20

Microarray Data Mining for Biological Pathway Analysis

Miyoung Shin and Jaeyoung Kim
School of Electrical Engineering and Computer Science, Kyungpook National University, Korea

1. Introduction

In recent years, microarray gene expression studies have been actively pursued for extracting significant biological knowledge hidden under a large volume of gene expression profiles accumulated by DNA microarray experiments. Particularly great attentions have been paid to a variety of data mining schemes for gene function discovery [Eisen et al., 1998], disease diagnosis [Saiki et al., 2008], pathway analysis [Werner et al., 2008], pharmaceutical target identification [Corn et al., 2007], and etc. Out of these, the pathway analysis is one of the most significant problems in current bioinformatics researches. Pathway analysis concerns about identifying significant pathways, which are the groups of genes actively involved in some biological processes, based on the gene expression profiles. By doing so, our objective is to understand the role of such biological processes in a given experiment condition and their associated gene activities. In this chapter, we investigate several computational techniques that are often used in a variety of contexts for pathway analysis. Specifically, we first give a brief overview of microarray gene expression profile data and some biological resources available to be used for pathway analysis. Then we examine three different approaches, i.e. clustering-based methods, gene-based methods, and Gene set-based methods, all of which can be employed for understanding biological pathways in various environments. Subsequently we perform some case studies with the leukemia disease data and finally conclude this chapter with some remarks and discussions.

1.1 Overview of microarray gene expression profile data

DNA microarray is generally a glass or plastic substrate, or silicon chip, onto which tens of thousands of DNA molecules (*probes*) are deposited in a regular grid-like pattern [Zhang, 2006; Draghici, 2003]. Each grid spot corresponds to a DNA sequence of a specific gene. The idea of a microarray is to detect the presence and abundance of specific DNA molecules (*targets*) in biological samples of interests. For this purpose, from two mRNA samples (a test sample and a control sample), cDNAs are obtained and labeled with fluorescent dyes and the solutions including the labeled targets are hybridized on the surface of the chip. Then the chips are scanned to read the expression intensities emitted from the labeled and hybridized targets. Thus, by doing so, the microarray enables us to monitor the expression levels of tens of thousands of genes simultaneously. Once the raw microarray gene expression profiles are obtained in this way, several pre-processing steps are usually

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca, ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

performed which include data transformation, data filtering, missing value imputation, data normalization, and etc. Consequently, for the analysis of microarray data, the gene expression profiles are typically employed in a $p \times n$ matrix form as follows:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ & & \vdots & \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix}$$

where x_{ij} denotes the expression intensity of the ith gene (i=1, …, p, p is the number of genes) in the jth sample (j=1, …, n, n is the number of experiment conditions). Thus, the procedure of obtaining gene expression data matrix from a collection of raw data obtained by microarray experiments can be summarized as in Fig. 1.

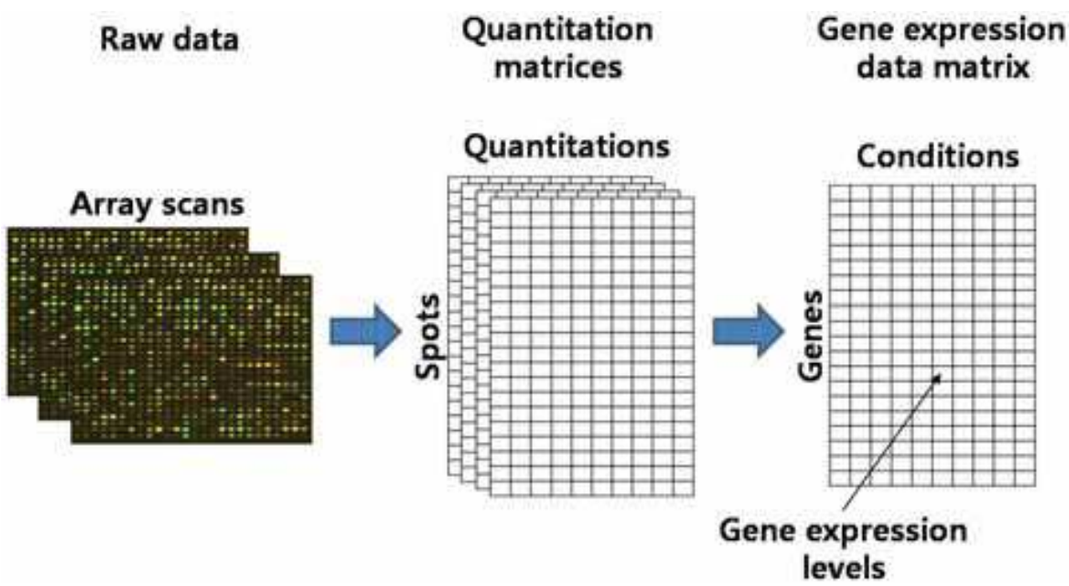

Fig. 1. The procedure of obtaining gene expression data matrix from a collection of raw microarray data (adapted from [Brazma et al., 2001]).

1.2 Biological resources for pathway analysis

There exist several types of biological resources which can be used for pathway analysis, such as pathway databases, gene annotation databases, Gene Ontology, and etc. In particular, approximately 179 kinds of pathway databases are currently available and some of them are given in Table 1. These databases are easily accessible through a web and Include a collection of pathway maps already known for various organisms. The pathway maps represent the knowledge on molecular interaction and reaction networks in metabolic pathways, genetic networks, signaling pathways, and complexes. Some of the pathway databases provide application program interfaces (APIs) enabling us to access pathway data in various downloadable formats. Among them, the KEGG(Kyoto Encyclopedia of Genes and Genomes) database is one of the most well-known extensive pathway databases

including the information about the genes of 181 organisms, their associated pathways and graphical diagrams, which can be downloadable in KGML and XML formats. Similarly, there are some other useful pathway databases such as BioCyc, GenMapp, BioCarta, and etc.

Pathway DB	Organisms	Pathway Types	Downloadable Formats
KEGG	181(varied)	metabolic, genetic, signaling, complexes	KGML, XML
BioCyc	E.Coli, human(20 others)	metabolic and complexes	BioPax, SBML
GeneMAPP	human, mouse, rat, fly, yeast	metabolic, signaling, complexes	MAPP format
Reactome	human, rat, mouse, chicken, fugu, zebrafish	metabolic, signaling, complexes	SBML, MySQL
BioCarta	human, mouse	metabolic, signaling, complexes	Just Images
TransPATH	human, mouse	Signaling, genetic	XML

Table 1. Some examples of currently available pathway databases (adapted from http://bioinformatics.ca/)

These pathway databases are also useful for evaluating and interpreting the analysis results of microarray gene expression profiles from the biological aspects. Fig. 2 shows an example of the acute myeloid leukaemia pathway map in KEGG pathway database.

