Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 140819, 10 pages
http://dx.doi.org/10.1155/2015/140819

Research Article

MIrExpress: A Database for Gene Coexpression Correlation in Immune Cells Based on Mutual Information and Pearson Correlation

Luman Wang,[1,2] Qiaochu Mo,[1] and Jianxin Wang[1,3]

[1]*School of Information, Beijing Forestry University, Beijing 100083, China*
[2]*Department of Natural Science in Medicine, Peking University Health Science Center, Beijing 100191, China*
[3]*Center for Computational Biology, Beijing Forestry University, Beijing 100083, China*

Correspondence should be addressed to Jianxin Wang; wangjx@bjfu.edu.cn

Received 29 May 2015; Accepted 9 November 2015

Academic Editor: Francesco Pappalardo

Most current gene coexpression databases support the analysis for linear correlation of gene pairs, but not nonlinear correlation of them, which hinders precisely evaluating the gene-gene coexpression strengths. Here, we report a new database, MIrExpress, which takes advantage of the information theory, as well as the Pearson linear correlation method, to measure the linear correlation, nonlinear correlation, and their hybrid of cell-specific gene coexpressions in immune cells. For a given gene pair or probe set pair input by web users, both mutual information (MI) and Pearson correlation coefficient (r) are calculated, and several corresponding values are reported to reflect their coexpression correlation nature, including MI and r values, their respective rank orderings, their rank comparison, and their hybrid correlation value. Furthermore, for a given gene, the top 10 most relevant genes to it are displayed with the MI, r, or their hybrid perspective, respectively. Currently, the database totally includes 16 human cell groups, involving 20,283 human genes. The expression data and the calculated correlation results from the database are interactively accessible on the web page and can be implemented for other related applications and researches.

1. Introduction

In recent years, the advance of microarray technology has provided amounts of information for us to observe the expression levels of genes together. Based on the increasing availability of gene expression data, public gene expression repositories were successfully constructed, such as the GEO [1] database and ArrayExpress [2]. These supply more opportunities to study gene expressional correlation (gene coexpression). Gene coexpression may reveal general functional tasks and regulatory mechanisms; moreover, it may identify novel genes to be involved in certain diseases. In addition, in the related fields of biology, many studies illustrated that the dependencies of gene expression can reflect the normal and dysfunctional biological processes and furthermore make us understand the underlying molecular mechanisms [3, 4]. It is difficult, however, for biologists without bioinformatics background to retrieve the gene coexpression information effectively and efficiently. For such, there in the field of plant biology are many coexpression databases, such as PLANEX [5], ATTED-II [6], Cop [7], TEGD [8], and PlantCART [9], the information in which was derived from large-scale gene expression data. Besides those, several coexpression databases peculiarly for mammals recently have been established and widely used by researchers and have thus accelerated the coexpression analysis process in the field of bioinformatics. COXPRESdb [10] was constructed with gene expression data from 63 human tissues and it utilizes the correlation rank to compare the coexpression strengths among multiple species. The database GeneFriends [11] adopted the same approach as COXPRESdb to construct coexpression maps based on transcriptome sequencing (RNA-seq) gene data instead of microarray gene data. HGCA [12] was constructed based on gene expression data from about two thousand samples of various cells and tissues. The overall correlation in gene expression was identified in

this database across multiple tissues, or mixed tissues and cells, without meeting the necessity of coexpression in the same cell type. Immuco [13] is a cell-specific database in which gene expression values in each cell type across various conditions are provided, as well as gene coexpression and correlation information. Though these databases have been constructed successfully and are able to meet users' needs to some extent, they capture only the linear coexpression relationships between different genes by the Pearson correlation coefficient (value r). In fact, r with small absolute value of two genes does not necessarily mean that the two genes are independent, since nonlinear relationship may exist in the gene coexpression data [14]. In particular, two variables with a vanishing correlation coefficient may be heavily dependent, as illustrated in the later example in this paper (see Figure 2). The mutual information (MI) is able to measure the mutual dependence of two random variables, particularly in terms of positive, negative, and nonlinear correlations [15], and in comparison with Pearson correlation coefficient, it may provide a criterion better and more general to investigate gene coexpression. And in recent years, the mutual information is regarded as a common way to detect dependencies between different genes. Steuer et al. initiated the mutual information approach [16] for one specific gene dataset to analyze intergene dependencies.

Bioconductor is an open source software which provides the key function in Affymetrix array analysis in the R software environment (http://www.r-project.org/) [17], and Meyer et al. [18] developed a package "minet" in Bioconductor, in which a powerful tool is provided to calculate the mutual information between different gene pairs. Based on a publicly available dataset *Saccharomyces cerevisiae* [19] including 2,467 genes, Butte and Kohane applied the mutual information to measure gene-gene interaction and obtained the result that the mode of MI was about 0.7. Consequently, 22 relevance networks were constructed when the threshold of information (TMI) was set to 1.3 [20]. With gene expression data from various environments, the mutual information approach [21] was employed to reconstruct regulatory networks of relationships.

In spite of the many researches and applications mentioned above about mutual information for gene correlations, few publications related to mutual information focus on immune cells. Since the mutual information should be calculated for each gene thoroughly connected to every other gene for correlation [20], the amount of correlation coefficients is tremendous and grows significantly with increasing number of genes. Thus, most publications applied the mutual information algorithms to measure coexpression on public sample datasets or testing datasets that includes much fewer genes than initial datasets.

