Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 726413, 12 pages
doi:10.1155/2010/726413

Research Article

Neighborhood Rough Set Reduction-Based Gene Selection and Prioritization for Gene Expression Profile Analysis and Molecular Cancer Classification

Mei-Ling Hou,[1, 2] Shu-Lin Wang,[1, 3] Xue-Ling Li,[1] and Ying-Ke Lei[1, 4]

[1] Intelligent Computing Laboratory, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China
[2] Department of Biology, University of Science and Technology of China, Hefei, Anhui 230027, China
[3] School of Computer and Communication, Hunan University, Changsha, Hunan, Anhui 410082, China
[4] Department of Information, Electronic Engineering Institute, Hefei 230037, China

Correspondence should be addressed to Shu-Lin Wang, jt_slwang@hotmail.com

Received 11 October 2009; Accepted 7 April 2010

Academic Editor: Ying Xu

Selection of reliable cancer biomarkers is crucial for gene expression profile-based precise diagnosis of cancer type and successful treatment. However, current studies are confronted with overfitting and dimensionality curse in tumor classification and false positives in the identification of cancer biomarkers. Here, we developed a novel gene-ranking method based on neighborhood rough set reduction for molecular cancer classification based on gene expression profile. Comparison with other methods such as PAM, ClaNC, Kruskal-Wallis rank sum test, and Relief-F, our method shows that only few top-ranked genes could achieve higher tumor classification accuracy. Moreover, although the selected genes are not typical of known oncogenes, they are found to play a crucial role in the occurrence of tumor through searching the scientific literature and analyzing protein interaction partners, which may be used as candidate cancer biomarkers.

1. Introduction

DNA microarray technology, a powerful tool in functional genome studies, has yet to be widely accepted for extracting disease-relevant genes, diagnosis, and classification of human tumor [1–3]. Generally, genes are ranked according to their differential expression by analysis of combination of normal and tumor samples, and genes above a predefined threshold are considered as candidate genes for the cancer being studied [4]. However, this method may produce a vast number of false, positives. In addition to the false-positive problem, the imbalance between the number of samples and genes may potentially degrade the classification accuracy and it can lead to possible overfitting and dimensional curse or even to be a complete failure in the analysis of microarray data [2]. An efficient way to solve these problems is gene selection. In fact, a good gene-selection method that can identify key tumor-related genes is of vital importance for tumor classification and identification of diagnostic and prognostic signatures for predicting therapeutic responses [5, 6].

Identifying minimum gene subsets means discarding most noise and redundancy in dataset to the utmost extent, resulting in not only classification accuracy improvement but also tumor diagnosis cost decrease in clinical application, which is still a key challenge in gene expression profile- (GEP-) based tumor classification. Rough set theory has been successfully used in feature selection [7, 8]. However, it is difficult to directly and effectively deal with real-valued attributes of microarray dataset [9]. Dataset discretization is usually adopted to tackle the problem, but the pretreatment may lose some useful information. To combat this problem, Hu et al. [10] first presented the basic concepts on neighborhood rough set (NRS) model and designed a novel feature selection method called forward attribute reduction based on neighborhood model (FARNeM) to select a minimal reduct, which avoided the preprocess of data discretization and hence decreased the information lost in pretreatment. But the reduct which satisfies criterions of higher classification performance and fewer gene numbers is not unique and full of chance. Obviously, it is not appropriate to use only a gene subset (a reduct) to train classifier, which necessitates it to select numerous minimal gene subsets with the highest or near highest dependence on training set to avoid the selection bias problem. Breadth-First Search (BFS) [11], a basic graph search algorithm that begins at the root node and explores all the neighboring nodes, were adopted to implement our goals for selecting any number of optimal and minimum gene subsets. However, for n nodes, there are 2^n combinations of gene subsets in total. It is not practical to search all of the gene subsets in 2^n combinations. The computational complexity is too high. To circumvent these problems, we proposed a breadth-first heuristic search algorithm based on neighborhood rough set (HBFSNRS) to select numerous gene subsets. The dependence function of NRS was selected as the heuristic information.

To prioritize the numerous selected genes, a parameter sig was introduced. Previous studies showed that significant class predictor genes whose expression profile vector show remarkable discrimination capability among different class samples of specific cancer maybe play a crucial role in the development of cancer [4]. We hypothesized that the occurrence probability of genes in the final selected gene subsets may reflect the power of tumor classification and the significance of them to some extent. To probe our hypothesis, several publicly available microarray datasets were applied. HBFSNRS method was also compared with four related methods: PAM, ClaNC, Kruskal-Wallis rank sum test (KWRST), and Relief-F to demonstrate its good performance, efficiency, and effectiveness in gene selection, prioritization and cancer classification.

2. Materials and Methods

2.1. The Framework of Our Analysis Method. Our proposed method is different from the traditional gene selection strategies: Filters and Wrappers. The Filter methods are based mostly on selecting genes using between-class separability criterion [12], and they do not use feedback information from predictor performance in the process of gene selection, such as relative entropy, information gain, KWRST, and t-test. The wrapper methods select genes by using a predictor performance as a criterion of gene subset selection such as GA/SVM [13] and GA/KNN [14]. Our method is a combination of Filter and Wrapper methods. A novel HBFSNRS-based cancer classification framework is illustrated as Figure 1. Four major steps of the designed method are described as follows.

2.2. Gene Pre-Selection Based on KWRST. All of the microarray datasets, without respect to training and test dataset, were normalized per gene by subtracting the minimum expression measurements and dividing by the difference between the maximal and minimum values of that gene. The expression levels for each gene were scaled on [0, 1].

Gene preselection can improve the classification performances since it may reduce the noise, which is also the common procedure for most classification application [15]. We applied gene preselection on training dataset to reduce the noise. All of the genes on the arrays of training data were sorted according to KWRST which is suitable for multiclass problem. In this study, the p top ranking genes (the initial informative gene set G^*) were used for finding minimum gene subsets for constructing ensemble tumor classifier with HBFSNRS. Generally speaking, more than 1% of genes in the human genome are involved in oncogenesis [16], so we set the number of the selected top-ranked gene $p = 300$.

2.3. Neighborhood Rough Set Reduction. The basic concepts of neighborhood rough set (NRS) have been introduced by Hu et al. [10]. In our proposed algorithm, the dependence function of NRS was introduced to evaluate the goodness of selected gene subsets. Here, we presented only the basic notation from NRS approach used in the paper.

Assume there are c subclasses of cancers, let $D = \{d_1, d_2, \dots, d_m\}$ denotes the class labels of m samples, where $d_i = k$ indicates the sample i being cancer k, where $k = 1, 2, \dots, c$. Let $S = \{s_1, s_2, \dots s_m\}$ be a set of samples and $G^* = \{g_1, g_2, \dots, g_n\}$ be a set of genes, the corresponding gene expression matrix can be represented as $X = (x_{ij})_{m\times n}$, where x_{ij} is the expression level of gene g_i in sample s_j, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$, and usually $n \gg m$.

Given an information system for classification learning $\mathrm{NDT} = \langle S, G^* \cup D, V, f \rangle$, where S is a nonempty sample set called sample space, G^* is a nonempty set of genes also called condition attributes to characterize the samples, D is a set of output variable called decision attribute (class labels of tumor samples), V_a is a value domain of attribute $a \in G^* \cup D$, f is an information function $f : S \times (G^* \cup D) \to V$, $V = \cup_{a \in G^* \cup D} V_a$, a reduction is a minimal set of attributes $B \subseteq G^*$.

Given for all $s_i \in S$ and $B \subseteq G^*$, the neighborhood $\delta_B(s_i)$ of s_i in the subspace B is defined as

$$\delta_B(s_i) = \left\{ s_j \mid s_j \in S, \Delta_B\left(s_i, s_j\right) \le \delta \right\}, \tag{1}$$

where δ is the threshold and $\Delta_B(s_i, s_j)$ is the metric function in subspace B. There are three common metric functions that are widely used. Let s_1 and s_2 be two samples in n-dimensional space $G^* = \{g_1, g_2, \dots, g_n\}$. $f(s, g_i)$ denotes the value x_{is} of g_i in the sample s. Then Minkowsky distance is defined as

$$\Delta_p(s_1, s_2) = \left(\sum_{i=1}^{n} \left| f(s_1, g_i) - f(s_2, g_i) \right|^p \right)^{1/p}, \tag{2}$$

where (1) if $p = 1$, it is called Manhattan distance Δ_1; (2) if $p = 2$, it is called Euclidean distance Δ_2; (3) if $p = \infty$, it is called Chebychev distance. Here, we use the Manhattan distance.

