Using Hamming Distance as Information for Clustering SNP Sets and Testing for Disease Association

Description

This document contains procedures to perform a Hamming distance-based clustering algorithm and a Hamming distance-based association test. This clustering algorithm is a heirarchical clustering method to cluster SNP sets or other categorical data. The association test is used to test whether a SNP-set is associated with the disease of interest. This test statistic assesses, based on Hamming distance, whether the similarity between a diseased and a normal individual differs from the similarity between two individuals of the same disease status.

Five functions are described in the followings:

1. `hd.prop()` and `hd.count()`: *Calculate Hamming distance between two strings of equal length, or between paired row vectors of a matrix*
2. `do.cluster.apw()` : *Cluster SNP sets or categorical data sets*
3. `cluster.tree.object()`: *Generate an object for dendrogram plot*
4. `gen.snp.set.list()`: *Select the number of clusters and determine the cluster for each SNP*
5. `HDAT()`: *Hamming distance-based association test*

Reference

Wang, C., Kao, W.H. and Hsiao, C.K. (2013). Using Hamming Distance as Information for Clustering SNP Sets and Testing for Disease Association. *submitted*.

Maintainer

Charlotte Wang <d99849002@ntu.edu.tw>

Last Updated

November 6, 2013

`hd.prop & hd.count` *Calculate Hamming Distance*

Description

Calculate Hamming distance between two strings of equal length, or between a matrix and a string.

Usage

```
hd.prop(x, y)
hd.count(x, y)
```

Arguments

`x`	A vector of a string or a matrix
`y`	A vector of a string

Details

`hd.prop()` calculates the proportion of elements that are dissimilar between two strings. `hd.count()` calculates the number of elements that are dissimilar between two strings. If `x` is a matrix, then `y` needs not be specified, and the Hamming distance is calculated between every pair of row vectors in `x`. If `x` is a vector, `y` must be specified.

Value

Return Hamming distance between two strings.

Example

```
x <- c(0, 1, 1, 0, 0)
y <- c(0, 0, 1, 1, 0)
hd.prop(x, y)


z <- c(1, 1, 0, 0, 0)
xx <- rbind(x, y, z)
hd.count(xx)
```

`do.cluster.apw` *Clustering Algorithm for SNP Sets or Categorical Data Sets*

Description

Perform a heirarchical clustering algorithm based on Hamming distance for SNP sets or categorical data sets.

Usage

```
do.cluster.apw(raw.data, obs.id = NULL)
```

Arguments

`raw.data`	Data matrix, rows are strings to be clustered. For instance, if SNP sets are investigated for clustering, then the *i*-th row contains the SNP genotypes of SNP *i* of all subjects.
`obs.label`	The id's of observations, also the indices of rows.

Details

`do.cluster.apw()` performs the Hamming distance-based clustering algorithm for SNPs, SNP sets, or categorical data.

Value

`cluster.id`	Matrix, the *i*-th column contains the cluster id's that the *i*-th SNP belongs to along the clustering procedures.
`distance`	Vector, the minimum distance in each clustering procedure.
`merge.SNP.list`	A list of two members. The first member indicates the order of clustering procedures. The second member provides the updated clusters in the current clustering procedure, that is, the id's of SNPs in the newly updated cluster.

Example

```
# cluster Soybean data (dataset "soybean-small.data"
downloadable)
test.data <- soybean[,-36]
HD.cluster <- do.cluster.apw(test.data, obs.label =
    rownames(test.data))
```

```
# cluster SNP data (dataset "cluster_SNPdata.csv"
downloadable)
test.data <- t(cluster.SNPdata[,-1]) # transpose the input
matrix
HD.cluster <- do.cluster.apw(test.data, obs.label =
    rownames(test.data))
```

`cluster.tree.object` *Generate an Object for Dendrogram Plot*

Description

Generate an object which can be used to plot a dendrogram.

Usage

```
cluster.tree.object(raw.data, cluster.result, obs.label =
NULL)
```

Arguments

`raw.data`	Data matrix, rows are strings to be clustered. For instance, if SNP sets are investigated for clustering, then the *i*-th row contains the SNP genotypes of SNP *i* of all subjects.
`cluster.result`	An object returned from `do.cluster.apw()`
`obs.label`	The id's of observations, also the indices of rows.