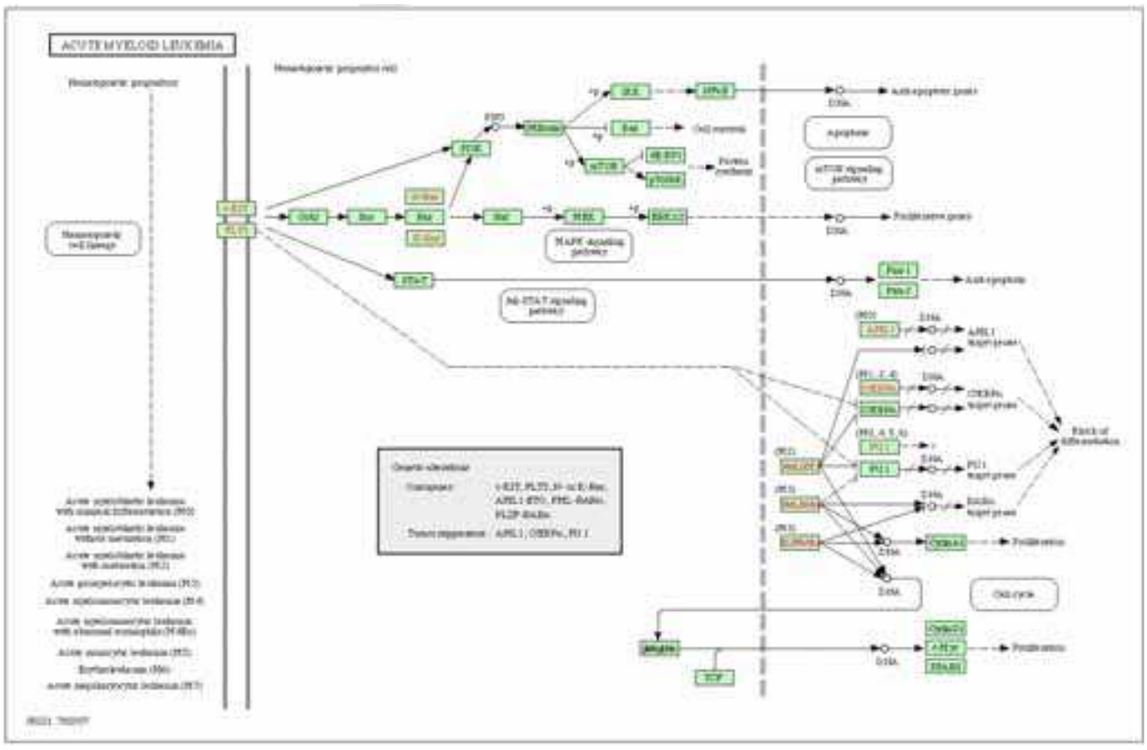

Fig. 2. An example of acute myeloid leukemia Pathway map in KEGG pathway database.

2. Computational methods for pathway analysis

There are several computational approaches for pathway analysis which employ microarray gene expression profiles and *a priori* known biological knowledge such as Gene Ontology, KEGG pathway database, and etc. According to the aim of microarray experiments and analyses, an appropriate computational method can be chosen. Here three different approaches are detailed. First, the clustering-based methods are based on the assumption that the genes having similar expression profiles would have similar biological functions. Thus, the genes showing similar expression Patterns under a series of conditions are grouped into a cluster and its corresponding common functional pathways are identified. By doing so, the functional pathways of uncharacterized genes in a cluster may be possibly conjectured from the pathway categories of characterized genes in the same cluster. Second, the gene-based methods are the ones that can be applied for two-grouped samples data (e.g., treatment vs. control, cancer vs. normal). That is, their main purpose is to identify the differentially expressed genes (DEGs) between two groups and find out which functional pathways these DEGs are significantly involved in. Third, the gene set-based methods are intended to identify significant pathways (i.e. gene-sets) showing good differential expression between two groups from the candidate gene-sets each of which is generated by taking all the genes belonging to a certain pathway. The approach of gene set enrichment analysis (GSEA) by Subramanian et al. (2005) belongs here. Unlike earlier two approaches, this enables us to identify the most significant pathways in a unified analytical framework by employing *a priori* known biological knowledge along with gene expression profiles for the analysis.

2.1 Clustering-based methods

Conventionally cluster analysis for identifying groups of objects having similar characteristics. In our context, clustering methods are employed to generate groups of genes, i.e. gene clusters, showing similar expression profiles. The clustering-based approach for pathway analysis is based on the assumption that the co-expressed genes are of similar biological functions. Thus, once gene clusters are generated, the corresponding common functional pathways for each cluster are identified. For this purpose, Fisher's exact test [Trajkovski et al., 2008] can be used with some biological resources like pathway databases or gene annotation databases. Also, for cluster generation, any clustering methods are possibly used. Here we describe one of the most popular methods for microarray data analysis, which is hierarchical clustering.

2.1.1 Cluster generation by hierarchical clustering

To generate gene clusters, two types of hierarchical clustering methods can be used, top-down and bottom-up approaches. The top-down approach is to start with one large cluster consisting of all genes and keep dividing them into smaller clusters until they become singleton clusters. On the other hand, the bottom-up approach is to start with as many singleton clusters as the gene size and keep grouping together two closest genes or clusters until they reach a single large cluster consisting of all genes. For gene expression data analysis, the bottom-up approach is generally used and the clustering results are Given in a tree-like graph, called *dendrogram*.

According to the way of defining the distance between two clusters in hierarchical clustering, which is called *linkage method*, some variations [Han et al., 2000] are possible as follows.

