GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Prediction of core cancer genes using a hybrid of feature selection and machine learning methods

Y.X. Liu[1,3], N.N. Zhang[2], Y. He[1,3] and L.J. Lun[4]

[1]School of Basic Medical Science, Harbin Medical University,
Harbin, Heilongjiang, China
[2]Modern Laboratory Centre, Harbin Normal University, Harbin, China
[3]Network & Information Centre, Harbin Medical University,
Harbin, Heilongjiang, China
[4]College of Computer Science and Information Engineering,
Harbin Normal University, Harbin, China

Corresponding author: Y.X. Liu
E-mail: liuyixin@ems.hrbmu.edu.cn

Genet. Mol. Res. 14 (3): 8871-8882 (2015)
Received December 4, 2014
Accepted April 6, 2015
Published August 3, 2015
DOI http://dx.doi.org/10.4238/2015.August.3.10

ABSTRACT. Machine learning techniques are of great importance in the analysis of microarray expression data, and provide a systematic and promising way to predict core cancer genes. In this study, a hybrid strategy was introduced based on machine learning techniques to select a small set of informative genes, which will lead to improving classification accuracy. First feature filtering algorithms were applied to select a set of top-ranked genes, and then hierarchical clustering and collapsing dense clusters were used to select core cancer genes. Through empirical study, our approach is capable of selecting relatively few core cancer genes while making high-accuracy predictions. The biological significance of these genes was evaluated using systems biology analysis. Extensive functional pathway and network analyses have confirmed findings in previous studies and can bring new insights into common cancer mechanisms.

Key words: Feature selection; Machine learning; Core cancer gene; Microarray data; Classification method

INTRODUCTION

Various machine learning methods have been applied for cancer diagnostic research (Cruz and Wishart, 2007; Lisboa and Taktak, 2006). From the medical perspective, cancer comprises more than 100 distinct diseases specific to cell type and tissue origin (Stratton et al., 2009). For these diseases, diagnosis is to identify one disease by its signs and symptoms while prognosis is to predict the outcome of the disease and status of the patient. Therefore, it is widely believed that a number of cancers share a common pathogenesis (Stratton et al., 2009). Elucidating common cancer mechanisms will certainly enhance our ability to devise effective therapeutics against the disease responsible for one in eight deaths worldwide (Khalil and Hill, 2005). Researchers have proven that machine learning methods could generate more accurate diagnoses or prognoses as compared to traditional statistical methods (Cruz and Wishart, 2007; Cheng and Cheng, 2009).

In some studies, the researchers have attempted to identify the core cancer genes, or the meta-signatures across a wide range of cancer types by analyzing genome-wide gene expression profiles from multiple-microarray data sets (Rhodes et al., 2004; Segal et al., 2004; Chuang and Yang, 2009; Gao et al., 2013). The research of common cancer mechanisms are part of an emerging biological domain termed cancer systems biology (Kreeger and Lauffenburger, 2010). In order to discovering the common cancer mechanisms from a genome-wide gene expression profiles, a feature selection technique is one better way for selecting a small subset of genes as features for classification. In this research, we focus on this gene selection problem and attempt to discover core cancer genes using a hybrid approach (Wang et al., 2005). In a microarray dataset FHCCancer9 including 9 sub-datasets, each represented a binary classification of cancer *vs* normal samples. We applied feature filtering algorithms on the whole set of genes using training data, pre-select top-ranked genes from the whole set, and finally applied pre-filtering approaches to select core cancer genes. The pre-filtering approaches had two levels, namely hierarchical clustering (HC) and collapsing dense clusters. In the HC level, HC was applied to this set of pre-selected genes, while collapsing dense clusters was used to reduce the gene redundancy and extract core cancer genes.

As a result, 41 Affymetrix probe sets were identified of a total of 22,277 sets found in all samples, as core features of 9 cancer types. The effectiveness of 41 features was cross-validated on the training dataset FHCCancer9_train and was further tested on an independent dataset FHCCancer9_test. Systems biology analysis for these 41 genes is largely consistent with previous studies and brings new insights into possible common mechanisms of cancers.

