Package 'partitionMetric'

February 20, 2015

Type Package

Title Compute a distance metric between two partitions of a set

Version 1.1

Date 2014-03-01

Author David Weisman, Dan Simovici

Maintainer David Weisman <David.Weisman@acm.org>

Depends R (>= 2.10.1)

Description partitionMetric computes a distance between two partitions of a set.

LazyLoad yes

Repository CRAN

Date/Publication 2014-03-02 14:03:31

NeedsCompilation no

R topics documented:

AhRs *Sample data for partitionMetric*

Description

This small dataset contains aligned protein sequences for seven alleles of the aryl hydrocarbon receptor (AhR).

Usage

```
data(AhRs)
```

Format

The format is a character matrix in which column i represents the i'th position in the alignment, and contains an amino acid code or "-" indicating an indel. Row names contain the animal species.

Details

A DNA or protein sequence has an associated index set $\{1, 2, \ldots, n\}$ that labels the n positions of the nucleotides or amino acids (AA). This index set can be partitioned such that all members referring to the same AA share a homogeneous partition. For example, given the sequence ATGTA and its index set $\{1, 2, \ldots, 5\}$, the "A" partition contains the subset $\{1, 5\}$, the "T" partition contains $\{2, 4\}$, and so on.

Given two aligned sequences and their respective partitions of the index set, a metric distance between these partitions can be computed. See `partitionMetric` for such a metric, along with an example of clustering this AhR dataset.

Source

This dataset was derived from NCBI HomoloGene:1224.

References

Mark Hahn, Aryl hydrocarbon receptors: diversity and evolution. *Chem Biol Interact*, 2002, **141**, 131-160

partitionMetric *Compute a distance metric between two partitions of a set*

Description

Given a set partitioned in two ways, compute a distance metric between the partitions.

Usage

```
partitionMetric(B, C, beta = 2)
```

Arguments

B	B and C are vectors that represents partitions of a single set, with each element representing a member of the set. B_i corresponds to C_i, and the two vectors must be the same length. The data types of B and C must be identical and convertable to a factor data type. See examples below for more information.
C	See B above.
beta	β is the nonlinear parameter used to compute the distance metric. See the publication referenced below for full details.

Value

The return value is a nonnegative real number representing the distance between the two partition of the set. Full details are in the paper referenced below.

Author(s)

David Weisman, Dan Simovici

References

David Weisman and Dan Simovici, Several Remarks on the Metric Space of Genetic Codes. *International Journal of Data Mining and Bioinformatics*, 2012(6).

See Also

as.dist, hclust

Examples

```
## Define several partitions of a 4-element set
gender <- c('boy',   'girl', 'girl',   'boy')
height <- c('short', 'tall', 'medium', 'tall')
age    <- c(7,       6,      5,        4)

## Compute some distances
```

```
(dGG <- partitionMetric (gender, gender))
(dGH <- partitionMetric (gender, height))
(dHG <- partitionMetric (height, gender))
(dGA <- partitionMetric (gender, age))
(dHA <- partitionMetric (height, age))

## These properties must hold for any metric
dGG == 0
dGH == dHG
dGA <= dGH + dHA

## Note that the partition names are irrelevant, and only need to be
## self-consistent within each B and C.  It follows that these two set
## partitions are identical and have distance 0.
partitionMetric (c(1,8,8), c(7,3,3)) == 0


## Use the set partition to measure amino acid acid sequence differences
## between several alleles of the aryl hydrocarbon receptor.

data(AhRs)
dim(AhRs)
AhRs[,1:10]

distanceMatrix <-
  matrix(nrow=nrow(AhRs), ncol=nrow(AhRs), 0,
         dimnames=list(rownames(AhRs), rownames(AhRs)))

for (pair in combn(rownames(AhRs), 2, simplify=FALSE)) {
  d <- partitionMetric (AhRs[pair[1],], AhRs[pair[2],], beta=1.01)
  distanceMatrix[pair[1],pair[2]] <- distanceMatrix[pair[2],pair[1]] <- d
}

hc <- hclust(as.dist(distanceMatrix))
plot(hc,
     sub=sprintf('Cophenentic correlation between distances and tree is %0.2f',
       cor(as.dist(distanceMatrix), cophenetic(hc))))
```

Index