This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2618927, IEEE/ACM Transactions on Computational Biology and Bioinformatics

GSEH: A novel approach to select prostate cancer-associated genes using gene expression heterogeneity

Hyunjin Kim, Sang-min Choi, Sanghyun Park*

Department of Computer Science, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea

* Corresponding author. Tel.:+82 2 2123 5714; fax:+82 2 365 2579; e-mail: sanghyun@yonsei.ac.kr

Abstract— When a gene shows varying levels of expression among normal people but similar levels in disease patients or shows similar levels of expression among normal people but different levels in disease patients, we can assume that the gene is associated with the disease. By utilizing this gene expression heterogeneity, we can obtain additional information that abets discovery of disease-associated genes. In this study, we used collaborative filtering to calculate the degree of gene expression heterogeneity between classes and then scored the genes on the basis of the degree of gene expression heterogeneity to find "differentially predicted" genes. Through the proposed method, we discovered more prostate cancer-associated genes than ten comparable methods. The genes prioritized by the proposed method are potentially significant to biological processes of a disease and can provide insight into them.

Index Terms— Gene selection, Gene prioritization, Disease-associated genes, Prostate cancer-associated genes, Gene expression heterogeneity

1 Introduction

THE average life expectancy of human has increased throughout the world on account of advancements in medical science [1]. With large gains in life expectancy, a rising interest exists in disease management. If the diagnosis and prognosis are precisely predicted, the correct therapeutic methods can be used and significant disease damage can thereby be avoided. Indicators (biomarkers), such as genes or proteins, are typically used in predicting the diagnosis and prognosis of a disease [2-3]. Many biologists must choose which genes or proteins to investigate; therefore, gene prioritization has become increasingly important. Four computational strategies for gene prioritization exist [4]: filtering, text mining, similarity profiling and data fusion, and network-based. In the filtering strategy, filters are defined by properties of the ideal candidate gene. In the text-mining strategy, disease-relevant keywords are employed to retrieve disease-relevant literature, which is mined to identify candidate genes. In the similarity profiling and data fusion strategy, similarities between the candidate genes and known genes from various data sources are considered. In the network-based strategy, candidate genes in a gene network are selected based on the distance between the candidate genes and known disease genes. The proposed method is categorized as a filtering strategy because it employs a filter defined by heterogeneous gene expression characteristics.

Genes that are differentially expressed between two different conditions (i.e., malignant and benign) have received considerable attention because they are expected to predict the diagnosis and prognosis of the disease [5-6]. Feature selection methods can be used to identify genes that are differentially expressed between the two different conditions. In bioinfor-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2618927, IEEE/ACM Transactions on Computational Biology and Bioinformatics

matics field, they are generally used in classification problems since they can increase prediction performance and reduce data dimensions [7]. The features chosen by feature selection methods increase prediction performance, which indicates that those features have the characteristics that can distinguish between the conditions. This supports the fact that feature selection can identify differentially expressed genes.

A typical approach for feature selection is a method using conditions (class labels) of samples. The worth of an attribute can be evaluated by calculating the value of the chi-square statistic [8] with respect to classes of data. A chi-square value is calculated using the difference between the observed frequency and expected frequency between an attribute and a class. The larger the chi-square value is, the more interdependent the attribute and class become. This can imply that there is a strong connection between the attribute and class. Several methods of feature selection based on information theory exist. Information gain [9] chooses an attribute by comparing information before and after the classification. Assuming that the total information is given, the information gain is the amount of decreased information after being classified into the attribute. The larger the gain, the better the attribute is. However, information gain is biased towards choosing attributes which have various and diverse types of values. Gain ratio [10] has been suggested as a remedy for the problem. Although gain ratio is similar to information gain, gain ratio overcomes the bias of information gain by normalizing the information gain using the split information. Symmetrical uncertainty [11], which divides the information gain by the sum of the information of variables, can compensate for the weakness of information gain. The imbalance resulting from not normalizing information gain can be solved if the information gain is divided by the sum of the information of the attribute and class. The Relief-A [12] feature selection method does not analyze the correlation between attributes and classes; rather, it analyzes the characteristics of attributes. It assumes that, if an attribute is useful, the attribute values of samples belonging to the same class become similar, but the attribute values of samples belonging to the different class have a different pattern. Relief-A finds the nearest neighbor (nearest hit) within the same class in terms of the Euclidean distance, and the nearest neighbor (nearest miss) in other classes; it then evaluates the importance of the attribute. However, because the method becomes vulnerable to noise on account of finding only the one nearest neighbor and produces the wrong outcome, it finds *k* neighbors (*k* is a number the user selects) and it utilizes the average value of the neighbors as an attribute weighting.

In gene expression analysis field, some statistical methods are widely used. The simplest statistical method for discovering differentially expressed gene is the t-test [13]. The t-test examines whether two conditions of data are significantly different from each other or not based on an assumption that the data is normally distributed. Limma [14] is one of the most powerful models for detecting differential gene expression. Limma is especially good for data with small number of samples because limma is similar to the t-test but it pools information across other genes to moderate the standard errors.

Gene selection is a method for identifying differentially expressed genes, which can play the role of biomarkers in predicting the diagnosis and prognosis of a disease. However, the degree of differential expression is not necessarily biologically meaningful [15]. Therefore, it is important to identify genes relating to biological processes of a disease rather than differentially expressed genes that are helpful for classifying disease conditions. CV (Correlation Vector) method [16] is used to identify differentially correlated genes under different conditions, but not differentially expressed genes under different conditions. For example, CV is used to identify genes that differ in their degree of correlation with other genes between Class 1 and Class 2. CV can identify potentially important genes that have not been identified by traditional methods using the differentially correlated approach.

We propose a novel gene selection method GSEH (Gene Selection using Expression Heterogeneity) that employs gene expression heterogeneity to identify genes relating to biological processes of a disease. The gene expression heterogeneity signifies that samples in the same class can have dissimilar gene expression levels. Specifically, Gene expression levels from one class can have various gene expressions while gene expressions from the other one have similar values. The class indicates a label of the samples (e.g., tumor, normal). The concept of gene expression heterogeneity can be used as beneficial information to discover disease-associated genes. The collaborative filtering method, which is often used in recommendation systems, is employed in the proposed method to estimate gene expression heterogeneity. The greater the degree of heterogeneity, the more difficult is the prediction task. Therefore, it can be estimated that the greater the difference in predicted levels, the more closely the gene relates to a disease. GSEH uses the “predictability” of gene expressions between two conditions to select disease-associated genes. GSEH is not intended to replace pre-existing methods; rather, it is intended to provide additional information for discovering genes that are related to diseases.

There are some methods which consider gene expression heterogeneity [17-20]. Tomlins et al. [17] proposed cancer outlier profile analysis (COPA) because in many of cancer datasets, heterogeneous patterns of oncogene activation have been observed. The COPA employs a simple approach based on the median and median absolute deviation of gene expression datasets. They also implemented COPA as part of Oncomine database (www.oncomine.org). MacDonald et al. [18] implemented the COPA in an R package because COPA on Oncomine is not extensible and limited to analyze significance of a specific outlier. Leek et al. [19] introduced surrogate variable analysis to capture gene expression heterogeneity. They treated the heterogeneity as a noise and tried to reduce the heterogeneity to obtain surrogate variable without the heterogeneity. Wang et al. [20] proposed modified cancer outlier profile analysis (mCOPA) because original COPA considers only up-regulated outliers. They considered both up-regulated outliers and down-regulated outliers to accurately identify gene expression heterogeneity.

As mentioned in the similar studies above, gene expression heterogeneity is a crucial factor in gene expression analysis.

We investigated related methods to compare with GSEH (Table 1).

Table 1. A summary of GSEH and similar methods

Methods	Background	Rationale	Characteristic	Result
Chi-square	Chi-squared statistic	Dependency between attributes and classes	Calculates correlation between attributes and classes	Genes correlated with classes
Information Gain	Information theory	Information before and after classification	Bias problem	Genes related with classification
Gain Ratio			Normalized version of information gain	
Symmetrical Uncertainty				
Relief-A	Nearest neighbor	Distance between target attribute and neighbors	Vulnerable to noise	Differentially valued genes
CV	Correlation vector	Difference in the degree of correlation with other genes	Utilizes correlation information as a new selection criterion	Differentially correlated genes
t-test	t-statistic	Statisical difference between two groups	Vulnerable to small sample size	Differentially expressed genes
Limma			Utilizes empirical Bayes method to moderate the standard error	
COPA	Outlier profile analysis	Gene expression heterogeneity	Numerical transformation based on median and median absolute deviation	Highly overexpressed genes
GSEH	Recommendation system		Uses collaborative filtering	Differentially predicted genes

GSEH employs two steps to select genes (Figure 1). The first step is to create a new matrix using collaborative filtering. Collaborative filtering selects samples with a similar pattern utilizing their correlation; it then calculates expected scores of the target genes of the samples using the selected samples. The second step is to calculate each gene's prioritization score by comparing the data produced in the first step with the original data; it then selects genes based on their scores. The greater the difference in the predicted degree level between the two conditions, the higher the score becomes. Finally, the genes with high scores are selected.

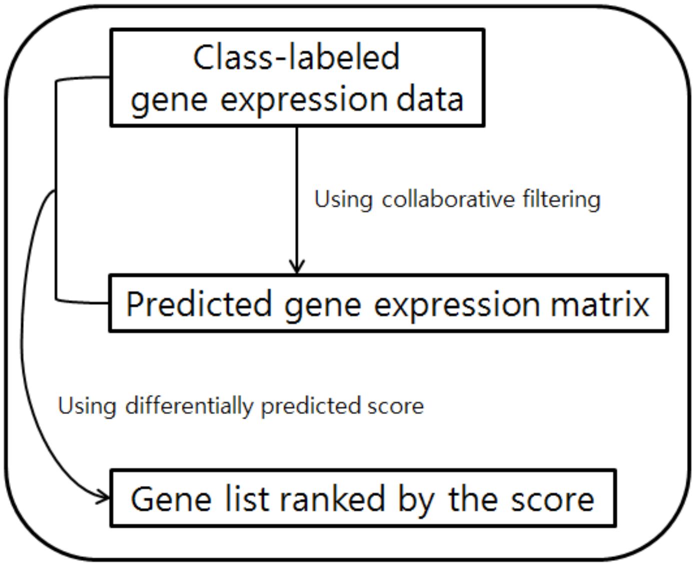

Class-labeled gene expression data
Using collaborative filtering
Predicted gene expression matrix
Using differentially predicted score
Gene list ranked by the score

Figure 1. Flow of the GSEH algorithm

2 METHODS

GSEH employs collaborative filtering to select biologically meaningful candidate genes. The GSEH process is generally divided into two phases. The first phase involves constructing a predicted gene expression matrix using collaborative filtering; the second phase involves calculating the rank scores of the genes using a comparison between the predicted gene expression matrix and original gene expression matrix. When the calculation of the scores is complete, the genes can be ranked in order and k top-ranking genes can be selected. A formal description of GSEH is outlined in Algorithm 1.

2.1 Materials

Datasets applied in this study were Singh [21], GSE15484, and TCGA_PRAD (The Cancer Genome Atlas: Prostate Adenocarcinoma). The Singh dataset is comprised of 52 prostate cancer samples and 50 benign samples; each sample has 8,828 gene expression levels. GSE15484 is also a gene expression dataset from prostate cancer patients; it is registered in the GEO (Gene expression Omnibus) database of the NCBI (National Center for Biotechnology Information). GSE15484 contains 25 samples with a Gleason score of 6, 27 samples with a Gleason score of 8 through 10, and 13 benign samples. We performed GSEH with GSE15484 dataset with two conditions; (high risk vs low risk) and (cancer vs benign). The 6-Gleason-scoring samples are considered to be low risk (non-aggressive) and the 8, 9, and 10-Gleason-scoring samples are considered to be high risk (aggressive). TCGA _PRAD is prostate adenocarcinoma RNA-Seq data from TCGA[1]. It contains 297 tumor samples and 50 normal samples which are normalized by RSEM (RNA-Seq by Expectation Maximization). Because the number of samples in each class in TCGA_PRAD is largely dissimilar and a lot of samples cause high time complexity, we ran-

[1] The Cancer Genome Atlas (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm)

domly chose the same number of samples as smaller number in the larger number of class (50:50), iteratively tested 10 times and computed the average as a result in TCGA_PRAD dataset. Singh and TCGA_PRAD datasets are related to prostate cancer diagnosis and GSE15484 dataset is related to prostate cancer diagnosis and prognosis. The datasets and programming code of GSEH are available at (http://embio.yonsei.ac.kr/files/hjkim/gseh.zip).

```
Input: Gene expression data $OM(i \times j)$, Pearson correlation coefficient threshold $t$
Output: List of ranked genes

For each sample $s$ from data $OM(i \times j)$, Do
  For each sample $s'$ from data $OM(i \times j)$ except $s$, Do
    Calculate Pearson correlation coefficient $p$ between $s$ and $s'$ in the same class
    If $p \geq t$, Then add sample $s'$ to neighbor list of sample $s$
  End For
  For each gene $g_i$ from sample $s$, Do
    Calculate predicted gene expression of $g_i$
  End For
  Construct predicted gene expression matrix $PM(i \times j)$ with predicted gene expressions
End For
For each class $l$, Do
  For each gene $g_i$, Do
    For each sample $s_j$ in class $l$, Do
      Compute matrix difference $d$ of gene $g_i$ for each class $l$
    End For
  End For
End For
Calculate rank score $r_i$ of gene $g_i$ by using $d$ of each class
```