Given a neighborhood decision table NDT, $X_1, X_2, \dots, X_c$ are the sample subsets with decisions 1 to c, $\delta_B(x_i)$ is the neighborhood information granules including x_i, and is generated by gene subset $B \subset G^*$, then the lower and upper

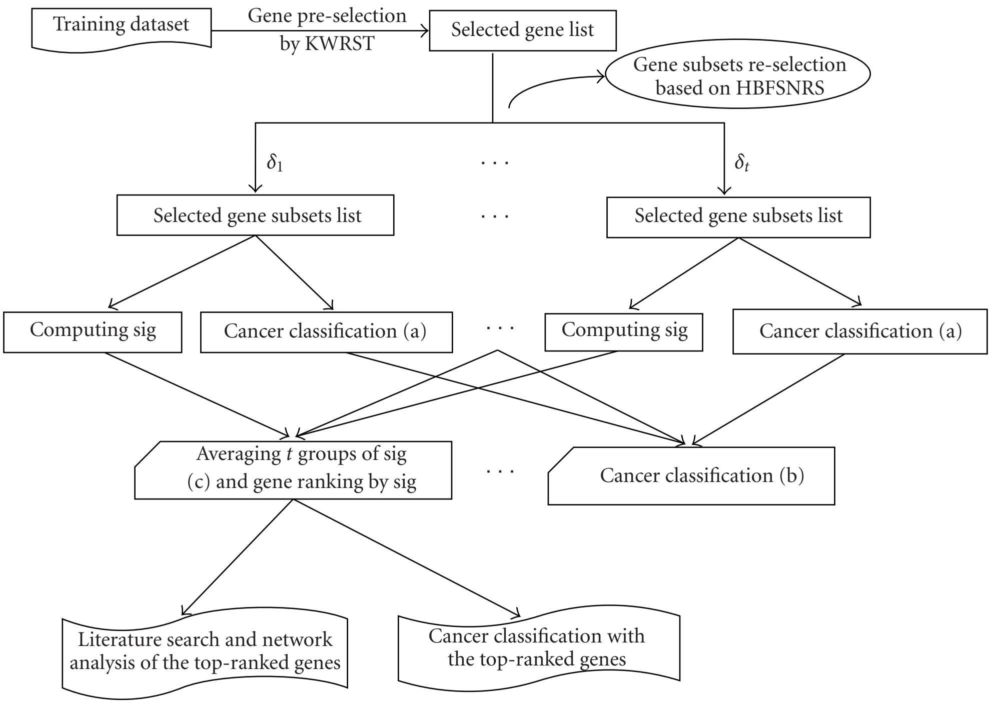

Training dataset
Gene pre-selection by KWRST
Selected gene list
Gene subsets re-selection based on HBFSNRS
δ_1
δ_t
Selected gene subsets list
Selected gene subsets list
Computing sig
Cancer classification (a)
Computing sig
Cancer classification (a)
Averaging t groups of sig (c) and gene ranking by sig
Cancer classification (b)
Literature search and network analysis of the top-ranked genes
Cancer classification with the top-ranked genes

Figure 1: The framework of our analysis method. (a) An ensemble classifier was constructed on the basis of the selected genes subsets by HBFSNRS with a specific threshold value δ. (b) Another ensemble classifier was constructed based on classification results of each δ value. (c) sig denotes the significance of genes, which is defined as (6).

approximations of the decision D with respect to gene subset B are, respectively, defined as

$$\text{Lower}_B(D) = \bigcup_{i=1}^{c} \text{Lower}_B(X_i),$$
$$\text{Upper}_B(D) = \bigcup_{i=1}^{c} \text{Upper}_B(X_i), \tag{3}$$

where $\text{Lower}_B(X) = \{x_i \mid \delta_B(x_i) \subseteq X,\ x_i \in S\}$ is the lower approximations of the sample subset X with respect to gene subset B, and is also called positive region denoted by $\text{Pos}_B(D)$ which is the sample set that can be classified into one of the classes without uncertainty with the gene subset B. $\text{Upper}_B(X) = \{x_i \mid \delta_B(x_i) \cap X \neq \phi,\ x_i \in S\}$ denotes the upper approximations, obviously $\text{Upper}_B(X) = S$. The decision boundary region of D to B is defined as

$$BN_B(D) = \text{Upper}_B(D) - \text{Lower}_B(D). \tag{4}$$

The neighborhood model divides the samples into two groups: positive region and boundary region. The decision boundary is the sample set with neighborhoods from more than one class. Through these neighborhood information, we cannot completely be sure that these samples can be classified into the class. The samples in different gene subset subspaces will have different boundary regions and positive regions. The size of the boundary region reflects the discriminability of the classification problem in the corresponding subspaces. It also reflects the recognition power or characterizing power of the condition attributes. The greater the positive region is, the smaller the boundary region will be, and the stronger the characterizing power of the condition attributes will be. So we use the dependency degree of D to B to characterize the power of the selected gene subsets, which is defined as the ratio of consistent objects

$$\gamma_B(D) = \frac{\text{Card}(\text{Pos}_B(D))}{\text{Card}(S)}, \tag{5}$$

where $\text{Card}(S)$ and $\text{Card}(\text{Pos}_B(D))$ denotes the cardinal number of sample set S and $\text{Pos}_B(D)$, respectively. If $\gamma_B(D) = 1$ we say that D depends totally on B, and if $\gamma_B(D) < 1$, we say that D depends partially. Here we define $\gamma_{\varnothing}(D) = 0$, and our goal is to find the gene subset B which $\gamma_B(D)$ is equal to the set value.

2.4. Gene Reduction Based on HBFSNRS. Informative gene selection involves evaluating the quality of the selected gene subsets and searching for good gene subsets quickly. Here, the dependence function of NRS is used to measure the goodness of the selected gene subset. Here, the computational cost problem is addressed as below.

Initially, let $\text{RED} = \{\{g_1\}, \{g_2\}, \ldots, \{g_p\}\}$ be a set of gene subsets where each subset only has an informative gene. Then, for $\forall\ \text{red}_i \in \text{RED}$, $\text{red}_i = \{g_i\}$ is expanded to $(p-1)$ subsets by adding a different genes $\{g_l \mid g_l \in G^*, g_l \notin \text{red}_i\}$ into each red_i, where we set $\text{tempory}_i = \{\{g_i g_1\}, \ldots, \{g_i g_{i-1}\}, \{g_i g_{i+1}\}, \ldots, \{g_i g_p\}\}$, we will get $p^*(p-1)$ subsets in total. Among these subsets, we select the ω top-ranked gene subsets by the dependence function that need to be expanded in the next iteration to reconstruct the set RED, and now each element of RED has 2 genes. Similarly, in the next search layer, for ∀ $\text{red}_x \in \text{RED}$, $\text{red}_x = \{g_i g_j\}$ is extended to $(p - 2)$ subsets excluding the genes have listed in the red_x, where we set $\text{tempory}_x = \{\{g_i g_j g_1\}, \ldots, \{g_i g_j g_{i-1}\}, \{g_i g_j g_{i+1}\}, \ldots, \{g_i g_j g_{j-1}\}, \{g_i g_j g_{j+1}\}, \ldots, \{g_i g_{\,j}\ g_p\}\}$, $i < j$, and we will get $w{*}(p - 2)$ subsets. Among these subsets, ω top-ranked gene subsets were selected to be expanded in next layer as the above method. Now, the element of RED has 3 genes. The search process continues following the above method until meeting the stop criteria. In each layer, we expend to $w{*}(p - \text{card}(\text{red}))$ subsets and only ω top-ranked gene subsets were selected to reconstruct the set RED from the total subsets, so that the search time will not increase exponentially with the increase of search depth. Here, card(red) denotes the cardinal gene number of the gene subset. In the virtue of the minimum construction idea, one of the techniques for the best feature selection could be based on choosing minimal gene subsets that fully describe classes of tumor classification in a given data set. Therefore, when the maximal dependence of the elements of RED (e.g., $r\_\text{Max} = 0.9999$) is obtained, the increment between the maximal dependence of two adjacent search levels is less than θ (e.g., $\theta = 0.0001$) or the number of iterative steps is equal to the set value Depth (e.g., Depth $=$ 20), the searching process ends at that level. Otherwise, we continue to search genes in this way until meeting the stopping criterions. The pseudocode of HBFSNRS is shown in Algorithm 1.