- Single Linkage: The distance between two clusters C_A and C_B is defined as the minimum distance between any two genes each belonging to C_A and C_B, respectively. This method has the characteristics that the between-cluster distances are relatively small while the within-cluster distances are relatively large.

$$d\left(C_A, C_B\right) = \min_{i \in C_A, j \in C_B} d_{ij} \tag{1}$$

- Complete Linkage: The distance between two clusters C_A and C_B is defined as the maximum distance between any two genes each belonging to C_A and C_B, respectively. This method has the characteristics that the within-cluster distances are relatively smaller than other linkage methods, since the maximum distance between two genes within a cluster is minimized. Because of this reason, it is the most popularly used.

$$d\left(C_A, C_B\right) = \max_{i \in C_A, j \in C_B} d_{ij} \tag{2}$$

- Average Linkage: The distance between two clusters C_A and C_B is defined as an Average of the distances between any of two genes each belonging to C_A and C_B, respectively. Here n_A is the number of genes in a cluster C_A and n_B is the number of genes in a cluster C_B. This method is considered as a compromise between single linkage and complete linkage.

$$d\left(C_A, C_B\right) = \frac{1}{\left(n_A + n_B\right)} \sum_{i \in C_A} \sum_{j \in C_B} d_{ij} \tag{3}$$

Also, several types of distance or similarity measures can be used for cluster generation, which include Euclidean distance, Manhattan distance, correlation coefficient, and etc. Each of these measures is described as below.

Distance measures	Formula
Euclidean distance	$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2}$
Manhattan distance (city block)	$d_{ij} = \sum_{k=1}^{n} \left\lvert x_{ik} - x_{jk} \right\rvert$
Correlation Coefficient	$\rho_{ij} = \frac{\sum_{k=1}^{n} \left(x_{ik} - \bar{x}_{i\cdot}\right)\left(x_{jk} - \bar{x}_{j\cdot}\right)}{\sqrt{\sum_{k=1}^{n} \left(x_{ik} - \bar{x}_{i\cdot}\right)^2} \sqrt{\sum_{k=1}^{n} \left(x_{jk} - \bar{x}_{j\cdot}\right)^2}}, \quad -1 \le \rho_{ij} \le 1$ $d_{ij} = 1 - \rho_{ij}$

2.1.2 Pathway analysis for gene clusters

Once gene clusters showing similar expression profiles are obtained by clustering method, for each cluster, the most representative common functional pathways need to be found. For this purpose, Fisher's exact test [Trajkovski et al., 2008] can be applied to estimate the probability of having *at least x* genes in a given cluster annotated by a specific functional pathway. The probability of having exactly x genes, out of k genes in a cluster, annotated by a specific functional pathway S is given as follows.

$$P(X = x \mid N, M, k) = \frac{\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}} \tag{4}$$

Here N is the number of total genes on a microarray, M is the number of genes *a priori* known as belonging to a specific functional pathway S, N-M is the number of genes not included in S out of total N genes. The Fisher's score p-value is thus calculated by using Eq. (5). For each of gene clusters, the corresponding significant pathways are extracted based on these p-values. In usual, the significance of functional pathways for a cluster are judged with p-value < 0.05.

$$p = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{k-1}}{\binom{N}{k}} \tag{5}$$

2.2 Gene-based methods

In analyzing the gene expression profiles consisting of two groups of samples (e.g. normal vs. cancer), some genes are expected to have significant difference in expression levels between two groups. We call these genes *differentially expressed genes* (DEGs), which are taken as the significant ones in a given experiment. Thus, the gene-based approach for pathway analysis is intended to identify DEGs first and then discover in which functional pathways the DEGs are actively involved. Prior to finding DEGs, some pre-processing like data filtering, log-transformation and normalization is generally performed on microarray expression profiles.

2.2.1 Identifying DEGs

When two-grouped microarray sample data are given, the differentially expressed genes can be identified by using one of the following methods, such as k-fold change, signal-to-noise ratio and t-test.

- k-Fold Change [Draghici, 2003; Schena et al., 1996]: This method is to find significant genes by calculating the ratio of the averaged expression intensities over each group of samples as follows.

$$\phi(i) = \frac{\mu_A(i)}{\mu_B(i)} \tag{6}$$

Here $\mu_A(i)$ and $\mu_B(i)$ denote the averaged expression intensities of the ith gene over the samples each belonging to group A and group B, respectively. If such ratio for a gene is larger than the threshold, the gene is considered as being significantly changed in expression levels between two groups. In usual, by taking an arbitrary threshold k=2 or 3, the DEGs are identified.

- Signal-to-Noise Ratio (SNR) [Golub et al., 1999]: This method is to find significant genes by calculating the difference between the averaged expression intensities of two groups divided by the sum of their standard deviations. The SNR is computed as follows.

$$\rho(i) = \frac{\mu_A(i) - \mu_B(i)}{\sigma_A(i) + \sigma_B(i)} \tag{7}$$

In Eq. (7), $\mu_A(i)$ and $\mu_B(i)$ denote the averaged expression intensities of the ith gene over the samples each belonging to group A and group B, respectively. Also, $\sigma_A(i)$ and $\sigma_B(i)$ denote the standard deviations of group A and group B, respectively. Once the SNRs for entire genes are computed and arranged in a decreasing order, the corresponding graph is typically shown as in Fig. 3. The DEGs are taken from the both ends of the graph.

Fig. 3. A typical form of SNR graph for the entire gene-list where the genes are arranged in a decreasing order of SNR values

- t-test [Tusher et al., 2001; Dudoit et al., 2002]: This method is to find significant genes by calculating t-statistic and using t-distribution to obtain p-value. The t-statistic for a specific gene is computed as in Eq. (8).

$$t(i) = \frac{\mu_A(i) - \mu_B(i)}{\sqrt{\frac{\sigma_A(i)^2}{n_A} + \frac{\sigma_B(i)^2}{n_B}}} \tag{8}$$

Here $\mu_A(i)$ and $\mu_B(i)$ denote the averaged expression intensities of the ith gene over the samples each belonging to group A and group B, respectively. Similarly, $\sigma_A(i)$ and $\sigma_B(i)$

denote the standard deviations of group A and group B, respectively, while n_A and n_B denote the numbers of samples in group A and group B, respectively. As the absolute value of *t*-statistic for a specific gene is larger, the corresponding gene would have smaller p-value which implies higher significance in a statistical sense. Once the *t*-statistics and p-values of all genes are obtained, we can take the genes having p-values less than a given significance level (generally, 0.01 or 0.05) as the DEGs.

