**---------------------------------------------------------------------------**

**ReadMe File**

**---------------------------------------------------------------------------**

**Integration of MicroRNA, mRNA, and Protein Expression Data   
for the Identification of Cancer-Related MicroRNAs.**

**Jiyoun Seo, Daeyong Jin, Chan-Hun Choi, Hyunju Lee**

Data Mining & Computational Biology Lab

Gwangju Institute of Science & Technology

Gwangju, South Korea

http://combio.gist.ac.kr/

**---------------------------------------------------------------------------**

**I. List of input files: “Input\_files.zip”**

**1. "Normalization\_input\_files"**

**2. "Ranking\_input\_files"**

**3. "Module\_input\_files"**

**II. List of Source Codes: “Source\_code.zip”**

**1. "Normalization\_source\_codes"**

**2. "Ranking\_source\_codes"**

**3. "Module\_source\_codes"**

**---------------------------------------------------------------------------**

**Description**

**----------------------------------------------------------------------------------**

**Ⅰ. Normalizing the files**

**Source codes**: “TN\_normalization.r”

**Input and output example files**: “TN\_normalization”

1. “TN\_normalization.r”

: Normalize the expression files according to the comparison between tumor and normal samples. Note that the values in expression files are already log-based transformed, therefore, the ratio of each expression value of tumor sample and mean expression of normal samples can be computed by subtraction. Protein expression data do not require this step since they are already normalized.

(input)

a) “GBM\_mRNA\_tumor.txt” and “GBM\_mRNA\_normal.txt”: mRNA expression data from tumor/normal samples

b) “GBM\_miRNA\_tumor.txt” and “GBM\_miRNA\_normal.txt”: miRNA expression data from tumor/normal samples

(output)

a) “GBM\_mRNA\_normalized.txt”: Normalized mRNA expression data according to the ratio between the values from tumor samples and the averaged one from normal samples

b) “GBM\_miRNA\_normalized.txt”: Normalized miRNA expression data according to the ratio between the values from tumor samples and the averaged one from normal samples

2. T-test and selection (codes are not required)

: Select the mRNAs or miRNAs having significant difference (p-value<0.05) between the tumor and normal samples that can be investigated by t-test, using the normalized expression data. We also did Bonferroni correction with multiplying the number of the samples to original p-value and the threshold for the selection, lower than 0.05, was applied to these corrected p-values.

**Ⅱ. Constructing the ranking lists**

**Source codes**: “Ranking\_source\_codes”

**Input and output example files**: “Ranking\_input\_files”

1. “1\_Protein\_name\_convert.r”

: Change the protein names into gene symbols before selecting the common genes.

(input)

a) “GBM\_protein\_list.csv”: The list of every protein in protein expression data.

b) “protein\_name\_convert\_table.txt”: A table of presenting protein names with their antibody names (1st and 2nd column) that were used for profiling and gene symbols (3rd column) to convert.

(output)

a) “GBM\_protein\_name\_changed\_list.txt”: Protein names are converted to gene symbols while the ones that are not included in the converting table are marked with “NA”.

2. “2\_Selecting\_genes.r”

: Select the common genes from both expression data of proteins and mRNAs.

(input)

a) “GBM\_genelist.txt”: The list of every mRNAs in mRNA expression data.

b) “GBM\_mRNA\_tumor.txt”: Expression matrix of every mRNA. Each row presents each mRNA while each column indicates each sample.

c) “GBM\_protein\_name\_changed\_list.txt”: Protein list with gene symbols.

d) “GBM\_protein\_tumor.txt”: Expression matrix of every protein. Each row presents each protein while each column indicates each sample.

(output)

a) “GBM\_mRNA\_exp\_select.txt”: mRNA expression matrix of common genes.

b) “GBM\_gene\_list\_select.txt”: Common gene list.

c) “GBM\_protein\_exp\_select.txt”: Protein expression matrix of common genes.

3. “3\_SCC\_and\_normalization.r”

: Calculate SCC values between mRNA-miRNA and protein-miRNA and normalize each matrix.

(input)

a) “GBM\_mRNA\_exp\_select.txt”: mRNA expression matrix of common genes. Each row presents each mRNA while each column indicates each sample.

b) “GBM\_miRNA\_tumor.txt”: Expression matrix of miRNAs. Each row presents each miRNA while each column indicates each sample.

c) “GBM\_protein\_exp\_select.txt”: Protein expression matrix of common genes. Each row presents each protein while each column indicates each sample.