The dependence function of NRS is chosen as the objective function for evaluating the goodness of the selected gene subset mainly because it is computationally fast in that it does not use the feedback information of test data in the training process. To optimize the parameter δ in NRS that control the size of the neighborhood, different values for δ from 0 to 1 with step 0.01 were tested by running forward attribute reduction based on neighborhood model (FARNeM). δ values were sorted according to the classification accuracy by 3-KNN classifier using the corresponding gene subset selected by FARNeM. The 5 top-ranked δ values were used in the next step. But for ALL (a multiclass dataset), the gene number of the selected minimal and optimal reduct set reach 20 or even more for some of the top five δ values. Considering that a large gene subset with an excessive number of genes may contain much noise and redundancy, which may bias and negatively influence the tumor classification and gene prioritization, we discarded such top-ranked δ values and reselected five top-ranked δ values that produced reduct set with less than 20 genes.

2.5. Evaluation Criterion for the Selected Gene Subsets. We adopted 3-KNN classifier to evaluate the classification performance of the selected gene subsets. To improve prediction accuracy and stability, an ensemble classifier was constructed on the basis of the selected gene subset. For each δ, a simple majority voting strategy was applied to integrate the w individual classifier that is constructed from the selected gene subsets obtained by HBFSNRS only on training set. Then, another ensemble classifier was built based on the above classification results with each δ value in the similar way.

Here, we hypothesized that genes with higher occurrence frequency are more likely to be important and cancer-related genes. Therefore, we count the occurrence frequency of each gene in all the selected gene subsets to measure its significance. But for a specific cancer, different δ value may select different sizes of the minimum gene subset. In this case, only counting the occurrence frequency is not appropriate for measuring the significance of genes. To avoid the selection bias, the significance of genes is measured by occurrence of probability, which is defined as

$$\text{sig}_j = \frac{1}{t}\sum_{i=1}^{t}\frac{f_{ij}}{n_i * \omega}, \tag{6}$$

where f_{ij} is the occurrence frequency of gene j in all the gene subsets which are selected by HBFSNRS with δ_i; t is the total number of neighborhood values (we set $t = 5$); n_i is the number of genes in a selected gene subset with δ_i; ω is the number of the final selected gene subsets by HBFSNRS (we set $\omega = 500$).

In order to further investigate the significance of the selected gene, two main methods were used: (1) the selected genes were regarded as predictor set or classification model; (2) literature search and protein-protein interaction (PPI) network analysis.

2.6. Dataset. To evaluate the performance of the proposed method, seven gene expression datasets were used in this study: Acute Lymphoblastic Leukemia (ALL) [17], Breast cancer 30 (GSE5764) [18], Breast cancer 22(GSE8977) [18], Colon cancer [19], Prostate cancer 102 [20], and Prostate cancer 34 [21]. The two pairs of cross-platform datasets were used to evaluate the generalization performance for our cross-platform classification model. Datasets of Breast cancer, Colon cancer, and Prostate cancer are two-class classification systems that contain normal and tumor samples. ALL dataset is a multiple-class classification system. The dataset contains six subtypes of ALL: BCR-ABL, E2A-PBX1, Hyperdip >50, MLL, T-ALL, TEL-AML1. For Breast-cancer datasets, there are too many (54675) affymetrix probe identifiers, therefore the raw data were processed following these steps: affymetrix probe identifier was converted to entrez identifier. When multiple probes corresponded to the same entrez ID, we averaged over these probe intensities. The division of training set and test set is shown in Table 1.

3. Results

3.1. Redundant and Irrelevant Genes Potentially Degrade the Classification Accuracy. To avoid overfitting problem and improve classification accuracy and stability, an ensemble classifier was constructed on the basis of the selected gene subsets. We observed that the final integrated results (Table 2) were not satisfactory and no higher classification accuracy obtained compared to some individual classifiers. The main reason may be that our methods used all the selected gene subsets as classification model, which contain

```
Input $\langle S, G, D\rangle$, $\delta$, $\theta$, $p$, $\omega$, $r\_\mathrm{Max}$, and Depth//$\delta$ is the threshold to control the size of the neighborhood, $\theta$ is the
threshold of increment, $p$ is the number of the preselected genes, $\omega$ is the search breadth, $r\_\mathrm{Max}$ is a given maximal
dependency function value and Depth is the upper bound of searching depth.
Output RED is the pool to contain the selected gene subsets red.
Step 1: For each $g_i \in G$//Compute p-value by KWRST
            $P_i = \mathrm{KWRST}(g_i)$;
        End
Step 2: $gg = \mathrm{sort}(P, \text{"ascend"})$; //Rank genes by $P$ in ascending order
Step 3: $G^* = \mathrm{Select}(G, gg, p)$; //Select he $p$ top-ranked genes as the initial gene set $G^*$ by $P$
Step 4: For each $g_i \in G^*$//Let RED $= \{\{g_1\}, \{g_2\}, \ldots \{g_p\}\}$ be a set of gene subsets where each
            $g_i \rightarrow \mathrm{red}_i$; //gene subset only has an informative gene.
            $\mathrm{red}_i \rightarrow$ RED;
        End
Step 5: $iter = 1$; //The times of iteration.
Step 6: For each $\mathrm{red}_j \in$ RED
            For each $g_k \in G^* - \mathrm{red}_j$
                $\mathrm{red}_j \cup g_k \rightarrow$ RED; //Adding genes not listed in $\mathrm{red}_j$ to $\mathrm{red}_j$ and save it as elements of RED
                $\gamma_{\mathrm{red}_j \cup g_k}(D) = \mathrm{Card}(\mathrm{Pos}_{\mathrm{red}_j \cup g_k}(D))/\mathrm{Card}(S)$; //Compute dependence degree of $D$ to $\mathrm{red}_j \cup g_k$.
            End
        End
Step 7: $rr = \mathrm{sort}(r, \text{"descending"})$; //Rank gene subsets by$r$in descending order
        RED = Select(RED, $rr$, $w$); //Select $\omega$ top-ranked gene subsets to reconstruct RED.
Step 8: If $(\max_{\mathrm{iter}}(\gamma) >= r\_\mathrm{Max})$ or $\mathrm{abs}(\max_{\mathrm{iter}}(r) - \max_{\mathrm{iter}-1}(r)) < \theta$ or (iter = Depth)
            Break;      //here, we define $\max_0(r) = 0$
        Else
            iter = iter + 1;
            Go to step 6;
        End
```

Algorithm 1: A heuristic breadth-first search algorithm based on neighborhood rough set (HBFSNRS).

Table 1: The division of training set and test set in our experiments.

No.	Dataset	Training set	Test set	No. of gene	No. of class
1	ALL	148	100	12625	6
2	Breast cancer	Breast30	Breast22	19802	2
3	Colon	42	20	2000	2
4	Prostate cancer	Prostate102	Prostate34	12600	2

many redundant and tumor-unrelated genes and may potentially degrade the classification performance. Figure 2 shows the classification accuracy with different numbers of the top-ranked genes sorted according to the significance of genes defined as (6), from which we found that only a few top-ranked genes were enough to obtain higher classification accuracy. Meanwhile, when more genes were used as predictor set, there was only a little increase or even decrease in the classification performance. Therefore, we inferred that too many selected genes involve much more redundancy and irrelevancy, which degrades the classification accuracy.

