

BUS Vignette

Yin Jin, Hesen Peng, Lei Wang, Christine Nardini

1 Introduction

GOAL: The BUS package allows the computation of two types of similarities (correlation [Sokal, 2003] and mutual information [Cover, 2001]) for two different goals: (i) identification of the similarity among the activity of molecules sampled across different experiments (we name this option Unsupervised, U), (ii) identification of the similarity between such molecules and other types of information (clinical, anagraphical, etc, we name this option supervised, S).

Unsupervised Option. The computation applies to data in tabular form (MxN) where rows represents different molecules (M), columns represents experiments or samples (N) and the content of the tables' cells the abundance of the molecule in the sample. Microarray experiments are the data of choice for this application, but the method can be applied to any data in the appropriate format (miRNA arrays, RNA-seq data, etc.). The results are in the form of an MxM adjacency matrix, where each cell represents the association computed among the corresponding molecules. This matrix has associated also a p-value matrix and a corrected p-value matrix (see below for details). Based on the cutoff selected, the adjacency matrix can be trimmed and lead to a predicted network of statistically significant interactions (**pred.network**). This output can be used as-is to represent a gene association network ([Margolin, 2004, Basso, 2005]), or can be further elaborated to cluster genes based on a shared degree of similarity (hence the Unsupervised label). Mutual information (from now on MI) is computed using the minet package [Meyer, 2008], all the options can be found in the corresponding vignette. Here argument `net.trim` decides which function (`mrnet/clr/aracne`) in MINET package is used to give the similarity based on mutual information matrix. Correlation is computed using the R built-in `cor` function.

Supervised Option. For the S option a second dataset is necessary, a TxN table, where T represents the number of external traits of interest. The result is an association MxT table where each cell indicates the association between the molecule and the external trait. This matrix has associated also a p-value matrix and a corrected p-value matrix (see below for details). As this can be used to associate samples to clinical classes we call this option Supervised (this type of approach was used in [Diehn, 2008]).

Statistical Significance. The package offers the possibility to evaluate the statistical significance of the computed similarity measures in two steps, a summary of the options is given in Table 1.

Option	<i>p</i> -value			
	single		multiple	
	ρ	<i>MI</i>	ρ	<i>MI</i>
S	Exact	<i>beta</i> distribution	MM-correction or permutations (3 options)	
U			MM-correction or permutations	

Table 1. Summary of the available options for statistical validation in BUS. ρ indicates correlation.

First, it allows the computation of the "single" p-value, i.e. the p-value relevant for the assessment of the statistical significance of the similarity of a given gene as if it was the only one tested.

For correlation this relies on the R built-in `cor.test` and it then computes the exact p-value.

For MI it is obtained from permutations and this method estimates the extreme p-values (close to 0) by fitting a beta distribution, whose analytical expression is obtained by the estimate of 2 shape parameters ($\hat{\alpha}$ and $\hat{\beta}$) using the method of the moments.

Second, the correction of the p-value for multiple hypothesis testing can be computed either based on permutations or on meta analysis (MM-correction, [Nardini,2008]). These approaches are the same for MI and correlation, however they offer 3 types of correction:

- S analysis option `method.permut = 1` corrects for all genes
- S analysis option `method.permut = 2` corrects for all traits
- S analysis option `method.permut = 3` corrects for all traits and genes

Missing Data Treatment. Data are pre-processed to cope with missing information (both in the MxN and in the TxN table) using (smooth) bootstrapping [Silverman, 1987].

The main function BUS has arguments for:

- the type of analysis (supervised/unsupervised)
- the distance metric (correlation/MI)
- the correction method for statistical significance on multiple hypothesis testing (permutation/MM-correction)

Expected computation times. The anticipated time for a 50*10 matrix is 3 minutes when running on an ordinary personal computer (with 256k memory).

The functions' dependencies scheme of the BUS package is illustrated below.