(output)

a) “GBM\_mRNA\_miRNA\_norm\_mat.txt”: Normalized mRNA-miRNA correlation matrix. Each row presents each mRNA while each column indicates each miRNA.

b) “GBM\_protein\_miRNA\_norm\_mat.txt”: Normalized protein-miRNA correlation matrix. Each row presents each protein while each column indicates each miRNA.

4. “4\_Ranking.r”

: Rank the miRNAs according to their significance score; significance score for each miRNA is calculated by the number of paired genes which their absolute correlation values from both matrices (mRNA-miRNA, protein-miRNA) satisfy the conditions (thresholds; 1% or 5%). Note that output file also presents whether each miRNA is included in HMDD or not.

(input)

a) “GBM\_mRNA\_miRNA\_norm\_mat.txt”: Normalized mRNA-miRNA correlation matrix. Each row presents each mRNA while each column indicates each miRNA.

b) “GBM\_protein\_miRNA\_norm\_mat.txt”: Normalized protein-miRNA correlation matrix. Each row presents each protein while each column indicates each miRNA.

c) “GBM\_miRNA\_class.txt”: List of miRNAs and the second column shows whether each miRNA is included in HMDD for GBM related (“O”) or included in HMDD but not related to GBM (“N”) or not included in HMDD (“X”).

(output)

a) “GBM\_ranklist\_norm\_abs1.txt” or “GBM\_ranklist\_norm\_abs5.txt”: The list of ranked miRNAs. Second column presents whether each miRNA is included in HMDD or not while the third column shows the significance score.

**Ⅲ. List of source codes to construct the modules**

**Source codes**: “Module\_source\_codes”

**Input and output example files**: “Module\_input\_files”

1. “1\_Z\_normalization.r”

: Normalize the mRNA expression values with Z-normalization method.

(input)

a) “GBM\_mRNA\_normal.txt”: Expression matrix of mRNAs of normal samples. Each row presents each mRNA while each column indicates each sample.

b) “GBM\_mRNA\_tumor.txt”: Expression matrix of mRNAs of tumor samples. Each row presents each mRNA while each column indicates each sample.

(output)

a) “GBM\_mRNA\_normalized\_Z.txt”: Z-normalized mRNA expression matrix.

2. Bi-clustering (codes are not required)

: Bi-cluster the mRNAs and samples based on the idea of bipartite graph, using SAMBA.

(input)

a) “GBM\_mRNA\_normalized\_Z\_samba.txt”: Each row presents each mRNA while each column indicates each sample.

(output)

a) “GBM\_mRNA\_samba\_result.txt”: Results from SAMBA bi-clustering. First column presents the number of clusters while the second column distinguishes whether the one in the third column is mRNA (“1”) or sample (“0”). The third column shows the name of each mRNA or sample.

3. “2\_Correlation\_Permutation\_test.r”

: To select the clusters (modules) that are having significant correlation among them, we constructed the random clusters to compare and through permutation tests for 1000 times, the p-values were calculated. Note that manual step is also needed to select the clusters having p-value less than 0.05.

(input)

a) “GBM\_biclustering\_mRNA.csv”: Clustered mRNAs are presented. Each row presents the mRNAs contained in each module which indicates the number of elements for each row is the number of each module.

b) “GBM\_biclustering\_sample.csv”: Clustered samples are presented. Each row presents the samples contained in each module.

c) “GBM\_genelist.txt”: The list of every mRNAs in mRNA expression data.

d) “GBM\_mRNA\_normalized.txt”: Normalized mRNA expression data according to the ratio between the values from tumor samples and the averaged one from normal samples

(output)

a) “GBM\_cluster\_p\_value.txt”: P-values from the permutation tests for each module.

4. “3\_Expanding.r”

: Expanding the modules including new mRNAs that are previously known, from PPI information, to correlate significantly with the mRNAs in the modules.

(input)

a) “GBM\_biclustering\_mRNA.csv”: Clustered mRNAs are presented. Note that the modules used here are the selected ones from the previous step.

b) “GBM\_biclustering\_sample.csv”: Clustered samples are presented. Note that the modules used here are the selected ones from the previous step.

c) “GBM\_genelist.txt”: The list of every mRNAs in mRNA expression data.

d) “GBM\_mRNA\_normalized.txt”: Normalized mRNA expression data.

e) “PPI.txt”: Information of presenting gene-gene interaction. Each row with two genes are the pairs that are known to be correlated each other.