3.2. Comparison with Other Related Methods. In order to elaborate the effectiveness of HBFSNRS, we compared the accuracy of our approach with other common filter methods including t-test, information gain, KWRST, and Relief-F. The experimental results indicate that our method is significantly superior to t-test and information gain, and slightly outperforms KWRST and Relief-F in the aspect of tumor classification. For simplicity, we only present KWRST and Relief-F results here (Figure 2). We found that only a few top-ranked genes could achieve higher accuracy in the classification of tumor samples of different classes by our proposed search algorithm. For ALL dataset, the prediction accuracy by HBFSNRS is superior to other methods regardless of the much fewer genes used in cancer classification. For breast-cancer dataset, using one active gene could test outcome with the accuracy of 22.73% by Relief-F, 63.64% by KWRST, whereas 100% test accuracy was obtained using one gene by the proposed HBFSNRS method. For colon-cancer dataset, using one, six active genes could get the prediction accuracy of 80% and 85% by our method, 65%, 70% by Relief-F, and 65%, 75% by KWRST, respectively. For prostate-cancer dataset, when using more than ten genes for tumor classification, KWRST significantly outperformed our method and Relief-F, but our method performs as well as the

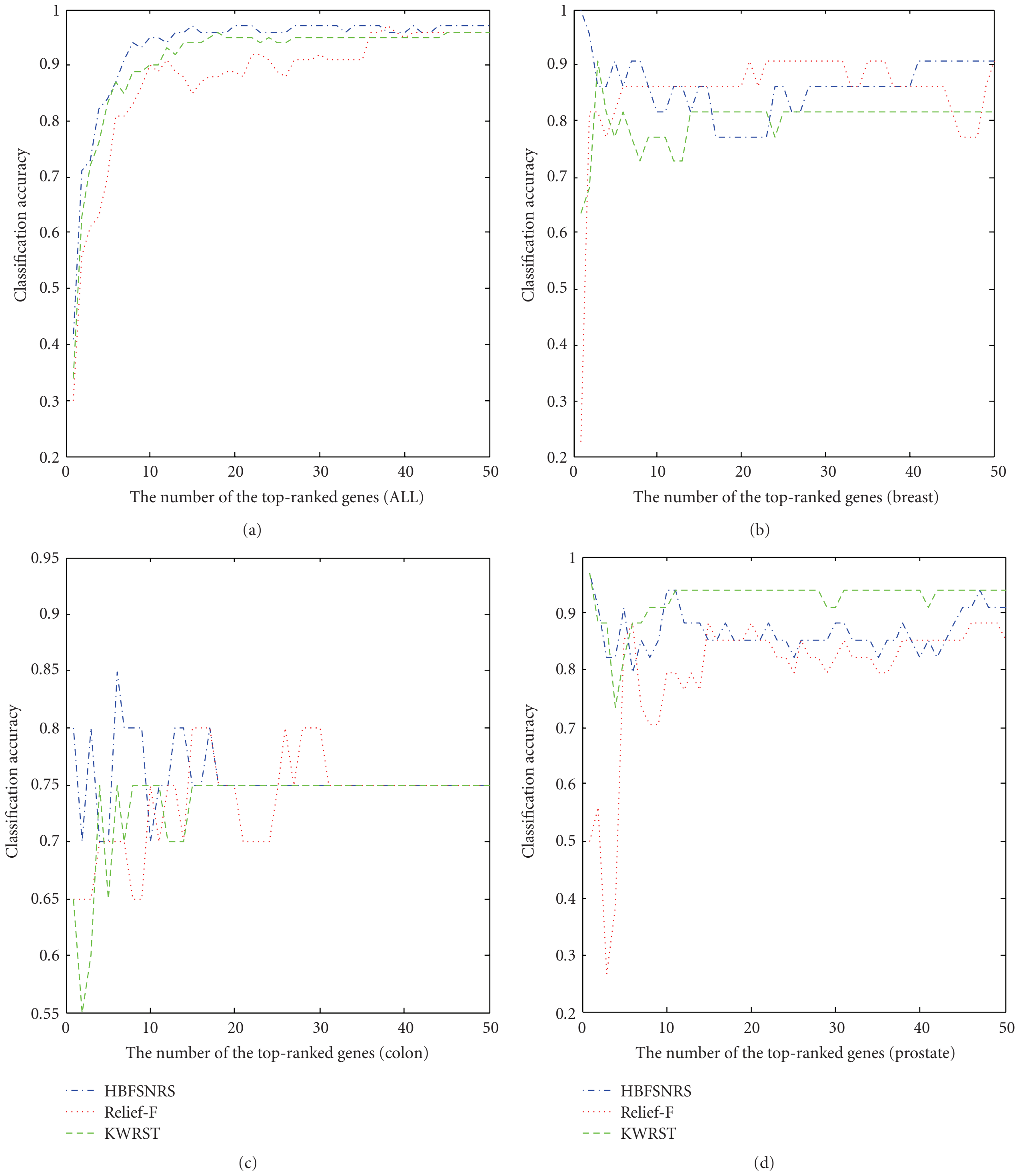


Figure 2: Comparison of classification accuracy with different numbers of top-ranked genes on the four test datasets by HBFSNRS, Relif-F, and KWRST.

KWRST when only using the few top-ranked genes (both of our method and KWRST could get 97.06% accuracy using one gene). What is more, we compared our method with other statistical methods PAM and ClaNC. PAM, a statistical technique for class prediction from gene expression data that uses nearest shrunken centroids, was used to identify class predictor genes [22]. ClaNC ranks genes by standard t-statistics, which does not shrink centroids and uses a class-specific gene selection procedure [23]. In our context, ClaNC slightly outperformed PAM, so we only present the comparison with ClaNC here (Table 3). In comparison with ClaNC, our method could obtain higher classification accuracy when using a few top-ranked genes. The one-gene model by our method provides the classification accuracy of 100%, 80%, and 97.06% for Breast-cancer, Colon-cancer, and Prostate-cancer dataset, respectively, whereas ClaNC requires more genes to get the same accuracy. In ALL dataset, the test accuracies on independent test dataset are 87% with six genes, 94% with 12 genes, and 97% with 18 genes by our method. Using the same six, 12, 18 active genes could test outcome with the accuracy of 86%, 95%, and 97% by ClaNC, respectively, which indicates our method was comparable for ALL dataset. As a comparison, the minimum genes with the highest accuracy can be obtained in the classification process by HBFSNRS. In addition, results show that our method is obviously better than ClaNC in colon-cancer and breast-cancer cross-platform datasets. It is likely that ClaNC is not suitable for cross-platform datasets. We proposed that these few genes whose expression profile vector showed remarkable discrimination capability may closely correlated to cancer and could be seen as possible disease signatures.

3.3. Analysis of the Top-Ranked Genes (Case Studies). Mining genes that give rise to ontogenesis is one of key challenges in the area of cancer research. Biologically the experimental results proved that the selected genes with high classification accuracy are functionally related to carcinogenesis or tumor histogenesis, so we could infer that the few top-ranked genes may be very important for tumor diagnosis. The 10 top-ranked genes according to the sig score for each tumor that were regarded as the candidate cancer genes listed in Table 4. To demonstrate our method's ability in uncovering known cancer genes and predicting novel cancer biomarkers, the breast-cancer dataset was employed to this study as the method of [24].

First, we checked whether our method can uncover known famous cancer genes. We downloaded a list of 25 breast cancer biomarkers that have been annotated in the OMIM database [25]. Unfortunately, our used dataset (the 300 top-ranked genes selected by KWRST) does not include the 25 known breast cancer genes. Therefore our method cannot be evaluated with it in terms of uncovering known cancer genes. From another point of view, it is verified that higher differential expression of a gene does not necessarily reflect a greater likelihood of the gene being related to cancer. In other words, important genes might not be necessarily differentially expressed. But it is undeniable that higher differential expressions of genes are inevitably important in the cancer diagnosis and development.

