Data Mining of Differentially Expressed Genes Based on Gene Expression Profiling Microarray

Zhiping Miao, Yiduo Qi, Airong Qian*
Bone Metabolism Lab, Key Laboratory for Space Bioscience & Biotechnology, School of Life Sciences, Northwestern Polytechnical University, Xi’an 710072, China.
*Corresponding author(E-mail: qianair@nwpu.edu.cn)

Tuanmin Yang
Hong-Hui Hospital, Xi’an Jiaotong University College of Medicine, Xi’an 710054, Shaanxi, China

Abstract
Gene expression profiling microarray has been proved extremely valuable for the study of various functions and mechanisms for organism at the gene level. However, gene expression profiling experiments usually generate huge amounts of data points. In many cases, expression profiling data analysis takes far more effort than performing the initial experiments. Therefore, suitable methods for data analysis must be performed to obtain useful information from gene expression profiling. This article comprehensively describes the methods to analyze differentially expressed genes (DEGs), including gene annotation, the Gene Ontology (GO) and pathway ontologies, functional annotation and clustering, pathway analysis, gene-gene interactions analysis and regulatory networks analysis. To highlight the key steps, we take the analysis of gene expression profiling in MLO-Y4 cells (an osteocyte-like cell line) under the treatment of diamagnetic levitation (DL) or random position machine (RPM) for 48 hours as an example. Common bio-statistical methods to get the data of DEGs are represented. Potential challenge for the prospects of microarray data analysis is discussed. This paper is intended to offer some suggestions for biologists and biomedical scientists engaged in data-mining of biologic significance based on gene expression profiling data.

Key words: data mining, gene expression profiling, microarray, GO, clustering, gene interactions analysis, gene regulatory network.

1. INTRODUCTION
Gene chip technology has been successfully applied in many fields due to its high throughput, high sensitivity, and speed advantages, for example, detection of gene expression (Loeser and Olek, 2012), gene function studies (Huang and Sherman, 2008), the identification of new genes, drug target screening (Smith and Tsalenko, 2007), guide medication, drug screening, clinical diagnosis and preventive medicine (Miller and Adam, 2010) etc. Gene expression profiling microarray is the most commonly used technology among them and becomes increasingly mature. Gene expression profiling contributes a lot for classification, prognostication, and prediction in breast cancer research (Khoury and Gwinn, 2007) and discovery of prostate cancer biomarkers, etc. This technology can save users a lot of time and research funds, and help them get access to a wide range of relevant results. Gene expression profiling microarray is primarily applied to analyze comprehensive information on a large scale of gene expression changes about a specific biological object under certain conditions. This application provides guidance for subsequent analysis and experiments, and also greatly accelerated the pace of life sciences research.

In recent years, along with the maturity of microarray technology, large numbers of research results have produced vast amounts of biological data. Improper ways for data analysis can lead to the missing of important information, thus hindering the subsequent excavation of the reality, and the advantage of high throughput may become disadvantage. So, how to gain the results with biological significance from these numerous gene expression data has become a new problem for researchers (Reis-Filho, 2011). In fact, biological workers need to know generally about the process of data-processing to have a clear understanding of the credibility of the data so as to plan for subsequent data-analysis and experiments. Useful results with biological significance are obtained only if the data is combined with biological information. There are lots of different kinds of programs for microarray data analysis online and offline. What biologists need to do is to make the best use of these resources. So, how to start efficient data analysis and extract meaningful results from massive data, and what are the common methods have become problems with vital importance.

With the microarray being applied more and more frequently, there are many literatures using the method of gene expression profiling. But what is the integrated analysis process? This paper describes the analysis process in a relatively comprehensive aspect, and then further illustrates how these methods are applied to actual data
analysis combined with the author's experimental data. It will be useful to reduce the workload of biologists, help them to understand methods of microarray data analysis in short order, and provide ideas and references for subsequent in-depth study.

2. BIO-STATISTICAL METHODS USED IN SCREENING DIFFERENTIALLY EXPRESSED GENES (DEGS)

2.1. Discrimination Analysis

In order to help biologists interpret the biological meaning from gene chips, bio-statistical methods are widely practiced. From the view of machine learning, strategies for gene chip data-analysis can be broadly grouped into two categories: discriminant analysis and cluster analysis (Sørensen, 2010).

