Computers in Biology and Medicine 43 (2013) 1363–1373

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

ICP: A novel approach to predict prognosis of prostate cancer with inner-class clustering of gene expression data

Hyunjin Kim [a], Jaegyoon Ahn [a], Chihyun Park [a], Youngmi Yoon [b], Sanghyun Park [a,*]

[a] Department of Computer Science, Yonsei University, South Korea
[b] Department of Computer Engineering, Gachon University, South Korea

ARTICLE INFO

Article history:
Received 4 August 2012
Accepted 19 June 2013

Keywords:
Classification
Clustering
Bioinformatics
Microarray data analysis
Prognosis
Prostate cancer

ABSTRACT

Prostate cancer has heterogeneous characteristics. For that reason, even if tumors appear histologically similar to each other, there are many cases in which they are actually different, based on their gene expression levels. A single tumor may have multiple expression levels with both high-risk cancer genes and low-risk cancer genes. We can produce more useful models for stratifying prostate cancers into high-risk cancer and low-risk cancer categories by considering the range in each class through inner-class clustering. In this paper, we attempt to classify cancers into high-risk (aggressive) prostate cancer and low-risk (non-aggressive) prostate cancer using ICP (Inner-class Clustering and Prediction). Our model classified more efficiently than the models of the algorithms used for comparison. After discovering a number of genes linked to prostate cancer from the gene pairs used in our classification, we discovered that the proposed method can be used to find new unknown genes and gene pairs which distinguish between high-risk cancer and low-risk cancer.

1. Introduction

Prostate cancer is a malignant tumor in the prostate gland. It is one of the most common cancers among men. Other than skin cancer, prostate cancer is the most prevalent cancer in American men. Since prostate cancer is a slow progressing cancer, it has low risk of metastasizing in most cases. Therefore, patients with prostate cancer who are over 70 years old are more likely to die from other causes than from prostate cancer over the 15 years following prognosis. Because prostate cancer may not cause severe pain or have any abnormal signs, it is hard for a patient to know if he has prostate cancer unless the prostate cancer has metastasized to other organs. Therefore, there is a high chance that the cancer has spread to other parts of the body once the patient detects its symptoms. If the prostate cancer has spread to other parts of the body, the metastasized cancer is more dangerous than original prostate cancer, which is a slow-growing cancer. Metastatic cancer that has spread to other areas of the body can grow rapidly and affect vital organs. For that reason, the most important factor related to prostate cancer is not whether 'it is' or 'it is not' a prostate cancer, but its prognosis, likely progression and probability of metastasis.

Generally, a patient who has cancer can predict his prognosis using clinical stage. The clinical stage is determined by the state of progress, the size and the range of the tumor together with whether or not the cancer has metastasized. The higher the stage number, the bigger the tumor and the more progress it has made. According to a related research study, however, differentiation in cancer cells has a greater effect on prognosis than diagnosed stage [1]. Differentiation refers to an operation or a process of cells specializing in structure and function. Hence, if cells are well differentiated, they are normal cells, and if cells are poorly differentiated, then they are immature and disorganized cells. The results of the research study show that if cancer cells are poorly differentiated, prostate cancer death is more probable even when the tumor is at a lower clinical stage. In a research study on watchful waiting, the two primary risk factors are age at the time of diagnosis and Gleason score [2]. The Gleason score is a means of measuring the aggressiveness of prostate cancer [3–7]. It is obtained by adding the two Gleason scale grades together. Each cell is given a Gleason scale grade according to the degree of differentiation of the cell. The scale grade is determined by examining cells from the prostate under a microscope during a biopsy. Each cell is then given a grade from 1 to 5. The higher the degree of differentiation, the lower the grade number it is given. Once the two most common types of cancer cells are identified in the prostate, the two grades of these two types are then added together to produce a Gleason score. Therefore, a Gleason score ranges from 2 to 10. The lower the score, the slower the cancer is

[*] Corresponding author. Tel.: +82 2 2123 5714; fax: +82 2 365 2579.
E-mail address: sanghyun@cs.yonsei.ac.kr (S. Park).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.06.014

growing, and the higher the score, the faster the cancer is likely to be growing and the more aggressive it is. In general, a Gleason score of 7 is considered intermediate and a score of 6 or less has a good prognosis. A score of 8 or higher has a poor prognosis.

Since the Gleason score is determined by examining cells from the prostate under a microscope, it cannot be considered as an absolute index of the prognosis for prostate cancer [8]. For that reason, we used two kinds of data in the experiment. For the first data set (GSE 15484), we considered a Gleason score of 7 or less (2–7) to be non-aggressive and a score of 8 or more (8–10) to be aggressive. We classified using this method in this paper. The second data set (GSE 21034) consists of the aggressive and non-aggressive prostate cancer samples obtained from clinical examination without reference to Gleason score. We did experiments using the same method as the one applied to the first data.

Prostate cancer has heterogeneous characteristics, which means samples in the same class do not necessarily have similar gene expression levels [9–11]. Classification algorithms for handling prostate cancer gene expression levels have to reflect that heterogeneity. The key to successfully overcoming this heterogeneity is capturing the distinctive gene expression level groups in each class and using these groups when performing classification. We propose an efficient classification method ICP (Inner-class Clustering and Prediction) based on the heterogeneity of gene expression levels to classify prostate cancer into two categories, high-risk and low-risk prostate cancer. ICP can distinguish several different gene expression level groups by using inner-class clustering. It reduces false positives and false negatives. Most of the other methods do not consider the different types in the same class. The classification method has 5 major phases (Fig. 1).

The first phase is a gene selection phase that reduces the number of genes to be used in the analysis, because if we experiment with a large number of genes, the time complexity is too large. In this phase, we sort out n top-ranked genes using relief-A and symmetrical uncertainty algorithms, which are verified feature selection methods. In the second phase, by making use of inner-class clustering, we calculate the cluster information for each gene pair, which is carried out in the first phase. In the third phase, we measure the degree of dispersion using the cluster information from the gene pairs we obtained in the second phase, and rank the gene pairs from highest to lowest according to the degree of dispersion. If there are multiple gene pairs which have the same score, we use variance-based secondary score to select a unique gene pair. Phases 4 and 5 are the phases to select a class. By using vote sets from the phase 3, we execute the prediction in each voter, and then select the class with the most votes.

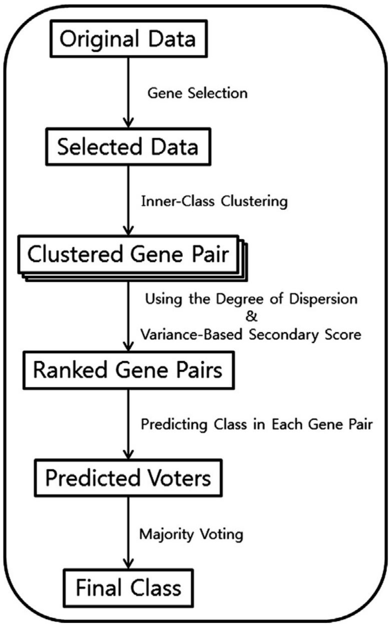

Fig. 1. Flow chart of ICP algorithm.

The results in distinguishing between the high-risk and the low-risk prostate cancers with the proposed classification method show that the proposed classification method is more efficient than other existing classification methods. Moreover, looking into the frequently appearing genes and gene pairs, which are ranked by the degree of dispersion, informed us that those genes and gene pairs are closely related to biological processes or to prostate cancer. Classification by making use of inner-class clustering is novel and is of great value because it can be applied to multi-class classifications.

2. Related works

Almost all classification problems of cancer diagnosis and prognosis can be solved by machine learning methods. These methods develop classifiers with training samples which are already classified and predict the class of test samples based on those classifiers.

The most popular cancer-related classification method among the machine learning methods is SVM (Support Vector Machine) [12]. SVM finds the linear optimal hyper plane which separates gene expression data samples into two groups and uses that plane to classify the given samples. After applying the transform function, the non-linear data can be handled in the same way as the linear data in SVM. The transform function is called the kernel function and there are many types of kernel functions. There have already been many studies on when the kernel function should be used and what type of function should be used [13–16]. A few regression versions of the SVM [17,18] also exist but methods which use SVM for gene expression data usually focus on which genes are to be selected to form a hyper plane rather than how to change the main algorithm of SVM to be more efficient. If genes are closely correlated, we can apply SVM-RFE (Recursive Feature Elimination) [19], one of the methods that focuses on gene selection. When SVM is finding a hyper plane and using it on the classification, it is important to obtain the maximum margin between two classes. The L1-norm penalty is helpful to obtain the soft maximum margin [20]. The L1-norm SVM does not choose all the genes which have a high correlation among themselves. To solve this problem for the L1-norm SVM, Wang [21] proposed HHSVM (Hybrid Huberized Support Vector Machine) making use of the huberized hinge loss function and elastic-net penalty.