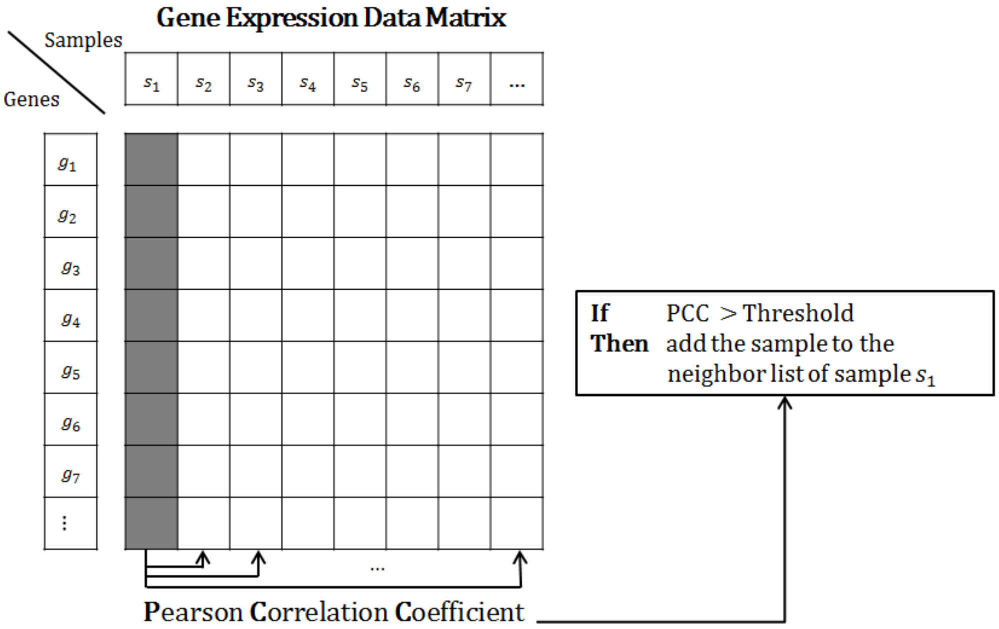
Algorithm 1. The algorithm of GSEH

2.2 Predicted Gene Expression Matrix Construction

Gene expression data can be reconstructed by collaborative filtering [22-24], which is commonly used in recommendation systems. Collaborative filtering takes various forms; in this study, we employed user-based collaborative filtering. User-based collaborative filtering is comprised of two steps. The first step involves selecting neighbor samples for a given sample. A neighbor sample indicates a sample that has characteristics similar to the given sample and the Pearson correlation coefficient is used as a selection criterion in this method. Pearson correlation coefficient P_{xy} can be described as follows:

$$P_{xy} = \frac{\mathrm{cov}(X,Y)}{\sigma_x \cdot \sigma_y} = \frac{\sum[(X_i - \overline{X}) \cdot (Y_i - \overline{Y})]}{\sqrt{\sum(X_i - \overline{X})^2} \cdot \sqrt{\sum(Y_i - \overline{Y})^2}} \qquad (1)$$

In the equation, X is the given sample for prediction, and $\overline{X}$ is the average of gene expressions for X. σ_x is the standard deviation of the average gene expression for X. X_i indicates the i-th gene expression of sample X. Y is the remaining samples excluding sample X. The Pearson correlation coefficient should be a real number between -1 to 1. The value closest to 1 can be considered a similar sample, and the value closest to -1 is a negatively similar sample. The value closest to 0 is a dissimilar sample. To predict gene expression levels using samples that have similar gene expression patterns, the samples that have a positive relation with the given sample as neighbors should be chosen using the Pearson correlation coefficient (Figure 2). The neighbors of the target sample are selected among the other samples in the same class. To be more specific, the correlations among the target sample and the other samples from the same class are calculated; samples with a Pearson correlation equal to or greater than threshold c are chosen as neighbors. If we repeat this process for all cells in the dataset, we can determine the neighbors of all the cells.



If PCC > Threshold
Then add the sample to the neighbor list of sample s_1

Figure 2. Process of selecting neighbor samples using the Pearson correlation coefficient

The second step is to predict the gene expression level of a given cell based on the neighbor's gene expression levels. The gene expression level of the given target cell is predicted based on the gene expression levels of the samples that have a larger Pearson correlation coefficient than threshold *c*. Collaborative filtering approach produces a prediction with sum of weighted average of neighbor samples. A predicted gene expression level for *i*-th row and *j*-th column V_{ij} is described as follows:

$$V_{ij} = \overline{S_j} + \frac{\sum_{S_n \in Neighbor}((E_{in} - \overline{S_n}) \cdot P_{S_j S_n})}{\sum_{S_n \in Neighbor} |P_{S_j S_n}|} \quad (2)$$

In the equation, V_{ij} is a predicted expression value for *i*-th row and *j*-th column. $\overline{S_j}$ is the average expression in all the genes of the sample S_j, and *Neighbor* is a set of neighbor samples of the sample S_j. S_n is one of the neighbor samples, and E_{in} is the *i*-th gene expression level of the neighbor sample S_n. $\overline{S_n}$ is the average expression of all the genes of the neighbor sample S_n, and $P_{S_j S_n}$ indicates the Pearson correlation coefficient between the sample S_j and neighbor sample S_n. If the sample S_j has no neighbors, $V_{ij} = E_{ij}$, which indicates that the predicted value has the same gene expression as the original value and that there is no prediction. We can predict the expression level for a certain gene of a given sample using this equation. For example, we can predict the gene expression level of gene g_1 and sample S_1 using gene expression levels of the neighbor samples of S_1 (Figure 3).

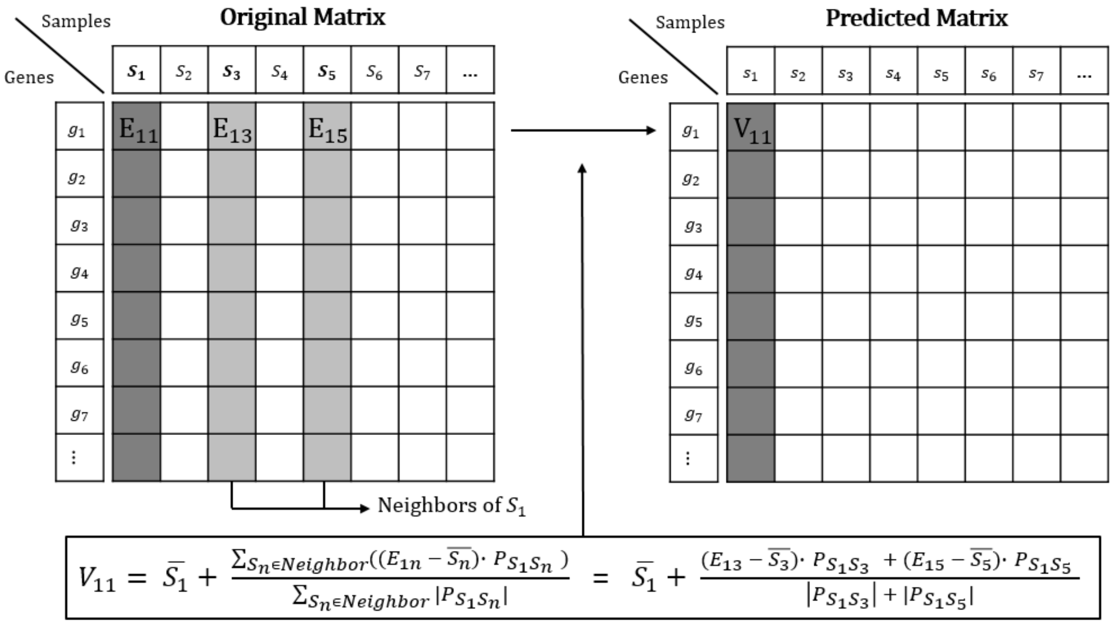

Original Matrix
Samples
Genes
s_1 s_2 s_3 s_4 s_5 s_6 s_7 ...
g_1 g_2 g_3 g_4 g_5 g_6 g_7 ⋮
E_{11}
E_{13}
E_{15}
Neighbors of S_1
Predicted Matrix
Samples
Genes
s_1 s_2 s_3 s_4 s_5 s_6 s_7 ...
g_1 g_2 g_3 g_4 g_5 g_6 g_7 ⋮
V_{11}
$V_{11} = \overline{S_1} + \frac{\sum_{S_n \in Neighbor}((E_{1n} - \overline{S_n}) \cdot P_{S_1 S_n})}{\sum_{S_n \in Neighbor} |P_{S_1 S_n}|} = \overline{S_1} + \frac{(E_{13} - \overline{S_3}) \cdot P_{S_1 S_3} + (E_{15} - \overline{S_5}) \cdot P_{S_1 S_5}}{|P_{S_1 S_3}| + |P_{S_1 S_5}|}$

Figure 3. An example of calculating a predicted value and constructing predicted matrix

Expression levels of all the genes and all the samples can be predicted by employing the collaborative filtering described above. If we apply the equation to all the cells in the original matrix, we will have a predicted matrix with predicted gene expression levels. The predicted gene expression levels are used to prioritize genes in the second phase of GSEH.

2.3 Gene Prioritization

After the construction of the predicted gene expression matrix, we can compute a prioritization score of genes using a comparison of the predicted gene expression matrix and original gene expression matrix (Figure 4).

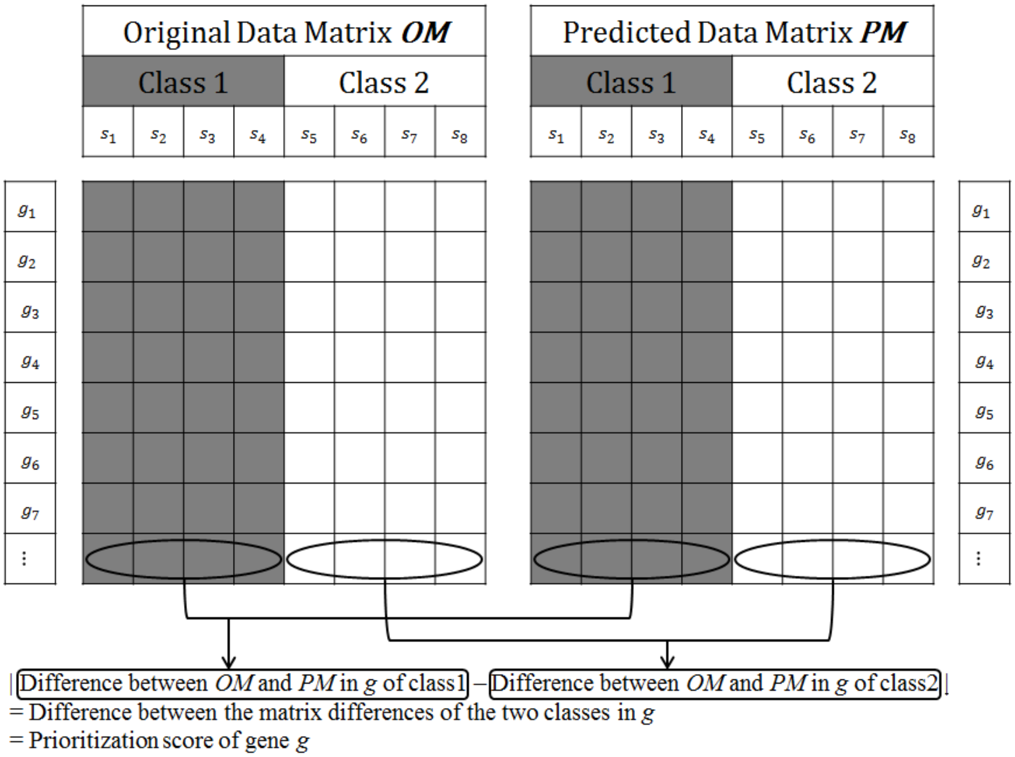

Original Data Matrix OM
Class 1
Class 2
s1 s2 s3 s4 s5 s6 s7 s8
Predicted Data Matrix PM
Class 1
Class 2
s1 s2 s3 s4 s5 s6 s7 s8
g1 g2 g3 g4 g5 g6 g7 ⋮
g1 g2 g3 g4 g5 g6 g7 ⋮
|Difference between OM and PM in g of class1 − Difference between OM and PM in g of class2|
= Difference between the matrix differences of the two classes in g
= Prioritization score of gene g

Figure 4. Calculation process of the gene prioritization score

The predicted gene expression matrix, which is constructed by using collaborative filtering, has "predicted values", which can be compared to the original values of the same region in the original gene expression matrix. In this study, if the difference between the predicted expression levels of a gene and the original expression levels of the gene in a class, as well as the difference between the predicted expression levels of the gene and the original expression levels of the gene in the other class, are dissimilar, we assume that the gene has a high possibility of having biological meaning with respect to the disease. Prioritization score R_i of the i-th gene can be described as follows:

$$R_i = \left| \left(\frac{\sum_{j=1}^{m} | OM_{1ij} - PM_{1ij} |}{m} - \frac{\sum_{j=1}^{n} | OM_{2ij} - PM_{2ij} |}{n} \right) \right| \quad (3)$$

* OM_{1ij} = Expression level of the *i*-th gene and the *j*-th sample of Class 1 in Matrix OM
* OM_{2ij} = Expression level of the *i*-th gene and the *j*-th sample of Class 2 in Matrix OM
* PM_{1ij} = Expression level of the *i*-th gene and the *j*-th sample of Class 1 in Matrix PM
* PM_{2ij} = Expression level of the *i*-th gene and the *j*-th sample of Class 2 in Matrix PM
* m = Number of samples in Class 1, n = Number of samples in Class 2

The equation calculates the extent of dissimilarity between the two matrices between the two classes. In short, R_i indicates the difference between the matrix differences of the two classes. The greater the difference between the two classes, the greater the prioritization score is. A large difference between the two classes of a gene indicates that the gene expression prediction in the given class and the gene expression prediction in the other class are dissimilar. Therefore, we can assume there is a possibility that the gene has a relation with biological processes of the disease. Accordingly, if genes are ranked in order with respect to the prioritization score, the top-scoring genes are potentially significant to biological processes and can be efficiently used as biomarkers.