In order to investigate the expression correlation of immune genes, we constructed a database named MIrExpress (http://wjx.bjfu.edu.cn/MIrExpress) including 41,477 probe sets for 20,283 human genes with each of the 16 cell types in immune cells to reflect the linear and nonlinear correlation of cell-specific gene coexpression profiles across multiple experimental conditions, aided by both Pearson correlation coefficient (r) and mutual information value (MI). Through a web interface, the database exhibits the scatter plot of the cooccurrence signal values of any two probe pairs to illustrate the extent and strength of correlation. For a given gene pair, not only is the MI given through the web interface, but its rank expressed in percentage is also presented in all the gene pairs, that is, about 8.6×10^8 pairs for each dataset. Besides, it is the same case for the Pearson correlation value r. Both the values and ranks of MI and r are displayed and contrasted graphically. In the querying web pages, the top 10 most relevant genes of an input gene can be listed with the perspective of Pearson correlation, mutual information, and their hybrid, respectively.

2. Materials and Methods

2.1. Data Preparation and Preprocessing. Gene Expression Omnibus (GEO) founded by National Center for Biotechnology Information (NCBI) in July 2000 is the largest public database to date for gene expression data (http://www.ncbi.nlm.nih.gov/geo/) [22]. In this paper, the SOFT format annotation files in GEO database were downloaded from the platform GPL570 for human cells. According to the SOFT files, samples related to immune cells were screened and sorted by cell types. Based on cell-specific sample ID, the raw gene expression data in CEL format files were downloaded from the GEO database using the GEOquery package [23] in R language environment, each expression data containing a single value describing the signal intensity for each probe set on the array.

In order to help improve the efficiency of the data analyzing process, the functions in the packages of Bioconductor were performed on the gene expression data. Firstly, package "simpleaffy" was used to discard the samples with extreme values in order to control the quality of raw data including scale factor, background level, percentage of genes which are called present, and $3'/5'$ ratios as the QC metrics [24]. After quality control, 6,909 human samples for 293 GEO series were selected as they were done in Immuco database [13]. Secondly, in the package "affy," MAS 5.0 algorithms including background correction, normalization, and summary were applied to generally process qualified gene expression data which are allowed for comparison among the gene expression data of samples from different experiments [25, 26]. After that, 41,477 probe sets for human organism were retained for later gene coexpression correlation analysis while about 15% of the samples were discarded due to quality control.

2.2. Calculation of Pearson Correlation Coefficient. Pearson correlation coefficient (r) is a measure of the linear correlation between two probe sets X and Y, which can be denoted by $r_{X,Y}$ and calculated as follows:

$$r_{X,Y} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}}, \quad (1)$$

where n is the number of samples from different experiments and X_i and Y_i are the expression profiles' values of probe sets X and Y in the ith sample, respectively.

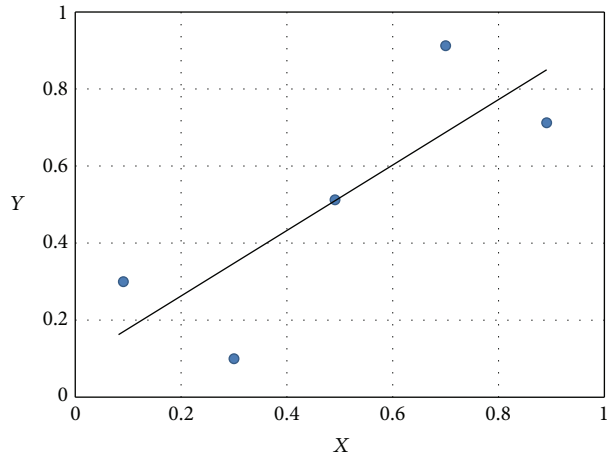


Figure 1: A scatter plot shows expression data of probe sets X and Y for dataset $[(0.1, 0.3), (0.3, 0.1), (0.5, 0.5), (0.7, 0.9), (0.9, 0.7)]$.

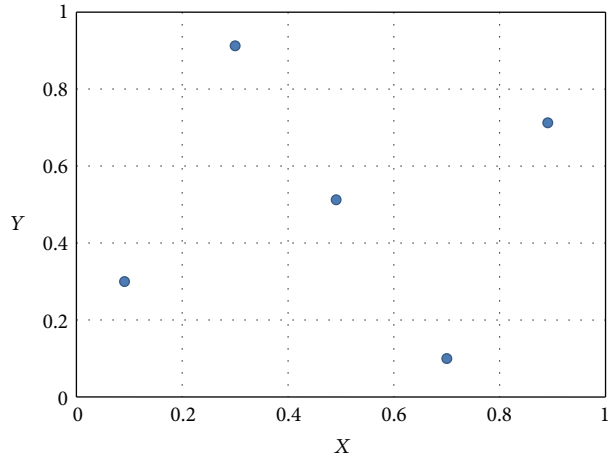


Figure 2: A scatter plot about expression data of probe sets X and Y with fixed intervals to divide the axes into discrete bins. Dataset = $[(0.1, 0.3), (0.3, 0.9), (0.5, 0.5), (0.7, 0.1), (0.9, 0.7)]$.

The r values range between −1 and +1, in which 1 implies total positive correlation, −1 total negative correlation, and 0 no correlation between the probe set pairs. The simple example in Figure 1 is a scatter plot about expression data of probe sets X and Y for a dataset, and the corresponding Pearson correlation coefficient r computed using (1) is 0.8. It indicates that there is strongly linear correlation between the probe set pairs.

