Gene expression

Identifying the biologically relevant gene categories based on gene expression and biological data: an example on prostate cancer

D. Huang* and Tommy W. S. Chow

Department of Electrical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR

Received on July 10, 2006; revised on April 2, 2007; accepted on April 6, 2007
Advance Access publication April 26, 2007
Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Most gene-expression based studies aim to identify genes with the capability of distinguishing different phenotypes. Although analysis at the genomic level is important, results of the molecular/cellular level are essential for understanding biological mechanisms. To deliver molecular/cellular-level results, a two-stage scheme is widely employed. This scheme just evaluates biological processes/molecular activities individually, totally overlooking the relationship between processes/activities. This treatment conflicts with the fact that most biological processes/molecular activities do not work alone. In order to deliver improved results, this shortcoming should be addressed.

Results: We design a selection model from a novel perspective to directly detect important gene functional categories (each category represents a cellular process or a molecular activity). More importantly, the correlations between gene categories are considered. Contributed by this capability, the proposed method shows its advantages over others.

Availability: the source code in Matlab is accessible via http://www.ee.cityu.edu.hk/~tweschow/category_selection/category_selection.htm

Contact: ifkorf@ucdavis.edu

1 INTRODUCTION

Gene is the basic functional and physical unit controlling and regularizing biological phenomena and human diseases. The studies at the gene level are thus essential and important. Microarray techniques enable researchers to record the expression profiles of thousands of genes simultaneously (Ekins, 1999). With the advent of these techniques, the gene-level studies have been greatly boosted. The studies include the annotation/function prediction of genes (Zhou, 2002), the analysis of gene signaling pathways (Segal, 2003), the diagnosis and prognosis of heterogeneous diseases (Lee, 2003; Yeung, 2005) and the identification of genes responsible for complex diseases (Golub, 1999; Gui, 2005; Guyon, 2002; Zhang, 2006), etc. The latter two issues, which are related to disease understanding and prediction, have been dealt with by using various supervised machine-learning methods.

In a gene expression dataset used for disease studies, each sample has a label indicating its phenotypical status (e.g. being diseased or normal or belonging to which type of disease or how to respond to a drug). According to the principle of machine learning, we generalize label of different types as response. Obviously, the genes that vary systematically with the change of response are influential and should be paid attentions on. Through discarding the unrelated and redundant genes and just focusing on the influential genes, analysis on gene expression data can be greatly simplified into a low-dimensional data domain, which is helpful to deliver precise, reliable and interpretable results. Also, focusing on the marker genes can reduce the cost of biological experiment and decision. These are the main reasons that gene selection becomes essential or even necessary.

In a gene selection scheme, genes are considered individually, either to evaluate the gene–response association or to assess the gene–gene relationship. On the other hand, most genes cannot work alone. Instead, the functionally associated genes always collaborate with each other through interaction and reaction to regularize a cellular process at a molecular level, via a signaling pathway or under a common mechanism. A series of interplays among cellular processes or molecular activities finally determine most biological phenomena and human diseases. Given this observation, there is a need to extend disease research from a gene level to a cellular/molecular level. Also the cellular/molecular results are able to offer more insights and a better understanding of the cellular mechanisms, and thus are favored by biologists.

The functional roles of genes have been studied in a huge amount of biological experiments. The confirmed results have been summarized into databases from various perspectives. In the Kyoto Encyclopedia of genes and genomes (KEGG) database, each gene is associated with the molecular signaling pathways in which it participates (Kanehisa, 1997). In the SWISS-PORT databases, a gene is labeled with a set of keywords, such as pathways and general biological processes.
(Boeckmann, 2003). Recently, the Gene Ontology consortium develops an excellent platform to categorize genes in the context of their cellular and molecular functions (Ashburner, 2000). In the Gene Ontology (GO), gene functions are organized into three hierarchical trees. These trees are parallel to each other and respectively stand for biological process, molecular function and cellular component. A node of tree represents a functional annotation of gene. Along the direction of top-down, the range covered by annotation is gradually narrowed. For example, in the biological process tree, the root node is named as biological process. One of its descendants is labeled with growth. Down along this direction, the annotations of nodes are in turn cell growth, regulation of cell growth, etc.