Details

With the Hamming distance-based clustering results from `do.cluster.apw()`, this function prepares an object that can be used to plot a dendrogram with `plot()`.

Value

This function returns an object with the class attribute "dendrogram" and therefore can used later to plot a dendrogram.

Example

```
# cluster Soybean dataset (dataset "soybean-small.data"
downloadable)
test.data <- soybean[,-36]
HD.cluster <- do.cluster.apw(test.data, obs.label =
    rownames(test.data))
HD.tree <- cluster.tree.object(test.data, HD.clustser,
    obs.label = rownames(test.data))
plot(HD.tree, ylab = "height", ylim = c(0,1))

# cluster SNP data (dataset "cluster_SNPdata.csv"
downloadable)
test.data <- t(cluster.SNPdata[,-1]) # transpose data matrix
HD.cluster <- do.cluster.apw(test.data, obs.label =
    rownames(test.data))
HD.tree <- cluster.tree.object(test.data, HD.cluster,
    obs.label = colnames(test.data))
plot(HD.tree, ylab = "height", ylim = c(0,1))
```

`gen.snp.set.list` *Select the Number of Clusters and Determine the Cluster for Each SNP*

Description

The number of clusters is determined based on a rule-of-thumb described in the reference. In brief, this rule is based on the maximum difference of heights between two successive nodes in a dendrogram. Then, select all SNPs whose corresponding maximum difference of height are greater than the *X*% percentile and the resulting number of clusters for these SNPs is the final number of clusters.

Usage

```
gen.snp.set.list(cluster.result, percentile.cut,
min.cluster.size)
```

Arguments

`cluster.result`	An object from `do.cluster.apw()`
`percentile.cut`	A percentile for the rule-of-thumb to determine the number of clusters after a dendrogram. Must be defined by the user.
`min.cluster.size`	Minimum size of a cluster which will involve Hamming distance-based association test. The value must be >= 1. Must be defined by the user.

Details

Based on the results from `do.cluster.apw()`, `gen.snp.set.list()` can be used to determine the number of clusters and the components inside each cluster. The latter information can be used for the Hamming distance-based association test. If the number of clusters is predetermined, then this function is not needed.

Value

`max.dist`	Vector, containing the maximum relative height in the dendrogram for each SNP
`snp.set`	List, containing the selected SNPs per every SNP set
`snp.set.dist`	Vector, containing the maximum relative height corresponding to each SNP set

Example

```
# cluster SNP data (dataset "cluster_SNPdata.csv"
downloadable)
test.data <- t(cluster.SNPdata[,-1]) # transpose data matrix
HD.cluster <- do.cluster.apw(test.data, obs.label =
    rownames(test.data))
output.snp.set <- gen.snp.set.list(HD.cluster, 0.95, 3)
```

`HDAT` *Hamming Distance-based Association Test*

Description

Hamming distance association test examines susceptibility to the disease of interest. This test assesses, based on Hamming distance, whether the similarity between a diseased and a normal individual differs from the similarity between

two individuals of the same disease status.

Usage

```
HDAT(test.data, disease.status, n.permu)
```

Arguments

`test.data`	The input genotype data matrix for SNP-set association test. The dimension is n by p, where n is the number of samples, and p is the number of SNPs in the set.
`disease.status`	Input vector of length n. The value 1 for cases and 0 for controls.
`n.permu`	Number of permutations. Must be defined by the user.

Details

`HDAT()` performs the SNP-set association test based on Hamming distance. This test statistic compares the "distance" (or difference) in SNP genotypes among individuals. If the SNP-set is not associated with the disease, then the expected distance between a case and a normal subject should be the same as that between two individuals of the same disease status. This function returns an object containing four items described in "Value".

Value

`info`	Vector containing three values, the numbers of cases and controls, respectively, and the number of SNPs used for analysis
`dissimilarity`	Two dissimilarity scores, U for the within-group comparisons and T for the between-group comparisons
`test.statistic`	Test statistic (the difference T-U)
`p.value`	P-value based on permutation tests

Example

```
# test SNP data (dataset "test_SNPdata.csv" downloadable)
HD.test <- HDAT(test.SNPdata[,-c(1,2)], disease.status =
    test.SNPdata[,"disease.status"], 5000)
```