Logistic regression [22] is similar to linear regression because a function is created based on the shape of the data so the class of a sample can be predicted. But the difference between logistic and linear regression is that logistic regression's prediction result is binomial, not continuous. Logistic regression method can be applied to other models, so extensibility is the one of the merits of this method. For instance, a logistic regression method combined with a parametric bootstrap model for the gene expression data classification problem was proposed by Liao [23] in 2007.

Another method called decision tree induction is a classification method which uses flowchart-like tree structure. Each internal node denotes a test on an attribute, each branch describes a result of the test, and each leaf node represents a class label. The attribute values of a test sample are tested with the internal nodes in the decision tree. A path can be traced from the root node to a leaf node and the leaf node's class label indicates the predicted class of the test sample. ID3 [24], C4.5 [25], and CART [26] are different versions of the decision tree which have different attribute selection measures. The attribute selection measure is used to select the splitting criterion that efficiently separates a given data for the testing in each internal node.

Methods which classify data with gene expression levels use differentially expressed genes between two classes. Since the proposed algorithm ICP uses expression levels of gene pairs, there are similar studies which use differentially expressed gene pairs [27–31]. The most notable algorithm which uses gene pairs is TSP (Top Scoring Pair) [27]. The TSP extracts most informative gene pair based on the relative frequencies of occurrences of $R_a < R_b$ within profiles where R_i denotes the rank of the i-th gene in the profile and use the gene pair to classify test samples. k-TSP [30] is an extended version of TSP classifier on the basis of k disjoint top scoring pairs of genes that hold the best combined score.

Profiling gene expression levels between high-risk cancer and low-risk cancer can identify markers that signify the aggressiveness of a given cancer. The method suggested by Pressinotti [32] is one of them. The SAM (Significance analysis of microarray) test [33] is given on high-risk prostate cancer samples and low-risk prostate cancer samples, then it can identify differentially expressed genes. Pressinotti carried out a quantitative real time PCR (Polymerase Chain Reaction) and IHC (Immunohistochemistry) analysis of these differentially expressed genes, and he distinguished 20 genes. Eleven out of the 20 genes are associated with an apoptotic process.

Leinonen's method [34] is similar to Pressinotti's method because it also used IHC analysis and fluorescence in situ hybridization on prostate cancer gene expression data. His research identified the expression level for SPINK1 and TMPRSS2:ERG fusion to predict the prognosis of prostate cancer.

There is another method proposed by Minner [35] in which the gene expression level of Her2 has an influence on prostate cancer prognosis. A relationship may exist between the gene expression level of Her2 and prostate cancer prognosis as determined by IHC analysis and fluorescence in situ hybridization on tissue microarray data. The experiment showed that the higher the Her2 expression level, the worse prostate cancer prognosis is.

It is not easy to determine the prognosis of cancer because distinguishing between tumors with high-risk and those with low-risk is difficult. There are few differences among samples in different classes so a novel approach to efficiently distinguish high-risk cancer and low-risk cancer is needed. For these reasons, we propose a novel method, ICP, which reflects the heterogeneity of prostate cancer data by using inner-class clustering.

3. Methods

As we mentioned in previous sections, prostate cancer has heterogeneous characteristics that are difficult to associate with gene pairs (Fig. 2). Fig. 2 represents gene expression levels of famous marker gene pair TMPRSS2:ERG fusion [36,37] within profiles. The samples in Fig. 2 are from GSE 15484 and from GSE 21034 data which are stored in the GEO (Gene Expression Omnibus) database from NCBI (National Center for Biotechnology Information). The samples even in the same class do not always have similar gene expression levels. Therefore, we propose ICP for this problem. It has 5 major phases including gene selection, inner-class clustering, computing the rank score for gene pairs, predicting class in a voter, and majority voting.

3.1. Gene selection

A microarray [38] is a useful tool to measure gene expression levels of a genome. The gene expression data used in this paper is generated from microarrays. It is created in the form of a matrix. Rows represent genes and columns represent samples such as high-risk cancers or low-risk cancers. Each spot represents the expression level of the given spot gene from the given spot sample (Fig. 3).

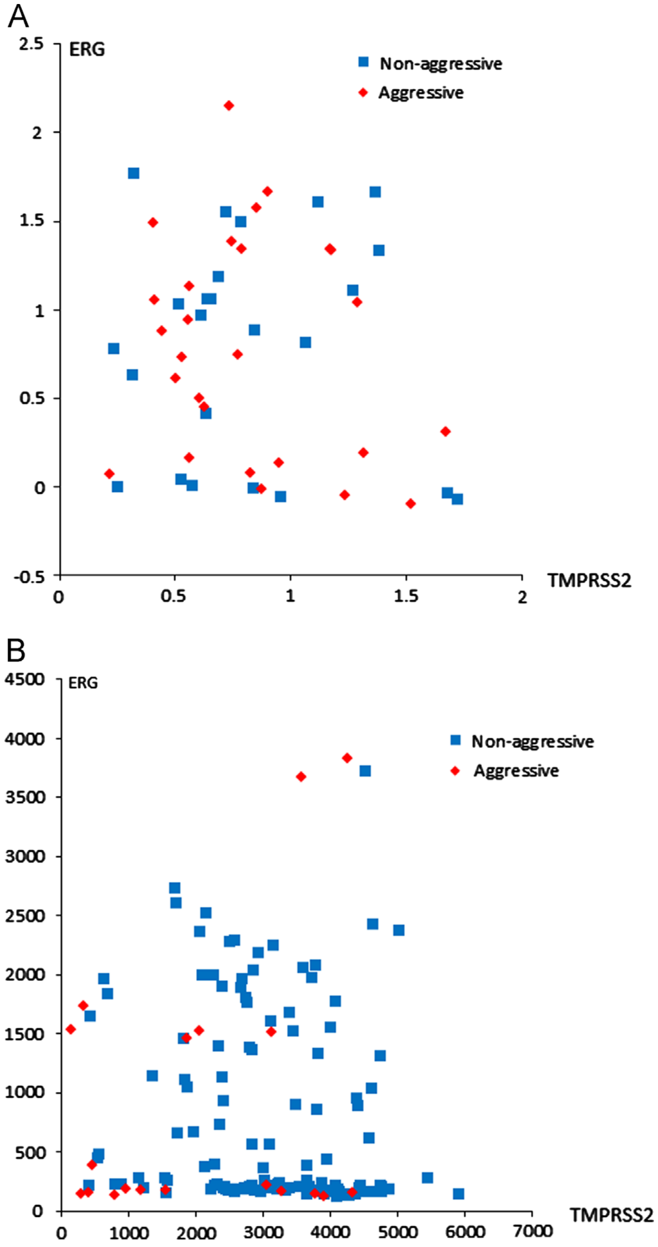

Fig. 2. Gene expression levels of TMPRSS2:ERG gene fusion. The x-axis represents the expression level of TMPRSS2, and the y-axis represents the expression level of ERG. (A) Gene expression levels of TMPRSS2:ERG in GSE 15484. It has 27 aggressive samples and 25 non-aggressive samples. (B) Gene expression levels of TMPRSS2:ERG in GSE 21034. It has 19 aggressive samples and 131 non-aggressive samples.

Gene expression data includes thousands or tens of thousands of genes. ICP classifies test samples into different classes using gene pairs, so if the number of genes is n, the number of gene pairs is n^2. Therefore, if we do not reduce the number of genes, we have to accept a high level of computational complexity. The ICP algorithm employs a feature selection method to minimize process time. Feature selection methods can choose a specific number of genes to decrease process time. The experiments in this paper used relief-A [39] and symmetrical uncertainty [40]. Feature selection methods can reduce not only the process time but also exclude genes which are useless in classification. Feature selection

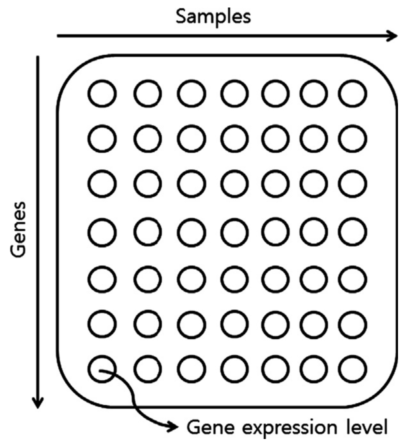

Fig. 3. Gene expression data set.

methods contribute to progress in accuracy because they classify by ignoring useless genes. Previous research already demonstrated that the feature selection techniques can be useful in cancer classification problems [41].