With the availability of these databases, and driven by biological needs, it has become a trend to change the focus of gene-expression-based studies from gene identification to the detection of relevant molecular activities/cellular processes, behind which are functional categories of genes. To extract molecular/cellular-level results, most existing studies adopt behind which are functional categories of genes. To extract gene-expression-based studies from gene identification to the obtaining an ordered gene list, Al-Shahrour to enhance performance, sophisticated strategies were designed. (Khatri, 2005; Kim, 2003; van’t Veer, 2002). More recently, the gene category–response association becomes very challenging. To cope with this difficulty, our method first estimates the probability distribution of each category in the gene space. This part is inspired by Yu et al. (2005). For understanding this point, please refer to Figure 1 in which several probability distribution histograms of gene categories are illustrated. Based on the obtained probability estimates, the category–category and category–response associations are measured by using a newly-introduced index. Then, a forward selection process is used to identify relevant categories. The core of this process is another newly developed evaluation criterion. With this criterion, we can identify the categories that are relevant to a given task but not redundant to the already-selected categories. The proposed methods include four sequential steps, as shown in Figure 1.

Step A. Evaluate the gene–gene and gene–response associations based on gene expression data. During this course, Pearson’s correlation and mutual information MI are used.


Step C. Evaluate category–category or category–response correlations based on the probabilities obtained in Step B.

Step D. Search the influential gene functional categories in a forward way according to the correlations calculated in Step C.

2 METHODS

In order to directly select relevant gene categories, the category–response and category–category associations should be evaluated. A category, generally consisting of up to thousands genes, is represented by a high-dimensional variable. Furthermore, different category variables, covering different numbers of genes, are with different dimensionalities. With these facts, estimating category-related associations becomes very challenging. To cope with this difficulty, our model first estimates the probability distribution of each category in the gene space. This part is inspired by Yu et al. (2005). For understanding this point, please refer to Figure 1 in which several probability distribution histograms of gene categories are illustrated. Based on the obtained probability estimates, the category–category and category–response associations are measured by using a newly-introduced index. Then, a forward selection process is used to identify relevant categories. The core of this process is another newly developed evaluation criterion. With this criterion, we can identify the categories that are relevant to a given task but not redundant to the already-selected categories. The proposed methods include four sequential steps, as shown in Figure 1.

Step A. Evaluate the gene–gene and gene–response associations based on gene expression data. During this course, Pearson’s correlation and mutual information MI are used.


Step C. Evaluate category–category or category–response correlations based on the probabilities obtained in Step B.

Step D. Search the influential gene functional categories in a forward way according to the correlations calculated in Step C.

2.1 Gene–gene and gene–response correlation

Given a gene expression dataset, the correlation between genes can be measured by using many indices, such as Pearson’s correlation, Kendall’s rank correlation (Hollander, 1999) and mutual information MI (Cover, 1994), etc. Among them, Pearson’s correlation, which is reliable and efficient enough, is widely used. Also, in most categories/pathways, induced and repressed genes exist at the same time. To account for variations in either direction, we follow other researchers (Yu, 2005; Zhou, 2002), using the absolute Pearson’s correlation to measure the similarity of genes.

To assess the gene–response associations, the popular ways include signal-to-noise ratio (Golub, 1999), t-test and MI (refered as information gain in some literatures). The signal-to-noise ratio and t-test make use of the averages and the SDs of samples in different classes to fulfill the evaluation task. These criteria are reliable when the sample size is large, but they may not be effective when the sample size is small. Compared with these criteria, MI is more flexible since it can reliably reflect the arbitrary relationship of variables. We thus use MI to measure the gene–response associations. Given two variables X and Y, Shannon’s MI is defined as

$$I(X; Y) = \int_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \, dx \, dy.$$  

The above equation shows that the calculation of MI requires the estimation of $p(x,y)$, $p(x)$ and $p(y)$. To estimate these probability density functions, histogram and kernel based models (e.g. Parzen window) are commonly used. In a histogram model,
the probability of a data point is determined based on the frequency of that point appearing in a given dataset. Based on histograms, the integration operation in MI (1) can be simplified as a summation operation. With this simplification, histogram model can make the computation of MI much efficient. On the other hand, histogram is argued to become infeasible in high dimensional space due to the sparseness of data samples (Moon, 1995). This shortcoming, however, does not concern us because we only need 2D MIs. With the above considerations, we employ histograms to estimate the probabilities required by MI (1).

