## alizadeh *Sample data from a lymphoma/leukemia gene expression study*

### Description

Sample data for the ISIS method

### Usage

```
data(alizadeh)
```

### Format

x A 2000 x 62 gene expression data matrix of log-ratio values. 2,000 genes with the highest variance across the samples have been selected from the set of 4,026 genes considered by the original authors.

type The annotations of the 62 samples with respect to the cancer types FL, CLL, DLBCL-A, DLBCL-G.

### Source

The data are from the lymphoma/leukemia study of A. Alizadeh et al., Nature 403:503-511 (2000), http://llmpp.nih.gov/lymphoma/data/figure1/figure1.cdt

## gotomax *Local search function*

### Description

Finds local maxima of the score function 'tscore' by greedy hill climbing.

### Usage

```
gotomax(dat, split, p = 50, p.offs = 0, min.split.size = ceiling(0.1*ncol(dat)))
```

### Arguments

| | |
|---|---|
| `dat` | The gene expression data matrix. Rows correspond to genes, and columns correspond to samples. |
| `split` | Either a binary vector encoding a bipartition of the set of samples, or a binary matrix with every row encoding a bipartition of the set of samples. |
| `p` | The number of best discriminating genes selected for computing the score of a bipartition. |

`p.offs` The number of very best discriminating genes that are ignored when computing the score of a partition. Thus only the genes with ranks p.offs+1, ..., p are used to compute the score.

`min.split.size`
The minimal size either class of a bipartition must have to be accepted as a local maximum.

### Value

A binary matrix where each column (except the last one) represents a sample (in the original order). The rows encode the found local maximum bipartitions. The last entry in each row gives the score of the partition.

### Examples

```
data(alizadeh)
split <- rep(0, ncol(alizadeh$x))
split[which(alizadeh$type=="FL")] <- 1
y <- gotomax(alizadeh$x, split)
```

---

## isis *Class discovery for microarray data*

---

### Description

Finds bipartitions of a set of tissue samples that show clear separation regarding the expression of a subset of genes.

### Usage

```
isis(x, p = 50, p.offs = 0, max.nr.cand.splits = 500)
```

### Arguments

`x` The gene expression data matrix. Rows correspond to genes, and columns correspond to samples. The values in x may be normalized logarithms of intensities or log-ratios.

`p` The number of best discriminating genes selected for computing the score of a bipartition.

`p.offs` The number of very best discriminating genes that are ignored when computing the score of a partition. Thus only the genes with ranks p.offs+1, ..., p are used to compute the score. A value >0 may improve the robustness of the score, sorting out possibly spurious partitions with very few genes that discriminate very well between the two groups.

`max.nr.cand.splits`
After the generation of candidate bipartitions, only the max.nr.cand.splits partitions with highest scores are further investigated. This is used in order to control the running time of the program.

### Details

It may be useful to first discard genes from the data matrix that display consistently low variance across the samples or consistently low expression intensity. For the generation of candidate bipartitions, the data matrix is enlarged by adding average expression profiles of all clusters of genes resulting from a centroid linkage hierarchical clustering based on the correlation coefficient as similarity measure. In order to reduce computation time, similar candidate partitions are merged through hierarchical clustering before applying the local search to find local maximum partitions (see 'gotomax').

### Value

A matrix where each column (except the last one) represents a sample (in the original order). The rows represent the found bipartitions with highest scores, coded as 0/1-vectors. The last entry in each row gives the score of the partition. Only partitions with each group containing at least 10

### Author(s)

Anja von Heydebreck and Wolfgang Huber http://www.molgen.mpg.de/~heydebre

### References

Identifying Splits with Clear Separation: A New Class Discovery Method for Gene Expression Data. Anja von Heydebreck, Wolfgang Huber, Annemarie Poustka, and Martin Vingron, ISMB 2001, Bioinformatics 17, Suppl. 1 (2001) p. 107-114.

### Examples

```
data(alizadeh)
cputime <- system.time(
                       y <- isis(alizadeh$x)
                         )
cat("Time used:", cputime[1], "sec on CPU,", cputime[3], "sec on wallclock.\n")
```

---

## thresh — *Simulated quantiles of t-statistic for ordered normal random vectors.*

---

### Description

For samples of at most 150 i.i.d standard normal random variables, the simulated 0.9999-quantiles of the two-sample t-statistic comparing the data values below and above any cutpoint is given.

### Usage

```
data(thresh)
```

### Format

A 150 x 150 matrix with the row index indicating the sample size and the column index indicating the cutpoint.

---

`tscore` *Score function for bipartitions*

---

### Description

Computes a score for a given bipartition of the set of samples that measures how well the two classes are separated by the expression levels of a suitable subset of genes.

### Usage

```
tscore(dat, split, p = 50, p.offs = 0)
```

### Arguments

| | |
|---|---|
| `dat` | The gene expression data matrix. Rows correspond to genes, and columns correspond to samples. The values in x may be normalized logarithms of intensities or log-ratios. |
| `split` | Either a binary vector encoding a bipartition of the set of samples, or a binary matrix with every row encoding a bipartition of the set of samples. |
| `p` | The number of best discriminating genes selected for computing the score of a bipartition. |
| `p.offs` | The number of very best discriminating genes that are ignored when computing the score of a partition. Thus only the genes with ranks p.offs+1, ..., p are used to compute the score. |

### Details

First, the genes with ranks p.offs+1, ..., p according to the two-sample t-statistic are selected. Then, a discriminant axis for a naive Bayes classifier (ML classification rule for two normal distributions with identical, diagonal covariance matrix) based on these genes is computed, and the score is computed as the t-statistic for the coordinates of the sample vectors projected onto this axis.

### Value

A vector with the scores of the bipartitions encoded in 'split'.

### Examples

```
data(alizadeh)
split <- rep(0, ncol(alizadeh$x))
split[which(alizadeh$type=="FL")] <- 1
z <- tscore(alizadeh$x, split)
```

`ttesttwo` *two-sample t-statistic*

## Description

Computes the two-sample t-statistic for each row of a data matrix, according to a bipartition of the set of columns.

## Usage

```
ttesttwo(dat, split)
```

## Arguments

| | |
|---|---|
| `dat` | A numeric data matrix. |
| `split` | A binary vector encoding a bipartition of the set of columns of 'dat'. |

## Value

A vector with the t-statistic values for each row of 'dat', according to the bipartition.

## Examples

```
data(alizadeh)
split <- rep(0, ncol(alizadeh$x))
split[which(alizadeh$type=="FL")] <- 1
t <- ttesttwo(alizadeh$x, split)
```