BUS Vignette

Yin Jin, Hesen Peng, Lei Wang, Raffaele Fronza, Yuanhua Liu and 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. Mutual information is computed according to the empirical method proposed in MINET package. It is implemented with a external c function. 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	p-value		
	single		multiple
	ρ	MI	MI
S	Exact	*beta* distribution	permutations (3 options)
U			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, for the p-value of MI, correction for multiple hypothesis testing is computed based on permutations. 3 types of corrections are offered:

- S analysis option `method.permut` = 1 correction for multiple traits tested
- S analysis option `method.permut` = 2 correction for multiple genes tested
- S analysis option `method.permut` = 3 correction for both 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 types for statistical significance on multiple hypothesis testing based on permutations (genes, traits or both)

Expected computation times. In the unsupervised case, the anticipated time for a 50*12 matrix (gene expression data) is 30 seconds when running on an ordinary personal computer (with 1G memory). While in the supervised case, with 50*12 gene expression data and trait data involved, it is 2 minute when correction of both genes and traits is considered.

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

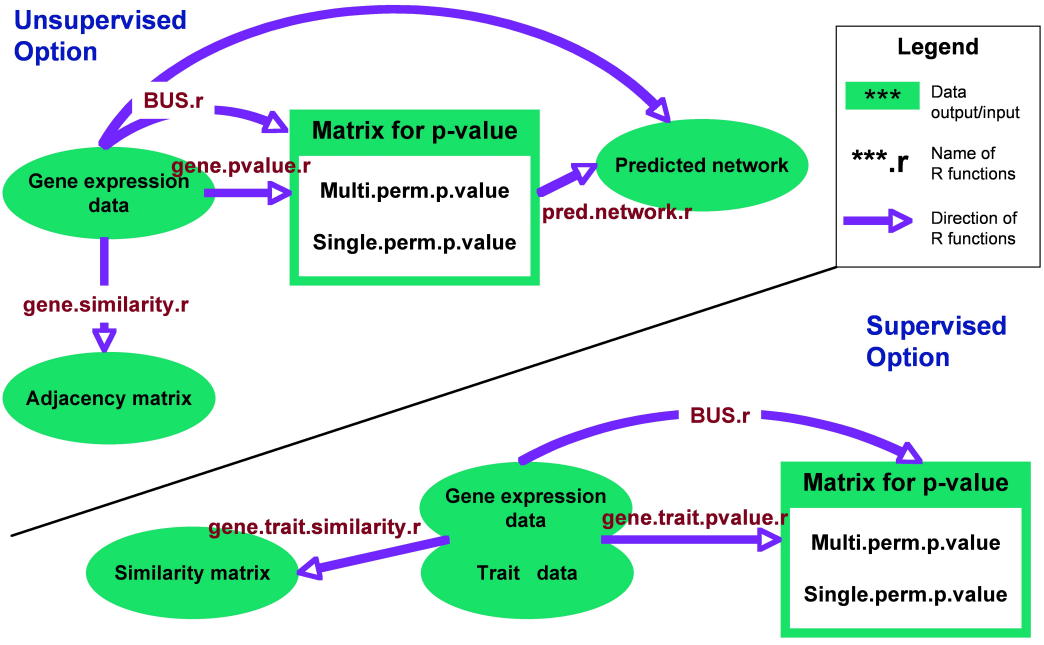

Figure 1. functions scheme

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 the distribution identified by the p permutation values across all genes (multi.perm.p.value). When correlation is used as matric, only exact p-value 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 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.

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

2 BUS Usage

```
> library(BUS)
> library(minet)
> data(copasi)
> mat = as.matrix(copasi)[1:5, ]
> rownames(mat) <- paste("G", 1:nrow(mat), sep = "")
> BUS(EXP = mat, measure = "MI", n.replica = 400,
+     net.trim = "aracne", thresh = 0.05, nflag = 1)

$similarity
          G1        G2        G3 G4        G5
G1 1.0000000 0.4682614 0.5126417  1 0.0000000
G2 0.4682614 1.0000000 0.0000000  0 0.7451189
G3 0.5126417 0.0000000 1.0000000  0 0.6168879
G4 1.0000000 0.0000000 0.0000000  1 0.0000000
G5 0.0000000 0.7451189 0.6168879  0 1.0000000

$single.perm.p.value
       G1     G2     G3     G4     G5
G1 0.0000 0.2225 0.2200 0.0000 0.4125
G2 0.2225 0.0000 0.4750 0.4225 0.1375
G3 0.2200 0.4750 0.0000 0.4625 0.1775
G4 0.0000 0.4225 0.4625 0.0000 0.4800
G5 0.4125 0.1375 0.1775 0.4800 0.0000

$multi.perm.p.value
          G1         G2         G3        G4         G5
G1 0.0000000 0.15249570 0.12993763 0.0000000 0.38828860
```

```
G2 0.1524957 0.00000000 0.40227508 0.3961938 0.05027739
G3 0.1299376 0.40227508 0.00000000 0.4026472 0.08656195
G4 0.0000000 0.39619377 0.40264716 0.0000000 0.38787666
G5 0.3882886 0.05027739 0.08656195 0.3878767 0.00000000

$net.pred.permut
   G1 G2 G3 G4 G5
G1  1  0  0  1  0
G2  0  1  0  0  0
G3  0  0  1  0  0
G4  1  0  0  1  0
G5  0  0  0  0  1
```