2.2 Probabilities of gene functional category in the gene space

Suppose that we have a gene set \( G = \{g_1, g_2, \ldots, g_M\} \), and a set of gene functional categories \( C = \{c_1, c_2, \ldots, c_K\} \). Each category consists of several genes in \( G \), i.e. \( c_i \subset G \) \((1 \leq i \leq K)\). Given a functional category \( c_i \), to evaluate its probability density in the gene space is to determine \( p(c_i|g_j) \) for \( j = 1, 2, \ldots, M \).

Given a gene, say \( g_k \), we evaluate the similarity of \( c_i \) to \( g_k \) in a way of

\[
\text{similarity} = \arg \max_{g_k \in G} s(g_k, c_i)
\]

where \( s(g_k, c_i) \), the similarity between the genes \( g_k \) and \( g_i \), is measured using the absolute Pearson’s correlation, as mentioned in the above section. Clearly, the conditional probability \( p(g_k|c_i) \) is positively related to \( s(g_k, c_i) \). That is, the more similarity between the gene \( g_k \) and the category \( c_i \), the greater the conditional probability \( p(g_k|c_i) \). We thus estimate \( p(g_k|c_i) \) according to

\[
p(g_k|c_i) \propto (1 + s(g_k, c_i))^r,
\]

where \( r \geq 1 \), for appropriately emphasizing strong correlations without enforcing a hard threshold. The similar idea has been used in other studies (Yu, 2005). Yu (2005) set \( r = 6 \) and delivered respectable results. Besides \( r = 6 \), we investigated different settings, such as \( r = 2, 10 \) and 18. \( r = 6 \) corresponded the best results in most cases and was thus used throughout this article. In the Supplementary Material, the comparative results of different values of \( r \) are presented for further reference. The probability of the category \( c_i \) at \( g_k \) is the normalized \( p(g_k|c_i) \), i.e.

\[
p(c_i|g_k) = \frac{p(g_k|c_i)}{\sum_{g_j \in G} p(g_j|c_i)}.
\]

The above probability estimation approach can be extended to the response variable \( y \) as

\[
p(y|g_k) = \frac{p(g_k|y)}{\sum_{g_j \in G} p(g_k|y)}.
\]

where \( p(g_k|y) \propto (1 + MI(g_k,y))^r \).

2.3 Category–Category Correlation and Category–response correlation

To evaluate the similarity between two categories \( c_i \) and \( c_j \), we define an index as

\[
\text{pcor}(c_i, c_j) = \sum_{g_k \in G} \sqrt{\text{MI}(g_k, c_i)\text{MI}(g_k, c_j)}.
\]

\( \text{pcor}(c_i, c_j) \) achieves the maximum when \( c_i \) and \( c_j \) have the same genome-wide distribution, that is, \( p(c_i|g_k) = p(c_j|g_k) \) for all \( k \). A small \( \text{pcor}(c_i, c_j) \) means the low similarity between \( c_i \) and \( c_j \). With these properties, \( \text{pcor}(c_i, c_j) \) are very similar to the Pearson’s correlation. We therefore call the proposed index as probability-correlation (pcor for short). As to the response \( y \) and a category \( c_i \), we have

\[
\text{pcor}(c_i, y) = \sum_{g_k \in G} \sqrt{\text{MI}(g_k, c_i)\text{MI}(y,g_k)}.
\]

Fig. 1. The block diagram of the proposed category selection method.
2.4 Forward selection process

Referring to the correlation-based feature selection index (Hall, 1999), we design the criterion \( rlv(TC, y) \) (7) to measure the classification capacity of a gene functional category set.

\[
rlv(TC, y) = \frac{\sum_{c \in TC} \text{pcor}(c, y)}{\sqrt{|TC| + \sum_{c \in TC} \sum_{c', c \neq c'} \text{pcor}(c, c')}} \tag{7}
\]

where TC is a category set and \( |TC| \) is the size of TC. \( rlv(TC, y) \) checks the quality of TC in a comprehensive way—the nominator measures the association of TC to the response, whilst the denominator evaluates the degree of the redundancy existing in TC. Obviously, a large \( rlv(TC, y) \) indicates a good TC in that the similarity of TC to the response is high and the redundancy in TC is low at the same time.