3 RESULTS

For the experimental environments, we used an Intel® Core™ i3 530 Dual 2.93 GHz, 8 GB RAM machine with the Windows 7 operating system. GSEH was implemented in the Java programming language with JDK 7. We performed an experiment to evaluate the proposed GSEH. The main purpose of GSEH is to provide additional information for discovering genes that are related to biological processes of a disease. Therefore, we compared how many disease-associated genes were discovered in top-ranking genes (Figure 5).

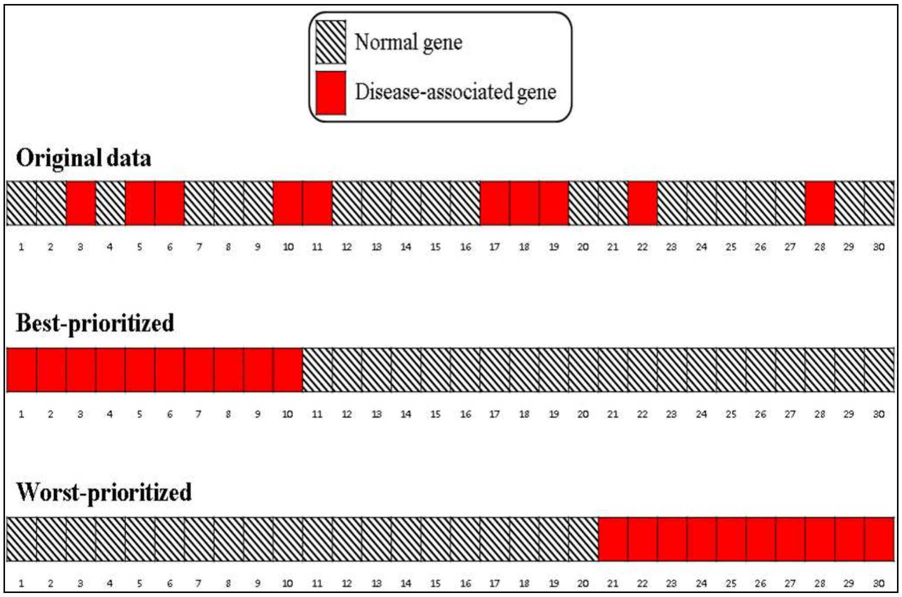


Figure 5. Examples of gene prioritization

For example, Singh dataset has 1021 prostate cancer-associated genes among 8828 genes. If we select the top 500 genes from the non-prioritized Singh dataset, we can expect there are 58 prostate cancer-related genes (500 x 1021 / 8828) among 500 genes. We can say our method is meaningful when our method finds more prostate cancer-associated genes among top-ranking genes than random selection approach or other methods. Ozgur [25] similarly performed a detailed evaluation of the 20 top-ranking genes by finding evidence of their association to the disease to prove the efficiency of his method. Moreau [4] introduced statistical benchmarking, which evaluates how well a method discovers known disease-gene associations for assessment of the gene prioritization. Because the Singh, GSE15484, and TCGA_PRAD datasets are prostate cancer-associated gene expression data, we analyzed how many prostate cancer-associated genes are included among top-scoring genes. The experiment was to find prostate cancer-associated genes among top-scoring genes based on two answer sets (Table 2). We searched the prostate cancer-associated genes in the top-ranking genes selected from GSEH for validation.

Table 2. Two answer datasets used in this study

	Data name	Number of prostate cancer genes	
Answer set 1	GeneRIF[2]	1,324	
Answer set 2	OMIM[3]	18	845
	DDPC[4]	703	
	PGDB[5]	124	

Before selecting prostate cancer-associated genes with GSEH, Pearson correlation coefficient threshold *c* should be determined. It is important to choose a Pearson correlation coefficient threshold because there is trade-off relation between low and high thresholds. Because our method employs a user-based collaborative filtering approach, if the correlation coefficient threshold is too low, the “dissimilar” neighbors can be chosen to predict gene expression levels and performance of GSEH can be worse. Otherwise, if the correlation coefficient threshold is too high, there can be no neighbor, which indi-

[2] Gene Reference Into Function (ftp://ftp.ncbi.nih.gov/gene/GeneRIF/generifs_basic.gz).
[3] Online Mendelian Inheritance in Man (http://www.omim.org/).
[4] Dragon Database of Genes Implicated in Prostate Cancer (http://www.cbrc.kaust.edu.sa/ddpc/).
[5] Human Prostate Gene DataBase (http://www.urogene.org/pgdb/).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2618927, IEEE/ACM Transactions on Computational Biology and Bioinformatics

cates that gene expression heterogeneity is not applied. When predicting a gene expression level of a cell in a matrix, precise prediction is required because the proposed method is based on the difference of predictability between two classes. For precise prediction, we should choose similar samples in collaborative filtering process. It is very natural that the higher the threshold *c* is, the more similar the selected neighbors are, the smaller the number of neighbors, and the better the performance is [24, 26-27]. But no neighbor for all cases in a gene indicates difference of predictability cannot be applied and prioritization score of the gene will be calculated as 0. Therefore, we analyzed neighbor numbers with varying the correlation coefficient threshold from 0.5 to 0.9 and chose the highest value for threshold c while avoiding the "no neighbor" situation (Table 3-6). The tables showed that Singh, GSE15484 (high risk vs low risk), and TCGA_PRAD datasets have some neighbors when using the threshold of 0.9 but the result of GSE15484 (cancer vs benign) dataset had no neighbor with 0.9 threshold in benign condition. The experiment of relationship between the threshold and the number of neighbors showed which is the proper threshold to use. The number of neighbors should be larger than 0 and at the same time, performance should be considerably good. We decided that 0.9 is the appropriate correlation coefficient threshold for Singh, GSE15484 (high risk vs low risk), and TCGA_PRAD datasets and 0.8 is the appropriate threshold for GSE15484 (cancer vs benign) dataset.

Table 3. Neighbor information of Singh dataset with varying correlation coefficient

Threshold *c*	Average number of neighbors		Percentage of no neighbor (%)	
	Cancer	Benign	Cancer	Benign
c < -1	51	49	0	0
0.5	46.92	36.48	0	0
0.6	44.92	33.84	0	0
0.7	41.58	28.96	0	0
0.8	33.81	21.56	0	0
0.9	15.62	9.16	7.69	6.00
c > 1	0	0	100	100

Table 4. Neighbor information of GSE15484 (high risk vs low risk) dataset with varying correlation coefficient

Threshold *c*	Average number of neighbors		Percentage of no neighbor (%)	
	High risk	Low risk	High risk	Low risk
c < -1	26	24	0	0
0.5	20.08	17.04	0	4.00
0.6	12.37	12.87	0	4.00
0.7	6.37	9.12	3.70	12.00
0.8	1.85	6.08	25.93	28.00
0.9	0.22	2.48	88.89	52.00

c > 1	0	0	100	100

Table 5. Neighbor information of GSE15484 (cancer vs benign) dataset with varying correlation coefficient

Threshold *c*	Average number of neighbors		Percentage of no neighbor (%)	
	Cancer	Benign	Cancer	Benign
c < -1	51	12	0	0
0.5	36.20	10.77	0	0
0.6	24.23	7.85	1.92	0
0.7	14.69	4.31	1.92	0
0.8	7.27	0.92	19.23	30.77
0.9	1.77	**0**	67.31	**100**
c > 1	0	0	100	100

Table 6. Neighbor information of TCGA_PRAD dataset with varying correlation coefficient

Threshold *c*	Average number of neighbors		Percentage of no neighbor (%)	
	Cancer	Benign	Cancer	Benign
c < -1	49	49	0	0
0.5	47.59	33.80	0	0
0.6	44.28	29.44	0.60	0
0.7	38.84	23.08	1.20	0
0.8	27.30	14.96	3.60	2.00
0.9	8.45	6.96	21.20	18.00
c > 1	0	0	100	100

To compare the performance of GSEH with the other related methods, we searched the prostate cancer-associated genes in the top-ranking genes selected using GSEH and ten comparable methods (Figure 6-9). The ten similar methods compared to GSEH in this study were chi-square statistic, information gain, gain ratio, Relief-A, symmetrical uncertainty, CV, t-test, DVE, Limma, and COPA. The DVE (Difference in Variance of Expression) is a simple method similar to COPA which is implemented by the authors for comparison. In DVE, gene expressions are normalized first and then the absolute differences in variance of gene expression between the two conditions are calculated. The genes are prioritized by the scores. The DVE uses variance difference between two conditions to tell which condition has more heterogeneous patterns. The other nine comparable similar methods are herein described in Introduction section.

In our experiment, chi-square, information gain, gain ratio, Relief-A, and symmetrical uncertainty were performed by Weka [28] software. We also performed CV[6] by using programming code. But in the experiments of the CV method, a running

[6] The code of CV is provided at (http://www.urmc.rochester.edu/biostat/people/students/hu.cfm).

error occurred and we did not get results from the CV. Therefore only in TCGA dataset, we did experiments with nine comparable methods. For experiments of t-test and Limma, we used *genefilter* and *limma* packages in R. We implemented COPA on our own because the COPA (Tomlins et al.) is not extensible and it can be only utilized on datasets of Oncomine database. Furthermore, we compared the result of GSEH with those of randomly selected genes. We could use proportions of the prostate cancer-associated genes to all the genes to estimate the result of randomly selected genes. The randomly selected genes were expected to have average information of the disease. In gene prioritization, it is important how many disease-associated genes are in high-ranking positions because the large number of disease-associated genes in low-ranking positions indicates bad prioritization.

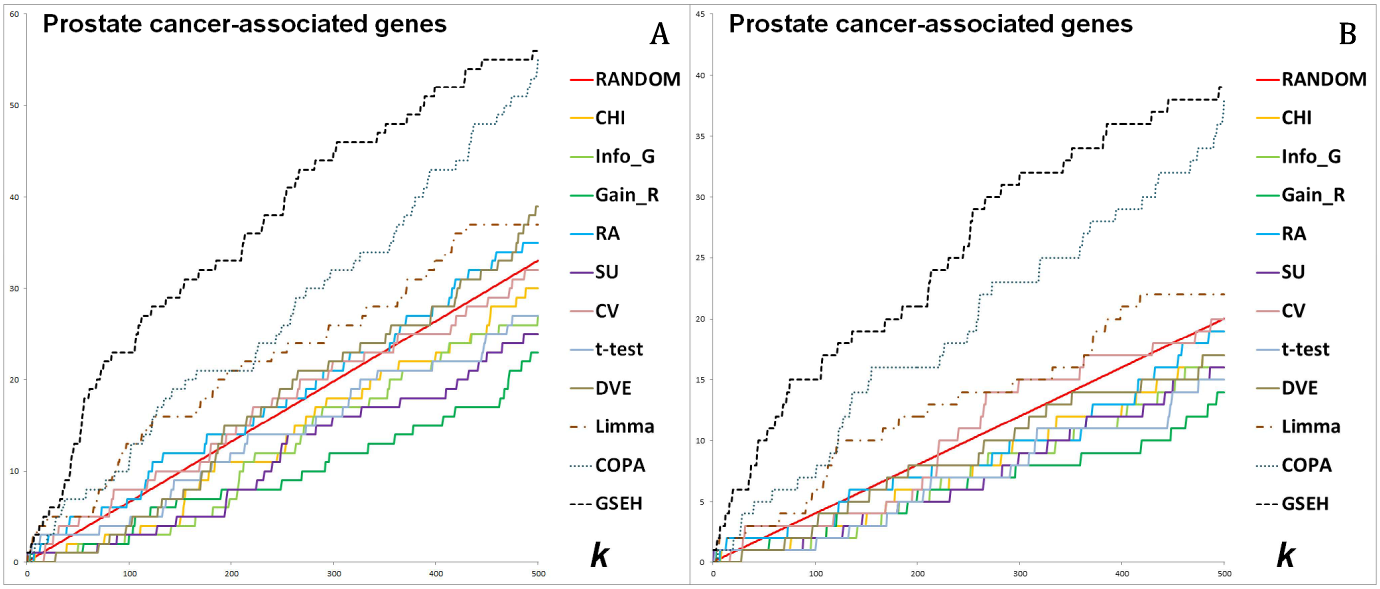


Figure 6. The number of discovered prostate cancer-associated genes of the methods with changing the number of *k* selected genes (CHI = Chi-Square, Info_G = Information Gain, Gain_R = Gain Ratio, RA = Relief-A, SU = Symmetrical Uncertainty). X-axis represents the number of selected genes *k*; Y-axis represents the number of prostate cancer-associated genes. **(A)** Result on GSE15484 (high risk vs low risk) dataset validated with Answer set 1. **(B)** Result on GSE15484 (high risk vs low risk) dataset validated with Answer set 2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2618927, IEEE/ACM Transactions on Computational Biology and Bioinformatics