2.2.2 Pathway analysis for DEGs

To identify functional pathways significantly involved by the DEGs, the Fisher's exact test Can be used in the same way as described earlier, except that the meanings of some parameters are different. Specifically, in Eqs (4) and (5), *k* corresponds to the total number of the identified DEGs and *x* is the number of genes included in a specific functional pathway S out of *k* DEGs. Once the Fisher's scores are computed for each of all the genes, significant functional pathways can be identified by taking the pathways having less than a specific *p*-value. Also, the specific roles of the DEGs in these functional pathways might be understood by further analysis of the inter-relationships among the gene expression profiles of the DEGs.

2.3 Gene set-based methods

As a gene set-based method for pathway analysis, the gene set enrichment analysis (GSEA) approach has been attaining lots of attentions lately. In particular, the GSEA employs microarray expression profiles and *a priori* known biological resources in a unified analytical framework to identify significant pathways. That is, it generates the candidate gene-sets of interest, where a gene-set consists of the genes belonging to a specific pathway, by using *a priori* known biological resources such as pathway databases, gene annotation databases, literatures, and etc. Then, the significance of each candidate gene-set (i.e. pathway) is evaluated by using microarray gene expression profiles. Specifically, for each gene-set, the enrichment score is computed and its statistical significance is estimated. The detailed steps of GSEA are summarized in the following [Subramanian et al., 2005; Taskesen, 2006]:

- *Step 1*: Computation of enrichment scores for gene-sets

This step is to compute an enrichment score of a given gene-set A gene-set consists of the genes *a priori* known as being involved in a specific pathway and many candidate gene-sets can be constructed by using pathway databases, Gene Ontology, and etc. To compute ES, the entire gene-list should be rearranged in the order of Ranking statistic such as SNR or Fisher's criterion [Kim et al., 2008]. Then, with the ordered gene-list, the Kolmogorov-Smirnov(KS) score is computed for each gene-set. For KS score, the empirical cumulative distribution functions For P_{hit} and P_{miss} are employed as shown in Eq. (11).

$$P_{hit}(i) = \sum_{j=1}^{i} \frac{E(j)}{N_H}$$
$$P_{miss}(i) = \sum_{j=1}^{i} \frac{(1 - E(j))}{N - N_H} \tag{11}$$

Here P_{hit} is the empirical cumulative distribution function of which cumulative sum becomes 1 when the first *i* genes in the ordered gene-list completely match the genes

included in a specific gene-set S. On the other hand, P_{miss} is the one of which cumulative sum becomes 1 when there is no match between them. In Eq. (11), E(j) is 1 if the jth gene in the ordered gene-set is included in a given gene-set (i.e. hit), or is 0 if the jth gene in the ordered gene-set is not included in a given gene-set (i.e. miss). Also N is the total number of genes in entire gene-list and N_H is the number of genes in a specific gene-set S. At times, the ranking statistic can be used for computing P_{hit} and P_{miss} [Taskesen, 2006]. For instance, assuming that the SNR-based gene ranking is used, E(j) Can be replaced by the SNR value of the *j*th gene in the ordered gene-list while N is the sum of the SNR values for total genes included in entire gene-list and N_H is the sum of the SNR values for the genes included in a specific gene-set S. In either way, by taking the maximum deviation between P_{hit} and P_{miss}, as shown in Eq. (12), the ES for a specific gene-set S is obtained.

$$ES(S) = \max_{i=1,\cdots,N} \left| P_{hit}(i) - P_{miss}(i) \right| \tag{12}$$

- *Step 2*: Estimation of statistical significance of ES

To estimate statistical significance of the ES obtained in step 1, *k* random permutations of a given microarray expression profile data on labels are generated and for each permutation, the corresponding ES is calculated. By using the computed ESs for *k* permuted data, we can obtain the null distribution of ES which will be used to calculate a *nominal* p-value of the ES.

- *Step 3*: Adjustment for multiple hypothesis testing

The estimated significance level is now adjusted to account for multiple hypothesis testing. First, the ES for each candidate gene-set is normalized for the gene-set size by dividing it with the mean of the *k* ESs for the permuted data obtained in Step 2. Then, for each normalized ES, the proportion of false positives is controlled by calculating the corresponding false discovery rate (FDR) or the family-wise error rate (FWER).

3. Case studies

For experiments, the leukemia dataset published by Golub et al. [1999] was used. The leukemia disease [Knudsen, 2006] is known as a cancer of the blood or bone marrow that affects blood-forming cells (usually white blood cells) in the bone marrow. The white blood cells are divided into granulocytes, monocytes, and lymphocytes. Leukemias starting in granulocytes or monocytes are called *myeloid leukemias*, and leukemias starting in lymphocytes are called *lymphocytic leukemias*. Further, leukemias are divided into acute and chronic. In acute leukemias, the cells cannot mature properly, whereas in chronic leukemias, the cells mature partly but do not obtain their full function. The Golub's leukemia dataset originally consists of 7129 human gene expression profiles shown in total 78 leukemia samples of two classes including 47 acute lymphoblastic leukemia(ALL) samples and 25 acute myeloid leukemia (AML) samples. Out of them, Golub et al. used 38 samples for training and 34 samples for testing to diagnose the subtypes of the leukemia. For our analyses, these 38 training samples including 27 ALL samples and 11 AML samples were used.

3.1 Pathway analysis with clustering

For cluster analysis, the data-filtering and data-transformation were applied for microarray expression profiles in the same way as in [Dudoit et al., 2002]. Specifically, the data-

transformation was done to replace the expression intensities less than 100 with 100 and expression intensities larger than 16000 with 16000. Also, the log-transformation with base 2 was applied for all the gene expression profiles. After such data-transformation, data-filtering was performed to remove the genes whose the difference between the maximum and the minimum of expression intensities is less than 500 or the ratio of the maximum to the minimum of expression intensities is less than 5. As a result, 3051 genes were remained and used for cluster generation. Clusters were generated by using complete-linkage hierarchical clustering with Euclidean distance and the results are shown in Fig. 6. By visual inspection of Fig. 6, the four was chosen to be an appropriate number of clusters. Thus, we generated four clusters and obtained the results as shown in Fig. 7, where four clusters include 666 genes, 656 genes, 427 genes, and 1302 genes, respectively.