Based on \( rlv(TC, y) \), we employ a forward searching process to identify relevant categories in a way of one by one. Below, this process is detailed.

1. (Initialization) Set SC, the selected gene category set, empty.
2. (Iteration) Repeat the following steps until SC reaches the desired size.
   a. For each unselected category, say \( c \), calculate \( rlv(SC + c, y) \).
   b. Choose the category with the maximal \( rlv(SC + c, y) \). Put that category into SC.
3. (Output) Output the selected set SC.

3 RESULTS ON SYNTHETIC DATA

We compared the proposed method with the related approaches, including significance analysis of functional categories (SAFE. Barry, 2005), gene set enrichment analysis (GSEA. Mootha, 2003) and the average-based method. The first two are a typical gene-list-based scheme, in which all given genes are ranked according to their classification capabilities. Then based on the obtained gene list(s), the biological relevancy of gene category is evaluated. Both SAFE and GSEA finally estimate the significance enrichment for each gene functional category. The smaller the enrichment value, the higher relevancy of the tested category has. For SAFE, we considered the cases with \( P \)-value < 0.1 as significantly associated. For GSEA, we regarded the categories with \( q \)-value < 0.25 as biologically relevant. The average-based method extracts a hyper-gene to represent a gene category. Given a category, the expression profile of its hyper-gene is the average of the expression profiles of the genes assigned to that category. Based on the extracted hyper-genes, the selection of gene functional categories is simplified as the selection of hyper-genes. In general, any gene selection method can be employed for this issue. We chose the MI based forward searching scheme proposed by Chow et al. (2005). This MI based scheme determines the salient hyper-genes in a way of one by one. The importance of each unselected hyper-gene (say, \( hg \)) is evaluated based on \( \text{MI}(S + hg, y) \) where \( S \) is the set of the selected hyper-genes. The hyper-gene with the largest \( \text{MI}(S + hg, y) \) is selected into \( S \). This selection process repeats until certain hyper-genes have been selected. In our Supplementary Material, the details of this MI based forward search scheme are given.

We first evaluated these methods under a synthetic scenario. The expression data we used is similar to the one mentioned by Weston et al. (2000). In this dataset, the samples are evenly distributed in the class 1 and -1. The gene variable \( X_1 \) and \( X_2 \) are drawn from a mixture of normal distributions: with the probability of 0.7, \( X_1 \) and \( X_2 \) respectively come from \( \mathcal{N}(3,1) \) and \( \mathcal{N}(0,1) \) where \( y \) is the class label, that is, \( y = 1 \) or \( -1 \); with the probability of 0.3, \( X_1 \) and \( X_2 \) are respectively generated according to \( \mathcal{N}(0,1) \) and \( y \mathcal{N}(3,1) \). It can be noted that \( X_1 \), which can determine the labels of 70% samples, is more important than \( X_2 \).

Generally, in a gene expression data, there are many irrelevant genes. To simulate this, we generated 20 genes according to the distribution \( \mathcal{N}(0,1) \). These variables are just background noise and irrelevant to distinguish the samples from different classes. Also, we generated redundant genes. In detail, given \( X_1 \), we drew five genes in a way of \( X_1 + 0.2 \mathcal{N}(0,1) \). These five genes are highly redundant to \( X_1 \). Similarly, around \( X_2 \), other five redundant genes were built. After adding the 10 redundant and the 20 irrelevant genes, we finally got a 32-gene dataset \( X \). Also, suppose that, in \( X \), the 32 genes are arranged in a way of

\[
X = \{X_1, X_2, \ldots, X_6, X_7, X_8, \ldots, X_8, X_9, X_{14}, \ldots, X_{32}\}
\]