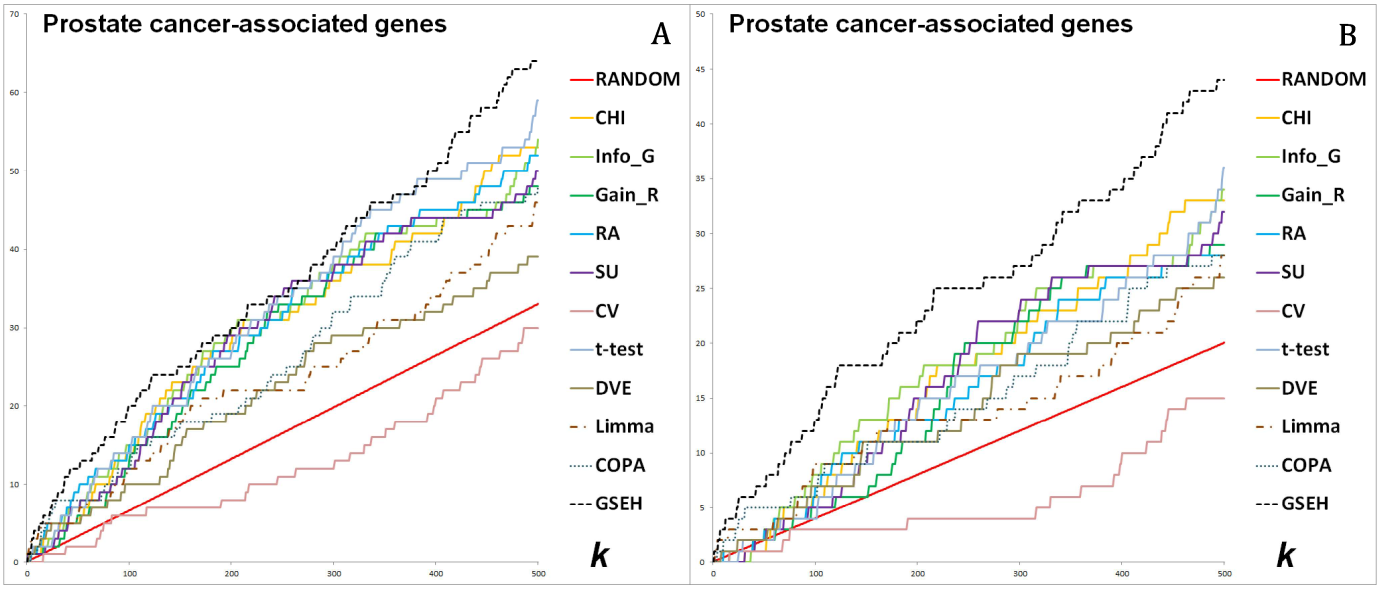


Figure 7. The number of discovered prostate cancer-associated genes of the methods with changing the number of *k* selected genes (CHI = Chi-Square, Info_G = Information Gain, Gain_R = Gain Ratio, RA = Relief-A, SU = Symmetrical Uncertainty). X-axis represents the number of selected genes *k*; Y-axis represents the number of prostate cancer-associated genes. **(A)** Result on GSE15484 (cancer vs benign) dataset validated with Answer set 1. **(B)** Result on GSE15484 (cancer vs benign) dataset validated with Answer set 2.

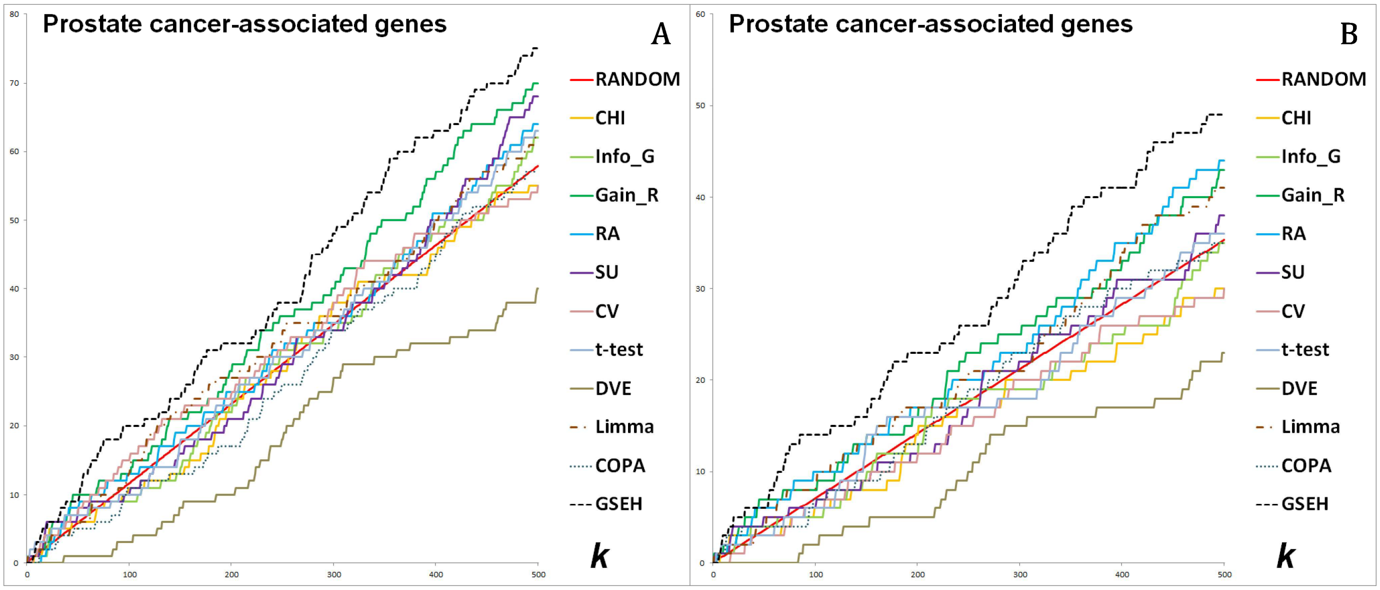


Figure 8. The number of discovered prostate cancer-associated genes of the methods with changing the number of *k* selected genes (CHI = Chi-Square, Info_G = Information Gain, Gain_R = Gain Ratio, RA = Relief-A, SU = Symmetrical Uncertainty). X-axis represents the number of selected genes *k*; Y-axis represents the number of prostate cancer-associated genes. **(A)** Result on Singh dataset validated with Answer set 1. **(B)** Result on Singh dataset validated with Answer set 2.

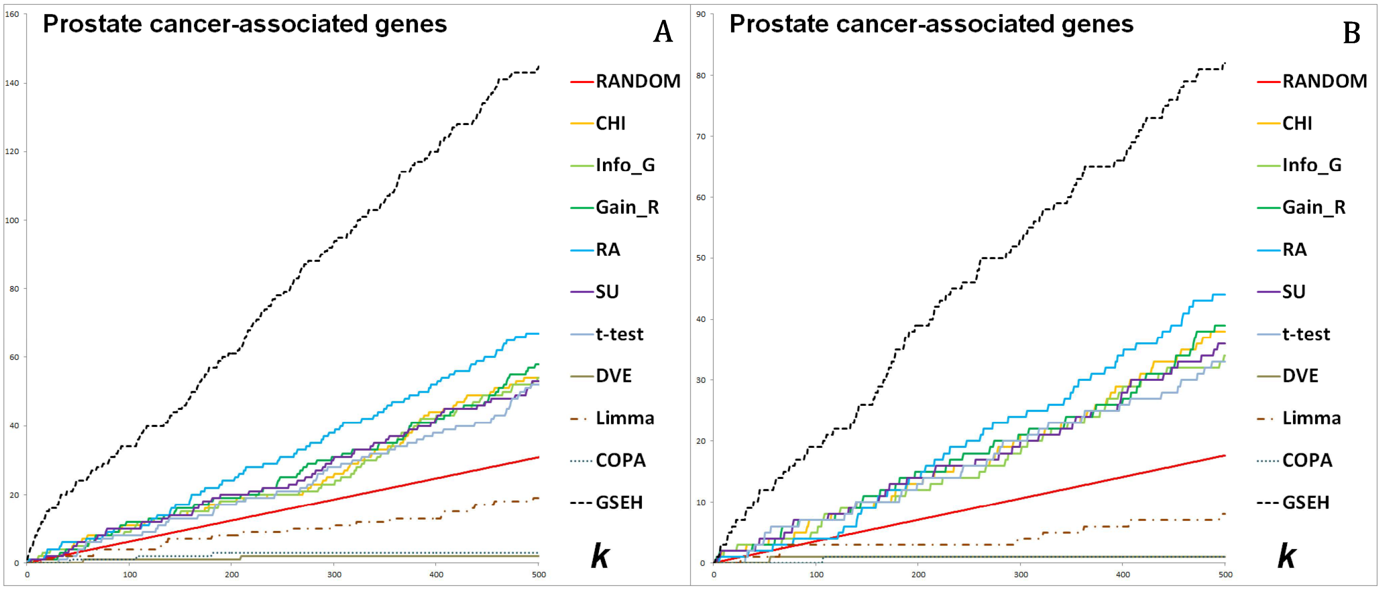


Figure 9. The number of discovered prostate cancer-associated genes of the methods with changing the number of *k* selected genes (CHI = Chi-Square, Info_G = Information Gain, Gain_R = Gain Ratio, RA = Relief-A, SU = Symmetrical Uncertainty). X-axis represents the number of selected genes *k*; Y-axis represents the number of prostate cancer-associated genes. **(A)** Result on TCGA_PRAD dataset validated with Answer set 1. **(B)** Result on TCGA_PRAD dataset validated with Answer set 2.

Since the GSEH aims to discover disease-associated genes, we analyzed influence of GSEH's disease-associated prioritization score. Gene selection methods prioritize genes based on their own scores and when the top-ranking genes are chosen, we count the number of disease-associated genes in the selected top-ranking genes using the answer sets. We searched the prostate cancer-associated genes in 1 through 500 top-ranking gene sets with varying the number of top-ranking genes to test gene set enrichment. As described in Figure 6-9, GSEH curves are trend up and left which indicates that prioritization of GSEH is better than the other methods and has more prostate cancer-associated genes in top 500 ranking genes. The results indicate that GSEH can efficiently discover disease-associated genes.

We also calculated false positive rate (FPR), false negative rate (FNR), p-value, hypergeometric test P(X >= *a*), and AUC (Table 7-10). The measures excluding AUC were computed on top 500 ranking genes. The p-values and the hypergeometric test P(X >= *a*) probabilities in the results were calculated by Fisher's exact test with 2x2 contingency table [29-30]. For instance, if there are 100 prostate cancer-associated genes out of 1,000 total genes and there are 50 prostate cancer-associated genes among the top 200 ranking genes prioritized by a gene selection method, a p-value calculated by Fisher's exact test with the contingency table (a = 50, b = 150, c = 50, d = 750) is 8.66 x 10^{-13} and P(X>=50) is also 8.66 x 10^{-13} as same as the p-value. The AUCs were calculated by using true positive rates and false positive rates with changing the number of selected top-ranking genes.