2.3. Statistical Analysis about Mutual Information. The concept of entropy originates in physics, which measures the disorder of a thermodynamic systems. Shannon [27] originally devised the entropy to study the amount of information in a transmitted message and constructed information theory. So far, entropy has wide applications in various fields. Based on the theory of entropy, mutual information is applied to measure the information contained in one probe set about the other. If the mutual information of two probe sets is high, it means that it is easy to predict the expression value of one probe set according to the expression value of the other, which indicates that there may be a close relationship between genes. On the other hand, if the mutual information of two probe sets is zero, it implies that the two variables (two genes) are independent and do not correlated [14]. Based on the entropy theory, we implemented the mutual information approach to study gene coexpression.

According to the concept of mutual information, we regard a probe set as a discrete random variable and calculate the mutual information of two probe sets as the following process [21]. Suppose that A is the value range of a probe set X and A is divided by the subinterval set $\{A_i\}$, $i = 1, 2, \ldots, M$, satisfying that $\bigcup_i \{A_i\} = A$ and that $A_i \cap A_k = \emptyset$ if $i \neq k$. The entropy $H(X)$ of the probe set X can be defined as

$$H(X) = -\sum_{i=1}^{M} p(A_i) \log_2 p(A_i), \tag{2}$$

where probabilities $p(A_i)$ are approximated by the corresponding relative frequencies of occurrence in A_i and can be calculated as

$$p(A_i) \longrightarrow \frac{k_i}{N}, \tag{3}$$

where k_i denotes the number of gene expression data in the subsection A_i and N is the total number of gene expression data for the probe set X [16]. When the probability $p(A_i)$ is 1 and all other probabilities $p(A_j)$ with $i \neq j$ are zero, we get the minimum of $H(X)$, zero. In contrast, if $p(A_i) = 1/M$ for each A_i, maximum of $H(X)$ can be reached as $\log_2 M$. The joint entropy $H(X, Y)$ of two probe sets X and Y is defined as

$$H(X, Y) = -\sum_{i=1}^{M_X} \sum_{j=1}^{M_Y} p(X_i, Y_j) \log_2 p(X_i, Y_j). \tag{4}$$

Here $p(X_i, Y_j)$ denotes the joint probability that X is in subinterval set $\{A_i\}$, $i = 1, 2, \ldots, M_X$, and Y is in subinterval set $\{B_j\}$, $j = 1, 2, \ldots, M_Y$, and $p(X_i, Y_j)$ can be computed approximately as

$$p(X_i, Y_j) \longrightarrow \frac{k_{ij}}{N}. \tag{5}$$

In the above equation, k_{ij} denotes the number of gene expression data when X lies in A_X and Y in B_Y. If the probe sets X and Y are statistically independent, we can get the joint entropy $H(X, Y)$ after factorizing the joint probabilities as the following formula [16]:

$$H(X, Y) = H(X) + H(Y). \tag{6}$$

The mutual information $I(X, Y)$ between the probe sets X and Y is then defined as

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \geqslant 0. \tag{7}$$

When the probe sets X and Y are statistically independent, the mutual information $I(X, Y)$ is zero according to (6) and (7). In sum, $I(X, Y)$ can be taken as measure of correlation

no matter whether the correlation is linear or nonlinear. According to (2) and (3), (7) can be rewritten as

$$\begin{aligned} I(X,Y) &= -\sum_{i=1}^{M_X} p(X_i)\log_2 p(X_i) \\ &\quad - \sum_{j=1}^{M_Y} p(Y_j)\log_2 p(Y_j) \\ &\quad + \sum_{i=1}^{M_X}\sum_{j=1}^{M_Y} p(X_i,Y_j)\log_2 p(X_i,Y_j) \\ &= \log_2 N + \frac{1}{N}\sum_{i=1}^{M_X}\sum_{j=1}^{M_Y} k_{ij}\log_2\frac{k_{ij}}{k_i k_j}. \end{aligned} \tag{8}$$

We now use the above formula of mutual information to estimate the value about MI for dataset A in Figure 1. The dataset consists of $N = 5$ data points divided into $M_X = M_Y = 5$ bins with fixed intervals and the resulted value of mutual information $I(X,Y)$ computed using (8) is 2.322.

If we change the positions of the data points in Figure 1 and rearrange them to a state as shown in Figure 2, we will find that Y is equally dependent on X as before, because an occurrence of X is equally capable of predicting the occurrence of Y as before. That is to say, the MI remains to be 2.322 without any change. The Pearson correlation coefficient r, however, changes dramatically from 0.8 to 0, which implies that X and Y are now not linearly correlated at all. This simple example indicates that MI can generally measure the dependency including both the linear and nonlinear correlation between two probe sets and overcome the drawback of Pearson correlation that takes only the linear correlation into account.

It is a simple approach to estimate the probabilities for gene expression data occurrence in each interval by (3), but it leads to overestimating the mutual information for finite-size datasets [28]. Instead of dividing the expression data range into equal intervals, we adopted an adaptive partitioning strategy to calculate mutual information between two variables (two probe sets here) [16, 29]. It means that the value range of each probe set is divided into M discrete nonoverlapping intervals, each containing approximately N/M data points. The width of each interval is thus various according to the density of data points and more occupied regions are covered with smaller intervals. For instance, let $M = 11$ and the entropy $H(X)$ and $H(Y)$ in (2) can be described as $H(X) = H(Y) = -\log_2 11$. Consequently, the mutual information $I(X,Y)$ between probe sets X and Y can be calculated as

$$I(X,Y) = 2\log_2 11 + \sum_{i=1}^{M_X}\sum_{j=1}^{M_Y}\frac{k_{ij}}{N}\log_2\frac{k_{ij}}{N}. \tag{9}$$