The arguments to the `BUS` function here are

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

The copasi dataset 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 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 results:

- `similarity`: the matrix for mutual information.
- `single.perm.p.value`: the single p-value matrix, i.e. the p-value matrix obtained by the simple purmutation 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.

- `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.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.

```
> data(tumors.mRNA)
> exp <- as.matrix(tumors.mRNA)[11:15, ]
> rownames(exp) <- rownames(tumors.mRNA)[11:15]
> data(tumors.miRNA)
> trait <- as.matrix(tumors.miRNA)[11:15, ]
> rownames(trait) <- rownames(tumors.miRNA)[11:15]
> BUS(EXP = exp, trait = trait, measure = "MI",
+     nflag = 2)

$similarity
            hsa-mir-132 hsa-mir-133a hsa-mir-135a
200017_at     0.6000000    0.4754888    0.4754888
200018_at     0.2000000    0.5509775    0.4754888
200022_at     0.4754888    0.5509775    0.0000000
200023_s_at   0.0000000    1.0000000    0.0000000
200024_at     0.4754888    0.5509775    0.0000000
            hsa-mir-135b hsa-mir-139
200017_at      0.4754888   1.0000000
200018_at      0.4754888   0.4754888
200022_at      0.0000000   0.4754888
200023_s_at    0.0000000   0.4754888
200024_at      0.4754888   0.4754888

$single.perm.p.value
            hsa-mir-132 hsa-mir-133a hsa-mir-135a
200017_at        0.2050        0.360       0.3375
200018_at        0.6150        0.305       0.4000
200022_at        0.3450        0.300       0.7250
200023_s_at      0.7050        0.000       0.6975
200024_at        0.3825        0.310       0.6900
            hsa-mir-135b hsa-mir-139
200017_at         0.3675      0.0000
200018_at         0.3575      0.3850
200022_at         0.7225      0.3700
200023_s_at       0.6725      0.3825
200024_at         0.3675      0.3650

$multi.perm.p.value
            hsa-mir-132 hsa-mir-133a hsa-mir-135a
200017_at        0.2220       0.3595        0.370
200018_at        0.6190       0.3150        0.370
200022_at        0.3515       0.3150        0.702
200023_s_at      0.6950       0.0000        0.702
200024_at        0.3515       0.3150        0.702
            hsa-mir-135b hsa-mir-139
200017_at          0.360       0.000
200018_at          0.360       0.373
```

```
200022_at          0.716       0.373
200023_s_at        0.716       0.373
200024_at          0.360       0.373
```

Here is a Supervised case, we use the tumor dataset from [Liu, 2007], the mRNA data as gene expression data and miRNA data as trait data. Gene expression data were obtained by microarray from human brain tumors, while miRNA data were obtained by RT-PCR. 12 brain tumors at different levels are analyzed for both mRNA and miRNA levels to study the correlation of any mRNA-miRNA pairs. Outputs are similar like that in the unsupervised case except the predicted network.

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] A. A. Margolin, I. Nemenman, K. Basso, U. Klein, C. Wiggins, G. Stolovitzky, R. Dalla. Favera, and A. 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.

[Liu, 2007] T. Liu, T. Papagiannakopoulos, K. Puskar, S. Qi, F. Santiago, W. Clay, K. Lao, Y. Lee, S. F. Nelson, H. I. Kornblum, F. Doyle, L. Petzold, B. Shraiman, and K. S. Kosik. Detection of a microRNA signal in an in vivo expression set of mRNAs. *Plos One*, 2(8): e804, 2007.