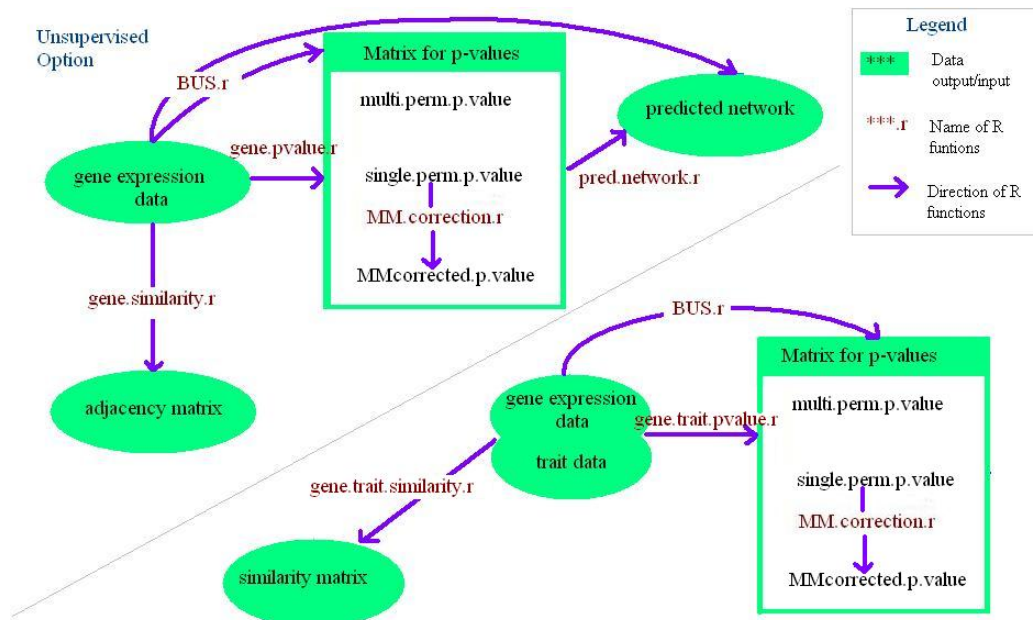


Figure 1. functions scheme

Functions Description

BUS: A wrapper function to compute (i) the similarity matrix (using correlation/MI as metric) and the single p-value matrix (each element is the p-value under the null hypothesis that the related row gene and column gene have no interaction), corrected p-values matrix (different levels of dependency are considered) and the predicted network matrix (predicted gene network, this output is effective for option U)

gene.similarity: Function for the computation of the adjacency matrix in the Unsupervised option (using correlation/MI as metric)

gene.trait.similarity: Function for the computation of the similarity matrix in the Supervised option (using correlation/MI as metric)

gene.pvalue: Function for the computation of the p-value matrix for the Unsupervised option. Single p-value (each element is the p-value under the null hypothesis that the related row gene and column gene have no interaction) is computed thanks to: (i) for MI the distribution identified by the p permutation values identified for each gene, with extreme p-values computed fitting a beta distribution; for correlation using the exact distribution provided by the built-in R cor function (single.perm.p.value). Corrected p-value is computed thanks to (i) the distribution identified by the p permutation values across all genes (multi.perm.p.value); (ii) the use of MM-correction (MMcorrected.p.value). When correlation is used as metric, only exact p-value and related MMcorrected.pvalue is output)

gene.trait.pvalue: Function for the computation of the p-value matrix for the Supervised option. Single p-value (each element is the p-value under the null hypothesis that the related row gene and column trait have no interaction) is computed thanks to: (i) for MI the distribution identified by the P permutation values identified for each gene, with extreme p-values computed fitting a beta distribution; for correlation using the exact distribution provided by the built-in R cor function (single.perm.p.value). Corrected p-value is computed thanks to (i) the distribution identified by the p permutation values across all genes (multi.perm.p.value); (ii) the distribution identified by the p permutation values across all traits; (iii) the distribution identified by the p permutation values across all genes and traits; (iv) the use of MM-correction (MMcorrected.p.value). P-values are calculated based on similarity for gene-gene interaction by calling function

pred.network: Function to predict the network from the selected corrected p-value matrix, only for the Unsupervised option.