Discriminant analysis, also known as supervised learning, is a machine algorithm inferring some functions from a training dataset (known law). Discriminant requires data consisting of two components. One is gene expression data from chips running on a set of samples. The other is from characterizing the samples (e.g., regulatory factor). This method aims at predicting sample characteristics using a mathematical model. There are a large number of statistical and computational approaches, including linear discriminant analysis, Support Vector Machines (Shannon, 2003), artificial neural network, etc. for discrimination. Wang et al. proposed a new process utilizing a series of artificial neural networks segmenting microarray image to identify which pixels within an image represent a certain gene. This design not only delivers results comparable and even superior to existing techniques but also has a faster run time.

2.2. Cluster Analysis

If the classification has already existed, supervised learning or discriminant analysis method will be an appropriate analytical method. But if the classification has not yet formed, the unsupervised learning or cluster analysis method will be more efficient than supervised learning.

Cluster analysis or unsupervised learning consists only of the gene expression data. The goal of this analysis is to classify similar genes in the same cluster. For unsupervised algorithm, clustering is completed based on the similarity (or non-similarity) of vectors, without a priori knowledge of the sample. It is a machine algorithm looking the composition law in the measurements from the cluster. Just like supervised learning, clustering also has plenty of approaches available in statistical and computational field. These include Hierarchical Clustering (Brown M P S, 2000), K-means Clustering, Self-organizing Map, and artificial neural network.

Cluster analysis is a powerful tool for data reduction and hypothesis generation. Even so, we should keep in mind that cluster analysis has two shortcomings. For one thing, it does not have a probabilistic foundation, that is, it runs no statistical test to guide the decision of where to cut the dendrogram. For another, there is not a generally accepted way to test the feasibility of the algorithms used in cluster analysis (Khan J, Wei J S, 2001). Methods are currently being developed to address the drawbacks of cluster analysis, and it will be a long way to go.

Cluster 3.0 has been developed by Stanford University and used widely nowadays. Analytical results can be observed through visualization software, TreeView. Cluster and TreeView are programs that provide a computational and graphical environment for analyzing data from DNA microarray experiments, or other genomic datasets, and the program Cluster can organize and analyze the data in a number of different ways (Wang, 2013). TreeView allows the organized data to be visualized and browsed. (Kanfi et al., 2012) used Cluster and found significant differences in gene expression between male Sirtuin-6-transgenic mice and male wild-type mice. The study showed the regulation of mammalian lifespan by a sirtuin family member and has had important therapeutic implications for age-related diseases.

After the input of gene chip data to the program, Cluster returns analysis results in some different formats. Generally, we use "Excel" to save the data and "heat map" (Figure 1) to visualize the results of the analysis. To fully understanding DNA microarray data, a set of visualization components able to interact with each other has been proposed, including parallel coordinates, cluster boundary genes, 3D cluster surfaces and DNA microarray visualizations as heat maps.

Figure 1 is an example taken from our recent research data using Cluster to analyze mechanics-sensitive genes differentially expressed in MLO-Y4 (an osteocyte-like cell line) after 48 hours of diamagnetic levitation (μg) treatment. As we can see, a heat map composed by a high number of colored grid points was shown and each grid point represents a gene expression value. The columns are different experimental conditions and the rows are different genes. In the heat map, colors at particular point (i.e. row by column coordinate) are endowed with the level of expression for genes (row) in different experimental conditions (column). Red means high expression, green indicates the opposite. Colors can be set in software before analysis.

The detailed steps for using Cluster are introduced in Figure 1. To begin with, text document including gene expression values are input to Cluster. It is an off-line software program and can be downloaded from websites of most biotech companies and authoritative biological research institutions. According to the data, the corresponding clustering algorithms are selected. After the procedure of data entry, "heat maps" through a built-
in JAVA program can give you a general idea of DEGs. So far genes have already been initially clustered by Cluster. Any heat maps with interesting genes can be selected and zoomed in or out for preview. The list of related genes and heat maps can be output to be saved.

**Figure 1.** Detailed steps for using Cluster

First, text document including gene expression values are input to Cluster; second, the corresponding clustering algorithms are selected according to the data; third, a general idea of genes differentially expressed can be given by "heat maps" through a built-in JAVA program. So far genes have already been initially clustered by Cluster. Any heat maps with interesting genes can be selected and zoomed in or out for preview purposes. The list of related genes and heat maps can be output to be saved.

### 3. DATA MINING METHODS BASED ON DEGS

Currently, the technique for application of high-throughput technologies to detect DEGs is very thriving. In order to introduce how to use this technique, details for data mining methods will be described respectively and some instances of analysis will be given below to facilitate its application. Main analysis steps refer to **Figure 2.** Methods discussed latter are included in the process of data analysis.