(output)

a) “changedinfo\_GBM\_clusters.txt”: This table shows the changed numbers of the elements in each module. Each row, divided by blank, shows the change of numbers from the first one to the second one.

b) “GBM\_new\_biclustering\_gene.csv”: The clusters with genes from PPI information

5. “4\_Protein\_inclusion”

5-1. “1\_Compute\_SCC\_Value.r”

: Construct a correlation matrix between mRNA and protein expression data.

(input)

a) “GBM\_mRNA\_tumor.txt”: Normalized mRNA expression data.

b) “GBM\_protein\_tumor.txt”: Protein expression data.

(output)

a) “GBM\_protein\_mRNA\_mat.txt”: mRNA-protein correlation matrix. Each row presents each mRNA while each column indicates each protein.

5-2. “2\_Construction\_of\_gene-protein\_Modules.r”

: Construct the modules with mRNAs and proteins. Note that since the input data presented here are just randomly made up files and therefore, the thresholds for the construction of example modules in this code are also modified. Please check the comments.

(input)

a) “GBM\_mRNA\_tumor.txt”: Normalized mRNA expression data.

b) “GBM\_protein\_tumor.txt”: Protein expression data.

c) “GBM\_genelist.txt”: The list of every mRNAs in mRNA expression data.

d) “GBM\_protein\_list.txt”: The list of every protein in mRNA expression data.

e) “GBM\_new\_biclustering\_gene.csv”: The modules of mRNAs and included genes from PPI.

f) “GBM\_protein\_mRNA\_mat.txt”: Correlation matrix of proteins and mRNAs.

(output)

a) “biclustering\_mRNA\_3module.csv”: The mRNAs included in the modules of mRNAs and proteins.

b) “biclustering\_protein\_3module.csv”: The proteins included in the modules of mRNAs and proteins.

6. “5\_miRNA\_inclusion”

6-1. “1\_Compute\_SCC\_Value\_with\_mirna.r”

: Construct a correlation matrix between mRNA and protein expression data.

(input)

a) “GBM\_mRNA\_tumor.txt”: Normalized mRNA expression data.

b) “GBM\_protein\_tumor.txt”: Protein expression data.

c) “GBM\_miRNA\_tumor.txt”: Normalized miRNA expression data.

(output)

a) “GBM\_mirna\_protein\_mRNA\_mat.txt”: mRNA-protein, mRNA correlation matrix. Each row presents each mRNA and protein while each column indicates each miRNA

6-2. “2\_ 2\_Construction\_of\_mRNA-protein-mirna\_Modules.r”

: Construct the modules with mRNAs and proteins and miRNAs. Note that since the input data presented here are just randomly made up files and therefore, the thresholds for the construction of example modules in this code are also modified. Please check the comments.

(input)

a) “GBM\_mRNA\_tumor.txt”: Normalized mRNA expression data.

b) “GBM\_protein\_tumor.txt”: Protein expression data.

c) “GBM\_genelist.txt”: The list of every mRNAs in mRNA expression data.

d) “GBM\_protein\_list.txt”: The list of every protein in protein expression data.

e) “biclustering\_mRNA\_3module.csv”: The mRNAs in the modules of mRNAs and proteins.

f) “biclustering\_protein\_3module.csv”: The proteins included in the modules of mRNAs and proteins.

g) “GBM\_miRNA\_protein\_mRNA\_mat.txt”: Correlation matrix of proteins, mRNAs and miRNAs.

h) “GBM\_miRNA\_tumor.txt”: Normalized miRNA expression data.

i) “GBM\_miRNA\_list.txt”: The list of every mRNAs in miRNA expression data.

(output)

a) “biclustering\_mRNA\_3module\_withmirna.csv”: The mRNAs in the modules of mRNAs, proteins and miRNAs.

b) “biclustering\_protein\_3module\_withmirna.csv”: The proteins in the modules of mRNAs, proteins and miRNAs.

c) “biclustering\_mRNA\_protein\_3module.csv”: Added version of showing both mRNAs and proteins in the modules.

d) “biclustering\_mirna\_3module\_withmirna.csv”: The miRNAs in the modules of mRNAs, proteins and miRNAs.