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

**ReadMe File**

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

**A Computational Approach to Identifying Gene-microRNA**

**Modules in Cancer**

**Daeyong Jin, Hyunju Lee**

Data Mining & Computational Biology Lab

Gwangju Institute of Science & Technology

Gwangju, South Korea

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

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

**List of source codes and input/output files: “GMC.zip”**

**1. "./1\_ Identifying\_DEG”**

Identify differentially expressed genes (DEG) using t-test.

**2. "./2\_Normalizing\_expression\_data” :**

Normalize gene and miRNA expression data.

**3. “./3\_Constructing\_GSM”**

Perform biclustering for DEG to construct gene-sample modules.

**4. “./4\_Correlation\_permutation\_test”**

Perform correlation permutation test for SAMBA bi-clusters.

**5. “./5\_Module\_gene\_expansion”**

Expand gene-sample modules.

**6. “./6\_Constructing\_GMM”**

Construct gene-miRNA modules using Bayesian network based on SCC.

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

**Description**

**1. Identifying differentially expressed genes using t-test**

**Directory**

**“./1\_Identifying\_DEG”** (codes are not required)

**Source codes**: codes are not required

**Inputs**

a) “./data/tumor\_gene\_exp.txt” b) “./data/normal\_gene\_exp.txt”

gene expression data from tumor/normal samples

**Output**

a) “./result/diff\_gene\_list.txt:

List of differentially expressed genes

**Description**

Select the genes 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-value.

**2. Normalizing differentially expressed genes using t-test**

**Directory**

**“./** **2\_Normalizing\_expression\_data”**

**1) Log2 (tumor/normal) normalization**

**Source codes**

“./src/1\_TN\_normalization.r”

**Inputs**

a) “./data/tumor\_gene\_exp.txt” b) “./data/normal\_gene\_exp.txt”

gene expression data from tumor/normal samples

a) “./data/tumor\_mirna\_exp.txt” b) “./data/normal\_mirna\_exp.txt”

miRNA expression data from tumor/normal samples

**Output**

a) “TN\_normalized\_gene\_exp.txt”:

Normalized gene expression data according to the ratio between the values from tumor samples and the averaged one from normal samples

b) “TN\_normalized\_miRNA\_exp.txt”:

Normalized miRNA expression data according to the ratio between the values from tumor samples and the averaged one from normal samples

**Description**

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.

**2) Z-score Normalization**

**Source codes**

“./src/2\_Z\_normalization.R”

**input**

a) “./data/TN\_normalized\_gene\_exp.txt”

TN normalized expression matrix of genes. Each row presents each gene while each column indicates each sample.

b) “./data/TN\_normalized\_miRNA\_exp.txt”

TN normalized expression matrix of miRNAs. Each row presents each miRNA while each column indicates each sample.

**output**

a) “Z\_TN\_normalized\_gene\_exp.txt”

Z-normalized gene expression matrix.

b) “Z\_TN\_normalized\_miRNA\_exp.txt”

Z-normalized miRNA expression matrix

**Description**

Normalize the gene expression values with Z-normalization method.

**3. Construct gene-sample modules using SAMBA biclustering using DEG**

**Directory**

**"./3\_SAMBA\_biclustering”**

**1) Perform SAMBA biclustering** (codes are not required)

**Input**

a) “./data/SAMBA\_Z\_TN\_normalized\_gene\_exp.txt”

Each row presents each gene while each column indicates each sample and first row is list of samples, first column is a list of genes, and second column is empty.

.

**Output**

a) “./result/bi\_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 gene (“1”) or sample (“0”). The third column shows the name of each gene or sample.

**Description**

Normalize the gene expression values with Z-normalization method.

**2) Convert result into gene and samples files**

**Source codes**: “./src/1\_Convert\_Bicluster\_To\_Gene\_Sample.R”

**Input**

a) “./result/bi\_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 gene (“1”) or sample (“0”). The third column shows the name of each gene or sample.

