Package 'bamboo'

April 3, 2020

Type Package

Title Protein Secondary Structure Prediction Using the Bamboo Method

Version 0.9.25

Date 2020-04-02

Description Implementation of the Bamboo methods described in Li, Dahl, Vannucci, Joo, and Tsai (2014) <DOI:10.1371/journal.pone.0109832>.

URL https://github.com/dbdahl/bamboo

BugReports https://github.com/dbdahl/bamboo/issues

Depends R (>= 3.5.0)

Imports rscala (>= 3.2.18)

LazyData true

License GPL-3

Encoding UTF-8

RoxygenNote 7.1.0

NeedsCompilation no

Author David B. Dahl [aut, cre]

Maintainer David B. Dahl <dahl@stat.byu.edu>

Repository CRAN

Date/Publication 2020-04-02 22:00:02 UTC

R topics documented:

bamboo.estimate *Bayesian Model of Protein Primary Sequence for Secondary Structure Prediction*

Description

These functions implement the methodology described in the paper "Bayesian Model of Protein Primary Sequence for Secondary Structure Prediction" cited below. The main function is `bamboo.estimate`, whose arguments are results from `bamboo.priorMSA`, `bamboo.priorNoHE`, and `bamboo.likelihood` functions. A `plot` method is provided to produce figures like those in the paper using results of the `bamboo.estimate` function. The `bamboo.prepare` function ensures that the necessary dependencies are loaded (and is automatically called by the other functions).

Usage

```
bamboo.likelihood(primary,secondary,countsDirectory="HETC",force=FALSE,warn=TRUE)
bamboo.priorMSA(countsMatrix,alpha=c(1,1,1,1))

bamboo.priorNonInfo()
bamboo.estimate(likelihood,prior,nSamples,dropFirst,initialState=NULL,
                doLeastSquaresEstimation=FALSE,dumpStates=FALSE)
## S3 method for class 'bamboo.estimate'
plot(x,ss=NULL,...)
```

Arguments

`primary`	A character vector (whose length is that same as *secondary*) that gives the amino acid sequences (using 1-letter amino acid codes) used to train the sampling model.
`secondary`	A character vector (whose length is that same as *primary*) that gives the secondary structure sequences (using 1-letter codes) used to train the sampling model and the MM prior.
`countsDirectory`	A name of the directory to use for storing count files.
`countsMatrix`	An L-by-4 matrices, where L is the protein length. Row l (where l=1,...,L) gives the number of times that the secondary structure of the aligned proteins is H, E, T, and C, respectively.
`alpha`	A numeric vector of four strictly-positive real values for the Dirichlet prior in the MSA prior.
`force`	A logical indicating that, in the case that the count files already exists, the counting should be performed again away. This is necessary if the *primary* or *secondary* arguments have changed since the last counting.
`warn`	A logical indicating that, in the case that the count files already exists, a warning should be displayed indicating that the counting is not performed again. Recounting is necessary if the *primary* or *secondary* arguments have changed since the last counting.

`likelihood` The result obtained by evaluating the function returned by `bamboo.likelihood` for an amino acid sequence encoded as a character vector of length 1 using 1-letter amino acid codes.

`prior` The result of a call to the `bamboo.priorMSA` function or the `bamboo.priorNonInfo` function.

`nSamples` An integer giving the number of MCMC samples after burnin to use for inference.

`dropFirst` An integer giving the number of MCMC samples to discard as burnin.

`initialState` A character vector of length 1 giving the secondary structure state to start the Markov chain Monte Carlo algorithm. If `NULL`, a reasonable default is used.

`doLeastSquaresEstimation`
A logical implementing an undocumented estimation method.

`dumpStates` A logical implementing whether secondary structure states should be printed to standard output (stdout). This feature is not intended for normal usage and the output is only likely to be seen when running R on a Linux terminal.

`x` The result from a call to the `bamboo.estimate` function.

`ss` A character vector of arbitrary length giving secondary structures to display above the marginal probabilities. The names of the elements of the vector is used to label each line. If `NULL`, nothing is plotted above the marginal probability plot.

`...` Extra arguments passed to the `par` function when plotting.

Value

The result of the `bamboo.estimate` function is a list. Some of the more interesting elements of the list are:

`mpState` The estimated secondary structure state using the marginal probability (MP) method that selects the most likely block form for each position.

`mapState` The estimated secondary structure state using the maximum a posterior (MAP) method that returns the visited state that maximizes the posterior probability.

`marginalProbabilities`
A matrix of marginal probabilities of each state for each position.

Author(s)

David B. Dahl <dahl@stat.byu.edu>

References

Q. Li, D. B. Dahl, M. Vannucci, H. Joo, J. W. Tsai (2014), Bayesian Model of Protein Primary Sequence for Secondary Structure Prediction, PLOS ONE, 9(10), e109832. <DOI:10.1371/journal.pone.0109832>

Examples

```
data(bamboo.training,
     bamboo.validation.casp9,
     bamboo.validation.astral30,
     bamboo.MSA.casp9,
     bamboo.MSA.astral30)


## Be patient, this example can take several seconds on a fast computer.
likelihood <- bamboo.likelihood(bamboo.training[,"primary"],bamboo.training[,"hetc"],force=TRUE)

prior.NonInfo <- bamboo.priorNonInfo()
bamboo.MSA <- c(bamboo.MSA.casp9,bamboo.MSA.astral30)

target <- "f3rvca_0"
aa <- bamboo.validation.astral30[bamboo.validation.astral30$name==target,"primary"]
fm.NonInfo <- bamboo.estimate(likelihood(aa),prior.NonInfo,5000,500)
fm.MSA   <- bamboo.estimate(likelihood(aa),bamboo.priorMSA(bamboo.MSA[[target]]),5000,500)

ss <- c(
  "Truth"=bamboo.validation.astral30[bamboo.validation.astral30$name==target,"hetc"],
  "NonInfo-MP"=fm.NonInfo$mpState,
  "MSA-MP"=fm.MSA$mpState
)
plot(fm.MSA,ss)
```

`bamboo.MSA.astral30` *MSA Information for the bamboo.validation.astral30 Test Dataset*

Description

This data provides the multiple sequence alignment (MSA) information for the `bamboo.validation.astral30` test dataset in the paper cited below. This MSA information list gives the count matrices for the 3,143 proteins in the bamboo.validation.astral30 test dataset that have MSA information. Each row in the matrix is the count vector for the number of times that each of the four secondary structure types is found in that position of the alignment output.