3.2. Inner-class clustering

Inner-class clustering performs clustering of samples in each class. When we use clustering methods to classify a class of test samples, it is typically presumed that samples in the same cluster are also included in the same class and samples in different clusters are included in different classes for the most part. Inner-class clustering's target samples are all in the same class. Inner-class clustering on classification allows various groups with different expression levels to be specified in only one class. Accordingly, inner-class clustering is very helpful when data have heterogeneous characteristics. Meanwhile, if a clustering method is used in classification to distinguish a class from other classes, target samples may be included in various classes. The purpose of the clustering method is to classify those samples into an appropriate class. In that classification situation, it performs clustering on the assumption that samples in the same class have similar expression levels which can be a problem due to the heterogeneity of each class. The problems are several false positives and false negatives (Fig. 4A). These problems can be solved by using inner-class clustering.

Some differences exist not only among classes but also within a class. Genes can have dissimilar appearances or expression patterns in the samples of the same class. Therefore, we propose an efficient classification method ICP (Inner-class Clustering and Prediction) which considers the heterogeneity of gene expression levels.

First of all, ICP performs 2-dimensional clustering on every selected gene pair from the data. At this point, if there are n genes in the data, the number of possible gene pairs is $n(n-1)/n$ and inner-class clustering is also performed $n(n-1)/n$ times in each class. The clustering algorithm used in this paper is k-means [42] that has fast running time and acceptable complexity even if there are a large number of genes. Typical k-means algorithm initiates with random cluster seeds and classification result can also be random. In order to prevent random results, we use deterministic initial seeds which are equally divided by sequential order based on the number of clusters k. Here, Δg denotes a gap between every two initial seeds where Δg=(the number of training data-1)/$(k-1)$, $k>1$. Then the initial seed for i-th cluster on 2-dimensional coordinate $S_i(X, Y)$ can defined as below where $X=\{x_1, x_2, \ldots, x_s\}$, $Y=\{y_1, y_2, \ldots, y_s\}$:

$$S_i(X,Y)=(x_{(i-1)*\Delta g}, y_{(i-1)*\Delta g})$$

After performing inner-class clustering with the k-means algorithm, each class has k clusters, so there must be $2k$ clusters when performing binary classification (Fig. 4B). In the end, results of inner-class clustering with k-means have $n(n-1)/2$ gene pairs, each of which has $2k$ clusters. A specific description of the inner-class clustering phase for binary classification is shown in Algorithm 1.

Algorithm 1. Algorithm of inner-class clustering process.

```
Input: Gene set $G$ consists of selected $n$ genes from a feature selection method in the previous phase, the number of clusters $k$
Output: Gene pair set $P$ that includes $n(n-1)/2$ inner-class clustered gene pairs
Make a gene pair set $P$ with every gene pair in gene set $G$
For each Gene pair($a$, $b$) from $P$ Do
    For each class C Do
        Cluster the gene pair($a$, $b$) in class C with $k$-means clustering method
    End For
    Depict the clustering results from each class together into 2-dimensional space to form a complete clustered gene pair
End For
```

3.3. Computing the rank score for every pair of genes

Before executing the majority vote process, ICP makes the top m voter sets and determines the class of sample which was most selected from the voters. The voters have clustering information of gene pairs, which are the outcomes from inner-class clustering and each voter predicts a class by using those clustering information. The prediction process can be executed in every gene pair, so if there are n genes, there exist $n(n-1)/2$ predicted classes for a given test sample because there is one predicted class per voter. This is why we have to choose the top m voters. We have to consider which gene pairs are the top m voters. In this situation, a powerful ranking method which determines the most efficient m voters for classification is needed.

In the ICP algorithm, an efficient voter refers to the gene pair with the most distributed clusters. When executing the prediction process, it is easy to find the nearest cluster if the clusters are highly distributed. In contrast, a gene pair that has clusters with high density (low distribution) makes prediction harder. Even if the algorithm finds a minimum distance cluster, it would not be reliable. Consequently, the longer the distance among clusters, the easier it is to predict a class and the more reliable the result of the voter.

The distance among clusters can be obtained by using the degree of dispersion in each voter. Degree of dispersion can be calculated using the sum of the Euclidean distance between all two clusters which are in different classes. If there are k clusters in each class and (x_{ij}, y_{ij}) stands for the mean point of samples in the i-th class and the j-th cluster on 2-dimensional coordinate, the degree of dispersion d can be represented as below in the binary classification. Empty clusters will be excluded during the calculation:

$$d=\sum_{a=1}^{k}\sum_{b=1}^{k}\{(x_{1a}-x)_{2b}\}^2+(y_{1a}-y_{2b})^2\}$$

when constructing the majority vote set, m gene pairs are selected which have a high degree of dispersion. However, it is possible for multiple gene pairs to have the same score. In order to rule out ties and to select reliable gene pairs, we use a secondary

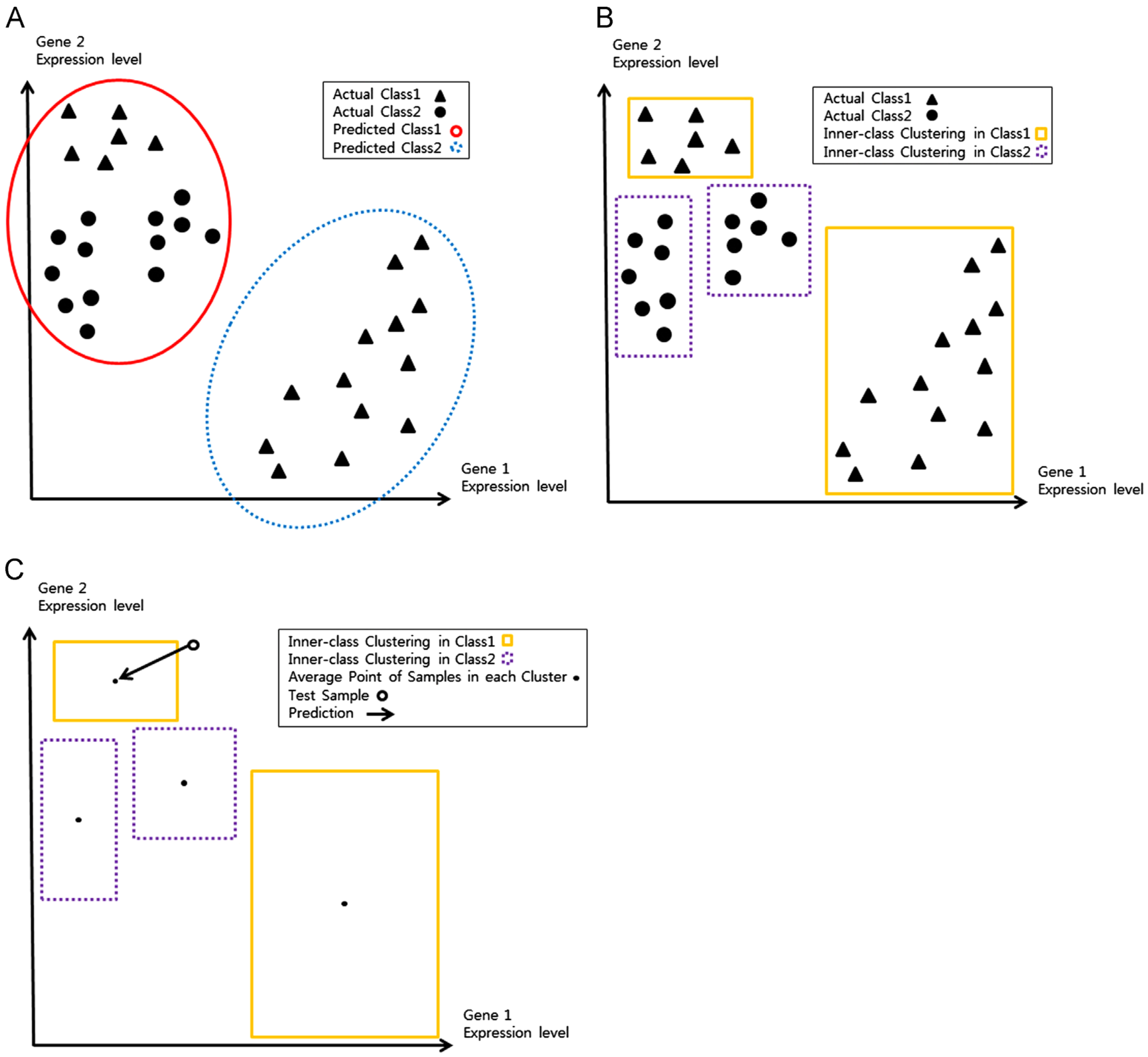

Fig. 4. The ground-truth examples for ICP algorithm. (A) An example of problem in binary classification using typical clustering method. In binary classification using a typical clustering method, actual class 1 samples in a solid line circle can be predicted as class 2 because they are closer to actual class 2 samples than to actual class 1 samples. Therefore, the actual class 1 samples in a solid line circle can be false positives or false negatives. (B) After inner-class clustering when performing binary classification. A result after performing inner-class clustering with the same data in (A). In this example, there are two clusters in each class, so there are four clusters in total. Inner-class clustering is independently executed in each class and unified into 2-dimensional space. (C) Prediction of test sample in ICP. Mark an expression level of a given test sample on 2-dimensional coordinates and find the nearest cluster. The nearest cluster's class is the predicted class of the given sample. If there are multiple clusters that have the same minimum distance values, the class with the largest number of samples is predicted.