MATERIAL AND METHODS

Dataset construction

The gene expression dataset FHCCancer9 used in this study was compiled from the web resource ONCOMINE (http://www.oncomine.org) (Rhodes et al., 2007) in April 2014. The primary filtering criteria were set to Differential Analysis and Cancer versus Normal Analysis, including three criteria: 1) each dataset must be specific to one cancer type, which represents one single task, 2) each dataset must contain the appropriate proportion of positive and negative samples, and 3) the largest one among datasets of the same cancer type was chosen. The platform filtering criterion was set to Affymetrix U133 to minimize the platform variation. Passing the filtering steps, we selected 9 datasets as the dataset FHCCancer9 from a total of 376 datasets.

Overall, FHCCancer9 covered 9 common cancer types: pancreas, vulva, prostate, leukemia, renal, lung, esophagus, colon, and breast cancer. In this study, 375 samples were used, with 193 cancer samples and 182 normal samples. The dataset FHCCancer9_train and the dataset FHCCancer9_test consisted of 249 and 126 samples, respectively, with a ratio of approximately 2:1 cancer samples to normal samples. The detailed description of FHCCancer9 can be seen in the **Supplementary File 1**.

Data preprocess and representation

Data preprocessing is very important for microarray data analysis. We performed data preprocessing as follows: 1) Affymetrix U133 platform includes three types: Human Genome U133 Plus 2.0 Array, Human Genome U133A 2.0 Array, and Human Genome U133A&B. These types differ by the number of probe sets presented in the chip. The shared genes of those three types of microarrays are represented by 22,277 Affymetrix identifiers, which are used as features to describe each sample. 2) Cancer samples are defined as positives and normal samples as negatives. 3) Samples in FHCCancer9 are normalized by the Robust Multi-Array Average (RMA) algorithm (Irizarry et al., 2003) individually. Finally, the sample is represented by N = 22,277 features in such form $X_i = (X_{i1}, X_{i2}, . . ., X_{ij}, . . ., X_{in})$.

Gene ranking

Pearson's correlation coefficient (*PCC*)

The PCC, also known as the product moment correlation coefficient, is represented in a sample by r. The coefficient is measured on a scale with no units and can take a value from -1 through 0 to +1. For a series of n variables of X and Y (denoted by x_i and y_i, respectively, where i = 1, 2,…, n), the sample correlation coefficient can be used to estimate the population Pearson's correlation r between X and Y, as shown in Equation 1, where x and y are the sample means of X and Y, accordingly, and s_x and s_y are the sample standard deviations of X and Y, accordingly. In this research, r is calculated and ranked for each of the feature input and the team with the highest r is selected.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \quad \text{(Equation 1)}$$

Relief-F

Relief-F (Kononenko, 1994) is a simple yet efficient procedure to estimate the quality of attributes in problems with strong dependencies between attributes. The basic idea of Relief-F is to draw instances at random, compute their nearest neighbors, and adjust a feature-weighting vector to give more weight to features that discriminate the instance from neighbors of different classes. This approach has shown good performance in various domains (Robnik and Kononenko, 2003). In this research, each feature input is ranked and weighted using the k-nearest neighbors classification, in which k = 5. The top features with large positive weights will be selected.

Information gain

Information gain (Liu et al., 2002) is used to select the features that hold the most information about each classification. This method has been used for gene selection by Liu et al. (2002) and Li et al. (2004). Note that information gain requires that numeric features be discretized. Li et al. (2003) indicated that mean-entropy discretized features are effective for classification using gene expression data. It measures the number of bits of information obtained for class prediction by knowing the value of a feature. The information gain of a feature *f* is defined to be:

$$G(f) = -\sum_{i=1}^{m} P(c_i)logP(c_i) + \sum_{v\in V}\sum_{i=1}^{m} P(f=v)P(c_i|f=v)logP(c_i|f=v) \quad \text{(Equation 2)}$$

Hierarchical clustering and the reduction of gene redundancy

After applying feature filtering algorithms on the whole set of genes using the dataset FHCCancer9_train, we preselected the top-ranked genes from the whole set. Typically, top-50, top-100, top-200, and top-300 genes were preselected. These genes can be used as classification characteristics of the sample. However, those genes may still contain redundancies. For reducing the dimension of the feature set and improving the classification accuracy of the classifier, we then presented the HCC gene approach, hierarchical clustering, and collapsing clusters for the core gene selection. Figure 1 shows the flowchart used for the HCC gene approach.