Table 7. Comparison results on GSE15484 (high risk vs low risk) dataset

Answer set 1	**Gene Count (*a*)**	**False Positive Rate**	**False Negative Rate**	**p-value**	**P(X >= *a*)**	**AUC**
Chi-square	30	0.0625	0.9435	0.6422	0.7364	0.4957
Information Gain	26	0.0630	0.9510	0.2259	0.9216	0.4953
Gain Ratio	23	0.0634	0.9567	0.0629	0.9787	0.4948
Relief-A	35	0.0618	0.9341	0.7095	0.3799	0.4867
Symmetrical Uncertainty	25	0.0632	0.9529	0.1620	0.9471	0.4951
CV	32	0.0622	0.9397	0.9260	0.5995	0.5009
t-test	27	0.0629	0.9492	0.3058	0.8882	0.4778
DVE	39	0.0613	0.9266	0.2638	0.1520	0.4923
LIMMA	37	0.0616	0.9303	0.4562	0.2515	0.5421
COPA	54	0.0593	0.8983	0.0003	0.0002	0.5368
GSEH	**56**	0.0590	0.8945	**0.0001**	0.0001	**0.5398**
Answer set 2	**Gene Count (*a*)**	**False Positive Rate**	**False Negative Rate**	**p-value**	**P(X >= *a*)**	**AUC**
Chi-square	17	0.0625	0.9472	0.5560	0.7921	0.4885
Information Gain	16	0.0626	0.9503	0.4093	0.8566	0.4880
Gain Ratio	14	0.0629	0.9565	0.1935	0.9433	0.4878
Relief-A	19	0.0622	0.9410	0.9064	0.6270	0.4719
Symmetrical Uncertainty	16	0.0626	0.9503	0.4093	0.8566	0.4879
CV	20	0.0621	0.9379	1.0000	0.5339	0.4955
t-test	15	0.0627	0.9534	0.2880	0.9068	0.4748
DVE	17	0.0625	0.9472	0.5560	0.7921	0.4882
LIMMA	22	0.0618	0.9317	0.6366	0.3512	0.5414
COPA	37	0.0599	0.8851	0.0002	0.0002	0.5420
GSEH	**39**	0.0596	0.8789	**4.79E-5**	3.95E-5	**0.5460**

Table 8. Comparison results on GSE15484 (cancer vs benign) dataset

Answer set 1	**Gene Count (*a*)**	**False Positive Rate**	**False Negative Rate**	**p-value**	**P(X >= *a*)**	**AUC**
Chi-square	53	0.0594	0.9002	0.0005	0.0003	0.5221
Information Gain	53	0.0594	0.9002	0.0005	0.0003	0.5214
Gain Ratio	48	0.0601	0.9096	0.0088	0.0050	0.5162
Relief-A	52	0.0596	0.9021	0.0010	0.0006	0.5159
Symmetrical Uncertainty	50	0.0598	0.9058	0.0028	0.0018	0.5164
CV	30	0.0625	0.9435	0.6422	0.7364	0.5016
t-test	59	0.0586	0.8889	9.49E-6	6.37E-6	0.5376
DVE	39	0.0613	0.9266	0.2638	0.1520	0.5254
LIMMA	46	0.0604	0.9134	0.0197	0.0125	0.5506
COPA	47	0.0602	0.9115	0.0118	0.0080	0.5321
GSEH	**64**	0.0580	0.8795	**1.92E-7**	1.34E-7	**0.5639**
Answer set 2	**Gene Count (*a*)**	**False Positive Rate**	**False Negative Rate**	**p-value**	**P(X >= *a*)**	**AUC**
Chi-square	33	0.0604	0.8975	0.0043	0.0029	0.5229
Information Gain	34	0.0603	0.8944	0.0020	0.0015	0.5219
Gain Ratio	29	0.0609	0.9099	0.0439	0.0271	0.5180
Relief-A	28	0.0611	0.9130	0.0755	0.0434	0.5128
Symmetrical Uncertainty	32	0.0605	0.9006	0.0089	0.0054	0.5181
CV	15	0.0627	0.9534	0.2880	0.9068	0.4954
t-test	36	0.0600	0.8882	0.0005	0.0004	0.5445
DVE	26	0.0613	0.9193	0.1570	0.1003	0.5308
LIMMA	28	0.0611	0.9130	0.0755	0.0434	0.5506
COPA	28	0.0611	0.9130	0.0755	0.0434	0.5256
GSEH	**44**	0.0590	0.8634	**6.71E-7**	4.83E-7	**0.5742**

Table 9. Comparison results on Singh dataset

Answer set 1	Gene Count (*a*)	False Positive Rate	False Negative Rate	p-value	P(X >= *a*)	AUC
Chi-square	55	0.0570	0.9461	0.7194	0.6795	0.5199
Information Gain	62	0.0561	0.9393	0.5644	0.2944	0.5176
Gain Ratio	70	0.0551	0.9314	0.0839	0.0491	0.5190
Relief-A	64	0.0558	0.9373	0.3873	0.2055	0.5095
Symmetrical Uncertainty	68	0.0553	0.9334	0.1496	0.0840	0.5213
CV	54	0.0571	0.9471	0.6149	0.7300	0.4798
t-test	63	0.0560	0.9383	0.4713	0.2477	0.5106
DVE	40	0.0589	0.9608	0.0094	0.9971	0.4779
LIMMA	62	0.0561	0.9393	0.5644	0.2944	0.4864
COPA	57	0.0567	0.9442	1.0000	0.5695	0.4979
GSEH	**75**	0.0544	0.9265	**0.0173**	0.0099	0.4994
Answer set 2	**Gene Count (*a*)**	**False Positive Rate**	**False Negative Rate**	**p-value**	**P(X >= *a*)**	**AUC**
Chi-square	30	0.0573	0.9518	0.3698	0.8515	0.5303
Information Gain	35	0.0567	0.9438	1.0000	0.5472	0.5267
Gain Ratio	43	0.0557	0.9310	0.1767	0.0996	0.5287
Relief-A	44	0.0556	0.9294	0.1258	0.0731	0.5234
Symmetrical Uncertainty	38	0.0563	0.9390	0.5903	0.3383	0.5322
CV	29	0.0574	0.9535	0.2814	0.8912	0.4801
t-test	36	0.0566	0.9422	0.8575	0.4755	0.5232
DVE	23	0.0581	0.9631	0.0245	0.9922	0.4636
LIMMA	41	0.0559	0.9342	0.3219	0.1734	0.4778
COPA	35	0.0567	0.9438	1.0000	0.5472	0.5041
GSEH	**49**	0.0550	0.9213	**0.0189**	0.0111	0.4942

Table 10. Comparison results on TCGA_PRAD dataset

Answer set 1	Gene Count (*a*)	False Positive Rate	False Negative Rate	p-value	P(X >= *a*)	AUC
Chi-square	54	0.0231	0.9573	0.0001	4.31E-5	0.5629
Information Gain	54	0.0231	0.9573	0.0001	4.31E-5	0.5630
Gain Ratio	58	0.0229	0.9542	3.13E-6	2.61E-6	0.5610
Relief-A	67	0.0225	0.9470	2.18E-9	1.56E-9	0.5758
Symmetrical Uncertainty	53	0.0232	0.9581	0.0001	0.0001	0.5625
t-test	52	0.0233	0.9589	0.0002	0.0002	0.5601
DVE	2	0.0258	0.9984	1.12E-11	1.0000	0.4897
LIMMA	19	0.0250	0.9850	0.0236	0.9931	0.5020
COPA	3	0.0258	0.9976	1.67E-10	1.0000	0.4893
GSEH	**144**	0.0185	0.8862	**2.89E-58**	2.89E-58	**0.6472**
Answer set 2	**Gene Count (*a*)**	**False Positive Rate**	**False Negative Rate**	**p-value**	**P(X >= *a*)**	**AUC**
Chi-square	38	0.0233	0.9475	1.08E-5	8.32E-6	0.5767
Information Gain	34	0.0235	0.9530	0.0003	0.0002	0.5766
Gain Ratio	39	0.0233	0.9461	3.68E-6	3.43E-6	0.5746
Relief-A	44	0.0230	0.9392	3.99E-8	2.71E-8	0.5929
Symmetrical Uncertainty	36	0.0234	0.9503	0.0001	4.49E-5	0.5761
t-test	33	0.0236	0.9544	0.0005	0.0005	0.5712
DVE	1	0.0252	0.9986	4.56E-7	1.0000	0.4718
LIMMA	8	0.0248	0.9890	0.0137	0.9970	0.5059
COPA	1	0.0252	0.9986	4.56E-7	1.0000	0.4721
GSEH	**82**	0.0211	0.8867	4.90E-32	4.90E-32	**0.6648**

In most of the cases, GSEH found the largest number of prostate cancer-associated genes, showed the lowest p-value among all the gene selection methods, and showed better AUCs than the other methods. Because GSEH is a disease-associated gene selection/prioritization method, an ability of giving high scores to disease-associated genes is important, and the ability can be evaluated by counting disease-associated genes in limited top-ranking genes. The above experiments showed that GSEH has a power of distinguishing disease-associated genes from normal genes. Moreover, another goal of GSEH is identifying disease-associated genes with a different kind of view from the other methods. GSEH prioritized different genes in the top ranks from the top-ranking genes of the other methods and it is presented in *Discussions* section.

4 DISCUSSIONS

GSEH is a filtering strategy of the four computational strategies mentioned in the introduction section, because it employs a filter defined by heterogeneous gene expression characteristics. As mentioned in Introduction section, the difference of gene expression heterogeneity between two conditions can provide information for finding disease-associated genes. We therefore used collaborative filtering to estimate the degree of being “differentially predicted” which indicates the difference of gene expression heterogeneity. GSEH uses the degree of being differentially predicted under different conditions to identify genes relating to the biological process of a disease.

The “Differential prediction” is the main concept of GSEH. If a data has heterogeneous gene expression characteristics, it is difficult to predict expressions. Because collaborative filtering is a method that can predict unfilled information in a recommendation system, the significant challenge of prediction is indicative of the great differences between the original gene expressions and collaboratively filtered gene expressions. If the difference is large, gene expressions from one class are poorly predicted, whereas gene expressions from the other class are accurately predicted. In other words, gene expressions from one class have heterogeneous gene expression patterns while gene expressions from the other class do not have heterogeneous gene expression patterns because we assume that it is difficult to predict expressions with heterogeneous characteristics.

When we devised the GSEH, we supposed about two cases. First, normal people’s gene expressions of a given gene are basically heterogeneous but when a disease affects the gene, the expressions of the gene show consistency. Second, it is the opposite case. Originally, gene expressions of a given gene are similar in normal people but when a disease affects, the gene expressions become heterogeneous. Therefore, the large prioritization score indicates high possibility to relate with a given disease in GSEH and it can be used to determine significance of genes when you consider differential predictability between two conditions in gene expression data. The top-ranking genes from GSEH were differentially predicted between two classes and we can conclude that differentially predicted genes can provide additional information for discovering disease-associated genes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2618927, IEEE/ACM Transactions on Computational Biology and Bioinformatics

GSEH discovered the largest number of prostate cancer-associated genes in high-ranking positions and showed the lowest p-values when compared to other similar methods. Moreover, we investigated prostate cancer-associated genes from the 20 top-ranking genes prioritized by GSEH. Singh, GSE15484 (high risk vs low risk), and GSE15484 (cancer vs benign) datasets have 5 prostate cancer-associated genes, and TCGA_PRAD dataset has 18 prostate cancer-associated genes among the 20 top-ranking genes. PTHLH, SERPINA1, JUN, GPX3, and KLK3 from the Singh dataset, OR51E1, ETV4, NPY, MT2A, and ID2 from the GSE15484 (high risk vs low risk) dataset, MSMB, TGM4, KLK11, EGFR, and ACPP from the GSE15484 (cancer vs benign) dataset, and SEMG1, SEMG2, KLK3, MYH11, TGM4, HSPA1A, ACPP, NPY, FLNA, SERPINA3, SPON2, LTF, TAGLN, MUC6, PLA2G2A, MYLK, KLK2, and TFF3 from TCGA_PRAD dataset were related to prostate cancer. We manually investigated functions of the prostate cancer-associated genes prioritized by GSEH (Table 11-14).

Table 11. Functions of prostate cancer-associated genes prioritized by GSEH in Singh dataset

Singh	Gene Symbol	Rank	Gene Functions
1	PTHLH	7	- Nuclear localization of PTHLH bestows prostate cancer cell resistance on anoikis, potentially contributing to metastasis of prostate cancer [31]. - PTHLH encourages prostate cancer cell growth [32]. - PTHLH has a role in tumorigenesis of prostate cancer; it is a key intermediary for communication and interactions between prostate cancer and the bone microenvironment [33]. - PTHLH expression engenders the skeletal progression of prostate cancer cells [34]. - PTHLH has a role in prostate tumor invasion and metastasis by influencing cell adhesion to the ECM (Extracellular Matrix) protein via up-regulation of specific integrin subdivisions [35].
2	SERPINA1	9	- Prostate cancer patients showed higher elevation in SERPINA1 serum levels compared to healthy controls [36]. - Men with prostate cancer had significantly higher SERPINA1 concentrations than those without prostate cancer [37].
3	JUN	10	- JUN activity in prostate cancer cells mediates EGF-R and PI3K signaling; it is crucial for their proliferation [38]. - Activation of JUN enhances apoptosis in prostate cancer cells [39]. - JUN plays a vital role in the pathway that links ligand-activated AR to elevated ETV1 expression, resulting in enhanced expression of matrix metalloproteinases and prostate cancer cell invasion [40].
4	GPX3	14	- A novel signaling pathway of GPX3-PIG3 is related to the regulation of cell death in prostate cancer [41].