score based on a variance of expression levels from each gene in a gene pair. Let *mean*(*G*) denote a mean value of inner-class clustering information from gene *G* in one dimension where *mean* $(G)=(\sum_{a=1}^{2}\sum_{b=1}^{k}(g_{ab}))/2k$. The variance-based secondary score d2 can examine the distributedness of a gene pair (X, Y), and it is defined as

$$d2=\sum_{a=1}^{2}\sum_{b=1}^{k}(x_{ab}-mean(X))^2+\sum_{a=1}^{2}\sum_{b=1}^{k}(y_{ab}-mean(Y))^2$$

Now we can select a unique gene pair with the largest variance score even if there are multiple gene pairs which have the same degree of dispersion.

If there are n genes, $n(n-1)/2$ gene pairs exist and only the top-scoring m gene pairs are chosen to form the majority vote set. A user can give a specific number for the variable m. In this paper, we used various variables m in each classifier which can elicit the best classification results.

3.4. *Predicting class in a voter*

After selecting top-scoring m gene pairs, prediction in each voter (gene pair) is needed. When test sample's expression point of the gene pair is projected onto the 2-dimensional coordinates, we can determine the class of a given test sample by detecting the nearest cluster. The nearest cluster's class is considered as the class of the given test sample (Fig. 4C). Here, $ud_{min}(C_i)$ denotes the smallest Euclidean distance between test sample and one of the class C_i's clusters. The distance is calculated with test sample's expression point of the gene pair and mean point of a cluster from class C_i. The predicted class of test sample s in a gene pair $PC(s)$ can be defined as below:

$$PC(s) = \begin{cases} \text{Class 1} & \text{if } ud_{\min}(C_1) < ud_{\min}(C_2) \\ \text{Class 2} & \text{if } ud_{\min}(C_1) > ud_{\min}(C_2) \end{cases}$$

There is a possibility that several clusters of different classes are at the same distance from the test sample when $ud_{min}(C_1) = ud_{min}(C_2)$. For that reason, we use a secondary score $ud2(C_i) = \sum_{a=1}^{k}((x_{ia}-t_1)^2 + (y_{ia}-t_2)^2)$ where test sample's expression point is (t_1, t_2) and (x_{ij}, y_{ij}) stands for the mean point of samples in the i-th class and the j-th cluster on 2-dimensional coordinate. The secondary score $ud2$ indicates the sum of Euclidean distance between test sample and every class C_i's clusters. The secondary score based prediction process is defined as

$$PC2(s) = \begin{cases} \text{Class 1} & \text{if } ud2(C_1) < ud2(C_2) \\ \text{Class 2} & \text{if } ud2(C_1) > ud2(C_2) \end{cases}$$

when both scores are the same which indicates $ud_{min}(C_1) = ud_{min}(C_2)$ and $ud2(C_1) = ud2(C_2)$, the class with the largest number of sample is selected, i.e. the number of samples of class 1 is 27, and the number of samples of class 2 is 25, the ICP predicts the test sample as class1.

3.5. Majority voting

If the process of predicting class in each voter is complete, there remains a process for determining the final class among the top m voters. The majority vote set has the top m voters and each voter has predicted classes from the prediction process. The majority voting process makes the final decision in the ICP algorithm. The process classifies a test sample into the most voted class from the top m voters. In order to prevent ties, an odd number is highly recommended for user-defined parameter m but when using an even number for m, if there are multiple classes that have the same maximum number of votes, the class with the largest number of samples is selected. The process of majority voting in binary classification can be expressed as below:

$$V_1(g) = \begin{cases} 1 & \text{if the test sample is predicted to be class 1 in gene pair } g \\ 0 & \text{otherwise} \end{cases}$$

$$V_2(g) = \begin{cases} 1 & \text{if the test sample is predicted to be class 2 in gene pair } g \\ 0 & \text{otherwise} \end{cases}$$

$$C(s) = \begin{cases} \text{Class 1} & \text{if} \sum_{i=1}^{m} V_1(g_i) > \sum_{i=1}^{m} V_2(g_i) \\ \text{Class 2} & \text{if} \sum_{i=1}^{m} V_1(g_i) < \sum_{i=1}^{m} V_2(g_i) \end{cases}$$

- g_i is i-th gene pair in the majority vote set.
- $C(s)$ is the predicted class for test sample s in the ICP algorithm.
- If the sum of $V_1(g)$ and the sum of $V_2(g)$ are equal, the class with the largest number of samples is selected.

The ICP algorithm consists of 5 phases described above. They are gene selection, inner-class clustering, computing the rank score for gene pairs, class prediction in a voter, and majority voting. A formal description of the ICP algorithm is shown in Algorithm 2.

Algorithm 2. Algorithm for ICP.

```
Input: Sample $s$, gene expression training data $D$, the number
of genes $n$, the number of clusters $k$, the number of majority
vote set $m$
Output: Class label $C$ of sample $s$
1. Make a new feature data set $D'$ with $n$ genes which are
   selected by feature selection (Relief-A or Symmetrical
   uncertainty) method
2. For each Gene pair$(a, b)$ from $D'$ Do
3.     Clustering the data $D'$ with $k$ clusters in each class by a
       $k$-means method
4.     Calculate the rank score of the gene pair$(a, b)$ using the sum
       of the Euclidean distance between clusters which are not
       included in the same class
5. End For
6. Construct the top $m$ voter set within all gene pairs which
   have the highest dispersion score in the data $D'$
7. If there are multiple gene pairs to have the same rank score,
   use variance-based second score to select $m$ voters
8. Predict the class of sample $s$ in each voter using the nearest
   cluster's class
9. If several clusters of different classes are at the same distance
   from the test sample, use Euclidean distance-based second
   score to predict the class of sample $s$
10. Do majority voting in the $m$ voter set and choose the class $C$
    which has the largest number of votes
```

4. Results

As experimental environments, we used Intel® Core™ i3 530 Dual 2.93 GHz, 4.00 GB RAM machine with Windows 7 operating system. We have implemented our algorithm using JAVA programming language with JDK 6.

Finding an optimal parameter values when using classification algorithms is a difficult problem. The ICP algorithm has 3 parameters. The first is the number of features. There are numerous genes which increase running time if they are all considered. In this paper, we used 300 features that had already been selected by a feature selection method, relief-A and symmetrical uncertainty. Note that we used 8052 features (GSE 15484) and 43 419 features (GSE 21034) for testing. Three hundred features are 3.73% of 8052 features and 0.69% of 43 419 features. The ICP process which employs 300 features showed sensible running time. The second parameter is the k value when performing inner-class clustering with k-means clustering. The k value which represents the number of clusters in the ICP algorithm is very substantial because it reflects the heterogeneity of the classes in the result. If the k value is too big, it reflects too many different clusters, making classification more difficult. On the contrary, for a small k value, it cannot reflect the heterogeneity of classes and that can be a reason for a high false positive rate. The parameter k is determined by cross-validation with the limitation that $k > 1$ and $k_{max} = 5$. The optimal value is selected after cross-validation for parameter k based on AUC (Area Under Curve). The third parameter is the number of voters in the majority vote set, which is involved with determining how many of the top m gene pairs have a high rank score for constructing the majority vote set. When there are more voters, the algorithm has better results in general. However, that is on the assumption that all voters have useful information for classification. If the number of voters is large, there is a chance of including useless voters. In this manner, the appropriate number of voters is needed. The number of voters m is determined by cross-validation with the restriction that $k_{max} = 10$ in this paper and is an odd number because of breaking ties in the majority voting phase. The parameter setting for the number of voters m is same as parameter setting for the number of voters k in k-TSP [30]. For validation, the LOOCV (Leave One Out Cross Validation) technique is employed in this paper. LOOCV is a cross-validation technique which uses one sample for testing and the other samples for training to construct a classifier.