Next, literature search method was used to check whether our method can predict novel cancer biomarkers. In the top 10 genes ranked by (6) for breast cancer, we found that these genes play an important role in the occurrence of breast cancer. The collagen triple helix repeat containing 1 (CTHRC1), ranked the first, whose aberrant expression is widely presented in human solid cancers including breast cancer and seems to be associated with cancer tissue invasion and metastasis [26]. The PDZ and LIM domain protein 4 (PDLIM4), ranked the second, was frequently methylated in breast cancers but not in normal breast tissues [27]. The keratin, type I cytoskeletal 17 (KRT17), ranked the third, was specifically overexpressed in basal-like subtypes of breast cancer [28]. The secreted frizzled-related protein 1 (SFRP1), ranked the fourth, was recently found to be associated with progression and poor prognosis in early stage of breast cancer [29]. The collagen alpha-1 (III) chain (COL3A1), ranked the fifth, was up-regulated in both invasive ductal and lobular carcinomas cells when compared with normal ductal and lobular cells [30]. The peptidase inhibitor 15 (PI15), ranked the sixth, was also differentially expressed but it was down regulated in lobular and ductal invasive breast carcinomas [30]. The actin gamma-enteric smooth muscle (ACTG2), ranked the seventh, is involved in the architecture and remodeling of cytoskeleton in basal medullary breast cancer [31]. The tissue factor pathway inhibitor 2 (TFPI2), ranked the eighth, whose aberrant hypermethylation with gene promoter was associated with metastasis in breast cancer [32]. The serpin B5 (SERPINB5), ranked the ninth, an epithelial-specific serine protease inhibitor, was a biomarker in disseminated breast-cancer cells [33].The fibronectin 1 (FN1), ranked the tenth, was recently suggested to be associated with the prognosis of patients with breast cancers [34].

Finally, we examined gene pathway that involved by the 10 top-ranked genes. The study is carried out using the software which can help the researchers to better understand the biological phenomenon understudied by pointing out significant cellular functions of the selected genes from the webpage "http://vortex.cs.wayne.edu/projects.htm" [35]. Results indicate that the pathways that the 10 top-ranked genes are involved in are ECM-receptor interaction (COL3A1, FN1), focal adhesion (COL3A1, FN1), vibrio cholerae infection (ACTG2), p53 signaling pathway (SERPINB5), Small cell lung cancer (FN1), wnt signaling pathway (SFRP1), regulation of actin cytoskeleton (FN1), pathways in cancer (FN1), which agree well with current knowledge on breast cancer [36]. Thus it can be seen that the selected genes that closely related to adhesion, motility, and metastasis may provide new insights in the underlying molecular mechanisms related to disease development, in designing therapy and in prognostication for patients with breast carcinoma. Thus, the analysis of existing biological experiment results of breast-cancer dataset well illustrates that our method has great power of identifying tumor-related genes.

Furthermore, another case study for prostate-cancer dataset was presented here. In the 10 top-ranked genes, six of them (HPN, MAF, GSTP1, WWC1, JUNB, and RND3) have been reported to be associated with prostate cancer. The hepsin (HPN), ranked the first, a cell surface serine protease that is markedly up-regulated in human prostate cancer, which is overexpression in prostate epithelium *in vivo* causes disorganization of the basement membrane and promotes primary prostate cancer progression and metastasis to liver, lung, and bone [37]. The transcription factor (MAF), ranked the second, was down-regulated in the tumors relative to normal prostate tissue and may be regarded as the candidate tumor suppressor gene [38]. The glutathione s-transferase P (GSTP1), ranked the fourth, whose CpG island hypermethylation is the most common somatic genome alteration

Table 2: Classification accuracy, sensitivity and specificity on all the test datasets by the ensemble classifier.

Dataset	δ value (the number of genes in the selected gene subset)					
ALL	0.32(8)	0.35(9)	0.44(13)	0.47(14)	0.66(20)	integration
Accuracy	89.00	92.00	93.00	94.00	93.00	95.00
Breast	0.04(2)	0.21(2)	0.29(2)	0.30(2)	0.69(3)	integration
Accuracy	86.36	90.91	90.91	90.91	95.45	90.91
Sensitivity	100.00	100.00	100.00	100.00	93.33	100.00
Specificity	57.14	71.43	71.43	71.43	100.00	71.43
Colon	0.03(2)	0.04(2)	0.82(6)	0.92(3)	0.13(2)	integration
Accuracy	70.00	75.00	75.00	80.00	75.00	75.00
Sensitivity	75.00	75.00	75.00	83.33	75.00	75.00
Specificity	62.50	75.00	75.00	75.00	75.00	75.00
Prostate	0.13(4)	0.20(5)	0.26(5)	0.57(5)	0.62(5)	integration
Accuracy	94.12	91.18	88.24	88.24	97.06	91.18
Sensitivity	100.00	100.00	100.00	100.00	100.00	100.00
Specificity	92.00	88.00	84.00	84.00	96.00	88.00

Table 3: The comparison with the ClaNC method in classification accuracy.

Method		Number of genes selected per subclass: n (all: $n \times c$)									
		1	2	3	4	5	6	7	8	9	10
ClaNC	ALL(×6)	86.00	95.00	97.00	99.00	98.00	99.00	99.00	99.00	99.00	98.00
	Breast(×2)	50.00	40.91	45.45	45.45	40.91	40.91	40.91	40.91	40.91	40.91
	Colon(×2)	65.00	65.00	65.00	70.00	70.00	75.00	75.00	75.00	75.00	75.00
	Prostate(×2)	73.53	85.29	79.41	76.47	76.47	79.41	79.41	76.47	76.47	79.41
Method		Number of all genes selected									
		1	2	3	4	6	8	12	18	24	30
HBFSNRS	ALL	41.00	71.00	73.00	82.00	87.00	94.00	94.00	96.00	96.00	97.00
	Breast	100.00	95.45	86.36	86.36	86.36	90.91	86.36	77.27	86.36	86.36
	Colon	80.00	70.00	80.00	70.00	85.00	80.00	75.00	75.00	75.00	75.00
	Prostate	97.06	91.18	82.35	82.35	79.41	82.35	88.24	85.29	85.29	88.24

described for human prostate cancer [39]. The gene WWC1, ranked the sixth, was found to interact with histone H3 via its glutamic acid-rich region and that such interaction might play a mechanistic role in conferring an optimal ER transactivation function as well as the proliferation of ligand-stimulated breast-cancer cells [40]. The transcription factor jun-B (JUNB), ranked the seventh, is an essential upstream regulator of p16 and contributes to maintain cell senescence that blocks malignant transformation of TAC. JUNB thus apparently plays an important role in controlling prostate carcinogenesis and may be a new target for cancer prevention and therapy [41]. The Rho-related GTP-binding protein RhoE (RND3), ranked the ninth, a recently described novel member of the Rho GTPases family, was regarded as a possible antagonist of the RhoA protein that stimulates cell cycle progression and is overexpressed in prostate cancer [42]. The remaining genes were not identified to correlate to prostate cancer previously. These genes need further analysis.

Genes related to a specific or similar disease phenotype tend to be located in a specific neighborhood in the protein-protein interaction network, and a protein is likely to be coexpressed with its interaction partners and those proteins that have similar function. Here, we applied a protein-network-based method to analyze the effect of neighborhood partners on the selected genes using all interactions in the Human Protein Reference Database [43]. Figure 3 indicates the protein-interaction network for each top-ranked gene of prostate cancer (KIAA0430 has no interaction partners in HPRD). The red-ellipse nodes represent the 10 top-ranked genes that were ranked by the sig score in (6), among which, those with an asteroid sign means known cancer genes. The diamond nodes indicate the direct interaction partners of the selected genes that were not cancer genes, and blue-octagon nodes show those partners that are identified as known cancer genes which were collected by querying the Memorial Sloan Kettering computational biology website, "Oncogene", "tumor suppressor", and "stability" are shown as [4, 44]. Among the 10 top-ranked genes for prostate-cancer dataset (Figure 3), 6 genes (ABL1, JUNB, MAP, P4HB, GSTP1, and RND3) that listed with an asteroid sign have been identified to be known cancer genes. Here, we mainly illustrate the three genes P4HB, PEX3, and ABL1 that we did not find

Table 4: The 10 top-ranked genes selected for the four datasets.