**Output**

a) “/result/biclustering\_gene.csv”

Clustered genes are presented. Note that the modules used here are the selected ones from the previous step.

b) “./result/biclustering\_sample.csv”

Clustered samples are presented. Note that the modules used here are the selected ones from the previous step.

**Description**

Convert results from SAMBA bi-clustering into two files (clustered genes and clustered samples)

**4. Correlation permutation test**

**Directory**

**“./4\_correlation\_permutation\_test”**

**1) Perform SAMBA biclustering** (codes are not required)

**Input**

a) “./data/biclustering\_gene.csv”

Clustered genes are presented. Each row presents the genes contained in each module which indicates the number of elements for each row is the number of each module.

b) “./data/biclustering\_sample.csv”

Clustered samples are presented. Each row presents the samples contained in each module.

c) “./data/gene\_list.txt”

The list of every genes in gene expression data.

d) “./data/TN\_normalized\_gene\_exp.txt”

Normalized gene expression data according to the ratio between the values from tumor samples and the averaged one from normal samples

.

**Output**

a) “./result/cluster\_p\_value.txt”

P-values from the permutation tests for each module.

**Description**

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 (or q-value) less than 0.05.

**5. Module gene expansion**

**Directory**

**“./** **5\_Module\_gene\_expansion”**

**Source codes**

“./src/1\_module\_gene\_expansion.r”

**Input**

a) “./data/f\_biclustering\_gene.csv”

Clustered genes after correlation permutation test are presented. Each row presents the genes contained in each module which indicates the number of elements for each row is the number of each module.

b) “./data/f\_biclustering\_sample.csv”

Clustered samples after correlation permutation test are presented. Each row presents the samples contained in each module.

c) “./data/gene\_list.txt”

The list of every genes in gene expression data.

d) “./data/TN\_normalized\_gene\_exp.txt”

Normalized gene expression data according to the ratio between the values from tumor samples and the averaged one from normal samples

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

.

**Output**

a) “./data/changedinfo\_modules.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) “./data/expanded\_biclustering\_gene.csv”

The clusters with genes from PPI information

**Description**

Expanding the modules including new genes that are previously known, from GGI information, to correlate significantly with the genes in the modules.

**6. Constructing GMM**

**Directory**

**“./** **6\_Constructing\_GMM”**

**6-1 Compute gene-miRNA correlation matrix**

**Source codes**

“./src/1\_compute\_SCC\_Value\_with\_miRNA.r”

**Input**

a) “TN\_normalized\_gene\_exp.txt”

TN-normalized gene expression data.

b) “TN\_normalized\_miRNA\_exp.txt”:

TN-normalized miRNA expression data.

**Output**

a) “/result/gene\_miRNA\_SCC\_mat.txt”

gene-miRNA correlation matrix (each row presents each gene while each column indicates each miRNA)

**Description**

Construct a correlation matrix (Spearman's rank correlation coefficient) between gene and miRNA expression data.

**6-2 Compute gene-miRNA correlation matrix**

**Source codes**

“./src/2\_Constructing\_gene-miRNA\_Modules.r”

**Input**

a) “./data/TN\_normalized\_gene\_exp.txt”

TN-normalized gene expression data.

b) “./data/TN\_normalized\_miRNA\_exp.txt”:

TN-normalized miRNA expression data.

c) “./data/gene\_list.txt”

The list of every genes in gene expression data.

d) “./data/mirna\_list.txt”

The list of every miRNA in miRNA expression data.

e) “./data/expanded\_biclustering\_gene.csv”

Clustered genes are presented. Each row presents the genes contained in each module which indicates the number of elements for each row is the number of each module.

**Output**

a) “final\_biclustering\_gene.csv”:

The genes in the modules of genes and miRNAs.

b) “final\_biclustering\_miRNA.csv”:

The miRNAs in the modules of genes and miRNAs.

c) “final\_biclustering\_sample.csv”:

The samples in the modules of genes and miRNAs.

**Description**

Construct the modules with genes 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