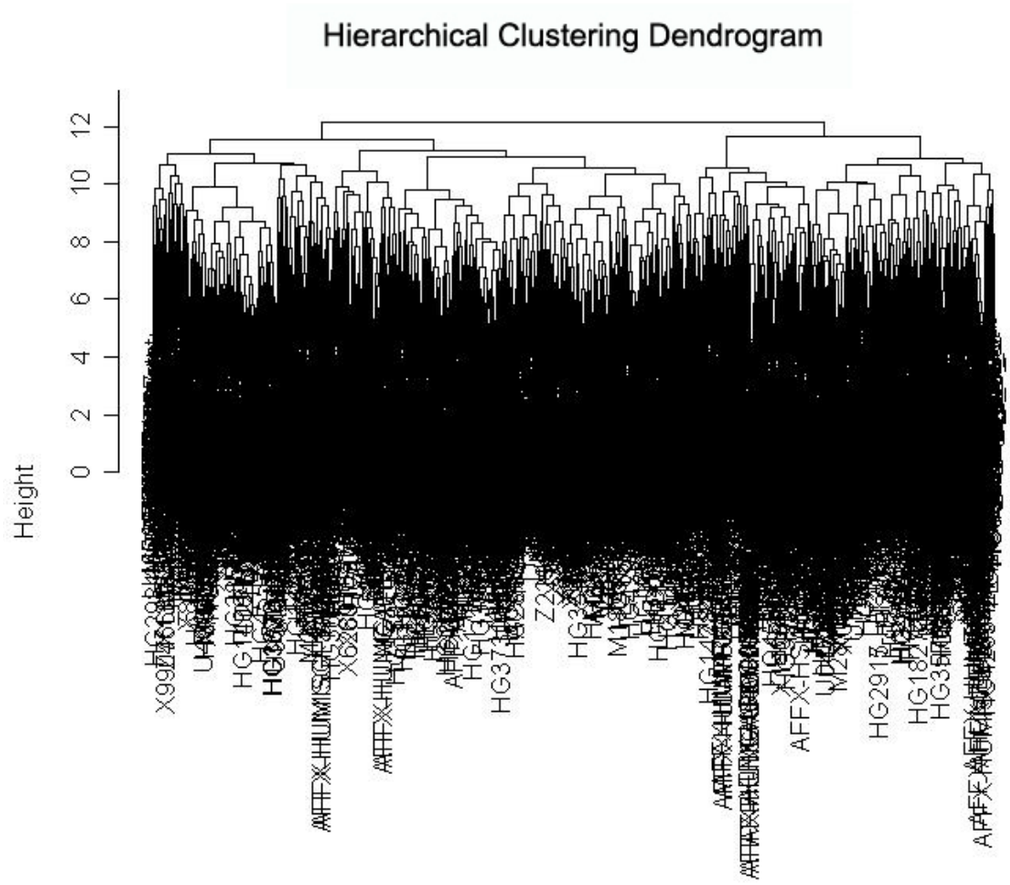

Fig. 6. Clustering result of Golub's leukemia dataset by complete-linkage hierarchical clustering with Euclidean distance.

To identify significant functional pathways involved in each cluster, we used KEGG pathway database to perform Fisher's exact test on the genes in a cluster. By using *p*-value≤0.05, the significant pathways for each of the four clusters were identified and shown in Tables 2-5.

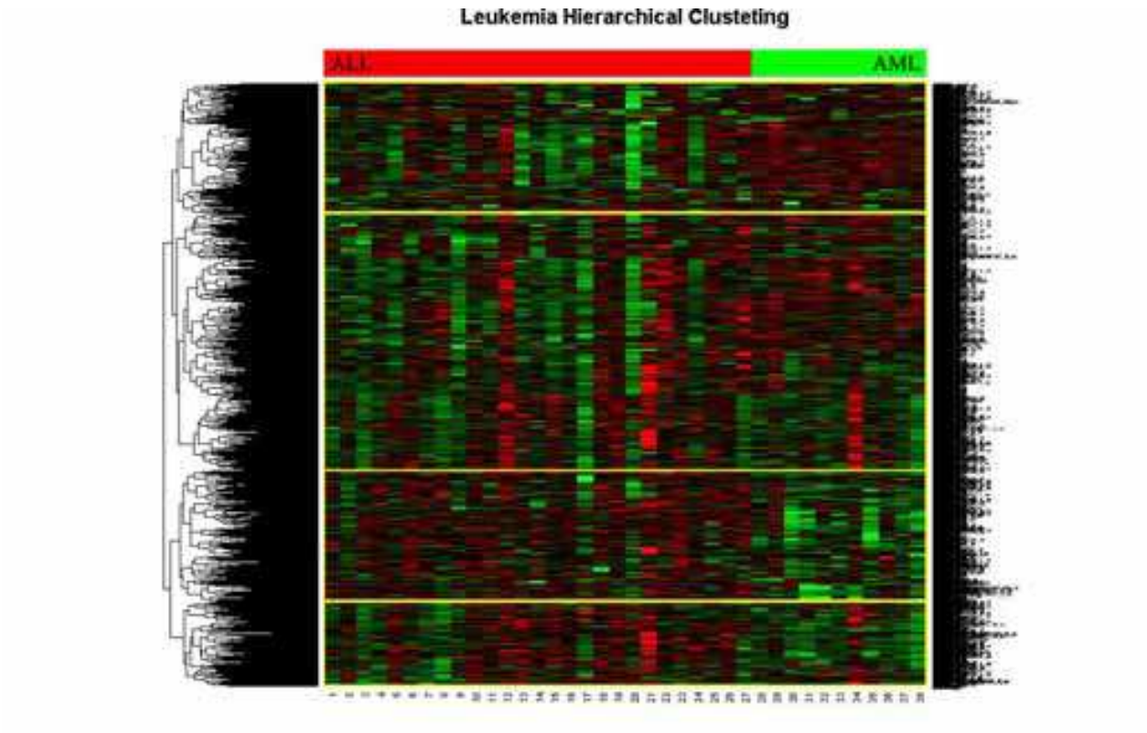

Fig. 7. Heatmap for four clusters of Golub's leukemia dataset obtained by complete-linkage hierarchical clustering with Euclidean distance measure

functional pathways	*p* -values
Hematopoietic cell lineage	3.7e-7
Type I diabetes mellitus	2.9e-5
Antigen processing and presentation	5.1e-5
Cytokine-cytokine receptor interaction	0.00013
Epithelial cell signaling in Helicobacter pylori infection	0.00042
Cell adhesion molecules (CAMs)	0.00067
Adipocytokine signaling pathway	0.00086
Toll-like receptor signaling pathway	0.00097
Porphyrin and chlorophyll metabolism	0.0012
Acute myeloid leukemia	0.0029
Aminosugars metabolism	0.035
Glutathione metabolism	0.042

