Rolemodel

October 14, 2013

ILP *Perform an integer linear programming*

Description

Maximum a posteriori (MAP) estimate via integer linear programming (ILP).

Usage

```
ILP(I, y, alpha, gamma, p)
```

Arguments

`I`	The incidence 0-1 matrix with unique row and column names, where rows are parts (genes) and columns are wholes (gene-sets).
`y`	Gene-level 0-1 data with the same names as the row names of I.
`alpha`	The false positive rate in role model, numeric value between 0 and 1. See reference.
`gamma`	The true positive rate in role model, numeric value between 0 and 1. See reference.
`p`	The prior active probability of wholes in role model, numeric value between 0 and 1. See reference.

Details

R package `Rglpk` is used to perform the integer linear programming. Generally, alpha and gamma can be estimated from the gene-level data by users themselves (see reference for examples), and alpha is less than gamma. p can be estimated via R package `MGSA` with alpha and gamma fixed. Since ILP is a complex problem in the optimization field, the running time might be very long. This function is invoked in `sequentialRM`.

Value

The output has the same structure as `Rglpk_solve_LP` in the `Rglpk` package, which is a list consisting of optimum, solution (in the order of wholes and parts) and status.

Author(s)

Zhishi Wang, Michael Newton and Subhrangshu Nandi.

References

Zhishi W., Qiuling H., Bret L. and Michael N.: A multi-functional analyzer uses parameter constaints to improve the efficiency of model-based gene-set analysis (2013).

See Also

`sequentialRM`

Examples

```
data(Type2D)
## Use 5 and 10 as the trimming parameters
newI <- subRM(I, 5, 10)
## the corresponding gene-level data
newy <- y[rownames(newI)]

## set the system parameters
alpha <- 0.00019
gamma <- 0.02279
p <- 0.00331
## perform the ILP
res <- ILP(newI, newy, alpha, gamma, p)
```

`optimalRM` *Extract wholes in the smaller incidence matrix*

Description

Extract wholes in the smaller incidence matrix, which might be active in the optimal solution with respect to the bigger incidence matrix in the sequential approach introduced in the reference.

Usage

```
optimalRM(xxbig, ysubbig, xx1, xx2, alpha, gamma, p)
```

Arguments

`xxbig`	The bigger incidence matrix in the sequential approach.
`ysubbig`	The corresponding gene-level data.
`xx1`	The smaller incidence matrix in the sequential approach.
`xx2`	The difference incidence matrix between xxbig and xx1.
`alpha`	The false positive rate in role model, numeric value between 0 and 1. See reference.

`gamma`	The true positive rate in role model, numeric value between 0 and 1. See reference.
`p`	The prior active probability of wholes in role model, numeric value between 0 and 1. See reference.

Details

In the sequential approach, we need to extract some other wholes in the smaller incidence matrix besides the active ones, which are already identified by the ILP calculation. See reference for details. This function will be invoked in `sequentialRM`.

Value

Return a logical vector with the length of the number of columns of xx1.

Author(s)

Zhishi Wang, Michael Newton, Subhrangshu Nandi

References

Zhishi W., Qiuling H., Bret L. and Michael N.: A multi-functional analyzer uses parameter constaints to improve the efficiency of model-based gene-set analysis (2013).

See Also

`sequentialRM`

`sequentialRM` *The sequential approach*

Description

Use the sequential approach introduced in the reference to speed up the running of integer linear programming (ILP).

Usage

```
sequentialRM(I, y, nupstart, by = 1, alpha, gamma, p)
```

Arguments

`I`	The incidence 0-1 matrix with unique row and column names, where rows are parts (genes) and columns are wholes (gene-sets).
`y`	Gene-level 0-1 data with the same names as the row names of I.
`nupstart`	The starting upper bound used in the sequential approach.

`by` The increment of the upper bound used in the sequential approach, default value 1.

`alpha` The false positive rate in role model, numeric value between 0 and 1. See reference.

`gamma` The true positive rate in role model, numeric value between 0 and 1. See reference.

`p` The prior active probability of wholes in role model, numeric value between 0 and 1. See reference.

Details

Generally, alpha and gamma can be estimated from the gene-level data by users themselves (see reference for examples), and alpha is less than gamma. p can be estimated via R package `MGSA` with alpha and gamma fixed.

We first perform the ILP on an initial incidence matrix (the smaller matrix) with lower bound equal lower bound of I and upper bound nupstart; then do another ILP, making use of the results obtained from the last ILP, on the bigger incidence matrix with upper bound equal nupstart + by. This process is repeated until the original incidence matrix I is reached. The suggested value for nupstart is 10. `sequentialRM` is our main function to perform the ILP calculation. `ILP`, `shrinkRM` and `optimalRM` are all invoked in this function.