3. Calculation

For improved measuring effect, the Pearson correlation coefficient and mutual information can be jointly applied to evaluate the strength of gene coexpression. The Pearson correlation coefficient r reflects the linear correlation between any two genes, while the mutual information MI generally measures the dependency of one gene on another, both linearly and nonlinearly. But the range and the distribution of values for these two measures are different ($\mathrm{MI} \in [0, +\infty)$, $r \in [-1, 1]$); thus it is not suitable to compare these two measures directly to quantify the linear and nonlinear correlation between any gene pairs [16, 30], and so we adopted the rank ordering of MI and r as coexpression measure in the MIrExpress database instead. So we need to compare ordering ranks of MI and r instead of their values; that is, we need to find the ranks of MI and r in more than 8.6×10^8 values, respectively. However, it is both space-inefficient and time-inefficient to find the rank of a given value in that large amount of values. In fact, ranks in percentage are sufficient because we do not need the rank ordering information more detailed than the percentage.

In order to get the MI rank of any gene pair, all we need now is a vector $V = \{v_i\}$, $i = 0, 1, \ldots, 100$, where v_0 is the minimum mutual information and v_{100} is the maximum one, and the mutual information of approximately 1 percent gene pairs resides between v_{i-1} and v_i, $i = 1, 2, \ldots, 100$. With the vector V, we can directly search the proper interval that a given MI resides in and thus find how many mutual information values in percentage are smaller than this MI. It is by this way that we reduce the memory consumption from about $O(8.6 \times 10^8)$ to $O(101)$.

It is easy for us to save the vector V, but difficult to obtain it, for we have no prior knowledge about the distribution of the mutual information values unless we scan the whole pairs and rearrange them. That means we have to record each information value of all the gene pairs for later use, which would consume large amount of memory. Fortunately we noticed that if the expression value range is divided into 11 intervals (as we did with MIrExpress database), the mutual information of any gene pair is between 0 and 3.5, and so we equally divided the MI value range $[0, 3.5]$ into 35,000 subintervals and counted the number of pairs whose mutual information resides in a given interval. After scanning all the gene pairs, we obtained another vector with integer values $U = \{u_j\}$, $j = 1, 2, \ldots, 35000$, where u_j is the number of gene pairs whose mutual information is in the interval $[(u_j - 1)/10000, u_j/10000)$. By using this compression technique, we reduced the memory consumption from more than $O(8.6 \times 10^8)$ to $O(35{,}000)$, yet still retaining high accuracy.

Besides the above two techniques of space-saving, we designed highly time-efficient algorithms to accelerate the coexpression analysis. Firstly, the initial values (expression data) were preprocessed and only the corresponding interval information was saved. Thus, the expression of each gene in about 1,000 samples was designated to one of the 11 intervals and the variable values needed in (9) are well-prepared. The second technique was constructing a table for the $(k_{ij}/N)\log_2(k_{ij}/N)$ part in (9). We noticed that the number of different values of $(k_{ij}/N)\log_2(k_{ij}/N)$ is no more than the number of samples that is less than 2,000. And so we saved the values in a table T indexed by the integer value k_{ij},

and thus we were able to search the table instead of computing the complex function.

With the vectors V, U and the table T well-prepared, it was relatively easy for us to calculate the MI for any given gene pair and its rank in all the MI values. Firstly, the expression value pair for each sample was divided into one element of an 11-by-11 matrix. Then, we could look up all the function values of $(k_{ij}/N)\log_2(k_{ij}/N)$ in the table T and add them up according to (9) to get the MI. Finally we would look up the MI value in the vectors V and U to get the knowledge of how many MI values are smaller than this one and at which percentage this value is located.

4. Results

4.1. The Rank of MI and r. The MIrExpress database displays a global view of cell-specific gene expression profile across different experiment conditions through two-dimensional scatter plots whose axes represent the signal values of two probe sets. The scatter plot provides database users significant intuition about the general coexpression level of two genes. As is mentioned above, it is not suitable for us to directly compare MI and r to find dependency level, linear correlation level, and linear component of the dependency relation, since their value ranges and distributions are widely different. However, it is much more reasonable if we compare their value ranks, that is, where the MI and r are located in all sorted MI values and r values, respectively. For example, if the rank of an MI is 70%, then it means that 70 percent of all the MI values are smaller than this MI value.

We denote the rank of an MI in all the sorted MI values by RoMI and that of r by Ror. In immune cells, there are 4 cases for two given probe sets when we have calculated their RoMI and Ror.

(1) Both RoMI and Ror are high. For example, the MI and r of probe sets ID 201577_at (Gene Symbol: NME1) and 1053_at (Gene Symbol: RFC2) in CD4+ T cells are 0.732516 and 0.821057, respectively, and the RoMI and Ror of these two measures are both 99%. This indicates that there exists strong linear and nonlinear correlation and coexpressed relationship between these two probe sets (Figure 3(a)).