Table 2. 1st cluster's significant functional pathways (*p* -value≤0.05)

functional pathways	*p* -values
Hematopoietic cell lineage	6.8e-7
B cell receptor signaling pathway	0.0048
Aminoacyl-tRNA biosynthesis	0.011
Insulin signaling pathway	0.016
Long-term potentiation	0.043
Alzheimer's disease	0.049

Table 3. 2nd cluster's significant functional pathways (*p* -value≤0.05)

functional pathways	*p* -values
Metabolism of xenobiotics by cytochrome P450	0.003
Linoleic acid metabolism	0.013
Colorectal cancer	0.013
VEGF signaling pathway	0.013
Small cell lung cancer	0.015
Pancreatic cancer	0.016
gamma-Hexachlorocyclohexane degradation	0.018
Non-small cell lung cancer	0.028
Neurodegenerative Diseases	0.029
Focal adhesion	0.032
Apoptosis	0.033
Acute myeloid leukemia	0.035
Prostate cancer	0.04
Chronic myeloid leukemia	0.047

Table 4. 3rd cluster's significant functional pathways (p -value≤0.05)

functional pathways	*p* -values
Glioma	2.0e-5
Chronic myeloid leukemia	5.7e-5
Prostate cancer	0.0001
Cell cycle	0.00024
Glycolysis / Gluconeogenesis	0.00074
Endometrial cancer	0.00078
T cell receptor signaling pathway	0.00092
Thyroid cancer	0.00098
Non-small cell lung cancer	0.0012
Small cell lung cancer	0.0018
Regulation of actin cytoskeleton	0.0027
Adherens junction	0.0035
Long-term potentiation	0.0051
Focal adhesion	0.0072
Oxidative phosphorylation	0.0073
Natural killer cell mediated cytotoxicity	0.0081
Gap junction	0.012
Proteasome	0.014
Acute myeloid leukemia	0.015
Colorectal cancer	0.016
ErbB signaling pathway	0.016
Renal cell carcinoma	0.016
Carbon fixation	0.018
Melanoma	0.021
Bladder cancer	0.022
Huntington's disease	0.022
Pancreatic cancer	0.033
Calcium signaling pathway	0.038
Insulin signaling pathway	0.041
Neurodegenerative Diseases	0.047

Table 5. 4th cluster's significant functional pathways (p -value≤0.05)

According to earlier Golub's study on this leukemia dataset, it was observed that their gene expression profiles can be grouped into AML related samples, ALL-T-cell related samples, ALL-B-cell related samples. To understand the biological meaning of our analysis results, thus, we analyzed whether significant functional pathways identified for each cluster are somewhat related to AML or ALL-T-cell or ALL-B-cell types and found some interesting observations. In Tables 2 and 4, it was observed that the genes in the 1st cluster and 3rd cluster are significantly involved in AML-related functional pathways such as hematopoietic cell lineage, apoptosis, and acute myeloid leukemia. Also, in Tables 3 and 5, it was observed that the genes in the 2nd cluster are significantly involved in ALL-B-cell related pathways such as hematopoietic cell lineage and B cell receptor signaling pathway, while the genes in the 4th cluster are significantly involved in ALL-T-cell related pathways such as T cell receptor signaling pathway and cell cycle. Consequently, from our experiments, it was found that each of the four clusters is closely related to the subtypes of the leukemia which include AML, ALL-T-cell, and ALL-B-cell types.

3.2 Pathway analysis with DEGs

To apply the gene-based method described earlier for pathway analysis, we need to find differentially expressed genes first. For this purpose, the data-filtering and data-transformation were applied for Golub's 7129 gene expression profiles in the same way as done for pathway analysis with clustering, and 3051 genes were obtained. Out of these genes, we extracted the most 50 differentially expressed genes by using SNR and the results are shown in Fig. 8. In this figure, we can see 25 DEGs showing relatively higher expression in ALL than AML, and the other 25 DEGs showing relatively higher expression in AML than ALL. This is why the DEGs were chosen by taking the genes having the 25 largest SNRs in positive region and the 25 smallest SNRs in negative region. Also, for comparisons, we applied *t*-test for 3051 gene expression profiles and selected 50 genes as the DEGs by having *p*-value≤0.000004, as shown in Fig. 9.. When the *t*-test is used for finding DEGs, the different number of DEGs can be chosen depending on the choice of *p*-value. In our case, we adjusted *p*-value in such a way to choose 50 genes, for comparative purpose with the SNR result. As seen in Figs 8 and 9, even if the same number of DEGs are chosen by SNR method and *t*-test, respectively, it is observed that the identified DEGs are quite different.

To understand significant functional pathways in which the DEGs identified by SNR method and *t*-test are actively involved, the Fisher's exact test was applied for the both cases. Table 6 shows the result Of significant functional KEGG pathways identified for the SNR DEGs while Table 7 shows the pathway analysis result for the *t*-test DEGS. As seen in Tables 6 and 7, the same pathways were identified for the two different sets of 50 DEGs chosen by SNR method and *t*-test, respectively. In particular, the pathway of hematopoietic cell lineage is known as being related to both ALL and AML types, and the B cell receptor signaling pathway is related to ALL-B-cell type. Thus, the DEGs showing significant expression difference between two groups of ALL and AML are empirically found to be actively involved in hematopoietic cell lineage and B cell receptor signaling pathway. In addition, it is observed that the DEGs identified by the *t*-test seem to be more meaningful in terms of *p*-values than the DEGs identified by the SNR.