In GO, the gene categories show a hierarchical tree structure in which each node covers the genes associated to itself and all its descendants. To simulate GO tree, we constructed six gene categories, as illustrated in Figure 2. \( S_6 \) is not a representative category and was rejected from our study. Details about representative category will be detailed later. Also, based on our generation mechanism, we observe that (1) of the five tested categories, \( S_2 \), \( S_3 \), \( S_4 \) and \( S_5 \) are useful since they contain
informative genes, while $S_1$, having no informative gene, is unrelated to this classification task; (2) $S_2$, $S_3$ and $S_4$ are redundant to each other because all of them contain $X_1$ or the genes close to $X_1$. Among these categories, $S_2$ is the best due to the highest purity, i.e. $S_2$ includes no irrelevant gene and the fewest redundant genes; (3) as $X_1$ is more influential than $X_2$, the categories including $X_1$ are more important than the ones having $X_2$. According to the above observations, the correct order of category selection is $(S_2/S_3/S_4) \rightarrow S_5$. Also, the result $S_2 \rightarrow S_3$ is the most desired since $S_2$ is better than $S_1$ and $S_4$.

We generated 1000 datasets, and run the compared methods on each of them. In Table 1, the statistical results across 1000 trails are summarized. It is noted that all the methods can put $S_2/S_3/S_4$ on the first place with a high probability. Also GSEA may have bias to large categories since the probability of $S_2$ being selected out in the first place is zero. Let us see the category selected in the second place. It is known that following $S_2/S_3/S_4$ should be $S_5$. $S_5$ is useful but slightly less important than $S_2/S_3/S_4$. Such $S_5$ poses great challenge to SAFE and GSEA, as indicated by the results presented in the third column of Table 1. Actually, in all trials, SAFE always assigned $S_4$ with $P$-value $>0.1$. In the estimates of GSEA, FDR $q$-values of $S_2$ were always larger than those of $S_2$, $S_3$ and $S_4$. These estimation results means that, in all trials, both SAFE and GSEA could not correctly consider $S_2$ as the second important category.

Furthermore, we added noise genes into $S_5$ such that $S_5 = \{X_8, X_9\}$, five genes randomly selected from $(X_{13}, \ldots, X_{32})$. As shown in Table 1 (b), with the noise-mixing $S_5$, the good performance of our method can be kept, whilst the average-based method is badly affected. Through the average operation, the useful genes will be masked by noise ones. This is the direct reason that, in the average-based method, the accuracy of selecting $S_2 \rightarrow S_3$ greatly decreases to 49.4% from 61.9% obtained in the previous example.

$S_3$ is the combination of $S_2$ and seven noise genes, $\{X_{13} \sim X_{20}\}$, as shown in Figure 2. In the estimates of SAFE, the significance $P$-values of $S_2$ and $S_1$ were 0.04 and 0.63, respectively. Also, the $P$-values of the original $S_5$ (in the first example) and the noise-mixing $S_5$ (of the second example) were respectively 0.26 and 0.47 on average. These comparisons indicate that irrelevant genes may greatly undermine the performance of SAFE. To sum up, we have the following conclusions. (1) All the methods except GSEA can successfully identify the categories with great importance and high purity, such as $S_2$. (2) The proposed method and the average-based method outperform SAFE and GSEA when one is dealing with useful but slightly unimportant categories, for instance, $S_3$. (3) The proposed method and GSEA can exhibit the advantage over others on identifying noise-mixing categories, e.g. noise-mixed $S_5$.

It is worth noting that, in a real dataset, ideal categories, like $S_2$, rarely exist. Thus, the capability of handling noise-mixing categories is very important, even crucial, for gene category selection.