Data sets applied in this paper are registered on the GEO (Gene expression Omnibus) database of NCBI (National Center for Biotechnology Information) as GSE 15484 and GSE 21034. These data sets are gene expression level data from prostate cancer patients. GSE 15484 contains 25 samples with Gleason score 6, 27 samples

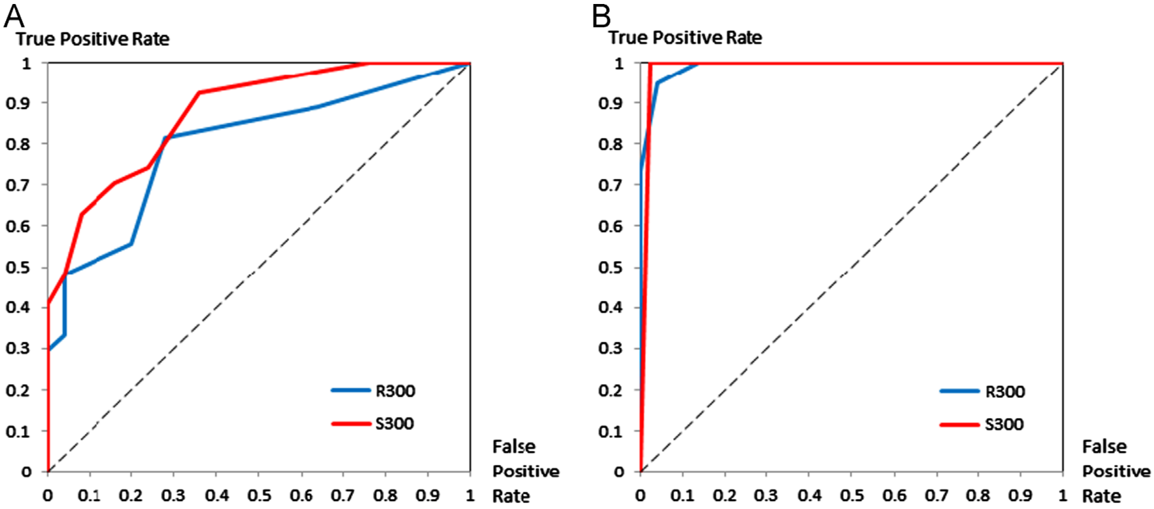

Fig. 5. ROC curve of ICP algorithm. R300 represents that the classifier used 300 features from Relief-A and S300 represents that the classifier used 300 features from symmetrical uncertainty. (A) ROC curve of ICP in GSE 15484. (B) ROC curve of ICP in GSE 21034.

Table 1
AUC of various algorithms in GSE 15484.

Classification algorithm	Parameters	300 features selected by relief-A (R300)	300 features selected by symmetrical uncertainty (S300)
Support Vector Machine	Polynomial kernel	0.846	0.827
	RBF kernel	0.826	0.826
k-means	$k=2$	0.715	0.773
Logistic regression	Linear	0.831	0.791
	Multinomial	0.827	0.806
Decision tree	C4.5	0.567	0.695
	CART	0.753	0.827
Top scoring pair	None	0.821	0.730
k-Top scoring pair	Optimal Parameters from Cross-validation	0.857	0.910
ICP	Optimal parameters from cross-validation	0.797	0.876>

Table 2
AUC of various algorithms in GSE 21034.

Classification algorithm	Parameters	300 feature selection by relief-A (R300)	300 feature selection by symmetrical uncertainty (S300)
Support vector machine	Polynomial kernel	0.833	0.917
	RBF kernel	0.895	0.812
k-means	$k=2$	0.966	0.790
Logistic regression	Linear	0.970	0.963
	Multinomial	0.868	0.987
Decision tree	C4.5	0.940	0.933
	CART	0.966	0.959
Top scoring pair	None	0.966	0.970
k-top scoring pair	Optimal parameters from cross-validation	0.973	0.973
ICP	Optimal parameters from cross-validation	0.991	0.989

with Gleason score 8–10, and 13 benign samples. We performed ICP with 6-Gleason-scoring samples considered low-risk (non-aggressive) and with 8, 9, and 10-Gleason-scoring samples considered high-risk (aggressive), excluding benign samples. GSE 21034 is made up of 185 samples that are gathered from individual prostate cancer patients who have been clinically observed for 5 years. One hundred and thirty-one samples representing primary tumors are used as low-risk cancer samples and 19 samples representing metastasis are used as high-risk cancer samples.

We draw ROC (Receiver Operating Characteristic) curves and brought out an AUC (Area Under Curve) value to estimate the efficiency of the ICP model (Fig. 5). Since AUC value is an indicator of the efficiency of the classification model, we compared the value from the ICP model with other classification algorithms in GSE 15484 and in GSE 21034 (Tables 1and 2). The comparing algorithms are SVM (Support Vector Machine) [43,44], *k*-means clustering algorithm [42], logistic regression [45,46], decision tree [25,26], TSP (Top Scoring Pair) [27], and *k*-TSP (*k*-Top Scoring Pair) [30]. The comparing algorithms are classification methods which are widely used in bioinformatics and in machine learning. TSP and *k*-TSP are the algorithms which use gene pairs for classification similar to the ICP.

For parameters, polynomial kernel and RBF (Radial Basis Function) kernel are employed for SVM, $k=2$ is used for *k*-means algorithm because the experiment deals with binary classification, linear and multinomial versions of logistic regression are selected for comparison and in decision tree, C4.5 and CART are chosen for attribute selection measure. TSP algorithm does not have user-defined parameters, on the other hand, *k*-TSP and ICP use optimal parameters from cross-validation. General descriptions for the comparing algorithms are represented in the 'related works' section. Relief-A and symmetrical uncertainty are used to select the features in comparing algorithms the same as in the ICP. In order to validate the comparing algorithms, LOOCV (Leave One Out Cross Validation) is used and it is the same technique which is used in the proposed algorithm ICP.

In GSE 15484 data, the AUC of the proposed algorithm is 0.797 in R300 and 0.876 in S300 where the R300 indicates using 300-feature-selection by relief-A and the S300 indicates using 300-feature-selection by symmetrical uncertainty. The ICP showed better performance than *k*-means, C4.5 decision tree, CART decision tree in the R300 and showed better performance than polynomial-kernel-SVM, RBF-kernel-SVM, *k*-means clustering, linear logistic regression, multinomial logistic regression, C4.5 decision tree, CART decision tree, and TSP in the S300. However, the ICP did not outperform *k*-TSP and a few algorithms with AUC in GSE 15484. On the other side, in GSE 21034, the ICP showed outstanding performance both in the R300 and in the S300. The AUC of the ICP is 0.991 in R300 and 0.989 in S300. There are no comparing algorithms that showed higher AUC than the ICP.

The ICP uses gene pairs for classification and it is possible that the ranked gene pairs have biological information. We searched the ranked gene pairs of the ICP on the PubMed compare to the search result of ranked gene pairs of the *k*-TSP which also uses ranked gene pairs. The PubMed is a database which contains abstracts of biomedical literatures and is maintained by NLM (The United States National Library of Medicine). We defined that if two genes of a gene pair are co-occurred in the same literature, the gene pair is informative. The top-scoring 9 gene pairs are chosen and searched on the PubMed (Table 3). Since, the ICP has one more parameter than the *k*-TSP and it is the number of clusters for inner-class clustering algorithm, We searched gene pairs with 2–5 clusters and get the average number of gene pairs which have PubMed search result, i.e. let c denote the number of clusters where $1<c<6$, $\#_c$ denotes the number of gene pairs which have PubMed search result with c clusters, and there are given values that $\#_2=2$, $\#_3=0$, $\#_4=1$, and $\#_5=1$. Then the average number of gene pairs which have PubMed search result is $(2+0+1+1)/4=(4/4)=1$. Based on our definition of informative gene pair, the *k*-TSP has no informative gene pair. However, the ICP has 1 informative gene pair from 9 total gene pairs in GSE 15484 using S300, 2 informative gene pairs from 9 total gene pairs in GSE 21034 using F300, and 6.25 informative gene pairs from 9 total gene pairs in GSE 21034 using S300. There is a list for the gene pairs of the ICP which are highly ranked from classifier and are searched on the PubMed (Tables 4 and 5). The listed gene pairs are worth researching because not only the informative gene pairs but also the other gene pairs can have biological meaning.