**Figure 2.** The process of gene expression profiling microarray data analysis

The diagram is a simple flow chart for gene expression profiling microarray data analysis. There are five steps for data analysis of DEGs, including: the Gene Ontologies (GO) and pathway ontologies, functional annotation and clustering, pathway analysis, gene-gene interactions analysis and regulatory networks analysis. These aspects constitute the main content of data analysis.

#### 3.1. Gene Annotation and Ontology towards Gene Function (Gene Ontology and Pathway Ontologies)

In order to know about the details of genes, gene annotation is required. As a major goal of functional genomics research, functional annotation of the genome means annotating biological functions of all encoded products from the genome. High-throughput with the application of Bioinformatics is adopted in this technique. Research in this field has become a hot point in post-genomic era.
In fact, the simple understanding of genes is far enough. More researches on functions genes should be carried out. This is necessary for functional genomics research. As the foundation of systems biology, ontologies, which are based on AmiGO database, provide a formal biological knowledge system that is amenable to computational and human analysis. Although ontology building and ontology annotation overlap in some parts, neither of them is dispensable to facilitate gene products mapping function in the genome. Ontology building consists of three knowledge domains: molecular function, biological process and cellular component.

Though there is no centralized resource like GO for pathway ontologies, pathway ontologies have even a longer history than GO. GO and pathway ontologies can tell protein function according to our current understanding of biology, and ontology annotation provides the actual information about genes. P-value is an important symbol that represents the relationship between GO categories and genes, considered as interesting and differentially expressed in the microarray.

The source of all ontology annotations ultimately derives from original research papers and one or more steps of inference, and not all ontology annotations are inferred by methods absolutely reliable. So whether or not an ontology annotation is reliable depends on the way it is used in these inferences. The development of experiment and inference, in return, makes the ontologies more perfect towards biology.

When it comes to the application, we can input gene names to AmiGO for GO and The Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/) for pathway ontologies. Besides, the results can be obtained from a variety of analysis software, and almost all analysis software is based on the GO annotations for gene expression microarray data analysis.

Table 1 gives a clear hint. GO annotations consists of three parts, namely, biological process, cellular component, and molecular function. When it comes to pathways, pathway annotation would be shown as the last two lines. The data come from the analysis of gene expression profiling in MLO - Y4 cells under RPM for 48 hours. The data were named “RPM data”.

GO annotation of genes consists of three parts: biological process, cellular component, and molecular function. When a gene is involved in the regulation of a certain pathway, information on pathway annotation would be given.

3.2. Classification and annotation Based on Gene Function

Plenty of distributed biological knowledge increases the difficulty of interpreting the data derived from microarray. Biological interpretation of large gene lists derived from high-throughput microarray includes: 1) large amounts of functional annotation for every gene; 2) summarizing which genes are associated with specific biological processes; 3) identifying functional biological modules consisting of related genes and terms; 4) and viewing inter-relationships between groups of genes and groups of biological terms. A number of publicly available bioinformatics tools have been developed to address these aspects. The most popular one is DAVID (The Database for Annotation, Visualization and Integrated Discovery) Gene Functional Classification Tool (http://david.abcc.ncifcrf.gov/), a widely used web-based application that provides a module-centric approach for functional analysis of large gene lists. It is characterized by gene functional classification and annotation clustering. The detailed steps for using DAVID can be found in the article published in Nature Protocols.

Table 1. An example of GO and pathway ontologies

<table>
<thead>
<tr>
<th>Gene Symbol: Name</th>
<th>Slfm: SAFB(scaffold attachment factor B )-like, transcription modulator</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO Biological Process ID</td>
<td>GO:0006350 // GO:0006915 // GO:0008150 // GO:0045449</td>
</tr>
<tr>
<td>GO Biological Process Term</td>
<td>Transcription // apoptosis // biological process // regulation of transcription</td>
</tr>
<tr>
<td>GO Cellular Component ID</td>
<td>GO:0005575 // GO:0005634</td>
</tr>
<tr>
<td>GO Cellular Component Term</td>
<td>Cellular component // nucleus</td>
</tr>
<tr>
<td>GO Molecular Function ID</td>
<td>GO:0000166 // GO:0003674 // GO:0003676 // GO:0003723</td>
</tr>
<tr>
<td>GO Molecular Function Term</td>
<td>nucleotide binding // molecular function // nucleic acid binding // RNA binding</td>
</tr>
<tr>
<td>Pathway Source</td>
<td>GenMAPP( <a href="http://www.genmapp.org/default.html">http://www.genmapp.org/default.html</a>)</td>
</tr>
<tr>
<td>Pathway Name</td>
<td>mRNA processing binding Reactome</td>
</tr>
</tbody>
</table>

3.3. Pathway Analysis

In many cases, genes participate in the regulation of the organism through the pathways they are involved in. So making sense of the pathway genes that are involved in is an important part of gene expression profiling microarray data analysis, which also provides clues for the understanding of the biological function genes involved in.