Four datasets							
ALL		Breast cancer		Colon cancer		Prostate cancer	
gene symbol	sig	gene symbol	sig	gene symbol	sig	gene symbol	sig
LRMP	0.0801	CTHRC1	0.1212	DES	0.0895	HPN	0.174
TCFL5	0.0569	PDLIM4	0.0476	MYH9	0.0834	MAF	0.1248
CD99	0.0526	KRT17	0.0321	C3	0.062	ABL1	0.0457
MPP1	0.0483	SFRP1	0.0292	FUCA1	0.0538	GSTP1	0.0225
CD72	0.0399	COL3A1	0.0261	CSRP1	0.0427	KIAA0430	0.0216
NONO	0.0377	PI15	0.0258	MT2A	0.0421	WWC1	0.0192
DNTT	0.0345	ACTG2	0.0241	TSPAN7	0.0346	JUNB	0.0164
PLXNB2	0.0329	TFPI2	0.0217	2-Sep	0.0294	PEX3	0.0153
ECM1	0.0325	SERPINB5	0.0203	FXN	0.0236	RND3	0.0151
SMARCA4	0.0296	FN1	0.0186	PMP22	0.0214	P4HB	0.0146

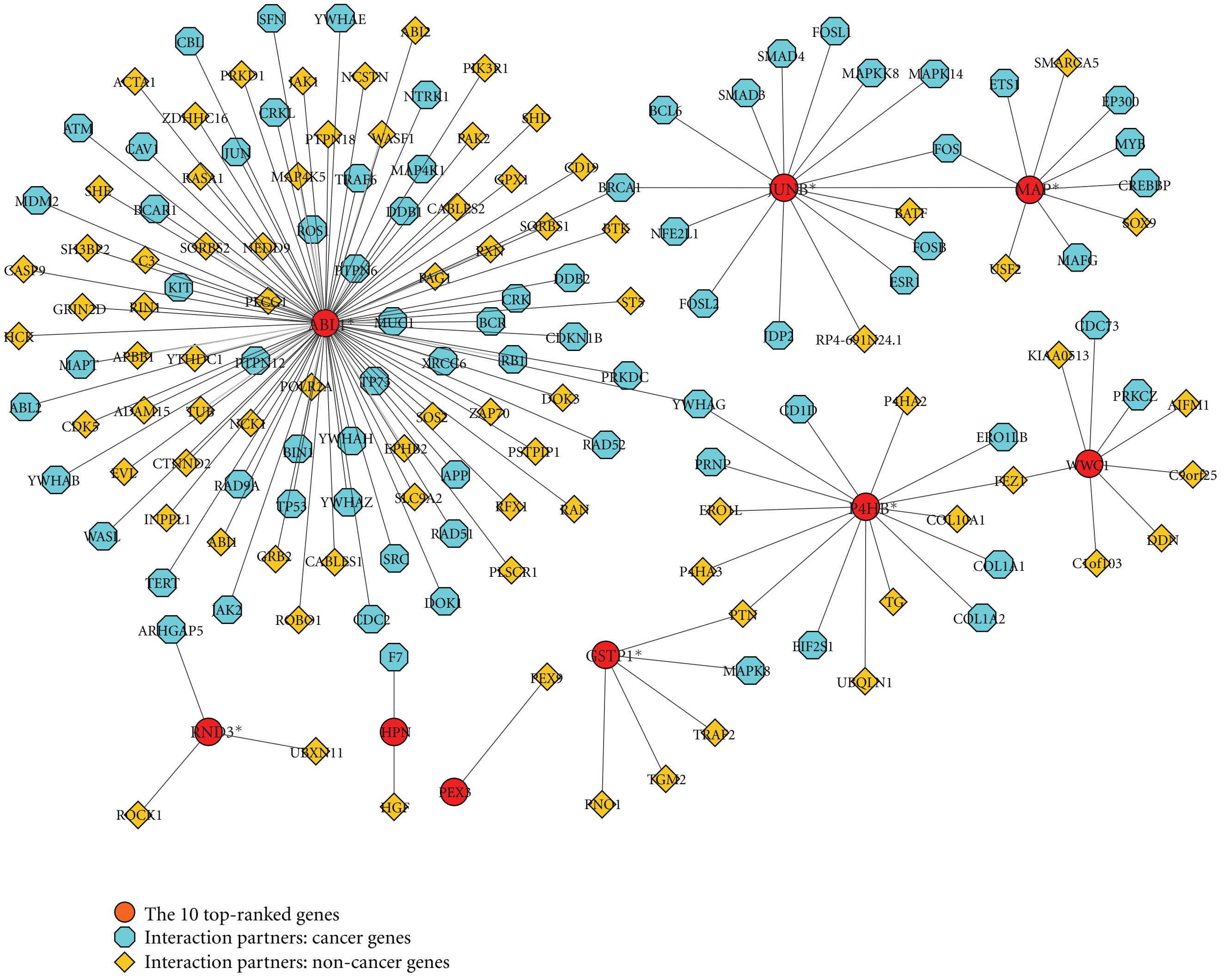


Figure 3: The protein-interaction network associated with the ten top-ranked genes for prostate cancer.

reports on their association with prostate cancer. In the three genes, P4HB and ABL1 have been known as cancer genes. PEX3 is also a famous disease gene which was the cause of peroxisome biogenesis disorder, complementation group 12, and zellweger syndrome. It can be seen that mutation in these genes can lead to many diseases and may have a close relationship with prostate cancer. In this sense, our method is effective on cancer-related gene selection. Recently, Aragues et al. [4] suggest that cancer linker degree (CLD) of a protein which was defined as the number of cancer genes to which a gene is connected is a good indicator of the probability of being a cancer gene. We analyzed the cancer linker degree (CLD) of 10 top-ranked genes on each of the four datasets. For prostate cancer, as is shown in Figure 4, most of the top-ranked genes have a direct interaction with known cancer genes excluding the gene PEX3, and the CLD of ABL1, JUNB, WWC1, MAF, P4HB, GSTP1, HPN, and RND3 is 46, 13, 2, 6, 7, 1, 1, and 1, respectively. In the 10 top-ranked genes of ALL (TCFL5 and LRMP have no interaction partners in HPRD), SMARCA4, DNTT, and NONO are known cancer genes, and the CLD of SMARCA4, DNTT, NONO, CD72, MPP1, and CD99 is 19, 3, 6, 1, 2, and 2, respectively. For breast cancer, CTHRC1, PI15, and SERPINB5 have no interaction partners in HPRD. In the remaining 7 genes of 10 top-ranked genes, SFRP1 and TFPI2 are known cancer genes, and SFRP1, TFPI2, FN1, COL3A1, and KRT17 have a direct interaction with known cancer genes, the CLD of which is 2, 1, 17, 2, and 1 respectively. For colon cancer, FUCA1 has no interaction partners in HPRD. In the remaining 9 genes, MYH9 is a known cancer gene, the CLD of DES, MYH9, C3, and 2-Sep is 4, 3, 1, and 1, respectively. These results show that besides a few selected genes that typically correspond to known specific cancer mutations, a considerable portion of the top-ranked genes have many direct interactions with cancer genes, which suggests that these genes should be very likely to be involved in cancer and may play a central role in the protein network by interconnecting many known cancer genes, and thus the top ranked genes can be regarded as reliable disease biomarkers.