			- GPX3 is a novel prostate cancer suppressor gene [42].
5	KLK3	16	- KLK3 may decrease or increase invasive properties of prostate cancer cells [43]. - Single-nucleotide polymorphisms in KLK3 are related with prostate cancer [44]. - Germline KLK3 variants could influence the diagnosis of nonaggressive prostate cancer by affecting the possibility of biopsy [45]. - The KLK3/free testosterone ratio may be considered a marker expressing different biology groups of prostate cancer patients; it is strongly associated with tumor extension and the Gleason sum [46]. - Polymorphisms in KLK3 genes may be regarded as potential biomarkers for prostate cancer [47]. - KLK3-RP2 is up-regulated in prostate cancer compared to benign prostatic hyperplasia tissues [48]. - The androgen response element of polymorphism on the KLK3 gene is related to prostate cancer [49]. - A novel splice variant of prostate specific antigen/human KLK3 is identified; it can be used to distinguish prostate cancer from benign prostate hyperplasia [50]. - KLK3 gene promoter variation may play a key role in the development of prostate cancer and benign prostatic hyperplasia [51]. - KLK3 has a functional role in the advancement of prostate cancer through their facilitation of tumor cell migration [52]. - Polymorphism of KLK3 gene promoter may be a significant biomarker for prostate cancer risk, especially an early outbreak of prostate cancer [53].

Table 12. Functions of prostate cancer-associated genes prioritized by GSEH in GSE15484 (high risk vs low risk) dataset

GSE15484	Gene Symbol	Rank	Gene Functions
1	OR51E1	1	- OR51E1 may be useful as a tissue marker and molecular target for the early detection and treatment of human prostate cancers [54]. - In some cases, expression of OR51E1 is substantially elevated in prostate cancer [55].
2	ETV4	5	- Increased expression of ETV4 is related to tumor aggression in prostate neoplasms [56]. - TMPRSS2-ETV4 gene fusions may cause an initiating event in prostate cancer development [57].
3	NPY	8	- A lower NPY expression level is highly related to the more aggressive clinical behavior of prostate cancer [58]. - Y1 receptor activation by NPY regulates the development of prostate cancer cells [59].
4	MT2A	13	- A strong relation between the rs28366003 genotype and MT2A expression level is found in prostate cancer patients [60]. - High MT2A expression is associated with prostate cancer [61]. - MT2A may have a role in prostate cancer [62].
5	ID2	17	- ID1 and ID2 proteins manage prostate cancer cell phenotypes and play roles as molecular markers of aggressive human prostate cancer [63].

Table 13. Functions of prostate cancer-associated genes prioritized by GSEH in GSE15484 (cancer vs benign) dataset

GSE15484	Gene Symbol	Rank	Gene Functions
1	MSMB	2	- A functional polymorphism in MSMB promoter contributes to genetic predisposition to prostate cancer [64].

			- A SNP in MSMB on chromosome 10q11 is a causal variant for prostate cancer risk [65-66] - High MSMB expression is associated with the progression of prostate cancer [67].
2	TGM4	5	- TGM4 plays a pivotal role in interaction between endothelial cells and prostate cancer cells [68]. - TGM4 can be a potential predictor of biochemical recurrence of prostate cancer [69]. - TGM4 is down-regulated in prostate cancer glands compared to normal glands [70].
3	KLK11	6	- KLK11 may be useful marker for distinguishing prostate cancer and benign samples [71]. - Down-regulation of KLK11 can be used as prognostic indicators for prostate cancer [72].
4	EGFR	8	- EGFR may have a role in disease relapse and progression to androgen-independence in prostate cancer [73]. - Down-regulation of EGFR plays an important role in pathogenesis of prostate cancer [74].
5	ACPP	13	- ACPP regulates prostate cancer cell growth [75]. - ACPP can be predictive indicator of prostate cancer diagnosis and prognosis [76-77].

Table 14. Functions of prostate cancer-associated genes prioritized by GSEH in TCGA_PRAD dataset

TCGA_PRAD	Gene Symbol	Rank	Gene Functions
1	SEMG1	1	- Overexpression of SEMG1 and SEMG2 are found in human prostate cancer and they can be used to predict prostate cancer progression after radical prostatectomy [78].
2	SEMG2	2	
3	KLK3	3	- KLK3 is realted with prostate cancer and already described in Table 11.
4	MYH11	4	- There is an evidence for a role of somatic MYH11 mutations in formation of prostate cancers [79].
5	TGM4	5	- TGM4 is associated with prostate cancer and already described in Table 13.
6	HSPA1A	6	- HSPA1A is overexpressed in human prostate cancer cells [80] - Down-regulation of HSPA1A suppresses ERK and NF-kappaB, which may be responsible for enhanced sensitivity of prostate carcinoma cells [81].
7	ACPP	7	- ACPP is related with prostate cancer and already described in Table 13.
8	NPY	8	- NPY is associated with prostate cancer and already described in Table 12.
9	FLNA	9	- FLNA may play important roles as a negative regulator to prostate cancer cell migration and invasion [82].
10	SERPINA3	10	- SERPINA3 is associated with increased risk of prostate carcinoma [83].
11	SPON2	11	- SPON2 is overexpressed in prostate cancer cell and it is a new diagnostic biomarker for prostate cancer [84].
12	LTF	12	- Silencing of the LTF may be causally linked to prostate cancer progression [85].
13	TAGLN	14	- Expression of TAGLN is decreased in prostate cancer [86]. - TAGLN acts as a suppressor to inhibit prostate cancer cell growth [87].

14	MUC6	15	- MUC6 is overexpressed in progression and lymphatic metastasis of prostate cancer [88].
15	PLA2G2A	17	- High level of PLA2G2A may serve as a tumor prognostic biomarker which is capable of distinguishing aggressive from indolent prostate cancers [89]. - PLA2G2A overexpression is associated with prostate development and progression [90-91]. - PLA2G2A expression is increased in prostate cancer but decreased in metastatic cancers [92].
16	MYLK	18	- MYLK is down-regulated by androgens in prostate cancer cells [93].
17	KLK2	19	- KLK2 promotes prostate cancer cell growth [94]. - Single nucleotide polymorphism in KLK2 is associated with prostate cancer [95]. - KLK2 enhances proliferation of prostate cancer cells [96].
18	TFF3	20	- TFF3 enhances oncogenic characteristics of prostate cancer cells [97]. - TFF3 is up-regulated in prostate cancer glands compared to the corresponding normal glands [98]. - Overexpression and promoter hypomethylation of TFF3 is associated with prostate cancer [99].

The genes prioritized by GSEH are potentially significant to prostate cancer. The 5 genes among the top 20 genes were identified as prostate cancer-associated genes in Singh and GSE15484 datasets, and 18 prostate cancer-associated genes were discovered among the top 20 genes in TCGA_PRAD dataset. However, this does not indicate that the other genes are meaningless. The genes in high-ranking positions have a high possibility of being associated with the prostate cancer and are worth researching (Table 15-18). Moreover, GSEH can discover disease-associated genes with different point of view: Gene expression heterogeneity. As described in the tables, GSEH provided different genes compared to other methods.

Table 15. Top 20 genes of GSEH and ranks of the other methods of the genes in Singh dataset

GSEH Rank	Gene symbol	CHI	Info_G	Gain_R	RA	SU	CV	t-test	DVE	Limma	COPA
1.	**EIF2AK2**	4303	4107	3996	4121	4339	5742	2221	8459	6111	644
2.	**NDUFB1**	472	575	1224	323	888	7014	316	4781	6334	61
3.	**GOLGA4**	2768	2860	1603	5635	1579	1021	7771	7834	5418	7379
4.	**HIST1H1C**	525	282	90	2306	192	4394	445	7515	6887	18
5.	**SLC5A1**	1380	1377	422	457	970	70	5030	1597	2502	6898
6.	**CIZ1**	649	761	1265	662	1194	8686	2871	5663	7321	2

7.	**PTHLH**	664	370	120	4808	231	2088	694	2795	7583	13
8.	**PPM1F**	7859	7859	7859	1657	7859	1549	4422	3095	4481	6831
9.	**SERP1NA1**	68	64	44	157	50	1926	125	1911	5600	229
10.	**JUN**	1012	592	196	4001	393	5618	1723	7769	6237	99
11.	**PI3**	5	9	8	6	7	8756	5	3349	5623	280
12.	**LCE2B**	1065	641	163	4605	407	1997	4226	6217	8581	505
13.	**EDN3**	4233	4288	3948	5139	4244	8617	4228	7274	7593	30
14.	**GPX3**	94	113	344	28	125	964	25	7697	2810	3543
15.	**TSPAN7**	527	509	663	471	465	8800	1024	613	7609	8
16.	**KLK3**	7535	7535	7535	591	7535	78	1625	7403	3647	5585
17.	**CLIC1**	3266	4515	3144	3172	3232	1069	7023	5959	5544	848
18.	**IGKV1OR15-118**	658	371	118	303	227	5072	591	7170	4847	8288
19.	**ATP6V1E1**	6	6	4	31	5	8670	22	4737	6002	153
20.	**FABP4**	224	242	684	175	323	3920	110	7893	6208	1571

Table 16. Top 20 genes of GSEH and ranks of the other methods of the genes in GSE15484 (high risk vs low risk) dataset

GSEH Rank	Gene symbol	CHI	Info_G	Gain_R	RA	SU	CV	t-test	DVE	Limma	COPA
1.	**OR51E1**	7641	7641	7641	2306	7641	979	4929	2264	1284	3564
2.	**IL17C**	5686	5686	5686	8000	5686	8047	3496	1694	5936	5976
3.	**UBE4B**	1187	1293	1254	3366	1275	3563	7324	6237	8038	7799
4.	**TAF7**	139	61	34	2353	38	7011	2683	8047	6908	3868
5.	**ETV4**	1688	1688	1688	3409	1688	1797	5447	4123	2171	1017
6.	**NAA11**	5351	5351	5351	2078	5351	7084	1657	3456	5978	6682
7.	**RPLP1**	2344	2344	2344	5509	2344	3923	4720	7021	5381	643
8.	**NPY**	7160	7160	7160	7638	7160	5257	7915	1229	1817	6076
9.	**RANBP2**	1646	1646	1646	5281	1646	1052	7785	3870	7505	3338
10.	**DSC2**	2393	2393	2393	146	2393	4523	584	6954	3535	7880
11.	**PLA1A**	2859	2859	2859	7682	2859	6628	6718	5688	1042	891
12.	**HLA-DRA**	3491	3491	3491	878	3491	3687	2271	6644	4507	6170
13.	**MT2A**	5965	5965	5965	5183	5965	2894	3365	2844	2097	1633
14.	**GGTL4**	448	312	77	52	178	6818	295	1541	5225	230

15.	**TMEM178A**	1473	1473	1473	882	1473	5287	862	4435	454	4511
16.	**MT1X**	5403	5403	5403	3384	5403	3874	2432	6775	4365	1923
17.	**ID2**	4631	4631	4631	3552	4631	2376	5113	1130	7741	7364
18.	**MT1H**	434	522	458	514	446	4131	445	6972	5681	828
19.	**VPS52**	5592	5592	5592	8048	5592	6736	4623	4716	6396	90
20.	**GDEP**	7515	7515	7515	3224	7515	969	6991	973	3877	41

Table 17. Top 20 genes of GSEH and ranks of the other methods of the genes in GSE15484 (cancer vs benign) dataset

GSEH Rank	Gene symbol	CHI	Info_G	Gain_R	RA	SU	CV	t-test	DVE	Limma	COPA
1.	**FCGBP**	31	47	28	8	24	7802	11	1141	2275	3
2.	**MSMB**	110	60	247	183	152	69	7682	4921	3823	1065
3.	**ORM2**	3413	3502	3527	1265	3527	5141	1239	1511	2831	127
4.	**ACSM1**	5997	5756	6110	4613	6110	7923	3233	6999	4040	145
5.	**TGM4**	7596	7715	7764	208	7764	2239	246	4764	3999	14
6.	**KLK11**	2242	2193	2196	7915	2196	2719	6204	4046	5548	11
7.	**SERTAD4**	1465	1232	1449	2080	1449	216	3460	2261	6593	4256
8.	**EGFR**	3061	2994	3029	1465	3029	4779	495	6890	4929	408
9.	**COL3A1**	2017	2095	2084	7467	2084	2658	2895	3246	401	1357
10.	**FBXL12**	965	959	779	7489	772	7166	7222	3701	5480	16
11.	**CYFIP2**	4613	4717	4574	1969	4574	6726	6011	6566	7833	1188
12.	**ABP1**	209	231	254	933	246	2333	141	772	4988	1068
13.	**ACPP**	7677	7787	7638	758	7638	361	7986	4744	5169	3003
14.	**SOBP**	157	247	42	535	98	3709	180	2998	1334	5428
15.	**IER3**	29	48	27	43	25	3792	82	2335	138	681
16.	**CACNA1D**	254	449	202	1537	288	2035	502	3480	805	4415
17.	**KRT5**	15	10	33	5	19	7900	4	1089	1613	15
18.	**B3GNT5**	193	316	190	172	228	6590	138	745	4797	5080
19.	**FAM208A**	2389	2000	2341	6460	2341	2935	7109	3958	831	436
20.	**CENPN**	5366	6114	6028	893	6028	2089	430	7144	3883	5789

Table 18. Top 20 genes of GSEH and ranks of the other methods of the genes in TCGA_PRAD dataset