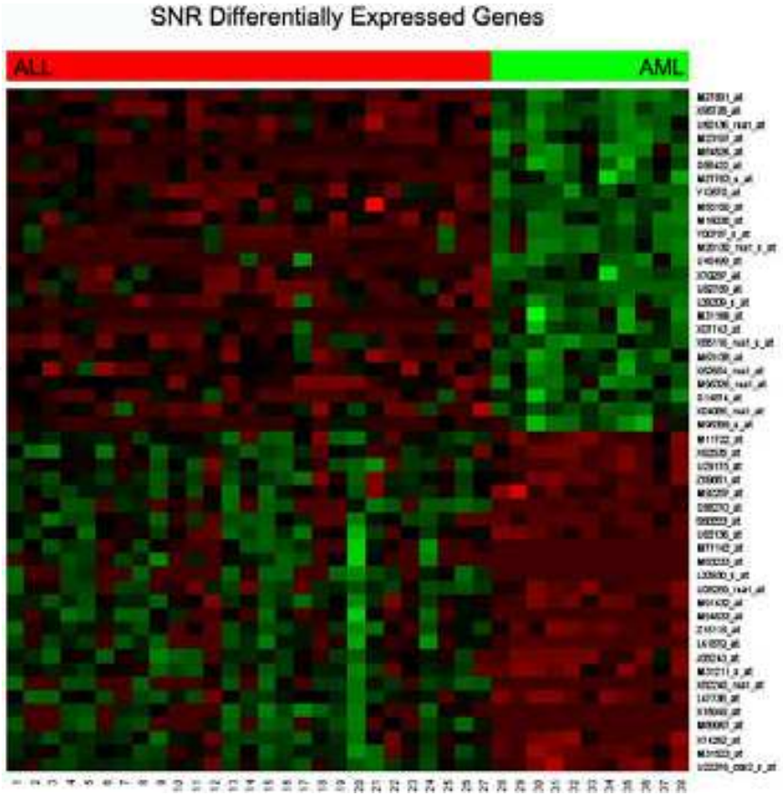

Fig. 8. Heatmap of 50 differentially expressed genes identified by SNR.

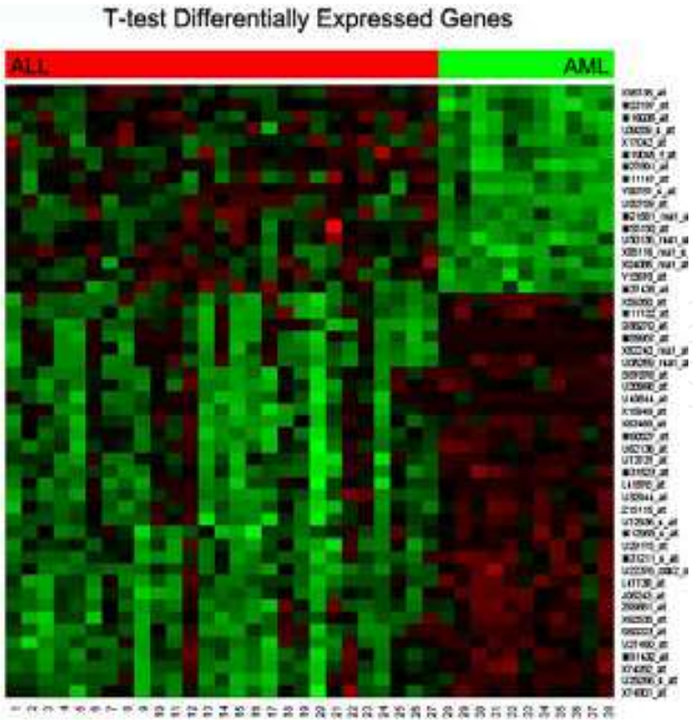

Fig. 9. Heatmap of 50 differentially expressed genes identified by *t*-test.

functional pathways	*p*-values
Hematopoietic cell lineage	0.0081
B cell receptor signaling pathway	0.041

Table 6. Significant functional KEGG pathways identified for 50 DEGs by SNR method (*p*-value≤0.05)

functional pathways	*p*-values
Hematopoietic cell lineage	0.005
B cell receptor signaling pathway	0.01

Table 7. Significant functional KEGG pathways identified for 50 DEGs by *t*-test method (*p*-value≤0.000004)

3.3 Pathway analysis by gene set enrichment analysis

To perform pathway analysis with GSEA, the *z*-score normalization was first applied for 7129 gene expression profiles. Also, 136 candidate gene-sets were generated from KEGG pathway database in such a way to take only the pathways each of which include at least 10 genes out of 7129 genes. Then, for each of the candidate gene-sets, the corresponding ES was computed, and its statistical significance level and the normalized ES were obtained with 1000 random permutations of 7129 gene expression profiles on class labels. In particular, the normalized ESs for the candidate gene-sets has a two-modal distribution as shown in Fig. 10. Thus, the most 40 significant pathways were identified as done in [Subramanian et al., 2005] by taking 20 gene-sets from the rightmost end in a positive region and 20 gene-sets from the leftmost end in a negative region. The results are shown as in Table 8.