4. Discussions and Conclusions

4.1. Better Performance on Tumor Classification and Gene Selection and Prioritization. An ongoing challenge is to identify new prognostic markers that are directly related to disease and that can more accurately predict the likelihood of gaining cancer in unknown samples. Results indicate that our proposed method of gene selection by HBFSNRS has the following advantages in trying to tack this challenge. (1) Our method could obtain the highest or near highest prediction accuracy of tumor classification with the minimum gene subset. (2) Lists of ranked potential candidate cancer biomarkers with a specific cancer are presented by our approach. (3) Our proposed method can obtain many optimal gene subsets in a short period of time, which is essential to the whole search process. (4) Compared to other gene ranking methods KWRST and Relief-F, our method is relatively stable and contains little chance factors. The success of our methods, gene selection by HBFSNRS, can be attributed to a combination of several aspects. First, we adopted the dependence function of NRS to evaluate the goodness of selected gene subsets. There are two main advantages for this point: time saving and tumor classification without the feedback and leaked information of the test dataset. Second and more importantly, the designed process of gene search by our method can select any number of optimal gene subsets in a comparatively short time, which is an optimization of best-first search. Finally, considering the selection of δ value in the evaluation of gene subsets has the problem that the genes with different δ value will have different ranked positions or relevance to cancer. To avoid this problem of selection bias, we defined a sig score to describe the significance of genes by combining five groups of results that obtained by each δ value. We presented two case studies on breast cancer and prostate cancer to illustrate the power of our method to identify tumor-related genes. Our method illustrates well its high power of tumor classification and gene prioritization.

4.2. Limitation and Extension. One limitation of our approach is in data quality: current high-throughput technologies remain error prone and may be far from complete. In a recent paper, Zhang et al. [45] held that the integration of microarray data gives us more analytical power and reduces the false discovery rate. Given a specific cancer, efficient ways to integrate multiple independent microarray data may be a good way to solve the issue of data quality. The other limitation is the optimization of the threshold value of neighborhood rough set. On one hand, we tried the neighborhood rough set reduction method to evaluate the goodness of the selected gene subsets to save time in tumor classification without using the feedback information of the test dataset. On the other hand, the threshold selection is obtained through the feedback information of the test set. In addition, different δ values may select different gene subsets, hence the genes with different δ value will have different positions in gene prioritization, so the selection of δ has become more critical for gene prioritization. Fortunately, the choice of δ is not so important for gene ranking because the change of gene position in different δ values is not significant. In our study, Spearman's rank correlation coefficient was used to determine whether there is a consistency between the results of gene prioritization with different δ values. Results indicate that there is high consistency among these results.

4.3. Future Work. Our proposed HBFSNRS method has improved the performance of tumor classification based on microarray and identified and prioritized lists of potential tumor-related genes from GEP, our future work will benefit further from integrating other sources. Recent high-throughput technologies have produced vast amounts of protein-protein interactions, which represent valuable resources for candidate-gene prioritization and give us new insights into the mechanism of disease. A great number of studies have shown that integration of multiple sources of data is more reliable for predicting cancer genes than the use of a single criterion [4, 46–48]. Thus, it is an efficient method to integrate GEP and protein interaction network for gene prioritization. Although gene expression data and protein interaction data have been integrated for gene prioritization [49, 50], the results are not satisfactory. Therefore, it is still a challenging problem in the area of cancer research.

Acknowledgments

This work was supported by the grants of the National Science Foundation of China, Nos. 60905023, 30900321, 60973153, 60873012 & 60805021, the grant from the National Basic Research Program of China (973 Program), No. 2007CB311002, the grant of the Guide Project of Innovative Base of Chinese Academy of Sciences (CAS), No. KSCX1-YW-R-30, the Knowledge Innovation Program of the Chinese Academy of Sciences (0823A16121), and the China Postdoctoral Science Foundation (Grant no. 20090450825).

References

[1] C.-H. Zheng, D.-S. Huang, X.-Z. Kong, and X.-M. Zhao, "Gene expression data classification using consensus independent component analysis," *Genomics, Proteomics and Bioinformatics*, vol. 6, no. 2, pp. 74–82, 2008.

[2] Q. Shen, W.-M. Shi, and W. Kong, "New gene selection method for multiclass tumor classification by class centroid," *Journal of Biomedical Informatics*, vol. 42, no. 1, pp. 3–9, 2009.

[3] H.-Q. Wang, H.-S. Wong, D.-S. Huang, and J. Shu, "Extracting gene regulation information for cancer classification," *Pattern Recognition*, vol. 40, no. 12, pp. 3379–3392, 2007.

[4] R. Aragues, C. Sander, and B. Oliva, "Predicting cancer involvement of genes from heterogeneous data," *BMC Bioinformatics*, vol. 9, p. 172, 2008.

[5] H.-Q. Wang and D.-S. Huang, "Regulation probability method for gene selection," *Pattern Recognition Letters*, vol. 27, no. 2, pp. 116–122, 2006.

[6] C.-H. Zheng, D.-S. Huang, and L. Shang, "Feature selection in independent component subspace for microarray data classification," *Neurocomputing*, vol. 69, no. 16–18, pp. 2407–2410, 2006.

[7] Z. Pawlak, "Rough set approach to knowledge-based decision support," *European Journal of Operational Research*, vol. 99, no. 1, pp. 48–57, 1997.

[8] S.-L. Wang, X. Li, S. Zhang, J. Gui, and D.-S. Huang, "Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction," *Computers in Biology and Medicine*, vol. 40, no. 2, pp. 179–189, 2010.

[9] R. Jensen and Q. Shen, "Fuzzy-rough data reduction with ant colony optimization," *Fuzzy Sets and Systems*, vol. 149, no. 1, pp. 5–20, 2005.

[10] Q. Hu, D. Yu, and Z. Xie, "Neighborhood classifiers," *Expert Systems with Applications*, vol. 34, no. 2, pp. 866–876, 2008.

[11] K. Mehlhorn and U. Meyer, "External-memory breadth-first search with sublinear I/O," in *Proceedings of the 10th Annual European Symposium on Algorithms*, vol. 2461 of *Lecture Notes in Computer Science*, pp. 723–735, 2002.

[12] R. O. Duda and P. E. Hart, *Pattern Recognition and Scene Analysis*, Wiley, New York, NY, USA, 1973.

[13] J. J. Liu, G. Cutler, W. Li et al., "Multiclass cancer classification and biomarker discovery using GA-based algorithms," *Bioinformatics*, vol. 21, no. 11, pp. 2691–2697, 2005.

[14] L. Li, T. A. Darden, C. R. Weinberg, A. J. Levine, and L. G. Pedersen, "Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method," *Combinatorial Chemistry and High Throughput Screening*, vol. 4, no. 8, pp. 727–739, 2001.

[15] D.-S. Huang and C.-H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, no. 15, pp. 1855–1862, 2006.

[16] P. A. Futreal, L. Coin, M. Marshall et al., "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3, pp. 177–183, 2004.

[17] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff et al., "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, 2002.

[18] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[19] U. Alon, N. Barka, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

[20] B. Furusato, C.-L. Gao, L. Ravindranath et al., "Mapping of TMPRSS2-ERG fusions in the context of multi-focal prostate cancer," *Modern Pathology*, vol. 21, no. 2, pp. 65–75, 2008.

[21] J. B. Welsh, L. M. Sapinoso, A. I. Su et al., "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer," *Cancer Research*, vol. 61, no. 16, pp. 5974–5978, 2001.

[22] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6567–6572, 2002.

[23] A. R. Dabney, "Classification of microarrays to nearest centroids," *Bioinformatics*, vol. 21, no. 22, pp. 4148–4154, 2005.

[24] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular Systems Biology*, vol. 4, no. 189, 2008.

[25] V. A. McKusick, "Mendelian inheritance in man and its online version, OMIM," *American Journal of Human Genetics*, vol. 80, no. 4, pp. 588–604, 2007.

[26] L. Tang, D. L. Dai, M. Su, M. Martinka, G. Li, and Y. Zhou, "Aberrant expression of collagen triple helix repeat containing 1 in human solid cancers," *Clinical Cancer Research*, vol. 12, no. 12, pp. 3716–3722, 2006.

[27] W. Feng, L. Shen, S. Wen et al., "Correlation between CpG methylation profiles and hormone receptor status in breast cancers," *Breast Cancer Research*, vol. 9, no. 4, pp. R57–R69, 2007.

[28] T. Sørlie, Y. Wang, C. Xiao et al., "Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms," *BMC Genomics*, vol. 7, article 127, 2006.

[29] E. Klopocki, G. Kristiansen, P. J. Wild et al., "Loss of SFRP1 is associated with breast cancer progression and poor prognosis in early stage tumors," *International Journal of Oncology*, vol. 25, no. 3, pp. 641–649, 2004.