GSEH Rank	Gene symbol	CHI	Info_G	Gain_R	RA	SU	t-test	DVE	Limma	COPA
1.	**SEMG1**	3931	4083	4885	9092	4308	2017	4788	17132	4719
2.	**SEMG2**	6350	6166	5777	11081	5968	2187	4787	17196	4720
3.	**KLK3**	7580	7591	6894	7737	7355	4680	11355	9166	11405
4.	**MYH11**	676	732	919	965	802	168	9120	11829	8877
5.	**TGM4**	19697	19697	19697	6355	19697	11175	2581	18152	2519
6.	**HSPA1A**	16093	16093	16093	7626	16093	5240	12393	15647	12929

7.	**ACPP**	8003	8002	8867	394	8438	4690	18553	12248	20191
8.	**NPY**	5777	5683	5142	17361	5428	13742	8458	16899	8360
9.	**FLNA**	1788	1927	2261	2961	1958	528	13999	11780	12217
10.	**SERPINA3**	18784	18784	18784	13857	18784	12928	4736	17145	4715
11.	**SPON2**	3162	2426	1418	10619	2075	4381	3361	14259	3466
12.	**LTF**	16959	16959	16959	8404	16959	12737	10136	16674	9942
13.	**ACTG2**	1054	1097	1320	3313	1170	1053	18583	13647	20185
14.	**TAGLN**	1606	1775	2171	5415	1830	1932	2885	13571	2751
15.	**MUC6**	13036	13012	12021	16023	12485	3250	9173	17530	8988
16.	**DES**	827	870	703	4842	773	2702	15687	14633	16353
17.	**PLA2G2A**	5407	5855	7407	16401	6521	9349	7013	16304	6924
18.	**MYLK**	91	114	102	847	105	27	9091	11613	8897
19.	**KLK2**	5326	5047	4796	2950	4992	2912	11354	7898	11408
20.	**TFF3**	3575	3727	4168	14610	3767	8822	2608	16516	43

GSEH can provide insight into gene expression heterogeneity of diseases; nevertheless, it has two limitations. First, because the method selects disease-associated genes based on gene expression heterogeneity, if the degree of gene expression heterogeneity between two conditions is low, the performance may not be good. For this reason, we used prostate cancer data for our experiment. Because prostate cancer has highly heterogeneous characteristics [21, 100-101], we expected that the degree of gene expression heterogeneity between two conditions in prostate cancer data is high, and it is suitable for an experiment that handles gene expression heterogeneity. Second, the calculation of correlation and prediction with collaborative filtering is time-consuming. It may take a long time to create a predicted matrix for data that includes a large number of samples. Addressing these limitations will comprise our future work on GSEH.

5 CONCLUSION

Most existing gene selection methods have focused on differentially expressed genes, which can help with classification, rather than on biologically meaningful genes. Our focus, on the other hand, is on discovering genes that relate to the biological processes of a disease. GSEH is not intended to replace those existing differentially expressed gene selection methods; rather, it serves to provide additional information for discovering genes that relate to the biological processes of a disease. The GSEH process is divided into two phases. The first phase involves constructing a predicted gene expression matrix using collaborative filtering. The second phase involves calculating the ranking scores of genes using a comparison between a predicted gene expression matrix and the original gene expression matrix. GSEH selects genes by scoring the difference of predicted expression levels by assuming that the predicted levels under the two different conditions within the same disease are different. The larger the heterogeneity, the more challenging is the prediction task. Therefore, it can be estimated that the greater the difference of the predicted expression level, the more closely related to a disease the gene is. GSEH discovered the largest number of prostate cancer-associated genes and showed considerably low *p*-value when compared to the other methods. However, GSEH has a limitation in this paper that the results are only from prostate cancer datasets. Applying GSEH to another various disease data sets is required to make more significant results. The genes prioritized by GSEH have high potential to be related with a disease. Moreover, they can provide a different insight into the biological processes of a disease compared to other methods.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIP) (NRF-2015R1A2A1A05001845). We appreciate Mr. Junsik Kim’s proofreading efforts.

References

[1] J. Salomon et al, “Healthy life expectancy for 187 countries, 1990–2010: a systematic analysis for the Global Burden Disease Study 2010”, The Lancet, vol. 380, issue 9859, pp. 2144-2162, 2012.
[2] Thomas Abeel et al, “Robust biomarker identification for cancer diagnosis with ensemble feature selection methods”, BIOINFORMATICS, vol. 26, no. 3, pp. 392-398, 2010.
[3] Jie Cheng et al, “Good practice guidelines for biomarker discovery from array data: a case study for breast cancer prognosis”, BMC Systems Biology, 7(Suppl 4):S2, 2013.
[4] Yves Moreau et al, “Computational tools for prioritizing candidate genes: boosting disease gene discovery”, Nature Reviews Genetics, vol. 13, pp. 523-536, 2012.
[5] Shuya Lu et al, “Biomarker detection in the integration of multiple multi-class genomic studies”, BIOINFORMATICS, vol. 26, no. 3, pp. 333-340, 2010.
[6] Marija Volk et al, “Expression Signature as a Biomarker for Prenatal Diagnosis of Trisomy 21”, PLOS ONE, 8(9), 2013.
[7] Yvan Saeys et al, “A review of feature selection techniques in bioinformatics”, BIOINFORMATICS, vol. 23, no. 19, pp. 2507-2517, 2007.
[8] X. Jin et al, “Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles”, Data Mining for Biomedical Applications, vol. 3916, pp. 106-115, 2006.
[9] H. Liu et al, “A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns”, Genome Informatics, vol. 13, pp. 51-60, 2002.
[10] A. Karegowda et al, “Comparative study of attribute selection using gain ratio and correlation based feature selection”, Int J Inf Technol Knowl Manage, vol. 2, no. 2, pp. 271-277, 2010.
[11] William H. Press et al, “Numerical recipes in C”, Cambridge University Press, 1988.
[12] Igor Kononenko et al, “Estimating attributes: analysis and extensions of RELIEF, Proceedings of European Conference on Machine Learning, pp. 171-182, 1994.
[13] Xiangqin Cui et al, “Statistical tests for differential expression in cDNA microarray experiments”, Genome Biology, vol. 4, issue 4, 210, 2003.
[14] Gordon K. Smyth, “Linear models and empirical bayes methods for assessing differential expression in microarray experiments”, Stat Appl Genet Mol Biol, vol. 3, Article 3, 2004.
[15] David R. Bickel et al, “Degrees of differential gene expression: detecting biologically significant expression differences and estimating their magnitudes”, BIOINFORMATICS, vol. 20, no. 5, pp. 682-688, 2004.
[16] Rui Hu et al, “Detecting intergene correlation changes in microarray analysis: a new approach to gene selection”, BMC Bioinformatics, 10:20, 2009.
[17] Scott A. Tomlins et al, “Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer”, SCIENCE, vol. 310, issue 5748, pp. 644-648, 2005.
[18] James W. MacDonald et al, “COPA – cancer outlier profile analysis”, BIOINFORMATICS, vol. 22, no. 23, pp. 2950-2951, 2006.
[19] Jeffrey T. Leek et al, “Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis”, PLoS Genetics, vol. 3, issue 9, e161, 2007.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2618927, IEEE/ACM Transactions on Computational Biology and Bioinformatics

[20] Chenwei Wang et al, “mCOPA: analysis of heterogeneous features in cancer expression data”, Journal of Clinical Bio-informatics, vol. 2, issue 1, 22, 2012.

[21] Dinesh Singh et al, “Gene expression correlates of clinical prostate cancer behavior”, Cancer Cell, vol. 1, issue 2, pp. 203-209, 2002.

[22] Daniel Billsus et al, “Learning collaborative information filters”, Proceedings of the 15th international conference on machine learning, pp. 46-54, 1998.

[23] Jonathan L. Herlocker et al, “An algorithmic framework for performing collaborative filtering”, Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, pp. 230-237, 1999.

[24] Badrul Sarwar et al, “Item-based collaborative filtering recommendation algorithms”, Proceedings of the 10th international conference on World Wide Web, pp. 285-295, 2001.

[25] Arzucan Ozgur et al, “Identifying gene-disease associations using centrality on a literature mined gene-interaction network”, BIOINFORMATICS, vol. 24, issue 13, pp. i277-i285, 2008.

[26] J. Herlocker et al, “An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms”, Information Retrieval Journal, vol. 5, issue 4, pp. 287-310, 2002.

[27] A. Bellogin et al, “Neighbor Selection and Weighting in User-Based Collaborative Filtering: A Performance Prediction Approach”, ACM Transactions on the Web, vol. 8, issue 2, 12, 2014.

[28] Mark Hall et al, “The WEKA data mining software: an update”, ACM SIGKDD Explorations, vol. 11, issue 1, pp. 10-18, 2009.

[29] Luca Abatangelo et al, “Comparative study of gene set enrichment methods”, BMC Bioinfomatics, vol. 10:275, 2009.

[30] Kimberly Glass et al, “Annotation Enrichment Analysis: An Alternative Method for Evaluating the Functional Proper-ties of Gene Sets”, Scientific Reports, vol. 4:4191, 2014.

[31] Serk I. Park et al, “Nuclear localization of parathyroid hormone-related peptide confers resistance to anoikis in prostate cancer cells”, Endocrine-Related Cancer, vol. 19, issue. 3, pp. 243-254, 2012.

[32] Tracy M. Downs et al, “PTHrP stimulates prostate cancer cell growth and upregulates aldo-keto reductase 1C3”, Cancer letters, vol. 306, issue 1, pp. 52-59, 2011.

[33] Jinhui Liao et al, “Tumor expressed PTHrP facilitates prostate cancer-induced osteoblastic lesions”, International Journal of Cancer, vol. 123, issue 10, pp. 2267-2278, 2008.

[34] Leonard J. Deftos, “Direct evidence that PTHrP expression promotes prostate cancer progression in bone”, Biochemical and Biophysical Research Communications, vol. 327, issue 2, pp. 468-472, 2005.

[35] Xiaoli Shen et al, “PTH-related protein modulates PC-3 prostate cancer cell adhesion and integrin subunit profile”, Molecular and Cellular Endocrinology, vol. 199, issues 1-2, pp. 165-177, 2003.

[36] Zeyad J. EI-Akawi et al, “Alpha-1 antitrypsin (alpha1-AT) plasma levels in lung, prostate and breast cancer patients”, Neuro endocrinology letters, vol. 29, issue. 4, pp. 482-484, 2008.

[37] Solo Kuvibidila et al, “Correlation between serum prostate-specific antigen and alpha-1-antitrypsin in men without and with prostate cancer”, Journal of Laboratory and Clinical Medicine, vol. 147, issue 4, pp. 174-181, 2006.

[38] Risto Kajanne et al, “Transcription factor AP-1 promotes growth and radioresistance in prostate cancer cells”, International Journal of Oncology, vol. 35, issue. 5, pp. 1175-1182, 2009.

[39] Xiaoping Zhang et al, ″Repression of NF-κB and activation of AP-1 enhance apoptosis in prostate cancer cells″, International Journal of Cancer, vol. 124, issue 8, pp. 1980-1989, 2009.

[40] Changmeng Cai et al, “c-Jun Has Multiple Enhancing Activities in the Novel Cross Talk between the Androgen Receptor and Ets Variant Gene 1 in Prostate Cancer”, Molecular Cancer Research, vol. 5, issue 7, pp. 725-735, 2007.

[41] Hui Wang et al, “p53-induced Gene 3 Mediates Cell Death Induced by Glutathione Peroxidase 3”, The Journal of Biological Chemistry, vol. 287, issue 20, pp. 16890-16902, 2012.

[42] Yan P. Yu et al, “Glutathione peroxidase 3, deleted or methylated in prostate cancer, suppresses prostate cancer growth and metastasis”, Cancer Research, vol. 67, issue 17, pp. 8043-8050, 2007.

[43] A. P. Cumming et al, “PSA affects prostate cancer cell invasion in vitro and induces an osteoblastic phenotype in bone in vivo”, Prostate Cancer and Prostatic Diseases, vol. 14, issue 4, pp. 286-294, 2011.

[44] Kathryn L. Penney et al, “Association of KLK3 (PSA) genetic variants with prostate cancer risk and PSA levels”, Carcinogenesis, vol. 30, issue 6, pp. 853-859, 2011.

[45] Hemang Parikh et al, “Fine mapping the KLK3 locus on chromosome 19q13.33 associated with prostate cancer susceptibility and PSA levels”, vol. 129, issue 6, pp. 675-685, 2011.

[46] A. B. Porcaro et al, “Investigative Clinical Study on Prostate Cancer Part II: On the Role of the Pretreatment Total PSA to Free Testosterone Ratio as a Marker Assessing Prostate Cancer Prognostic Groups after Radical Retropubic Prostatectomy”, Urologia Internationalis, vol. 85, no. 2, pp. 152-158, 2010.

[47] Hellen Kuasne et al, “Polymorphisms in the AR and PSA Genes as Markers of Susceptibility and Aggressiveness in Prostate Cancer”, Cancer Investigation, vol. 28, no. 9, pp. 917-924, 2010.

[48] Astrid K. Whitbread et al, “Expression of PSA-RP2, an alternatively spliced variant from the PSA gene, is increased in prostate cancer tissues but the protein is not secreted from prostate cancer cells”, Biological Chemistry, vol. 391, issue 4, pp. 461-466, 2010.