### 4 RESULTS ON PROSTATE DATA

#### 4.1 Data

We applied our method to a prostate cancer data. Prostate cancer is one of the most well-studied cancers. A lot of published clinical results can be used for evaluating our data analysis results. We used the gene expression dataset provided by Singh et al. (2002). This dataset, collected in the platform of Affymetrix HG U95A arrays, consists of 102 samples and records expression profiles of 9527 genes. As to gene-functional categories, we used GO annotation database (Ashburner, 2000) and KEGG gene pathway database (Kanehisa, 1997). Due to the hierarchical structure of GO, there must exists redundant GO terms. Here, a GO term is considered as redundant when most genes covered by it can be associated to their children. Redundant GO terms contain little exclusive information. To efficiently analyze GO terms, we filtered the redundant GO terms out and restricted our study to representative GO terms. We considered a term (say, $r$) as representative when, out of all genes associated to $r$, there exist genes that cannot be covered by the $r$’s descendants. Given a GO term (say $H$), assume that there are totally $w$ genes annotated either with it or with its descendants. Among these $w$ genes, $z$ genes are directly annotated with $H$ itself, while others are associated to its descendants. The uniqueness of $H$ can be calculated as $w/z$. A large uniqueness indicates that $H$ is a representative term in that most genes related to $H$ cannot be shared by $H$’s children. In this study, we considered the GO terms with uniqueness $>0.5$ as representative. Also, in order to avoid uncertainty, the gene-functional categories with less than 10 genes were out of our consideration.

![Table 1. The results on a synthetic example. A percentage value indicates the probability of the corresponding category being selected across 1000 trials.](http://bioinformatics.oxfordjournals.org/)

<table>
<thead>
<tr>
<th>(a) The results on the category setting shown in Figure 2</th>
<th>(b) The results when $S_2$ is noise mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The 1st selected gene</strong></td>
<td><strong>The 1st and 2nd selected genes</strong></td>
</tr>
<tr>
<td>Average-based method</td>
<td>$S_2$: 73.7%; $S_2$: 4.2%; $S_2$: 22.0%</td>
</tr>
<tr>
<td>SAFE</td>
<td>$S_2$: 97.0%; $S_2$: 0%; $S_2$: 0%</td>
</tr>
<tr>
<td>GSEA</td>
<td>$S_2$: 0%; $S_2$: 15%; $S_2$: 80%</td>
</tr>
<tr>
<td>Proposed method</td>
<td>$S_2$: 94%; $S_2$: 1.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>The 1st selected gene</strong></th>
<th><strong>The 1st and 2nd selected genes</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average-based method</td>
<td>$S_2$: 73.4%; $S_2$: 5.0%; $S_2$: 22.1%</td>
</tr>
<tr>
<td>SAFE</td>
<td>$S_2$: 96.0%; $S_2$: 0%; $S_2$: 0%</td>
</tr>
<tr>
<td>GSEA</td>
<td>$S_2$: 0%; $S_2$: 15%; $S_2$: 80%</td>
</tr>
<tr>
<td>Proposed method</td>
<td>$S_2$: 94.5%; $S_2$: 3.2%; $S_2$: 1.2%</td>
</tr>
</tbody>
</table>

1507
The given genes involve totally 3152 biological process GO terms. After filtering out the redundant and uncertain ones, we remained 621 representative GO terms for study. In KEGG dataset, totally 155 pathways are given. After rejecting the categories having no more than 10 genes, we finally left 124 pathways for analysis.

### 4.2 Results and evaluations

In the experiment, we first evaluated the classification capability of the selected gene functional categories. That is, based on a set of selected gene functional categories, we built the classifiers and evaluated the quality of the tested gene category set according to the performance of these classifiers. Considering that a category contains redundant or irrelevant genes, we did not employ all involved genes for building classifiers. Instead, from a selected category, we identified the best gene to represent it. The best gene here means the gene with the high classification capability. There are many ways to measure the classification capability of genes, such as signal-to-noise ratio, t-test and MI. We used MI to identify representative gene of a category (say, $C$) in a way of