MM.correction: Function to compute the MM-corrected p-values matrix from single p-value matrix using the MM-correction method, can be used as standalone function, on any set of p-values.

2 BUS Usage

```
> library(BUS)
> library(minet)
> data(copasi)
> mat = as.matrix(copasi)[1:5, ]
> BUS(EXP = mat, measure = "MI", method.correct = "both",
+     n.replica = 50, net.trim = "aracne", thresh = 0.05,
+     nflag = 1)
```

```
$single.perm.p.value
      [,1] [,2] [,3] [,4] [,5]
[1,] 0.00 0.56 0.52 0.54 0.20
[2,] 0.56 0.00 0.32 0.46 0.06
[3,] 0.52 0.32 0.00 0.36 0.16
[4,] 0.54 0.46 0.36 0.00 0.00
[5,] 0.20 0.06 0.16 0.00 0.00
```

```
$multi.perm.p.value
      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.00000000 0.41040462 0.4005525 0.4190751 0.03428571
[2,] 0.41040462 0.00000000 0.3629032 0.3792135 0.03611111
[3,] 0.40055249 0.36290323 0.0000000 0.3709677 0.16489362
[4,] 0.41907514 0.37921348 0.3709677 0.0000000 0.00000000
[5,] 0.03428571 0.03611111 0.1648936 0.0000000 0.00000000
```

```
$MMcorrected.p.value
      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.0000000 0.5600000 0.7172166 0.6640718 0.5454251
```

```
[2,] 0.5600000 0.0000000 0.5953991 0.6841103 0.1316550
[3,] 0.7172166 0.5953991 0.0000000 0.6319354 0.2521234
[4,] 0.6640718 0.6841103 0.6319354 0.0000000 0.0000000
[5,] 0.5454251 0.1316550 0.2521234 0.0000000 0.0000000
```

```
$net.pred.permut
      1      2 3 4      5
1 1.000000 0.000000 0 0 0.9152020
2 0.000000 1.000000 0 0 0.8608588
3 0.000000 0.000000 1 0 0.0000000
4 0.000000 0.000000 0 1 1.0000000
5 0.915202 0.8608588 0 1 1.0000000
```

```
$net.pred.MMcorrected
      1 2 3 4 5
1 1 0 0 0 0
2 0 1 0 0 0
3 0 0 1 0 0
4 0 0 0 1 1
5 0 0 0 1 1
```

The arguments to the BUS function here are

- a matrix (here **EXP**) for gene expression data.
- metric used to calculate similarity (here **measure**). There are two choices, MI and corr. We use MI here, applying the MINET package to output the similarity matrix with option of mrnet.
- a flag to indicate which method is used to correct permutation p-values (here **method.permut**). Here a default value (2) is used.
- method chosen (here **method.correct**) to calculate p-value: "MMcorrection" for MM-correction, "permutation" for corrected permutation, "both" for both MM-correction and permutation methods. Here both p-value matrices are output.
- number of permutations (here **n.replica**): default value is 400, for optimal precision in p-value computation.
- method chosen (here **net.trim**) to trim the network. Here aracne method is applied, where the least significant edge in each triplet is removed.
- threshold (here **threshold**), according to which significant association between genes are selected to construct the predicted network. This option is actually used in function **pred.network** for predicted network from p-value matrix.
- a flag for the type of analysis (here **nflag**). If Supervised nflag=2, if Unsupervised nflag=1. Here an Unsupervised option is considered.

The dataset copasi is taken from Copasi2 (Complex Pathway Simulator), a software for simulation and analysis of biochemical networks. The system generates random artificial gene networks according to well-defined topological and kinetic properties. These are used to run in silico experiments simulating real laboratory micro-array experiments. Noise with controlled properties is added to the simulation results several times emulating measurement replicates, before expression ratios are calculated. This series consists of 150 artificial gene networks. Each network consists of 100 genes with a total of 200 gene interactions (on average each gene has 2 modulators). All networks are composed of genes with similar kinetics, the only difference between networks is how the gene interactions are organized (i.e. which genes induce and repress which other genes). The networks belong to three major groups according to their topologies: RND stands for randomized network, SF for scale-free (many edges among few nodes) and SW for small world (edges exist between adjacent nodes). The data is given in the package

is an RND data. Actually, only first of five rows in the gene expression data is used to calculate to save the space here.