[30] G. Turashvili, J. Bouchal, K. Baumforth et al., "Novel markers for differentiation of lobular and ductal invasive breast

carcinomas by laser microdissection and microarray analysis," *BMC Cancer*, vol. 7, article 55, 2007.

[31] F. Bertucci, P. Finetti, N. Cervera et al., "Gene expression profiling shows medullary breast cancer is a subgroup of basal breast cancers," *Cancer Research*, vol. 66, no. 9, pp. 4636–4644, 2006.

[32] D. I. Rodenhiser, "Epigenetic contributions to cancer metastasis," *Clinical and Experimental Metastasis*, vol. 26, no. 1, pp. 5–18, 2009.

[33] M. Lacroix, "Significance, detection and markers of disseminated breast cancer cells," *Endocrine-Related Cancer*, vol. 13, no. 4, pp. 1033–1067, 2006.

[34] J. Helleman, M. P. H. M. Jansen, K. Ruigrok-Ritstier et al., "Association of an extracellular matrix gene cluster with breast cancer prognosis and endocrine therapy response," *Clinical Cancer Research*, vol. 14, no. 17, pp. 5555–5564, 2008.

[35] P. Khatri, S. Sellamuthu, P. Malhotra, K. Amin, A. Done, and S. Draghici, "Recent additions and improvements to the Onto-Tools," *Nucleic Acids Research*, vol. 33, no. 2, pp. W762–W765, 2005.

[36] S. Konstantinovsky, Y. Smith, S. Zilber, et al., "Breast carcinoma cells in primary tumors and effusions have different gene array profiles," *Journal of Oncology*, vol. 2010, 2010.

[37] O. Klezovitch, J. Chevillet, J. Mirosevich, R. L. Roberts, R. J. Matusik, and V. Vasioukhin, "Hepsin promotes prostate cancer progression and metastasis," *Cancer Cell*, vol. 6, no. 2, pp. 185–195, 2004.

[38] J. E. Vivienne Watson, N. A. Doggett, D. G. Albertson et al., "Integration of high-resolution array comparative genomic hybridization analysis of chromosome 16q with expression array data refines common regions of loss at 16q23-qter and identifies underlying candidate tumor suppressor genes in prostate cancer," *Oncogene*, vol. 23, no. 19, pp. 3487–3494, 2004.

[39] X. Lin, M. Tascilar, W.-H. Lee et al., "GSTP1 CpG island hypermethylation is responsible for the absence of GSTP1 expression in human prostate cancer cells," *American Journal of Pathology*, vol. 159, no. 5, pp. 1815–1826, 2001.

[40] S. K. Rayala, P. Den Hollander, B. Manavathi et al., "Essential role of KIBRA in co-activator function of dynein light chain 1 in mammalian cells," *Journal of Biological Chemistry*, vol. 281, no. 28, pp. 19092–19099, 2006.

[41] N. Konishi, K. Shimada, M. Nakamura et al., "Function of JunB in transient amplifying cell senescence and progression of human prostate cancer," *Clinical Cancer Research*, vol. 14, no. 14, pp. 4408–4416, 2008.

[42] J. Bektic, K. Pfeil, A. P. Berger et al., "Small G-protein RhoE is underexpressed in prostate cancer and induces cell cycle arrest and apoptosis," *Prostate*, vol. 64, no. 4, pp. 332–340, 2005.

[43] S. Peri, J. D. Navarro, T. Z. Kristiansen et al., "Human protein reference database as a discovery resource for proteomics," *Nucleic Acids Research*, vol. 32, pp. D497–D501, 2004.

[44] http://cbio.mskcc.org/cancergenes.

[45] Z. Zhang, D. Chen, and D. A. Fenstermacher, "Integrated analysis of independent gene expression microarray datasets improves the predictability of breast cancer outcome," *BMC Genomics*, vol. 8, article 331, 2007.

[46] C. Ortutay and M. Vihinen, "Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies," *Nucleic Acids Research*, vol. 37, no. 2, pp. 622–628, 2009.

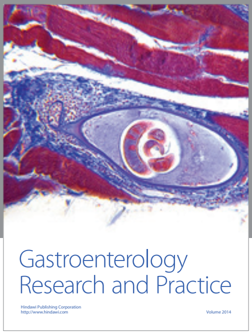
[47] P. Radivojac, K. Peng, W. T. Clark et al., "An integrated approach to inferring gene-disease associations in humans," *Proteins*, vol. 72, no. 3, pp. 1030–1037, 2005.

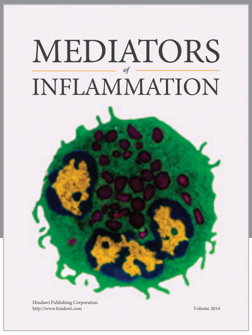
[48] S. Aerts, D. Lambrechts, S. Maity et al., "Gene prioritization through genomic data fusion," *Nature Biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.

[49] X. Ma, H. Lee, L. Wang, and F. Sun, "CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data," *Bioinformatics*, vol. 23, no. 2, pp. 215–221, 2007.

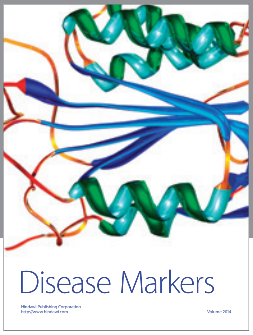
[50] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert, "GeneRank: using search engine technology for the analysis of microarray experiments," *BMC Bioinformatics*, vol. 6, article 233, 2005.

The Scientific World Journal

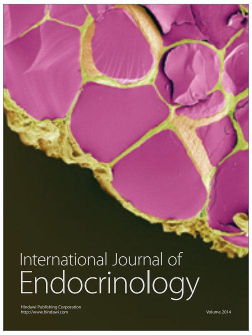

Gastroenterology Research and Practice


MEDIATORS of INFLAMMATION

Journal of Diabetes Research


Disease Markers

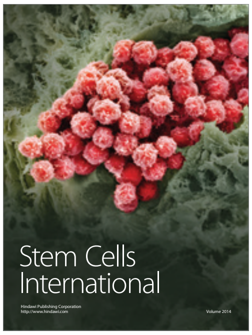
Journal of Immunology Research

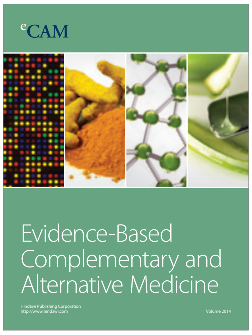

International Journal of Endocrinology

PPAR Research

BioMed Research International

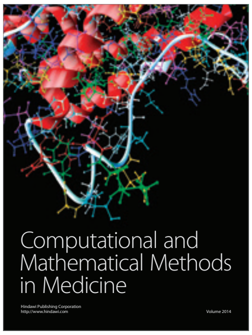
Journal of Ophthalmology

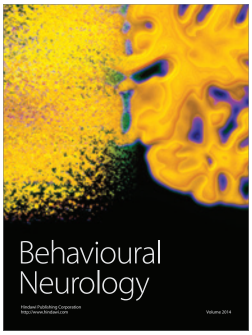

Stem Cells International


eCAM
Evidence-Based Complementary and Alternative Medicine

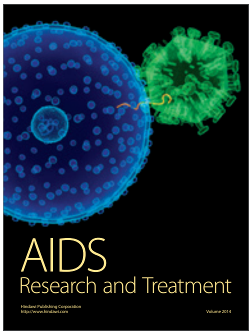
Journal of Obesity

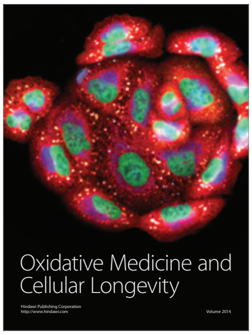
Journal of Oncology


Computational and Mathematical Methods in Medicine


Behavioural Neurology

Parkinson's Disease


AIDS Research and Treatment


Oxidative Medicine and Cellular Longevity