$$g_C = \arg \max_{g \in C} \text{MI}(g; y).$$

Based on the selected representative genes, classifiers were built. Good performance of these classifiers indicates that the examined gene categories set is respectable. We adopted the scheme of five cross-validation (5 CV) in which the total of 102 samples were equally divided into five parts. Four of them were used for identifying the influential categories and then for constructing classifiers, whereas the other part was for testing the classifiers. Five parts of samples were used for testing in turn. In our Supplementary Material, this 5 CV evaluation framework is stated in detail. We used three types of classifiers, including $k$ nearest neighbor rule ($k$NN, and we set $k = 1$), multi-layer perceptron neural network (MLP) and support vector machine (SVM), for evaluation. The MLP and the SVM models we used are respectively available at [http://www.ncrg.aston.ac.uk/netlab](http://www.ncrg.aston.ac.uk/netlab) and [http://www.isis.ecs.soton.ac.uk/resources/svminfo](http://www.isis.ecs.soton.ac.uk/resources/svminfo). In Figure 3, the comparative results on the GO terms are illustrated. There are 30 comparison cases (3 classifiers, 10 comparative results for each classifier). In 19 out of the 30 cases, the proposed methods enabled the classifiers to deliver the best results. Also, we used the student’s t-test to compare the classification results of the proposed method with those of others. The $P$-value of the proposed method against SAFE is $3 \times 10^{-8}$, and the $P$-value of the proposed method against GSEA is $6 \times 10^{-9}$. These results suggest that our method outperform the others.

Besides the view of machine learning, we evaluated the obtained results from the biological view. Prostate cancer has been studies for many years. Many conclusions have been published and clinically confirmed. SuperArray Bioscience Corporation presents a relatively comprehensive summary in which 263 genes are suggested to be (potentially) responsible for prostate cancer ([http://www.superarray.com/gene_array_product/HTML/OHS-403.html](http://www.superarray.com/gene_array_product/HTML/OHS-403.html)). Below, these prostate-cancer-related genes are called as PCR genes.

Based on PCR genes, we extracted the PCR annotations and PCR pathways by using a popular statistics framework. Concretely, given a functional annotation, say $C$, we recorded the number of genes given in the gene expression dataset are mentioned in $C$. We also counted the number of genes, out of the 263 PCR genes, could be associated to $C$. With these values and under the scheme of hyper-geometric distribution, the significance $P$-value of $C$ enriched by the PCR gene list was computed. A small $P$-value means that $C$ is significantly related to prostate cancer. We adjusted the obtained $P$-values to account for multiple hypothesis test using a false discovery rate (FDR) scheme (Benjamini, 2001). Then, with the condition that a PCR annotation has adjusted $P$-value $< 0.05$ and contains at least five genes, we obtained 58 PCR annotations in total. Similarly, we identified 13 PRC pathways. The 13 PCR pathways and the 58 PCR annotations are the basis of our evaluation. Our PCR annotation list includes many widely-recognized PCR cellular processes, for example, the apoptosis.

Fig. 3. Comparisons in terms of classification results.
related ones (GO:0006916 anti-apoptosis and GO:0008637 apoptotic mitochondrial changes), the insulin related ones (GO:0008286 insulin receptor signaling pathway), the cell growth/proliferation related ones (GO:0008284 positive regulation of cell proliferation, GO:0016049 cell growth). Our PCR pathway list also contains the pathways that have been stated to play important roles in prostate cancers, such as apoptosis (path:hsa04210), MAPK signaling pathway (path:hsa04010) and Focal adhesion (path:hsa04510). In the Supplementary Material, the full lists of PCR annotation and pathways are given.

To evaluate a selection result, we calculated the amount of the PCR annotation/pathway included. In the cases that GO annotations are involved, the hierarchical relationship between annotations was considered. That is, a category was marked as PCR, either when it is a PCR annotation or when it is the no-more-than-2-generations ancestor/descendant of a PCR annotation. Clearly, the more PCR items a selection result has, the better that result is. In Figure 4, the comparative results are illustrated. By these results, the advantages of our method can be clearly indicated. Also, in Tables 2 and 3, detailed results of our method are listed. The PCR items are marked by a symbol of asterisk. The obtained results are remarkably encouraging since many selected annotations/pathways are either in the PCR annotation/pathway list or supported by the published clinical results.