Value

Return a list consisting of onwholes: the active wholes, i.e., the MFA-ILP estimate, and sol: has the same structure with the output of `ILP`.

Author(s)

Zhishi Wang, Michael Newton and Subhrangshu Nandi.

References

Zhishi W., Qiuling H., Bret L. and Michael N.: A multi-functional analyzer uses parameter constaints to improve the efficiency of model-based gene-set analysis (2013).

Examples

```
data(Type2D)
## set the system parameters
alpha <- 0.00019
gamma <- 0.02279
p <- 0.00331
## use the sequential approach to get the MAP estimate on the Type 2 diabetes example
res <- sequentialRM(I, y, nupstart = 10, by =1, alpha, gamma, p)
```

`shrinkRM` *Shrink the size of the incidence matrix*

Description

Shrink the size of the incidence matrix by making a prejudgement about which wholes (gene-sets) and parts (genes) have to be zeros in the optimal solution of ILP. See reference for details.

Usage

```
shrinkRM(I, y, alpha, gamma, p)
```

Arguments

`I`	The incidence 0-1 matrix with unique row and column names, where rows are parts (genes) and columns are wholes (gene-sets).
`y`	Gene-level 0-1 data with the same names as the row names of I.
`alpha`	The false positive rate in role model, numeric value between 0 and 1. See reference.
`gamma`	The true positive rate in role model, numeric value between 0 and 1. See reference.
`p`	The prior active probability of wholes in role model, numeric value between 0 and 1. See reference.

Details

Generally, alpha and gamma can be estimated from the gene-level data by users themselves (see reference for examples), and alpha is less than gamma. p can be estimated via R package `MGSA` with alpha and gamma fixed.

The amount of shrinkage may be dramatic, but it depends on the observed data y, the system I and system parameters alpha, gamma and p. When alpha is small and gamma is large the effects may be minimal. This function is invoked in `sequentialRM`.

Value

Return a list consisting of newI: the incidence matrix after shrinking, and newy: the corresponding part-level data.

Author(s)

Zhishi Wang, Michael Newton, Subhrangshu Nandi

References

Zhishi W., Qiuling H., Bret L. and Michael N.: A multi-functional analyzer uses parameter constaints to improve the efficiency of model-based gene-set analysis (2013).

See Also

`sequentialRM`

Examples

```
data(Type2D)
## set the system parameters
alpha <- 0.00019
gamma <- 0.02279
p <- 0.00331
## shrink the matrix
new <- shrinkRM(I, y, alpha, gamma, p)
```

`subRM` *Trim the incidence matrix*

Description

Trim the incidence matrix using two parameters n.low and n.up.

Usage

```
subRM(I, n.low, n.up)
```

Arguments

`I` The incidence 0-1 matrix with unique row and column names, where rows are parts (genes) and columns are wholes (gene-sets).

`n.low` The lower bound of the sums of columns in the incidence matrix.

`n.up` The upper bound of the sums of columns in the incidence matrix.

Details

Trim the size of the incidence matrix for the computation via integer linear programming (ILP). The suggested values are n.low = 5 and n.up = 50.

Value

Return the new incidence matrix after trimming.

Author(s)

Zhishi Wang, Michael Newton and Subhrangshu Nandi.

References

Zhishi W., Qiuling H., Bret L. and Michael N.: A multi-functional analyzer uses parameter constaints to improve the efficiency of model-based gene-set analysis (2013).

Examples

```
data(Type2D)
## Use 5 and 10 as the trimming parameters
newI <- subRM(I, 5, 10)
```

`Type2D` *Data set*

Description

The data set includes an incidence matrix I and the corresponding gene-level data y.

Arguments

`I`	The incidence 0-1 matrix with unique row and column names, where rows are genes and columns are gene-sets. The size of I is 10626 x 6037.
`y`	Gene-level 0-1 data with the same names as the row names of I.

Details

From a large-scale genome-wide association study (GWAS) involving more than 34,000 cases and 114,000 control subjects, 77 human genes have been implicated as affecting T2D disease susceptibility (see reference). To assess the functional content of this gene list, we extracted 6037 gene ontology terms, each annotating between 5 and 50 genes. These 6037 terms annotate a total of 10,626 genes; among the 77 T2D-associated genes, 58 are in this moderately annotated class.

References

Zhishi W., Qiuling H., Bret L. and Michael N.: A multi-functional analyzer uses parameter constaints to improve the efficiency of model-based gene-set analysis (2013).

Andrew P. M. and others: Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes (2012). Nature Genetics, Volume 44-9.

Examples

```
data(Type2D)

str(I)
str(y)
```

Index