Table 3
The number of informative gene pairs which have PubMed search result.

Data	Parameters	*k*-TSP	ICP
GSE 15484	Relief-A (300), # of gene pairs in total=9	0	0
	Symmetrical uncertainty (300), # of gene pairs in total=9	0	1
GSE 21034	Relief-A (300), # of gene pairs in total=9	0	2
	Symmetrical uncertainty (300), # of gene pairs in total=9	0	6.25

Gene expression levels of the informative gene pairs can have interesting information. We evaluated 4 informative gene pairs, PTPN22:SPRED2, TBP:HIP1, MYH11:ACTA2, and ACTA2:AR on 2-dimensional coordinate (Fig. 6). PTPN22:SPRED2 has 1 PubMed search result, TBP:HIP1 has 2 PubMed search results, MYH11:ACTA2 has 16 PubMed search results, and ACTA2:AR has 10 PubMed search results. We found again that the data have heterogeneous characteristics. Gene expression levels are highly distributed even if the samples are in the same class except aggressive samples in MYH11:ACTA2. The ICP can catch heterogeneous characteristics of data, however, that does not indicate the ICP can only handle heterogeneous data. Because even if a data is highly clustered in each class which means the data does not have heterogeneous characteristics, the inner-class clustering results may be at the similar region on 2-dimensional coordinate.

5. Discussion

Data samples do not have similar values even when they are in the same class. We classify on the assumption that most classification algorithms rely on similar values in the same class and different values in the other classes. The ICP algorithm is useful for heterogeneous data because the algorithm performs the classification based on several different aspects of the data.

In the phase of majority voting, gene pairs are obtained by using genes which are selected through feature selection and ranks of the gene pairs are assigned based on their rank score. The gene pairs with high rank provide clearer division between aggressive cancer and non-aggressive cancer than the gene pairs with low rank. Moreover, the ranked gene pairs can have biological information. We investigated 7 gene pairs which have PubMed search result among top-ranking gene pairs and found 4 gene pairs are meaningfully related to biological processes (Table 6). They are TBP:HIP1, PTPN22:SPRED2, AR:MSMB, and MYH11:ACTA2. TBP:HIP1 is involved in transcription regulation process [47], PTPN22:SPRED2 is related to anti-cyclic citrullinated peptide antibodies [48], MSMB has indirect relationship with AR in transcription process [49], and MYH11:ACTA2 can cause TAAD (Thoracic Aortic Aneurysm/Dissection) [50,51].

From the research study on genes which are frequently appeared in top-ranking gene pairs of GSE 15484 and GSE 21034, we found that the genes are related to cancer and many of them are related to prostate cancer. Among genes associated with prostate cancer, HSD17B4, MSMB, PLA2G2A, ACPP, MYLK, and ANO7 are directly related to the prognosis of prostate cancer (Table 7). HSD17B4 is a gene closely related to the prostate cancer gene. It is an important index for determining patients with poor prognosis [52]. There are some studies which mention the relevance of MSMB to prostate cancer. For higher gene expression levels, the possibility of recurrence is lower after radical prostatectomy [53]. Also, the higher the gene expression level of PLA2G2A, the poorer the prognosis of the prostate cancer and the more metastatic it tends to be [54–56]. The expression level of ACPP may help in predicting the stage of prostate cancer and determining the cure [57]. One study suggests that the expression level of ACPP can be a sensitive tumor marker for diagnosing prostate cancer [58]. MYLK identified as a downstream target of the androgen signaling pathway in prostate cells [59]. Another gene ANO7 is a target for the therapies of prostate cancer because of because of its selective expression in prostate cancer [60,61]. The ICP algorithm is significant since the top-ranking gene pairs are related to biological processes, and the genes which are appeared in the top-ranking gene pairs are known to be associated with prostate cancer and its prognosis.

Table 4
The ranked gene pairs of ICP in GSE 15484.

Top-ranking gene pairs
(PTPN22 HSD17B4), (PTPN22, ANKH), (PTPN22, CCDC80), (PTPN22, CYCS), (PTPN22 DSC2), (PTPN22 MYLK), (PTPN22 MSMB), (DSC2 MSMB), (MYLK MSMB), (PTPN22 RPL10A), (PTPN22 SPARC), (URB MSMB), (ANKH MSMB), (ANO7 MSMB), (ADAR MSMB), (HSD17B4 MSMB), (RPL10A, MSMB), (SPARC MSMB), (NSEP1 MSMB), (SNX3 MSMB), (TTC37, MSMB), (NAP1L1 MSMB), (ERRFI1, MSMB), (VDAC1 MSMB), (UGDH MSMB), (ATP6V1A MSMB), (TBP IFIT1), **(TBP HIP1)**, (SEMA3D HIP1), (PTPN22 MFAP1), (PTPN22 GRINL1A), (TBP MFAP1), (PTPN22 PLA2G2A), (TBP URB), (TBP ANKH), (PTPN22 HSP90AA1), (PTPN22, XRN1), (PTPN22 TBP), **(PTPN22 SPRED2)**, (MFAP1 GRINL1A), (PTPN22 ADAR), (PTPN22 ATP13A3), (PTPN22 IGJ), (SYBL1, TBP), (MFAP1 PLA2G2A), (SYBL1 PTPN22), (TBP, PLA2G2A), (PLA2G2A RPL10A), (PLA2G2A SPARC), (PLA2G2A SYBU), (PTPN22 HSD17B4)

* Redundant gene pairs are removed and bolded gene pairs are informative gene pairs which have PubMed search result.

Table 5
The ranked gene pairs of ICP in GSE 21034.

Top-ranking gene pairs
(MYH11 MSMB), (ACTG2, MSMB), (ACTA2 MSMB), **(ACTA2 ACTG2)**, **(AR MSMB)**, (ACPP MSMB), (NRP1 MSMB), (ACTA2 ACPP), (ACTA2 MAOB), (ACTG2 MAOB), **(ACTA2 AR)**, **(MYH11 ACTA2)**, **(TAGLN ACTA2)**, (ACTA2 FLNA), (ACTA2 CSRP1)

* Redundant gene pairs are removed and bolded gene pairs are informative gene pairs which have PubMed search result.

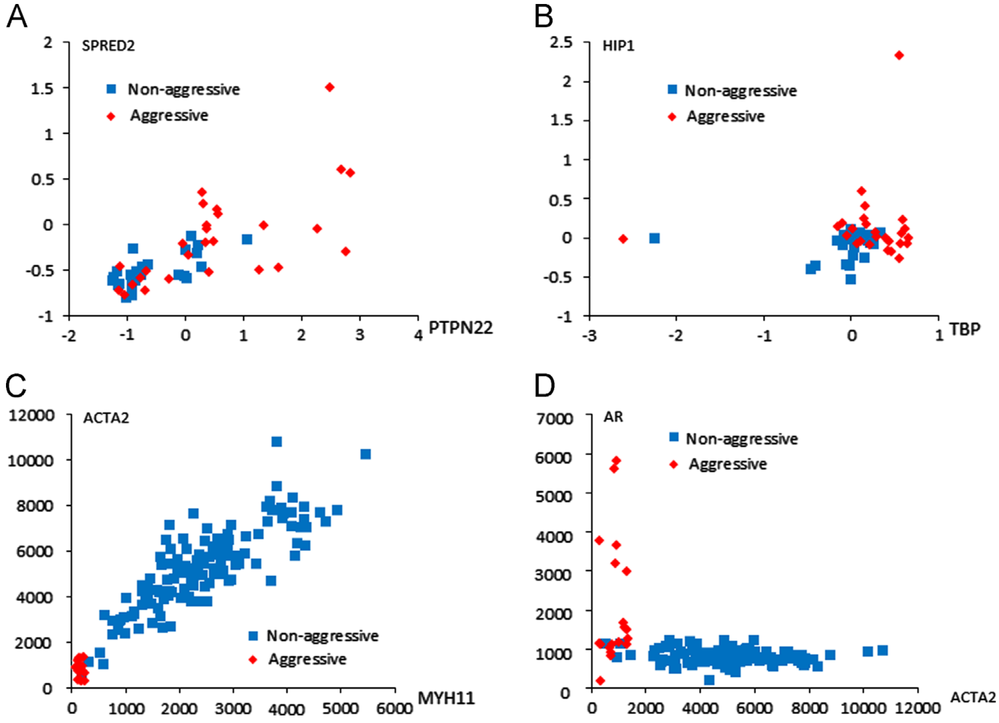

Fig. 6. Gene expression levels of the informative gene pairs. (A) Gene expression levels of PTPN22:SPRED2 gene fusion from GSE 15484. (B) Gene expression levels of TBP: HIP1 gene fusion from GSE 15484. (C) Gene expression levels of MYH11:ACTA2 gene fusion from GSE 21034. (D) Gene expression levels of ACTA2:AR gene fusion from GSE 21034.