### Table 2. The selection results of our method on GO annotations

<table>
<thead>
<tr>
<th>Selection</th>
<th>GO annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>* GO:0016049 cell growth</td>
</tr>
<tr>
<td>2</td>
<td>GO:0007262 STAT protein nuclear translocation; abnormal activity of certain STAT family members is associated with prostate cancers (Turkson, 2000)</td>
</tr>
<tr>
<td>3</td>
<td>* GO:0000209 Protein polyubiquitination</td>
</tr>
<tr>
<td>4</td>
<td>GO:0006457 protein folding</td>
</tr>
<tr>
<td>5</td>
<td>GO:0000097 sulfur amino acid biosynthesis</td>
</tr>
<tr>
<td>6</td>
<td>* GO:0016337 cell–cell adhesion</td>
</tr>
<tr>
<td>7</td>
<td>GO:0006334 nucleosome assembly</td>
</tr>
<tr>
<td>8</td>
<td>* GO:0051301 cell division</td>
</tr>
<tr>
<td>9</td>
<td>* GO:0001658 regulation of cell growth</td>
</tr>
<tr>
<td>10</td>
<td>* GO:0006916 anti-apoptosis</td>
</tr>
<tr>
<td>11</td>
<td>* GO:0015671 protein ubiquitination</td>
</tr>
<tr>
<td>12</td>
<td>* GO:0001501 skeletal development</td>
</tr>
<tr>
<td>13</td>
<td>GO:0006936 muscle contraction</td>
</tr>
<tr>
<td>14</td>
<td>GO:0015671 oxygen transport</td>
</tr>
<tr>
<td>15</td>
<td>GO:0015669 gas transport</td>
</tr>
<tr>
<td>16</td>
<td>* GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway</td>
</tr>
<tr>
<td>17</td>
<td>GO:0007076 mitosis</td>
</tr>
<tr>
<td>18</td>
<td>GO:0033497 chromatin assembly</td>
</tr>
<tr>
<td>19</td>
<td>* GO:0001658 regulation of cell growth</td>
</tr>
<tr>
<td>20</td>
<td>* GO:0016049 humoral defense mechanism (sensu Vertebrata)</td>
</tr>
<tr>
<td>21</td>
<td>* GO:0008015 circulation</td>
</tr>
</tbody>
</table>

### Table 3. The selection result of our method on KEGG pathways

<table>
<thead>
<tr>
<th>Selection order</th>
<th>Pathway name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>path:hsa00960 Alkaloid biosynthesis II</td>
</tr>
<tr>
<td>2</td>
<td>* path:hsa04510 Focal adhesion</td>
</tr>
<tr>
<td>3</td>
<td>* path:hsa0410 MAPK signaling pathway</td>
</tr>
<tr>
<td>4</td>
<td>path:hsa04220 Calcium signaling pathway</td>
</tr>
<tr>
<td>5</td>
<td>* path:hsa04530 Tight junction</td>
</tr>
<tr>
<td>6</td>
<td>* path:hsa04810 Regulation of actin cytoskeleton</td>
</tr>
<tr>
<td>7</td>
<td>path:hsa04910 Insulin signaling pathway</td>
</tr>
<tr>
<td>8</td>
<td>path:hsa05050 Dentatorubropallidoluysian atrophy (DRPLA)</td>
</tr>
<tr>
<td>9</td>
<td>path:hsa0230 Purine metabolism</td>
</tr>
<tr>
<td>10</td>
<td>path:hsa04630 Jak-STAT signaling pathway; the effects of IL-6 on prostate cancer cell growth are mediated through... (Lou, 2000)</td>
</tr>
</tbody>
</table>
5 CONCLUSIONS

In this article, we present a method to identify the relevant gene functional categories. In this method, the category–category relationship is taken into account. Contributed by this capability, the proposed approach can deliver the improved results. Our method currently employs a correlation-based index to evaluate a category set. In the further work, we will develop new evaluation indices based on sophisticated concepts, such as MI, in order to further enhance the selection performance.

ACKNOWLEDGEMENTS

The work described in this article was wholly supported by a grant by CityU of Project No. 7001998-570.

Conflict of Interest: none declared.

REFERENCES