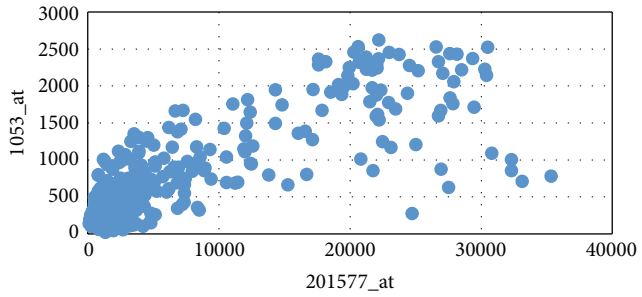
(2) The RoMI is high while the Ror is low. For example, there is strongly coexpressed relationship (total coexpression rate = 75.86%) between probe sets 1487_at (Gene Symbol: ESRRA) and 203176_s_at (Gene Symbol: TFAM) in DC cells (dendritic cells), which cannot be reflected through r value (−0.000118). If we employ MI value to measure the dependent relationship in MIrExpress database, then the MI is 0.59463 and the corresponding RoMI is 99%, a much higher rank than Ror (1%), which indicates a weak linear correlation but a strong nonlinear correlation between the probe sets (Figure 3(b)). Take CD4+ T cell, for instance, and there is the strong coexpressed relationship (total coexpression rate = 95.10%) between probe sets 219123_at (Gene Symbol: ZNF232) and 1552316_a_at (Gene Symbol: GIMAP1), which cannot be reflected by r value (−0.006148), either. But inquiring MIrExpress database, we get the RoMI and Ror as 99% and 1%, respectively (Figure 3(c)). Through these two examples we can observe that mutual information (MI) and the rank of it (RoMI) better interpret the coexpression relationship between probe sets than Pearson correlation coefficient (r) and the rank of it (Ror). So mutual information provides a more reliable and reasonable explanation of gene coexpression.

(3) Both RoMI and Ror are low. For example, the MI and r of probe sets ID 1320_at (Gene Symbol: PTPN21) and 1554627_a_at (Gene Symbol: ASCC1) in CD4+ T cells are 0.13114 and 0.00097, respectively, and the corresponding RoMI and Ror of these two measures are both 1%. It indicates that both linear and nonlinear correlation are quite weak between these probe sets (Figure 3(d)).

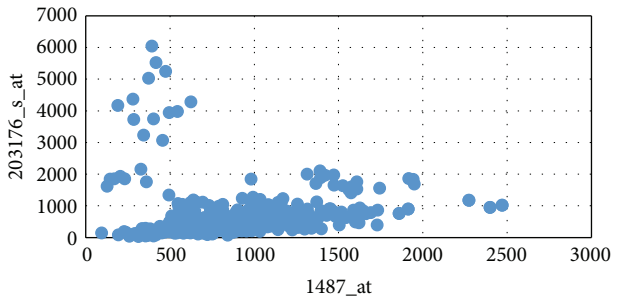
(4) The RoMI is low while the Ror is high. For example, the MI and r of probe set ID 1553169_at (Gene Symbol: LRRN4) and 234776_at (Gene Symbol: DMBX1) in CD4+ T cells are 0.13189 and 0.56614, respectively, and the corresponding RoMI and Ror are 1% and 99%, respectively (Figure 3(e)). But we notice that the absent-absent rate (AA) is 99.64% and present-present rate (PP) is 0.00%, which makes it seems true that the two probe sets are strongly linear-correlated. In fact, they are not indeed highly linear-correlated, because a high AA together with a low PP makes the Pearson correlation coefficient quite great. The low RoMI is consistent with the vanishing dependency between the two probe sets who both have low expression level in almost all samples.

4.2. Database Contents. We built the MIrExpress database (Browser/Server architecture) adopting Apache Tomcat as web server and MySQL as database server, and it provides users an easy-understanding web interface. All samples for the MIrExpress database are based on immune cells including 16 human cell groups, and the expression data of samples are chosen for Affymetrix Human Genome U133 plus 2.0 Array from GEO database. The web interface of MIrExpress database mainly includes three types of pages: page for pairwise correlation analysis (see Figure 4), page for most related genes, and page for cell-type-based overview of rank difference.

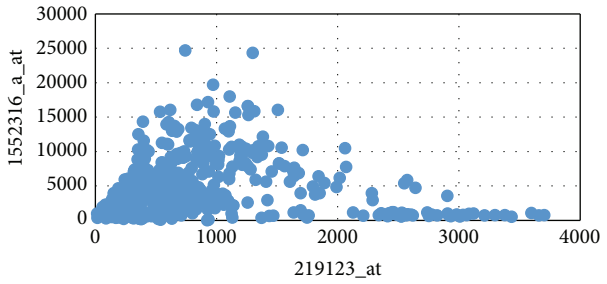
(1) Page for pairwise correlation analysis presents the general expression level of any two genes specified by users among 41,477 probe sets. Users only need to select the cell types and input two queried genes by symbol (e.g., DDR1 and RFC2) or probe sets (e.g., 1007_s_at and 1053_at) in the querying box and click the submitting button to acquire the two-dimensional scatter plot for these two genes. Meanwhile, the MI and r, together with the corresponding RoMI and Ror and their comparison, are displayed in the responding page.



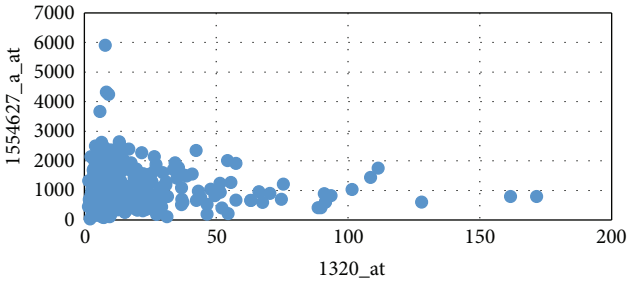
(a) CD4+ T cell; MI = 0.732516; r = 0.821057; RoMI = Ror = 99%; total coexpression rate = 88.38%