The hierarchical clustering algorithm is a hybrid between an agglomerative (bottom up) and a divisive (top down) algorithm. The dendrogram is built from the root node (all elements) down to the leaf nodes, the clusters in each level are ordered with a deterministic algorithm based on the same distance metric that is used in the clustering. In this way, the ordering produced in the final level of the tree does not depend on that of the data in the original data set (as can be the case with algorithms that have a random component in their ordering methods). We refer to our particular implementation with PAM (Partitioning around medoid) as the partitioning algorithm to produce a dendrogram.

A collapsing step can be applied at any level of the tree to unite similar clusters. By combining the strengths of two celebrated approaches with clustering, partitioning, and agglomerative methods, we create a more flexible algorithm for finding patterns in data. The Median (or Mean) Split Silhouette (MSS) criterium is used to determine the optimal number of children at each node, decide which pairs of clusters to collapse at each level, and identify the first level of the tree with maximally homogeneous clusters. In each case, the goal is to minimize MSS. Collapsing clusters output as the core genes the smallest set of representative genes.

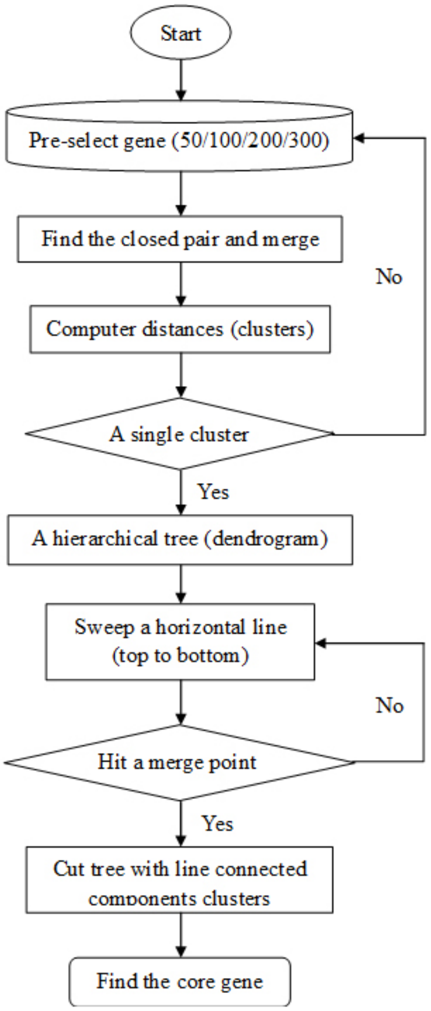
Start
Pre-select gene (50/100/200/300)
Find the closed pair and merge
No
Computer distances (clusters)
A single cluster
Yes
A hierarchical tree (dendrogram)
Sweep a horizontal line
(top to bottom)
No
Hit a merge point
Yes
Cut tree with line connected
components clusters
Find the core gene

Figure 1. HCC gene approach for the core gene selection flowchart.

Classification

K-nearest neighbor (*k*-NN)

The *k*-NN classifier (Dasarathy, 1991) is a well-known nonparametric classifier. In *k*-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k-nearest neighbors (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

Support vector machine (*SVM*)

Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns for classification and regression. SVMs can efficiently perform linear classification and a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. In this study, a widely used SVM tool, LIBSVM (Chang and Lin, 2011), was used. There are 2 steps involved in the LIBSVM: 1) the dataset was trained to obtain a model, and 2) the model was used to predict the information for the testing dataset. The details for LIBSVM can be found in the reference (Chang and Lin, 2011).