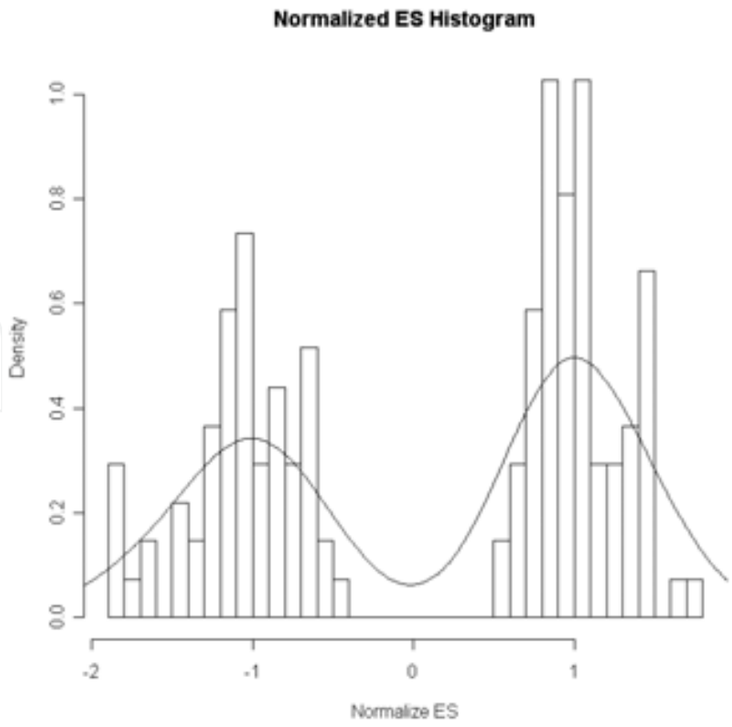

Fig. 10. The distribution of normalized ESs for the candidate gene-sets

functional pathways	Normalized ES
Pyruvate metabolism	1.759
Cell cycle	1.718
Galactose metabolism	1.511
Alanine and aspartate metabolism	1.500
Basal transcription factors	1.497
DNA polymerase	1.496
Aminoacyl-tRNA biosynthesis	1.481
Purine metabolism	1.463
Citrate cycle (TCA cycle)	1.445
Proteasome	1.440
Pentose and glucuronate interconversions	1.377
One carbon pool by folate	1.369
1- and 2-Methylnaphthalene degradation	1.360
Pyrimidine metabolism	1.327
Biosynthesis of steroids	1.311
Lysine degradation	1.296
Wnt signaling pathway	1.266
Folate biosynthesis	1.260
Butanoate metabolism	1.248
RNA polymerase	1.209
Apoptosis	-1.150
Neurodegenerative Disorders	-1.177
ECM-receptor interaction	-1.204
Metabolism of xenobiotics by cytochrome P450	-1.216
Glycosphingolipid biosynthesis - neo-lactoseries	-1.219
Cytokine-cytokine receptor interaction	-1.249
Methane metabolism	-1.280
Sphingolipid metabolism	-1.308
Leukocyte transendothelial migration	-1.323
Hematopoietic cell lineage	-1.378
Glycan structures - degradation	-1.438
Nitrogen metabolism	-1.477
Complement and coagulation cascades	-1.497
Toll-like receptor signaling pathway	-1.656
Glycosaminoglycan degradation	-1.682
Arachidonic acid metabolism	-1.754
Glutathione metabolism	-1.755
Epithelial cell signaling in Helicobacter pylori infection	-1.778
Porphyrin and chlorophyll metabolism	-1.805
Adipocytokine signaling pathway	-1.810

Table 8. The most 40 significant pathways identified by GSEA for Golub's 7129 gene expression profiles

As seen in Table 8, It is interesting that the pathways including an apoptosis pathway known as being related to cell death, or the hematopoietic cell lineage known as being related to generating an antibody in a blood were identified as significant pathways. Also, some other cancer-related pathways were found.

4. Concluding remarks

In this chapter we introduced several mining methods for biological pathway analysis with microarray gene expression profiles. For pathway analysis, our concern is how to identify significant functional pathways in which many genes showing differential expression between treatment and control groups are actively involved. In earlier approaches, the computational techniques only for microarray data had been more focused than biological interpretation of the results. On the other hand, the recent approaches concentrate more on the effective use of a variety of biological resources in analyzing large volume of microarray expression data to obtain more biologically meaningful results. By applying them for a variety of fields such as drug discovery [Bild et al., 2006] or disease diagnosis[Watters et al., 2006], pathway analysis with microarray expression data can play a key role in the path to new scientific discoveries and allows us to understand what biological phenomena causes the observed expression patterns. Furthermore, by making an attempt to use new types of biological resources along with expression profiles for the analysis, our understanding of the expression data could be enhanced in a deeper manner.

5. Acknowledgement

This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MEST) (No. R01-2008-000-11089-0)

6. References

Werner, T., "Bioinformatics applications for pathway analysis of microarray data.", Current opinion in biotechnology, 19(1):50-4.(2008)

Corn, P.G. et al., "Microarray analysis of p53-dependent gene expression in response to hypoxia and DNA damage," Cancer biology & therapy, 6(12):1858-66.(2007)

Saiki, T. et al., "Identification of marker genes for differential diagnosis of chronic fatigue syndrome," Molecular Medicine, 14(9-10):599-607.(2008)

Zhang, A., “Advanced analysis of gene expression microarray data,” World scientific publishing co. (2006).

Draghici, S., “Data Analysis Tools for DNA microarrays,” Chapman & Hall/CRC(2003)

Brazma, A. et al., "A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays", European Bioinformatics Institute, Draft. (2001)

Canadian Bioinformatics Workshops, http://bioinformatics.ca/, Protein Pathways and Pathway Databases

KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.ad.jp/kegg/

BioCyc, http://biocyc.org/

K. D. Dahlquist et al., “GenMAPP: A new tool for viewing and analyzing microarray data on biological pathways,” Nature genetics 2002 May;31(1):19-20.

BioCarta, http://www.biocarta.com/

Gene Ontology, http://www.geneontology.org/

Subramanian, A. et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.", Proc. Natl Acad Sci USA 102: 15545-50. (2005)

Eisen, M. B. et al., "Clustering analysis and display of genome-wide expression patterns," PNAS, 95, 14863-14868. (1998)

Han, J. et al., Data Mining : Concepts and Techniques, Academic Press. (2000)

Trajkovski, I. et al., "SEGs: Search for enriched gene sets in microarray data," Journal of biomedical informatics 41, 588-601. (2008)

Schena, M. et al., "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes," Proc Natl Acad Sci U S A, 93(20):10614-10619. (1996)

Golub, T. R. et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science (Wash. DC), 286: 531.537. (1999)

Tusher, V. G. et al., "Significance analysis of microarrays applied to ionizing radiation response," Proc Natl Acad Sci U S A, 98(9), 5116-5121. (2001)

Dudoit, S. et al., "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," Statistica Sinica, 12, 111-140. (2002)

E. Taskesen, "Sub-typing of model organisms based on gene expression data," Bioinformatics technical University of Delft Research Assignment. (2006)

Knudsen, S. "Cancer diagnostics with DNA microarrays," Wiley-Liss (2006)

Bild, A. H. et al., "Linking oncogenic pathways with therapeutic opportunities," Nature reviews. Cancer, 6(9):735-41. (2006)

Watters, J. W. et al., "Developing gene expression signatures of pathway deregulation in tumors", Molecular cancer therapeutics, 5(10):2444-9. (2006)

Kim, J.Y. et al., "Identifying biologically significant pathways by gene set enrichment analysis using Fisher's criterion," To appear in proc. of bioscience and biotechnology 2008.

Data Mining and Knowledge Discovery in Real Life Applications

Edited by Julio Ponce and Adem Karahoca

ISBN 978-3-902613-53-0
Hard cover, 436 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2009
Published in print edition January, 2009

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and some of the chapters have cases with offered data mining solutions. We hope that this book will be a useful aid in showing a right way for the students, researchers and practitioners in their studies.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Miyoung Shin and Jaeyoung Kim (2009). Microarray Data Mining for Biological Pathway Analysis, Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.), ISBN: 978-3-902613-53-0, InTech, Available from: http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/microarray_data_mining_for_biological_pathway_analysis

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen