tr>
<td><strong>print</strong></td>
<td>summary.hsmm_pred</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>summary.hsmm_pred_list</td>
<td>11</td>
</tr>
<tr>
<td>aaaggregation</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>benchmark_dat</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>connection</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>duration_viterbi</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>find_nhc</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>gui_signal.hmm</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>hsmm_pred</td>
<td></td>
<td>5, 6, 9-11</td>
</tr>
<tr>
<td>hsmm_pred_list</td>
<td></td>
<td>5, 7, 9, 11</td>
</tr>
<tr>
<td>is_protein</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>plot.hsmm_pred</td>
<td></td>
<td>5, 6</td>
</tr>
<tr>
<td>pred2df</td>
<td></td>
<td>5, 7</td>
</tr>
<tr>
<td>predict.sighsmm_model</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>read_txt</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>SeqFastaAA</td>
<td></td>
<td>2, 4, 8, 9</td>
</tr>
<tr>
<td>signal.hmm</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>signal.hsmm-package</td>
<td>(signal.hsmm)</td>
<td>10</td>
</tr>
<tr>
<td>summary.hsmm_pred</td>
<td></td>
<td>5, 10</td>
</tr>
<tr>
<td>summary.hsmm_pred_list</td>
<td></td>
<td>5, 11</td>
</tr>
<tr>
<td>train_hsmm</td>
<td></td>
<td>11</td>
</tr>
</tbody>
</table>