Explain the result:

- `single.perm.p.value`: the single p-value matrix, i.e. the p-value matrix obtained by the simple permutation method. We can see it is a 5*5 matrix here as we only use data for 5 genes.
- `multi.perm.p.value`: the corrected permutation p-value matrix, i.e. the p-value matrix obtained via corrected permutation method.
- `MMcorrected.p.value`: the MMcorrection p-value matrix, i.e. the p-value matrix obtained via MM-correction method.
- `net.pred.MMcorrected`: the network predicted based on the MM-correction p-value matrix. This network is based on MMcorrected p-values.
- `net.pred.permut`: the network predicted based on the corrected permutation p-value matrix. This network is based on multi-hypothesis-corrected p-values.

This is an Unsupervised case. We could see that a lower values in `single.perm.p.value/multi.perm.p.value` or a higher values in `net.pred.MMcorrected/net.pred.permut` indicate a strong link between the row and column genes. The value 0 in the p-value matrix or 1 in network matrix respectively infers a strong link.

```
> library(BUS)
> data(copasi)
> mat = as.matrix(copasi)[1:5, ]
> gt <- matrix(rnorm(200), 2, 100)
> BUS(EXP = mat, trait = gt, measure = "corr", method.correct = "both",
+      nflag = 2)
```

```
$single.perm.p.value
      [,1]      [,2]
[1,] 0.52628450 0.5447687
[2,] 0.92666021 0.2470282
[3,] 0.07283963 0.8387745
[4,] 0.95827370 0.9006039
[5,] 0.33162235 0.5461911
```

```
$multi.perm.p.value
NULL
```

```
$MMcorrected.p.value
      [,1]      [,2]
[1,] 0.00000000 0.9352725
[2,] 0.9934791 0.0000000
[3,] 0.6253483 0.9672966
[4,] 0.9582737 0.9901586
[5,] 0.8868599 0.9395435
```

Here is a Supervised case, we use copasi data as gene expression data and randomly generate for the trait data. It is shown that only p-value matrices are output in this case. And we apply the correlation metric this time; thus multi.p.value will not be output.

References

[Sokal, 2003] R.R.Sokal and F.J.Rohlf. *Biometry*. Freeman, New York, 2003.

- [Cover, 2001] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, 2001.
- [Margolin, 2004] Adam A. Margolin, Ilya Nemenman, Katia Basso, Ulf Klein, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, 2004.
- [Basso, 2005] K Basso, A A Margolin, G Stolovitzky, U Klein, R Dalla-Favera, and A Califano. Reverse engineering of regulatory networks in human b cells. *Nat Genet*, 37(4):382–390, Apr 2005.
- [Meyer, 2008] P E Meyer, F Lafitte, and G Bontempi. minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9:461–461, 2008.
- [Diehn, 2008] M. Diehn, C. Nardini, D. S. Wang, S. McGovern, M. Jayaraman, Y. Liang, K. Aldape, S. Cha, and M. D. Kuo. Identification of non-invasive imaging surrogates for brain tumor gene expression modules. *Proc. Natl. Acad. Sci.*, 105(13):5213–5218, 2008.
- [Silverman, 1987] B. W. Silverman and G. A. Young. The bootstrap: To smooth or not to smooth? *Biometrika*, 74(3):469–479, 1987.
- [Nardini,2008] C. Nardini, L. Wang,H. Peng, L. Benini,and M.D. Kuo. MM-Correction: Meta-analysis-Based Multiple Hypotheses Correction in Omic Studies. *Springer CCIS*, 25:242-255, 2008.