Gene annotation opens a window for biologists to know about the gene in genome from a biological sense. Pathways put emphasis on functional interactions between genes rather than processing the merely gene-centered GO analysis. It is a procedure of mapping the significantly regulated genes to pre-compiled pathways.
Then, biologists can observe the whole chains of events formed by a suite of genes in microarray experiments, rather than the only situation of the genes.

Pathway analysis is also known as functional enrichment, with the approach of mapping the differentially expressed genes in gene signal transduction pathways and metabolic pathways. This method connects the relevant gene to a network or pathway, on the basis of a known network or pathway and the function annotation classification of genes. Then, the normalized microarray data and their annotations will be integrated into some new category, taking metabolic pathways and GO functional classifications for example. The pathways genes involved in can be acquired from inputting gene names to related databases, such as KEGG, or some functional analysis software, like DAVID etc. In a recent research literature, pathway analysis was applied to the study of breast cancer, finding that many different signaling pathways were activated, strongly linked to invasion, metastasis development, proliferation, and with a significant cross-talking rate.

Many software programs are available to visualize gene expression data both on the pathways and GO. The international community has a lot of databases that can be shared on pathway gene networks, including BioCarta, EcoCye, MetaCye and KEGG, etc. KEGG is the most widely applied database and also familiar to biologists for its comparative enormous data volume, and always be the basis of other pathway analysis software.

Nowadays, a lot of microarray data analysis tools can acquire some simple function network in the way of integrating these databases, combining with gene expression values. Some software programs even provide tools for customers to construct or modify the specific gene network pathways according to the gene expression data, such as GenMAPP, a tool for analyzing and displaying the expression profile data on the basis of gene network pathways.

When conducting pathway analysis, pathway diagrams are often used to ascertain specific roles of genes in pathways. Pathway diagrams vividly show how genes are involved in pathways and participate in the regulation of the expression. These diagrams also tell the information for up and downstream genes and genes performing the same function. Figure 3 is a part of pathway analysis results from “RPM data”, a diagram of Growth Arrest and DNA Damage-inducible 45(GADD45) signaling pathway. GADD45 is one of the several known p53 target genes that is inducted by the stabilization and accumulation of p53 tumor suppressor protein, a major mediator in the cellular response to DNA damage. GADD45 signaling pathway contains four differentially expressed genes: CCNE2 (cyclin E2), ATR (ataxia telangiectasia and Rad3 related), BRCAl (breast cancer 1, early onset), ATM (ataxia telangiectasia mutated).

In this figure, pink markers represent genes differentially expressed which are detected in this experiment and are low expression genes in this pathway. The arrows indicate the regulatory relationship between genes. Meanings of related symbols in the pathway diagram refer to the block on the left of the image. Ellipse represents Transcription Regulator; circle corresponds to Group or Complex; inverted triangles mean kinase.

3.4. Gene-gene Interactions Analysis

Following the identification of various pathways by whole genome association analysis, the detection of interactions among the genes enriched is an increasingly recognized problem. In the elucidation of biological phenomena, interactions among genes may play important roles. Theoretical considerations suggest that gene–gene interactions are quite common in interpretation of biological mechanism. At present, detecting gene–gene interactions is an important part of genomics research. (Cordell, 2009) conducted the research on mechanism of human disease by detecting gene–gene interactions. Many microarray data analysis software could analyze gene–gene interactions to facilitate biologists obtaining underlying information. Figure 4 is a diagram of gene–gene interaction from “RPM data” processed by iReport. Each circle represents a gene, BLM (Bloom syndrome, RecQ helicase-like) and CUL3 (cullin 3) mean the upstream genes. ATM and TOP2A (topoisomerase (DNA) II alpha 170kDa) indicate the downstream genes. MKI67 (antigen identified by monoclonal antibody Ki-67), NCL (nucleolins), PRKDC (protein kinase, DNA-activated, catalytic polypeptide), HNRNPA3 (heterogeneous nuclear ribonucleoprotein A3), CKAP5 (cytoskeleton associated protein 5), SF3B1 (splicing factor 3b, subunit 1, 155kDa), PMS1 (PMS1 postmeiotic segregation increased 1 (S. cerevisiae)) and EIF2AK2 (eukaryotic translation initiation factor 2-alpha kinase 2) represent bi-directional genes. We can learn that gene TOP1 (topoisomerase (DNA) I) has two upstream genes, two downstream genes and eight bi-directional genes. TOP1, BLM, CUL3, ATM, TOP2A, MKI67, NCL, PRKDC, HNRNPA3, CKAP5, SF3B1, PMS1 and EIF2AK2 are gene names. Light pink means the upstream genes, green indicates the downstream genes, blue represents for bi-directional genes. TOP1 has two upstream genes, BLM and CUL3. ATM and TOP2A are downstream genes. MKI67, NCL, PRKDC, HNRNPA3, CKAP5, SF3B1, PMS1 and EIF2AK2 are bi-directional genes.

