

The aaMI Package

February 16, 2008

Version 1.0-1

Date 2005-07-08

Title Mutual information for protein sequence alignments

Author Kurt Wollenberg <kurt.wollenberg@gmail.com>

Maintainer Kurt Wollenberg <kurt.wollenberg@gmail.com>

Depends R (>= 2.0.0)

Description This package contains five functions. `read.FASTA` reads in a FASTA-format alignment file and parses it into a data frame. `read.CX` reads in a ClustalX .aln-format file and parses it into a data frame. `read.Gdoc` reads in a GeneDoc .msf-format file and parses it into a data frame. The alignment data frame returned by each of these functions has the sequence IDs as the row names and each site in the alignment is a column in the data frame. The program `aaMI` calculates the mutual information between each pair of sites (columns) in the protein sequence alignment data frame. The program `aaMIin` calculates the normalized mutual information between pairs of sites in the protein sequence alignment data frame. The normalized mutual information of sites *i* and *j* is the mutual information of these sites divided by their joint entropy.

License GPL version 2 or newer

R topics documented:

<code>aaMI</code>	2
<code>aaMIin</code>	3
<code>read.CX</code>	4
<code>read.FASTA</code>	5
<code>read.GDoc</code>	5
Index	6

Description

Calculate a matrix of pairwise mutual information values for a protein sequence alignment

Usage

```
aaMI(file)
```

Arguments

`file` a connection or character string giving the name of the file to load.

Details

This script calculates the mutual information between pairs of sites for a protein sequence alignment. The alignment must be in the form of a data frame with the sequence IDs as the row names and each site as a column. The program begins by calculating the amino acid frequencies at each site. These frequencies are then used to calculate a vector containing the Shannon entropy H for each site. Shannon entropy is calculated using the equation

$$H_i = \sum_i (P(X_i) \log_2(P(X_i)))$$

where $P(X_i)$ = frequency of amino acid X at site i of the alignment. Next the program calculates the joint probabilities $P(X_i, Z_j)$ of pairs of amino acids X and Z at sites i and j . The Shannon entropy and joint probabilities are used to calculate the mutual information MI with the formula

$$MI_{i,j} = H_i + H_j - \sum_{i,j} (P(X_i, Z_j) \log_2(P(X_i, Z_j)))$$

Value

For the analysis of a protein sequence alignment data frame "file", the output is an NxN upper-triangular matrix, where N is the number of sites in the alignment. Values along the diagonal of the matrix are the entropy values (H) for each site.

Author(s)

Kurt Wollenberg

References

Shannon, C. E. and W. Weaver. (1949) *The Mathematical Theory of Communication*, University of Illinois Press.

Wollenberg, K. R. and W. R. Atchley. (2000) Separation of phylogenetic from functional associations in biological sequences by using the parametric bootstrap. *Proceedings of the National Academy of Science* **97** 3288-3291.

Examples

```
## Read in a protein sequence alignment file, FastA format
## Not run: SeqDataFA <- read.FASTA("ProteinSeqFastA.txt")
## Read in a protein sequence alignment file, ClustalX .aln format
## Not run: SeqDataCX <- read.CX("ProteinSeq.aln")
## Read in a protein sequence alignment file, GeneDoc .msf format
## Not run: SeqDataGD <- read.Gdoc("ProteinSeq.msf")

## Calculate the mutual information matrix for one of these alignments.
## Not run: ProteinSeqmi <- aaMI(SeqDataGD)
```

aaMIn

Normalized Mutual Information for a Protein Sequence Alignment

Description

Calculate a matrix of pairwise normalized mutual information values for a protein sequence alignment

Usage

```
aaMIn(file)
```

Arguments

`file` a connection or character string giving the name of the file to load.

Details

This script calculates the normalized mutual information between pairs of sites for a protein sequence alignment. The normalization constant used is the joint entropy, as described by Gloor, et al. (2005). The alignment must be in the form of a data frame with the sequence IDs as the row names and each site as a column. The program begins by calculating the amino acid frequencies at each site. These frequencies are then used to calculate a vector containing the Shannon entropy H for each site. Shannon entropy is calculated using the equation

$$H_i = \sum_i (P(X_i) \log_2(P(X_i)))$$

where $P(X_i)$ = frequency of amino acid X at site i of the alignment. Next the program calculates the joint probabilities $P(X_i, Z_j)$ of pairs of amino acids X and Z at sites i and j . The joint probabilities are used to calculate the joint entropy with the formula

$$JH_{ij} = \sum_{i,j} (P(X_i, Z_j) \log_2(P(X_i, Z_j)))$$

Shannon entropy and joint entropy are used to calculate the normalized mutual information MI with the formula

$$MI_{ij} = (H_i + H_j - JH_{ij}) / JH_{ij}$$

Value

For the analysis of a protein sequence alignment data frame "file", the output is an NxN upper-triangular matrix, where N is the number of sites in the alignment. Values along the diagonal of the matrix are the entropy values (H) for each site.

Author(s)

Kurt Wollenberg

References

Gloor, G. B, L. C. Martin, L. M. Wahl, and S. D. Dunn. (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* **44** 7156-7165.

Shannon, C. E. and W. Weaver. (1949) *The Mathematical Theory of Communication*, University of Illinois Press.

Wollenberg, K. R. and W. R. Atchley. (2000) Separation of phylogenetic from functional associations in biological sequences by using the parametric bootstrap. *Proceedings of the National Academy of Science* **97** 3288-3291.

Examples

```
## Read in a protein sequence alignment file, FastA format
## Not run: SeqDataFA <- read.FASTA("ProteinSeqFastA.txt")
## Read in a protein sequence alignment file, ClustalX .aln format
## Not run: SeqDataCX <- read.CX("ProteinSeq.aln")
## Read in a protein sequence alignment file, GeneDoc .msf format
## Not run: SeqDataGD <- read.Gdoc("ProteinSeq.msf")

## Calculate the mutual information matrix for one of these alignments.
## Not run: ProteinSeqminorm <- aaMIn(SeqDataGD)
```

read.CX

Read in a ClustalX .aln alignment file

Description

Reads in a ClustalX .aln alignment file and parses it into a data frame.

Usage

```
read.CX(file)
```

Arguments

file a connection or character string giving the name of the file to load.

read.FASTA	<i>Read in a FASTA alignment file</i>
------------	---------------------------------------

Description

Reads in a FASTA alignment file and parses it into a data frame. The FASTA file must be in .txt format.

Usage

```
read.FASTA(file)
```

Arguments

file	a connection or character string giving the name of the file to load.
------	---

read.GDoc	<i>Read in a GeneDoc alignment file</i>
-----------	---

Description

Reads in a GeneDoc alignment file and parses it into a data frame. The GeneDoc file must be in .msf format.

Usage

```
read.GDoc(file)
```

Arguments

file	a connection or character string giving the name of the file to load.
------	---

Index

*Topic **file**

aaMI, [1](#)

aaMIn, [3](#)

read.CX, [4](#)

read.FASTA, [5](#)

read.GDoc, [5](#)

*Topic **univar**

aaMI, [1](#)

aaMIn, [3](#)

aaMI, [1](#)

aaMIn, [3](#)

read.CX, [4](#)

read.FASTA, [5](#)

read.GDoc, [5](#)