Table 6
Literature summary of informative gene pairs which are related to biological processes.

Gene pair	Literature summary
(TBP HIP1)	TBP is significantly enriched in the HIP1 protein interactor binding site which indicates that TBP might co-operate with HIP1 for transcription regulation
(PTPN22 SPRED2)	Associations with the HLA (Human Leukocyte Antigen) region, PTPN22 and SPRED2 improved in individuals positive for anti-cyclic citrullinated peptide antibodies when identifying susceptibility loci for rheumatoid arthritis
(AR MSMB)	Both MSMB and an adjacent gene, NCOA4, are subjected to transcriptional control via androgen response elements. And the androgen receptor (AR) is a member of the steroid receptor superfamily that interacts with androgen response elements to regulate target gene transcription. The NCOA4 interacts directly with the androgen receptor as a co-activator to enhance AR transcriptional activity
(MYH11 ACTA2)	MYH11 and ACTA2 mutations enhance TGF-beta signaling in familial TAAD (Thoracic Aortic Aneurysm/Dissection). Mutations in the vascular smooth muscle cell (SMC)-specific beta-myosin (MYH11) and alpha-actin (ACTA2) can cause the TAAD

Table 7
Literature summary of genes which are frequently appeared in voter sets.

Gene symbol	Literature summary
HSD17B4	Overexpression of HSD17B4 was associated with prostate cancer. Multivariate Cox proportional hazards analysis adjusted for known prognostic indicators revealed that expression of HSD17B4 is important index for poor prognosis. If the expression level of HSD17B4 is high, the prognosis of prostate cancer patient is poor
MSMB	Multivariate analysis adjusted for clinicopathological parameters demonstrated that MSMB expression is an independent predictor of decreased risk of recurrence. MSMB is a significant indicator, predicting outcome after radical prostatectomy for localized prostate cancer. If the expression level of MSMB is high, the possibility of cancer recurrence is low
PLA2G2A	Overexpression of PL2AG2A was observed in prostate cancer samples using quantitative real-time PCR. Its expression was higher in prostate cancer than in benign prostate. PLA2G2A was also related with prostate development and progression, since it is associated significantly with high Gleason score and advanced tumor stage
ACPP	Expression of ACPP showed an inverse correlation with tumor stage. If the expression level of ACPP is high, the tumor stage of the prostate cancer is low. ACPP expression can be potential biomarkers for prostate cancer diagnosis and prognosis and may be helpful for clinical decisions in terms of individual therapeutic management
MYLK	Analysis of gene expression profiles revealed MYLK mRNA levels are markedly down-regulated by the synthetic androgen R1881. And androgens play a major role in the growth and survival of primary prostate tumors. MYLK is a downstream target of the androgen signaling pathway in prostate cells
ANO7	ANO7 is detected only in prostate cancer and normal prostate. Because of its selective expression in prostate cancer, ANO7 is a target for the T cell-mediated immunotherapy and antibody-based therapies of prostate cancer

6. Conclusions

We showed that the method in this paper ICP is novel and showed better performance than other machine learning algorithms and related algorithms which use gene pairs. By researching the gene pairs with high rank score, we found gene pairs associated with biological processes and genes associated with prostate cancer. Many of the genes are already known to be associated with the diagnosis and prognosis of prostate cancer and it is worth researching other genes or gene pairs which have not been sufficiently studied. In addition, a clinical kit can be developed that predicts the prognosis of patients with prostate cancer using the suggested genes.

This method can be applied to multi-class classification. Since multi-class classification is different from binary classification, if we develop the method in this paper further, we will be able to use the ICP for not only binary classification but also for multi-class classification.

Conflict of interest statement

None declared.

Acknowledgements

Funding: This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012010775), and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (20120002887).

References

[1] J. Johansson, O. Andrén, S. Andersson, P. Dickman, L. Holmberg, A. Magnuson, H. Adami, Natural history of early, localized prostate cancer, J. Am. Med. Assoc. 291 (22) (2004) 2713–2719.

[2] P. Albertsen, J. Hanley, J. Fine, 20-year outcomes following conservative management of clinically localized prostate cancer, J. Am. Med. Assoc. 293 (17) (2005) 2095–2101.

[3] J. Baillar, G. Mellinger, D. Gleason, Survival rates of patients with prostatic cancer, tumor stage and differentiation- a preliminary report, Cancer Chemother. Rep. 50 (1966) 129–136.

[4] D. Gleason, Classification of prostate carcinomas, Cancer Chemother. Rep. 50 (1966) 125–128.

[5] D. Gleason, G. Mellinger, The Veteran's Administration Cooperative Urological Research Group, Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging, J. Urol. 111 (1974) 58–64.

[6] D. Gleason, The Veteran's Administration Cooperative Urologic Research Group, Histologic grading and clinical staging of prostatic carcinoma, Urologic Pathology: The Prostate, Lea & Febiger, Philadelphia, PA171–198.

[7] G. Mellinger, D. Gleason, J. Bailar III, The histology and prognosis of prostatic cancer, J. Urol. 97 (1967) 331–337.

[8] B. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B. Carver, V. Arora, P. Kaushik, E. Cerami, B. Reva, Y. Antipin, N. Mitsiades, T. Landers, I. Dolgalev, J. Major, M. Wilson, N. Socci, A. Lash, A. Heguy, J. Eastham, H. Scher, V. Reuter, P. Scardino, C. Sander, C. Sawyers, W. Gerald, Integrative genomic profiling of human prostate cancer, Cancer Cell 18 (1) (2010) 11–22.

[9] U. Chandran, C. Ma, R. Dhir, M. Bisceglia, M. Lyons-Weiler, W. Liang, G. Michalopoulos, M. Becich, F. Monzon, Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process, BMC Cancer 7 (2007) 64–84.

[10] J. Lapointe, C. Li, J. Higgins, M. Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, P. Ekman, A. DeMarzo, R. Tibshirani, D. Botstein, P. Brown, J. Brooks, J. Pollack, Gene expression profiling identifies clinically relevant subtypes of prostate cancer, Proc. Natl. Acad. Sci. U.S.A. 101 (3) (2004) 811–816.

[11] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, E. Lander, M. Loda, P. Kantoff, T. Golub, W. Sellers, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2) (2002) 203–209.

[12] B. Boser, I. Guyon, V. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the 5th Annual ACM Workshop on COLT, 1992, pp. 144–152.

[13] M. Hu, Y. Chen, J. Kwok, Building sparse multiple-kernel SVM classifiers, IEEE Trans. Neural Networks 20 (5) (2009) 827–839.

[14] S. Maji, A. Berg, J. Malik, Classification using intersection kernel support vector machines is efficient, in: Proceedings of Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[15] D. Simian, A model for a complex polynomial SVM kernel, in: Proceedings of the 8th Conference on Simulation, Modeling and Optimization, 2008, pp. 164–169.

[16] J. Sun, Fast tuning of SVM kernel parameter using distance between two classes, in: Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering, 2008, pp. 108–113.

[17] H. Drucker, C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, in: Proceedings of Advances in Neural Information Processing Systems 9, 1997, pp. 155–161.

[18] J. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (3) (1999) 293–300.

[19] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.

[20] P. Bradley, O. Mangasarian, W. Street, Feature selection via mathematical programming, INFORMS J. Comput. 10 (1998) 209–217.

[21] L. Wang, J. Zhu, H. Zou, Hybrid huberized support vector machines for microarray classification and gene selection, Bioinformatics 24 (3) (2008) 412–419.

[22] J. Friedman, T. Hastie, R. Tibshirani, Additive Logistic Regression: A Statistical View of Boosting, Stanford University, CA, USA, 1998.

[23] J. Liao, K. Chin, Logistic regression for disease classification using microarray data: model selection in a large p and small *n* case, Bioinformatics 23 (15) (2007) 1945–1951.

[24] J. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.

[25] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

[26] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth and Brooks/Cole Advanced Books and Software.

[27] D. Geman, C. d'Avignon, D. Naiman, R. Winslow, Classifying gene expression profiles from pairwise mRNA comparisons, Stat. Appl. Genet. Mol. Biol. 3 (2004), Article 19.