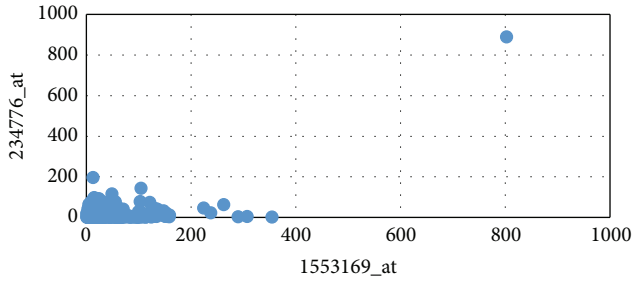
(b) DC; MI = 0.59463; r = −0.000118; RoMI = 99%; Ror = 1%; total coexpression rate = 75.86%



(c) CD4+ T cell; MI = 0.60838; r = −0.00615; RoMI = 99%; Ror = 1%; total coexpression rate = 95.10%



(d) CD4+ T cell; MI = 0.13114; r = 0.00097; RoMI = Ror = 1%; total coexpression rate = 1.09%



(e) CD4+ T cell; MI = 0.13189; r = 0.56614; RoMI = 1%; Ror = 99%; total coexpression rate = 99.64%

Figure 3: Sample applications for gene coexpression analysis. (a) NME1 and RFC2 in CD4+ T cells. (b) ESRRA and TFAM in DC cells. (c) ZNF232 and GIMAP1 in CD4+ T cells. (d) PTPN21 and ASCC1 in CD4+ T cells. (e) LRRN4 and DMBX1 in CD4+ T cells.

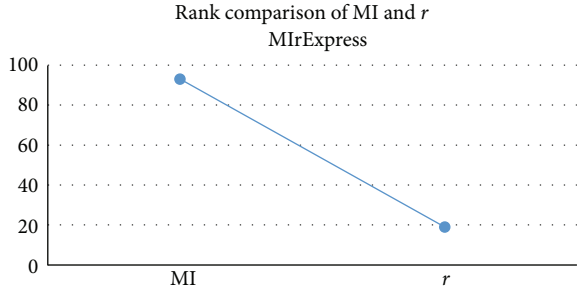
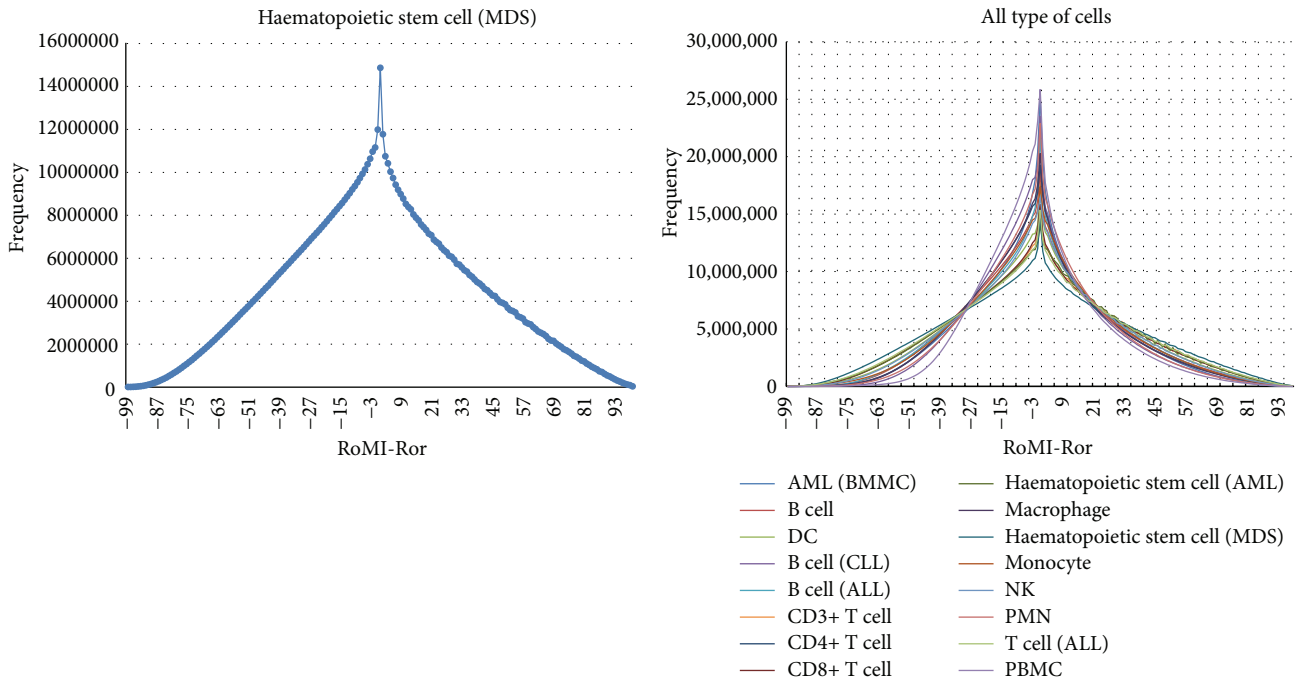


Figure 4: Page for pairwise correlation analysis. The scatter diagram of probe set pair is omitted which appears as Figures 3(a), 3(b), 3(c), 3(d), and 3(e). The species is "human"; the dataset is "CD3+ T cell"; for Gene A, the probe set ID is "1007_s_at" and the gene symbol is "DDR1"; for Gene B, the probe set ID is "1053_at" and the gene symbol is "RFC2." The Pearson's r value is −0.05538, the MI value is 1.04368, and the MIr value is 0.36477 for their hybrid.