Random decision forests (*RF*)

Random forests (Breiman, 2001) are an ensemble learning method for classification and regression that uses an ensemble of unpruned decision trees, each of which is built on a bootstrap sample of the training data using a randomly selected subset of variables. In summary, this algorithm possesses a number of properties, making it an attractive technique for the classification of microarray gene expression data.

RESULTS AND DISCUSSION

In this study, a hybrid method for classification was used to select the core cancer genes. Description of the algorithms is provided in the Material and Methods section. The research consists of three parts. First, three gene ranking methods were used to build classifiers with all 22,277 features to select the top-ranked genes. All of the top-ranked genes are directly used for classification. Second, the hierarchical clustering algorithm and collapsing clusters were applied to select the core cancer genes. The effectiveness of feature selection was cross-validated on the dataset FHCCancer9_train and was further tested on an independent testing dataset FHCCancer9_test. Third, these core genes were mapped and analyzed by functional annotations, clustering analysis, pathways, and networks.

Feature selection and validation

Each of the three gene ranking methods (Relief-F, information gain, and *HCC*) was used to select the top-50, top-100, top-200 and top-300 ranked genes. Each was cross-validated on the dataset FHCCancer9_train. Five-fold cross-validation of parameters is used to estimate the performance of all the classifiers. Then, the accuracy of all the top-ranked genes was compared and was directly used for classification. The top-ranked genes are further processed using the HCC gene approach. Note that the HCC gene approach selects different numbers of genes with different classifiers and then those selected genes are tested for the efficiency of classification on the dataset FHCCancer9_test. The results show that applying the HCC gene approach produced higher classification accuracy than applying the top-ranked genes directly (Tables 1-3). In all cases, the best feature selection method for FHCCancer9 is PCC with *k*-NN classification. Finally, all 22,277 features were ranked by PCC and 41 features (**Supplementary File 2**) were selected from the top-200 by the HCC gene approach as significant common features. As a classification method, the classifier with 41 genes reached greater classification accuracy (98.41%) than another by feature selection.

Table 1. Classification accuracy (Acc) on dataset FHCCancer9 using Relief-F.

Relief-F		*k*-NN	SVM	RF
50 Top-ranked genes	All 50 Acc	85.71	91.26	80.95
	Genes	27	27	27
	HCCgene Acc	90.47	93.65	88.88
100 Top-ranked genes	All 100 Acc	88.88	92.06	82.53
	Genes	25	25	25
	HCCgene Acc	90.47	93.65	90.47
200 Top-ranked genes	All 200 Acc	90.47	94.44	86.50
	Genes	33	33	33
	HCCgene Acc	94.44	96.82	92.06
300 Top-ranked genes	All 300 Acc	87.30	92.85	87.30
	Genes	37	37	37
	HCCgene Acc	92.06	93.65	90.47

Table 2. Classification accuracy (Acc) on dataset FHCCancer9 using PCC.

PCC		*k*-NN	SVM	RF
50 Top-ranked genes	All 50 Acc	93.65	90.48	88.89
	Genes	31	31	31
	HCCgene Acc	95.24	92.86	90.48
100 Top-ranked genes	All 100 Acc	94.44	90.48	89.68
	Genes	23	23	23
	HCCgene Acc	97.62	93.65	90.48
200 Top-ranked genes	All 200 Acc	95.24	93.65	90.48
	Genes	41	41	41
	HCCgene Acc	98.41	94.44	93.65
300 Top-ranked genes	All 300 Acc	92.86	91.27	90.48
	Genes	35	35	35
	HCCgene Acc	96.83	93.65	91.27

Table 3. Classification accuracy (Acc) on dataset FHCCancer9 using information gain.