Usage

```
bamboo.MSA.astral30
```

Format

A list containing 3,143 matrices.

Source

Q. Li, D. B. Dahl, M. Vannucci, H. Joo, J. W. Tsai (2014), Bayesian Model of Protein Primary Sequence for Secondary Structure Prediction, PLOS ONE, 9(10), e109832. <DOI:10.1371/journal.pone.0109832>

References

Q. Li, D. B. Dahl, M. Vannucci, H. Joo, J. W. Tsai (2014), Bayesian Model of Protein Primary Sequence for Secondary Structure Prediction, PLOS ONE, 9(10), e109832. <DOI:10.1371/journal.pone.0109832>

---

bamboo.MSA.casp9 *MSA Information for the bamboo.validation.casp9 Test Dataset*

Description

This data provides the multiple sequence alignment (MSA) information for the `bamboo.validation.casp9` test dataset in the paper cited below. This MSA information list gives the count matrices for the 109 proteins in the bamboo.validation.casp9 test dataset that have MSA information. Each row in the matrix is the count vector for the number of times that each of the four secondary structure types is found in that position of the alignment output.

Usage

```
bamboo.MSA.casp9
```

Format

A list containing 109 matrices.

References

Q. Li, D. B. Dahl, M. Vannucci, H. Joo, J. W. Tsai (2014), Bayesian Model of Protein Primary Sequence for Secondary Structure Prediction, PLOS ONE, 9(10), e109832. <DOI:10.1371/journal.pone.0109832>

---

bamboo.training *Training Dataset*

Description

This training dataset gives the names, the primary structure (amino acid sequences), and the secondary structure of 15,201 individual proteins from the ASTRAL SCOP 1.75 structure set filtered at 95% sequence identity as used in the paper cited below.

Usage

```
bamboo.training
```

Format

A data frame containing 15,201 observations on the following 3 variables.

1. name: protein name;
2. primary: protein primary structure (amino acid sequence) in 20 letters denoting the 20 amino acids;
3. hetc: secondary structure in 4 letters denoting the 4 structure types: helix (H), strand (E), turn (T) and coil (C).

Source

Chandonia JM, Hon G, Walker NS, Conte LL, Koehl P, et al. (2004) The astral compendium in 2004. *Nucleic Acids Research* 32: D189-D192

References

Q. Li, D. B. Dahl, M. Vannucci, H. Joo, J. W. Tsai (2014), Bayesian Model of Protein Primary Sequence for Secondary Structure Prediction, PLOS ONE, 9(10), e109832. <DOI:10.1371/journal.pone.0109832>

`bamboo.validation.astral30`

Validation (Test) Dataset named astral30

Description

This validation dataset gives the names, the primary structure (amino acid sequences), and the secondary structure of 3,344 individual proteins from the SCOPe 2.03 data set filtered at 30% sequence identity as used in the paper cited below.

Usage

```
bamboo.validation.astral30
```

Format

A data frame containing 3,344 observations on the following 3 variables.

1. name: protein name;
2. primary: protein primary structure (amino acid sequence) in 20 letters denoting the 20 amino acids;
3. hetc: secondary structure in 4 letters denoting the 4 structure types: helix (H), strand (E), turn (T) and coil (C).

Source

Fox NK, Brenner SE, Chandonia JM (2013) Scope: Structural classification of proteins extended, integrating scop and astral data and classification of new structures. *Nucleic Acids Research* 42: D304-D309. <DOI:10.1093/nar/gkt1240>

References

Q. Li, D. B. Dahl, M. Vannucci, H. Joo, J. W. Tsai (2014), Bayesian Model of Protein Primary Sequence for Secondary Structure Prediction, PLOS ONE, 9(10), e109832. <DOI:10.1371/journal.pone.0109832>

`bamboo.validation.casp9`

Validation (Test) Dataset Named casp9

Description

This validation dataset gives the names, the primary structure (amino acid sequences), and the secondary structure of 203 individual proteins from the targets used in CASP9 experiments as used in the paper cited below.

Usage

```
bamboo.validation.casp9
```

Format

A data frame containing 203 observations on the following 3 variables.

1. name: protein name;
2. primary: protein primary structure (amino acid sequence) in 20 letters denoting the 20 amino acids;
3. hetc: secondary structure in 4 letters denoting the 4 structure types: helix (h), strand (e), turn (t) and coil (c).

Source

Moult J, Fidelis K, Kryshtafovych A, Tramontano A (2011) Critical assessment of methods of protein structure prediction (casp) round ix. *Proteins: Structure, Function, and Bioinformatics* 79: 1-5. <DOI:10.1002/prot.23200>

References

Q. Li, D. B. Dahl, M. Vannucci, H. Joo, J. W. Tsai (2014), Bayesian Model of Protein Primary Sequence for Secondary Structure Prediction, PLOS ONE, 9(10), e109832. <DOI:10.1371/journal.pone.0109832>

Index