(2) Page for most related genes lists information about the 10 most strongly correlated genes to the queried one with 3 perspectives, namely, MI, r, and their hybrid, respectively. We use MIr to denote the hybrid measure of MI and r, calculated as the follows:

$$\mathrm{MIr} = \beta \frac{r\left(X_i, X_j\right)}{\max_{k \neq i}\left(r\left(X_i, X_k\right)\right)} + \left(1-\beta\right) \frac{\mathrm{MI}\left(X_i, X_j\right)}{\max_{k \neq i}\left(\mathrm{MI}\left(X_i, X_k\right)\right)}, \tag{10}$$

where X_i is the queried gene, X_j is any other one, and β is a coefficient often set to be round 0.5 for optimum effect. For example, if a user inputs a probe set 1007_s_at of CD3+ T cell in the selected page



(a) Haematopoietic stem cell (MDS) overview of rank difference

(b) Cell-type-based overview of rank difference for all the cell types

Figure 5: Page for cell-type-based overview of rank difference.

and submits the selection, then 30 probe sets and their gene information will be retrieved, in which 10 probe sets are most related to the queried probe set according to the r, another 10 to MI, and still another 10 to MIr (Table 1).

(3) The page for cell-type-based overview of rank difference provides for each cell type an overview of how all the RoMI-Ror values are distributed. For instance, the RoMI-Ror value distribution overview of haematopoietic stem cell (MDS) type is shown in Figure 5(a), in which the horizontal axis represents the difference between RoMI and Ror, namely, RoMI-Ror, and the vertical axis represents the frequency of a given difference RoMI-Ror which occurs. And the sum of all the frequency is 860,150,026. Specifically, MI and r have the same rank when RoMI-Ror = 0, and we observe that the frequency of this rank difference is the highest. As the difference RoMI-Ror gradually increases (or decreases) from this point, the frequency that the corresponding difference occurs gradually decreases. The rank difference (RoMI-Ror) distributions of all the 16 cell types are shown in Figure 5(b).

4.3. Managing and Expanding the Database. The website and the database are totally automatic in responding the users' query if there is no abnormity. However, human operators are required to be involved to expand the database. In fact, in order to increase the visit speed of the website, we have preprocessed all the data before they are mounted into the database, and thus all newly acquired expression data should be preprocessed by human operators with the preprocessing software, individually or in batch.

5. Conclusions and Discussions

The MIrExpress database provides an effective and novel method to observe linear and nonlinear dependencies for pairwise gene expression data under a series of experiment conditions in immune cells. To date, this cannot be achieved in other related databases about correlation of gene expression. Traditionally, standard methods, such as Pearson correlation, are used to identify gene coexpression and correlation relationships. However, in some cases, coexpression relationship exists obviously but the Pearson correlation coefficient cannot reflect the dependency, which indicates that there is nonlinear correlation between gene pairs. In this paper, we took into account the rank ordering of mutual information and Pearson correlation coefficient to generally measure the gene correlation in linear and nonlinear aspects, which better describes the gene coexpressions.

There is much room for the MIrExpress database to be improved. First, much more samples may also be incorporated to enrich the database content in order to more precisely measure the correlation in the future. Second, the more kinds of cells, especially those of animals, can be incorporated into a next version of MIrExpress to more extensively reveal coexpression relationship between gene pairs. Third, a pressing need from a variety of applications is to cluster the genes according to mutual information or its variations in order to find interesting gene groups within which the genes share common functional tasks and

Table 1: Page for most related genes.

The most relevant probe sets to 1007_s_at								
Most relevant probe sets to Gene A (according to MI)			Most relevant probe sets to Gene A (according to r)			Most relevant probe sets to Gene A (according to MIr)		
Probe set	Gene symbol	Pearson's r	Probe set	Gene symbol	Pearson's r	Probe set	Gene symbol	Pearson's r
225437_s_at	C7orf27	1.26033	223460_at	CAMKK1	0.72291	219071_x_at	C8orf30A	0.76873
203028_s_at	CYBA	1.2656	202182_at	KAT2A	0.73306	1570410_at	CYGB	0.77039
229348_at	UBIAD1	1.26796	40359_at	RASSF7	0.73899	48580_at	CXXC1	0.77492
227811_at	FGD3	1.27493	222674_at	C9orf114	0.74328	40359_at	RASSF7	0.77801
206138_s_at	PI4KB	1.2802	1555866_a_at	HEXDC	0.74495	36545_s_at	SFI1	0.78158
36545_s_at	SFI1	1.28182	43977_at	TMEM161A	0.74878	229348_at	UBIAD1	0.79097
1570410_at	CYGB	1.2902	221629_x_at	C8orf30A	0.75543	43977_at	TMEM161A	0.79156
211512_s_at	OGFR	1.29953	208779_x_at	DDR1	0.75607	213681_at	CYHR1	0.79345
203419_at	MLL4	1.35051	213681_at	CYHR1	0.76656	221629_x_at	C8orf30A	0.80694
210749_x_at	DDR1	1.56079	210749_x_at	DDR1	0.90993	210749_x_at	DDR1	1

regulatory mechanisms and thus offer insights into various transcriptional and biological processes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Luman Wang and Qiaochu Mo are equally contributors.

Acknowledgments

This study was supported by National Nature Science Foundation of China (no. 31470675) and Special Fund for Forest Scientific Research in Public Welfare of China (no. 201404102).

References

[1] T. Barrett, D. B. Troup, S. E. Wilhite et al., "NCBI GEO: archive for high-throughput functional genomic data," *Nucleic Acids Research*, vol. 37, no. 1, pp. D885–D890, 2009.