[28] D. German, B. Afsari, A. Tan, D. Naiman, Microarray classification from several two gene expression comparisons, in: Proceedings of the 7th International Conference on Machine Learning and Applications, 2008, pp. 583–585.

[29] E. Shin, Y. Yoon, J. Ahn, S. Park, TC-VGC: A tumor classification system using variations in gene's correlation, Comput. Methods Programs in Biomed. 104 (3) (2011) 87–101.

[30] A. Tan, D. Naiman, L. Xu, R. Winslow, D. Geman, Simple decision rules for classifying human cancers from gene expression profiles, Bioinformatics 21 (20) (2005) 3896–3904.

[31] A. Teschendorff, S. Gomez, A. Arenas, D. El-Ashry, M. Schmidt, M. Gehrmann, C. Caldas, Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules, BMC Cancer 10 (2010) 604–623.

[32] N. Pressinotti, H. Klocker, G. Schäfer, V. Luu, M. Ruschhaupt, R. Kuner, E. Steiner, A. Poustka, G. Bartsch, H. Sültmann, Differential expression of apoptotic genes PDIA3 and MAP3K5 distinguishes between low- and high-risk prostate cancer, Mol. Cancer 8 (2009) 130–141.

[33] V. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. USA 98 (9) (2001) 5116–5121.

[34] K. Leinonen, T. Tolonen, H. Bracken, U. Stenman, T. Tammela, O. Saramäki, T. Visakorpi, Association of SPINK1 expression and TMPRSS2:ERG fusion with prognosis in endocrine-treated prostate cancer, Clin. Cancer Res. 16 (10) (2010) 2845–2851.

[35] S. Minner, B. Jessen, L. Stiedenroth, E. Burandt, J. Köllermann, M. Mirlacher, A. Erbersdobler, C. Eichelberg, M. Fisch, T. Brümmendorf, C. Bokemeyer, R. Simon, T. Steuber, M. Graefen, H. Huland, G. Sauter, T. Schlomm, Low level Her2 overexpression is associated with rapid tumor cell proliferation and poor prognosis in prostate cancer, Clin. Cancer Res. 16 (5) (2010) 1553–1560.

[36] B. Barwick, M. Abramovitz, M. Kodani, C. Moreno, R. Nam, W. Tang, M. Bouzyk, A. Seth, B. Leyland-Jones, Prostate cancer genes associated with TMPRSS2-ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts, Br. J. Cancer 102 (2010) 570–576.

[37] S. Fine, A. Gopalan, M. Leversha, H. Al-Ahmadie, S. Tickoo, Q. Zhou, J. Satagopan, P. Scardino, W. Gerald, V. Reuter, TMPRSS2-ERG gene fusion is associated with low Gleason scores and not with high-grade morphological features, Mod. Pathol. 23 (2010) 1325–1333.

[38] D. Duggan, M. Bittner, Y. Chen, P. Meltzer, J. Trent, Expression profiling using cDNA microarrays, Nat. Genet. Suppl. 21 (1999) 10–14.

[39] I. Kononenko, Estimating attributes: analysis and extensions of RELIEF, in: Proceedings of European Conference on Machine Learning, 1994, pp. 171–182.

[40] W. Press, B. Flannery, S. Teukolsky, W. Vetterling, Numerical Recipes in C, Cambridge University Press, 1988.

[41] Y. Wang, I. Tetko, M. Hall, E. Frank, A. Facius, K. Mayer, H. Mewes, Gene selection from microarray data for cancer classification – a machine learning approach, Comput. Biol. Chem. 29 (1) (2005) 37–46.

[42] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1967, pp. 281–297.

[43] S. Keerthi, S. Shevade, C. Bhattacharyya, K. Murthy, Improvements to Platt's SMO algorithm for SVM classifier design, Neural Comput. 13 (3) (2001) 637–649.

[44] J. Platt, Fast training of support vector machines using sequential minimal optimization, Advances in Kernel Methods-Support Vector LearningMIT Press Cambridge, MA, USA, 1999.

[45] S. le Cessie, J. van Houwelingen, Ridge estimators in logistic regression, Appl. Stat. 41 (1) (1992) 191–201.

[46] N. Landwehr, M. Hall, E. Frank, Logistic model trees, Mach. Learn. 59 (1,2) (2005) 161–205.

[47] M. Datta, A. Choudhury, N. Bhattacharyya, Genome wide gene expression regulation by HIP1 protein interactor, HIPPI: predict and validation, BMC Genomics 12 (2011) 463–479.

[48] D. Herraez, et al., Rheumatoid arthritis in Latin Americans enriched for Amerindian ancestry is associated with loci in chromosomes 1, 12,13, and HLA class II region, Arthritis & RheumatismWiley Subscription Services, 2013.

[49] H. Lou, H. Li, M. Yeager, K. Im, B. Gold, T. Schneider, J. Fraumeni Jr, S. Chanock, S. Anderson, M. Dean, Promoter variants in the MSMB gene associated with prostate cancer regulate MSMB/NCOA4 fusion transcripts, Hum. Genet. 131 (9) (2012) 1453–1466.

[50] D. Milewicz, D. Guo, V. Tran-Fadulu, A. Lafont, C. Papke, S. Inamoto, C. Kwartler, H. Pannu, Genetic basis of thoracic aortic aneurysms and dissections: focus on smooth muscle cell contractile dysfunction, Annu. Rev. Genom. Hum. Genet. 9 (2008) 283–302.

[51] M. Renard, et al., Novel MYH11 and ACTA2 mutations reveal a role for enhanced TGFβ signaling in FTAAD, Int, J. Cardiol. 165 (2) (2013) 314–321.

[52] K. Rasiah, M. Gardiner-Garden, E. Padilla, G. Moller, J. Kench, M. Alles, S. Eggleton, P. Stricker, J. Adamski, R. Sutherland, S. Henshall, V. Hayes, HSD17B4 overexpression, an independent biomarker of poor patient outcome in prostate cancer, Mol. Cell Endocrinol. 301 (1,2) (2009) 89–96.

[53] A. Dahlman, E. Rexhepaj, D. Brennan, W. Gallagher, A. Gaber, A. Lindgren, K. Jirström, A. Bjartell, Evaluation of the prognostic significance of MSMB and CRISP3 in prostate cancer using automated image analysis, Mod. Pathol. 24 (5) (2011) 708–719.

[54] Z. Dong, Y. Liu, K. Scott, L. Levin, K. Gaitonde, R. Bracken, B. Burke, Q. Zhai, J. Wang, L. Oleksowicz, S. Lu, Secretory phospholipase A2-IIa is involved in prostate cancer progression and may potentially serve as a biomarker for prostate cancer, Carcinogenesis 31 (11) (2010) 1948–1955.

[55] T. Mirtti, V. Laine, H. Hiekkanen, S. Hurme, O. Rowe, T. Nevalainen, M. Kallajoki, K. Alanen, Group IIA phospholipase a as a prognostic marker in prostate cancer: relevance to clinicopathological variables and disease-specific mortality 117 (3) (2009) 151–161c-myc and phospholipase 2A in prostate cancer tissue samples obtained by needle biopsy. Pathol. Oncol. Res. 117 (3) (2009) 279–283.

[57] S. Gunia, S. Koch, M. May, M. Dietel, A. Erbersdobler, Expression of prostatic acid phosphatase (PSAP) in transurethral resection specimens of the prostate is predictive of histopathologic tumor stage in subsequent radical prostatectomies, Virchows Arch. 454 (5) (2009) 573–579.

[58] Q. Huo, Protein complexes/aggregates as potential cancer biomarkers revealed by a nanoparticle aggregation immunoassay, Colloids Surf. B Biointerfaces 78 (2) (2010) 259–265.

[59] N. Leveille, A. Fournier, C. Labrie, Androgens down-regulate myosin light chain kinase in human prostate cancer cells, J. Steroid Biochem. Mol. Biol. 114 (3–5) (2009) 174–179.

[60] T. Bera, S. Das, H. Maeda, R. Beers, C. Wolfgang, V. Kumar, Y. Hahn, B. Lee, I. Pastan, NGEP, a gene encoding a membrane protein detected only in prostate cancer and normal prostate, Proc. Natl. Acad. Sci. U.S.A. 101 (9) (2004) 3059–3064.

[61] V. Cereda, D. Poole, C. Palena, S. Das, T. Bera, C. Remondo, J. Gulley, P. Arlen, J. Yokokawa, I. Pastan, J. Schlom, K. Tsang, New gene expressed in prostate: a potential target for T cell-mediated prostate cancer immunotherapy, Cancer Immunol. Immunother. 59 (1) (2010) 63–71.