[49] Ebru Kalay et al, “ARE-I Polymorphism on PSA Gene in Prostate Cancer Patients of a Turkish Population”, Anticancer Research, vol. 29, issue 4, pp. 1395-1398, 2009.

[50] Georgios Pampalakis et al, “Novel splice variants of prostate-specific antigen and applications in diagnosis of prostate cancer”, Clinical Biochemistry, vol. 41, issues 7-8, pp. 591-597, 2008.

[51] Sezgin Gunes et al, “Prostate-Specific Antigen and 17-Hydroxylase Polymorphic Genotypes in Patients with Prostate Cancer and Benign Prostatic Hyperplasia”, DNA and Cell Biology, vol. 26, issue 12, pp. 873-878, 2007.

[52] T. L. Veveris-Lowe et al, “Kallikrein 4 (hK4) and prostate-specific antigen (PSA) are associated with the loss of E-cadherin and an epithelial-mesenchymal transition (EMT)-like effect in prostate cancer cells”, Endocrine-Related Cancer, vol. 12, issue 3, pp. 631-643, 2005.

[53] Rui Medeiros et al, “Linkage between polymorphisms in the prostate specific antigen ARE1 gene region, prostate cancer risk, and circulating tumor cells”, The Prostate, vol. 53, issue 1, pp. 88-94, 2002.

[54] Jinsheng Weng et al, “PSGR2, a novel G-protein coupled receptor, is overexpressed in human prostate cancer”, International Journal of Cancer, vol. 118, issue 6, pp. 1471-1480, 2006.

[55] Jianghua Wang et al, ″The prostate-specific G-protein coupled receptors PSGR and PSGR2 are prostate cancer biomarkers that are complementary to α-methylacyl-CoA racemase″, The Prostate, vol. 66, issue 8, pp. 847-857, 2006.

[56] Sugure Maruta et al, “E1AF expression is associated with extra-prostatic growth and matrix metalloproteinase-7 expression in prostate cancer”, APMIS, vol. 117, issue 11, pp. 791-796, 2009.

[57] Scott A. Tomlins et al, “TMPRSS2:ETV4 Gene Fusions Define a Third Molecular Subtype of Prostate Cancer”, Cancer Research, vol. 66, issue 7, pp. 3396-3400, 2006.

[58] Ai-jun Liu et al, “Quantitative analysis of a panel of gene expression in prostate cancer—with emphasis on NPY expression analysis”, Journal of Zhejiang University SCIENCE B, vol. 8, issue 12, pp. 853-859, 2007.

[59] Massimiliano Ruscica et al, “Activation of the Y1 Receptor by Neuropeptide Y Regulates the Growth of Prostate Cancer Cells”, Endocrinology, vol. 147, issue 3, pp. 1466-1473, 2005.

[60] Anna Krześlak et al, “Effect of metallothionein 2A gene polymorphism on allele-specific gene expression and metal content in prostate cancer”, Toxicology and Applied Pharmacology, vol. 268, issue 3, pp. 278-285, 2013.

[61] J. Gumulec et al, “Evaluation of alpha-methylacyl-CoA racemase, metallothionein and prostate specific antigen as prostate cancer prognostic markers”, Neoplasma, vol. 59, issue 2, pp. 191-201, 2012.

[62] Prema S Rao et al, ″Metallothionein 2A interacts with the kinase domain of PKCμ in prostate cancer″, Biochemical and Biophysical Research Communications, vol. 310, issue 3, pp. 1032-1038, 2003.

[63] Jean-Philippe Coppe et al, “Id-1 and Id-2 Proteins as Molecular Markers for Human Prostate Cancer Progression”, Clinical Cancer Research, vol. 10, issue 6, pp. 2044-2051, 2004.

[64] B. Xu et al., “A functional polymorphism in MSMB gene promoter is associated with prostate cancer risk and serum MSMB expression”, Prostate, vol. 70, issue 10, pp. 1146-1152, 2010.

[65] B. L. Chang et al., “Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk”, Human molecular genetics, vol. 18, issue 7, pp. 1368-1375, 2009.

[66] H. Lou et al., “Fine mapping and functional analysis of a common variant in MSMB on chromosome 10q11.2 associated with prostate cancer susceptibility”, PNAS, vol. 106, issue 19, pp. 7933-7938, 2009.

[67] L. W. Harries et al., “Alterations in LMTK2, MSMB and HNF1B gene expression are associated with the development of prostate cancer”, BMC Cancer, vol. 10:315, 2010.

[68] W. G. Jiang et al., “The prostate transglutaminase (TGase-4, TGaseP) regulates the interaction of prostate cancer and vascular endothelial cells, a potential role for the ROCK pathway”, Microvascular research, vol. 77, issue 2, pp. 150-157, 2009.

[69] Z. Cao et al., “Overexpression of transglutaminase 4 and prostate cancer progression: a potential predictor of less favourable outcomes”, Asian journal of andrology, vol. 15, issue 6, pp. 742-746, 2013.

[70] Z. Shaikhibrahim et al., “Analysis of laser-microdissected prostate cancer tissues reveals potential tumor markers”, International journal of molecular medicine, vol. 28, issue 4, pp. 605-611, 2011.

[71] T. Nakamura et al., “Alternative splicing isoforms of hippostasin (PRSS20/KLK11) in prostate cancer cell lines”, Prostate, vol. 49, issue 1, pp. 72-78, 2001.

[72] X. Bi et al., “Association of TMPRSS2 and KLK11 gene expression levels with clinical progression of human prostate cancer”, Medical oncology, vol. 27, issue 1, pp. 145-151, 2010.

[73] G. Di Lorenzo et al., “Expression of epidermal growth factor receptor correlates with disease relapse and progression to androgen-independence in human prostate cancer”, Clinical cancer research, vol. 8, issue 11, pp. 3438-3444, 2002.

[74] R. B. Shah et al., “Epidermal growth factor receptor (ErbB1) expression in prostate cancer progression: correlation with androgen independence”, Prostate, vol. 66, issue 13, pp. 1437-1444, 2006.

[75] T. D. Chuang et al., “Human prostatic acid phosphatase, an authentic tyrosine phosphatase, dephosphorylates ErbB-2 and regulates prostate cancer cell growth”, The Journal of biological chemistry, vol. 285, issue 31, pp. 23598-23606, 2010.

[76] S. Gunia et al., “Expression of prostatic acid phosphatase (PSAP) in transurethral resection specimens of the prostate is predictive of histopathologic tumor stage in subsequent radical prostatectomies”, Virchows Archiv, vol. 454, issue 5, pp. 573-579, 2009.

[77] Q. Huo, “Protein complexes/aggregates as potential cancer biomarkers revealed by a nanoparticle aggregation immunoassay”, Colloids and surfaces. B, Biointerfaces, vol. 78, issue 2, pp. 259-265, 2010.

[78] A. Canacci et al, “Expression of semenogelins I and II and its prognostic significance in human prostate cancer”, The Prostate, vol. 71, issue 10, pp. 1108-1114, 2011.

[79] P. Alhopuro et al, “Somatic mutation analysis of MYH11 in breast and prostate cancer”, BMC Cancer, vol. 17, issue 8, 263, 2008.

[80] H. Wang et al, “Ultrasound-targeted microbubble destruction combined with dual targeting of HSP72 and HSC70 inhibits HSP90 function and induces extensive tumor-specific apoptosis”, Int J Oncol, vol. 45, issue 1, pp. 157-164, 2014.

[81] V. Gabai et al, “Increased expression of the major heat shock protein Hsp72 in human prostate carcinoma cells is dispensable for their viability but confers resistance to a variety of anticancer agents”, Oncogene, vol. 24, issue 20, pp. 3328-3338, 2005.

[82] G. Sun et al, “Filamin A regulates MMP-9 expression and suppresses prostate cancer cell migration and invasion”, Tumour Biol, vol. 35, issue 4, pp. 3819-3826, 2014.

[83] F. Licastro et al, “Alpha 1 antichymotrypsin genotype is associated with increased risk of prostate carcinoma and PSA levels”, Anticancer Res, vol. 28, issue 1B, pp. 395-399, 2008.

[84] X. Qian et al, “Spondin-2 (SPON2), a more prostate-cancer-specific diagnostic biomarker”, PLoS ONE, vol. 7, issue 5, e37225, 2012.

[85] S. Shaheduzzaman et al, “Silencing of Lactotransferrin expression by methylation in prostate cancer progression”, Cancer Biol Ther, vol. 6, issue 7, pp. 1088-1095, 2007.

[86] P. Prasad et al, “Expression of the actin-associated protein transgelin (SM22) is decreased in prostate cancer”, Cell Tis-sue Res, vol. 339, issue 2, pp. 337-347, 2010.

[87] Z. Yang et al, “Transgelin functions as a suppressor via inhibition of ARA54-enhanced androgen receptor transactiva-tion and prostate cancer cell growth”, Mol Endocrinol, vol. 21, issue 2, pp. 343-358, 2007.

[88] P. Cozzi et al, “MUC1, MUC2, MUC4, MUC5AC and MUC6 expression in the progression of prostate cancer”, Clin Exp Metastasis, vol. 22, issue 7, pp. 565-573, 2005.

[89] L. Oleksowicz et al, “Secretory phospholipase A2-IIa is a target gene of the HER/HER2-elicited pathway and a poten-tial plasma biomarker for poor prognosis of prostate cancer”, The Prostate, vol. 72, issue 10, pp. 1140-1149, 2012.

[90] Z. Dong et al, “Secretory phospholipase A2-IIa is involved in prostate cancer progression and may potentially serve as a biomarker for prostate cancer”, Carcinogenesis, vol. 31, issue 11, pp. 1948-1955, 2010.

[91] B. Nagy et al, “Overexpression of CD24, c-myc and phospholipase 2A in prostate cancer tissue samples obtained by needle biopsy”, Pathol Oncol Res, vol. 15, pp. 279-283, 2009.

[92] T. Mirtti et al, “Group IIA phospholipase A as a prognostic marker in prostate cancer: relevance to clinicopathological variables and disease-specific mortality”, APMIS, vol. 117, issue 3, pp. 151-161, 2009.

[93] N. Leveille et al, “Androgensdown-regulatemyosinlightchain kinase inhumanprostatecancercells”, J Steroid Biochem Mol Biol, vol. 114, pp. 174-179, 2009.

[94] Z. Shang et al, “Human kallikrein 2 (KLK2) promotes prostate cancer cell growth via function as a modulator to pro-mote the ARA70-enhanced androgen receptor transactivation”, Tumor Biol, vol. 35, pp. 1881-1890, 2014.

[95] R. Klein et al, “Blood biomarker levels to aid discovery of cancer-related single-nucleotide polymorphisms: kallikreins and prostate cancer”, Cancer Prev Res, vol. 3, issue 5, pp. 611-619, 2010.

[96] G. Mize et al, “Prostate-specific kallikreins-2 and -4 enhance the proliferation of DU-145 prostate cancer cells through protease-activated receptors-1 and -2”, Mol Cancer Res, vol. 6, issue 6, pp. 1043-1051, 2008.

[97] O. Perera et al, “Trefoil factor 3 (TFF3) enhances the oncogenic characteristics of prostate carcinoma cells and reduces sensitivity to ionising radiation”, Cancer Letters, vol. 361, issue 1, pp. 104-111, 2015.

[98] Z. Shaikhibrahim et al, “Analysis of laser-microdissected prostate cancer tissues reveals potential tumor markers”, Int J Mol Med, vol. 28, issue 4, pp. 605-611, 2011.

[99] E. Vestergaard et al, “Promoter hypomethylation and upregulation of trefoil factors in prostate cancer”, Int J Cancer, vol. 127, issue 8, pp. 1857-1865, 2010.

[100] Uma R. Chandran et al, “Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process”, BMC Cancer, vol. 7, issue 1, pp. 64-84, 2007.

[101] Hyunjin Kim et al, “ICP: A Novel Approach to Predict Prognosis of Prostate Cancer with Inner-class Clustering of Gene Expression Data”, Computers in biology and medicine, vol. 43, no. 10, pp. 1363-1373, 2013.

Hyunjin Kim received a Bachelor of Engineering degree in computer science from Yonsei University, Seoul in 2010. He is currently a Ph. D. candidate in the Yonsei University supervised by Professor Sanghyun Park. He has interests in development of data mining algorithms with large-scale biomedical data to produce clinically actionable information. He would like to help disease patients and also would like to identify causes and development processes of serious and rare diseases. His ultimate research goal is constructing a medical decision support system with diagnosis and prognosis of various diseases.

Sang-min Choi received a Ph.D. degree in the Department of Computer Science at Yonsei University, Seoul. His research interests include recommendation system, algorithm design, and formal verification.

Sanghyun Park received his B.S. and M.S. degrees (computer engineering) from Seoul National University in 1989 and 1991, respectively, under the supervision of Prof. Sukho Lee. Then he received a Ph.D. degree (computer science) from University of California at Los Angeles (UCLA) in 2001 under the supervision of Prof. Wesley W. Chu. His thesis title is Indexing Techniques for Similarity Searches in Sequence Databases. His research interest includes database, data mining and bioinformatics, and I am currently supervising the Data Engineering Lab in Yonsei University.