[2] A. Brazma, H. Parkinson, U. Sarkans et al., "ArrayExpress—a public repository for microarray gene expression data at the EBI," *Nucleic Acids Research*, vol. 31, no. 1, pp. 68–71, 2003.

[3] A. Al-Qahtani, M. Al-Anazi, A. A. Abdo et al., "Correlation between genetic variations and serum level of interleukin 28B with virus genotypes and disease progression in chronic hepatitis C virus infection," *Journal of Immunology Research*, vol. 2015, Article ID 768470, 10 pages, 2015.

[4] N. Nagi-Miura, D. Okuzaki, K. Torigata et al., "CAWS administration increases the expression of interferon γ and complement factors that lead to severe vasculitis in DBA/2 mice," *BMC Immunology*, vol. 14, article 44, 2013.

[5] W. C. Yim, Y. Yu, K. Song, C. S. Jang, and B.-M. Lee, "PLANEX: the plant co-expression database," *BMC Plant Biology*, vol. 13, 83, 2013.

[6] T. Obayashi, K. Kinoshita, K. Nakai et al., "ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in *Arabidopsis*," *Nucleic Acids Research*, vol. 35, no. 1, pp. D863–D869, 2007.

[7] Y. Ogata, H. Suzuki, N. Sakurai, and D. Shibata, "CoP: a database for characterizing co-expressed gene modules with biological information in plants," *Bioinformatics*, vol. 26, no. 9, pp. 1267–1268, 2010.

[8] Z. Fei, J.-G. Joung, X. Tang et al., "Tomato functional genomics database: a comprehensive resource and analysis package for tomato functional genomics," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1156–D1163, 2011.

[9] M. Lescot, P. Déhais, G. Thijs et al., "PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences," *Nucleic Acids Research*, vol. 30, no. 1, pp. 325–327, 2002.

[10] T. Obayashi, S. Hayashi, M. Shibaoka, M. Saeki, H. Ohta, and K. Kinoshita, "COXPRESdb: a database of coexpressed gene networks in mammals," *Nucleic Acids Research*, vol. 36, no. 1, pp. D77–D82, 2008.

[11] S. van Dam, T. Craig, and J. P. de Magalhães, "GeneFriends: a human RNA-seq-based gene and transcript co-expression database," *Nucleic Acids Research*, vol. 43, no. 1, pp. D1124–D1132, 2015.

[12] I. Michalopoulos, G. A. Pavlopoulos, A. Malatras et al., "Human gene correlation analysis (HGCA): a tool for the identification of transcriptionally co-expressed genes," *BMC Research Notes*, vol. 5, article 265, 2012.

[13] P. Wang, H. Qi, S. Song et al., "ImmuCo: a database of gene co-expression in immune cells," *Nucleic Acids Research*, vol. 43, no. 1, pp. D1133–D1139, 2015.

[14] N. Gupta and S. Aggarwal, "MIB: using mutual information for biclustering gene expression data," *Pattern Recognition*, vol. 43, no. 8, pp. 2692–2697, 2010.

[15] I. Priness, O. Maimon, and I. Ben-Gal, "Evaluation of gene-expression clustering via mutual information distance measure," *BMC Bioinformatics*, vol. 8, article 111, 2007.

[16] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, supplement 2, pp. S231–S240, 2002.

[17] R. C. Gentleman, V. J. Carey, D. M. Bates et al., "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, article R80, 2004.

[18] P. E. Meyer, F. Lafitte, and G. Bontempi, "minet: A R/bioconductor package for inferring large transcriptional networks using mutual information," *BMC Bioinformatics*, vol. 9, article 461, 2008.

[19] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[20] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," *Pacific Symposium on Biocomputing*, vol. 5, pp. 418–429, 2000.

[21] J. Wang, B. Chen, Y. Wang et al., "Reconstructing regulatory networks from the dynamic plasticity of gene expression by mutual information," *Nucleic Acids Research*, vol. 41, no. 8, article e97, 2013.

[22] T. Barrett, S. E. Wilhite, P. Ledoux et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, no. 1, pp. D991–D995, 2013.

[23] D. Sean and P. S. Meltzer, "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor," *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, 2007.

[24] C. L. Wilson and C. J. Miller, "Simpleaffy: a BioConductor package for Affymetrix quality control and data analysis," *Bioinformatics*, vol. 21, no. 18, pp. 3683–3685, 2005.

[25] S. D. Pepper, E. K. Saunders, L. E. Edwards, C. L. Wilson, and C. J. Miller, "The utility of MAS5 expression summary and detection call algorithms," *BMC Bioinformatics*, vol. 8, article 273, 2007.

[26] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "Affy—analysis of Affymetrix GeneChip data at the probe level," *Bioinformatics*, vol. 20, no. 3, pp. 307–315, 2004.

[27] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[28] A. F. Villaverde, J. Ross, F. Morán, and J. R. Banga, "MIDER: network inference with mutual information distance and entropy reduction," *PLoS ONE*, vol. 9, no. 5, Article ID e96732, 2014.

[29] F. M. Giorgi, G. Lopez, J. H. Woo, B. Bisikirska, A. Califano, and M. Bansal, "Inferring protein modulation from gene expression data using conditional mutual information," *PLoS ONE*, vol. 9, no. 10, Article ID e109569, 2014.

[30] T. Obayashi and K. Kinoshita, "COXPRESdb: a database to compare gene coexpression in seven model animals," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1016–D1022, 2011.