3.5. Regulatory Networks Analysis of Gene Expression

After thoroughly getting across the biological function of genes involved, biologists may want to see whether there are some interactions between genes. Actually, biological processes usually involve more than one pathway which interconnects in a specific manner. These pathways compose into regulatory networks. Networks are not constant, but changeable with context. A significant feature of the regulatory networks is that it may be
stretched across a number of pathways, but targets only a few genes in each pathway. That is why sometimes we find that some genes are also included in the pathway we are studying but do not cause the entire up/down-regulation of complete pathway.

**Figure 3.** A diagram of gene-gene interaction. Each circle represents a gene.

At present, a lot of researches on regulatory networks, standing alone or combined with GO and pathways, have been carried out. Sequence-based approaches have evolved from the simple systems like yeast to more complex systems like mammalian. In addition to genes, the role of miRNAs in regulating gene expression networks has attracted more and more attention nowadays.

There are a lot of biological companies that can help us establish gene networks. STRING (URL: http://string-db.org/) is a good choice. It provides intuitive interaction network maps of proteins expressed by genes. The interaction between genes can be checked through this software.

Suppose we have obtained a set of genes through the analysis above, we want to know something about the regulatory networks. We can input them to STRING or other software to have a look. Let’s take this group of genes as an example: CUL3, HSP90AA1 (heat shock protein 90kDa alpha (cytosolic), class A member 1), BRCA1, ATM, NCL, PRKDC, CUL4B (cullin 4B), SMC3 (structural maintenance of chromosomes 3), BLM, SSB (Sjogren syndrome antigen B (autoantigen La)), EIF4A2 (eukaryotic translation initiation factor 4A2), TOP1, DDX3X (DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked) and RANBP2 (RAN binding protein 2), which are from “RPM data” analysis results too. As shown in figure 4, the gene network is obvious.

In this group, genes are CUL3, HSP90AA1, BRCA1, ATM, NCL, PRKDC, CUL4B, SMC3, BLM, SSB, EIF4A2, TOP1, DDX3X and RANBP2. Each sphere represents one gene. The corresponding gene name has been given aside. Lines link two spheres indicate genes are interacted. Different colors only make it easily distinguished.

**4. CONCLUSIONS**

Within the next period, high-throughput microarray data mining will still be the optimum selection for genomics research. With the further improvement in the gene chip methodology, microarray data would be more accurate. These methods of gene expression profiling microarray data analysis are very useful. They provide some references for biologists to extract information of biological significance from gene chip, and may play a guiding role in subsequent biological experiments. Nevertheless, the main task of the next step would still be how to improve the systematic and scientific analysis of data and extract information of the biological significance, which is also the fundamental purpose of microarray analysis.

Although biologists are keeping on exploiting new tools for pathway and network analysis, none of them ventures beyond identification and annotation of the bound genomic regions till date. The system for biological knowledge is gradually making the transition for life sciences research from molecular level to the system level, and research in this respect is still in their infancy. The current software and databases are based on the understanding of known biological networks, which is far from being enough compared with the entire complex biological networks. A great deal of effort to study unknown mechanism and function has left to be done thoroughly. The analysis results and predictions would be more accurate than ever only if we have more biology knowledge.
As a feature of biological research, the key to reveal biological networks is to combine as many data and the analysis as possible. Fast, reasonable analysis of data, finding the hidden information, is an important and difficult work which needs to combine with statistics, information science and other disciplines to provide new ideas and methods for extracting information from microarray data.

**ACKNOWLEDGEMENTS**

This work is partially supported by the National Natural Science Foundation of China (Grant No. U1501253, No. 51651901). Thanks for the help.

**REFERENCES**