Information gain		*k*-NN	SVM	RF
50 Top-ranked genes	All 50 Acc	88.10	83.33	90.48
	Genes	25	25	25
	HCCgene Acc	92.06	92.86	89.68
100 Top-ranked genes	All 100 Acc	88.10	86.51	87.30
	Genes	20	20	20
	HCCgene Acc	88.10	92.86	89.68
200 Top-ranked genes	All 200 Acc	86.51	84.13	88.10
	Genes	24	24	24
	HCCgene Acc	86.51	92.86	89.68
300 Top-ranked genes	All 300 Acc	86.51	87.30	88.10
	Genes	45	45	45
	HCCgene Acc	87.30	92.86	88.89

Cancer gene expression profile, clustering analysis and network

It is widely accepted that genes involved in the same networks and pathways likely share similar expression patterns. In order to elucidate possible mechanisms, we performed hierarchical clustering of these 41 core cancer genes across all 375 samples of FHCCancer9 (Figure 2). The clustering of the heat map rows shows the relationship between samples and the clustering of the heat map columns shows the relationship between genes (**Supplementary File 3**) (Figure 3). The comparison of the clustering results with Ingenuity Pathway Analysis (IPA, http://www.ingenuity.com) (Jimenez et al., 2009) may raise interesting questions or form new hypotheses. These complex relationships between core cancer genes may provide useful information for further investigation. For example, one cluster includes *NUSAP1*, *FEN1*, and *MCM2* with similar expression patterns. Among them, *FEN1* and *MCM2* have a very close relationship and both hubs are in the network. According to the Atlas of Genetics and Cytogenetics in Oncology and Haematology, *FEN1* is implicated in prostate cancer, pancreatic cancer, gastric cancer, lung cancer, brain cancer, breast cancer, and testicular cancer. Another gene related to *FEN1* in the network is *CDK1*. These two are connected by a solid line, which represents a very close relationship between the two genes. Indeed, it has been found that CDK1 can decrease FEN1 activity *in vitro* (Freedland et al., 2003).

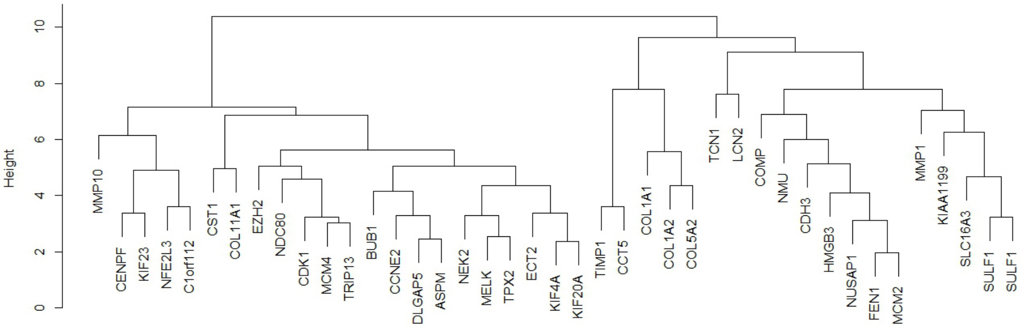

Figure 2. Clustering results of 41 genes by 375 samples.

Biological significance of selected core genes

The biological significance of the identified features must be determined by experimental validation. We attempted to address the biological significance of these genes using systems biology analysis. The rationale of this approach is that these core genes should be related to cancers and play significant roles in networks and pathways important to cancers. In addition, we want to investigate whether the selected genes can interact with each other and form networks.

We first mapped the 41 significant features (Affymetrix probe sets) to 41 known genes. We then compared these genes to annotated cancer genes and genes possibly implicated in cancers collected by the Atlas of Genetics and Cytogenetics in Oncology and Haematology (http://atlasgeneticsoncology.org/). Among these 41 genes, 15 are annotated cancer genes and 23 are possibly related to cancers (**Supplementary File 4**). Analysis of all genes using Ingenuity Pathway Analysis of all genes using IPA reveals that all of the genes are either directly related to cancers or indirectly through interactions with cancer-related biomolecules. Therefore, all selected genes are known to be or are possibly related to cancers.

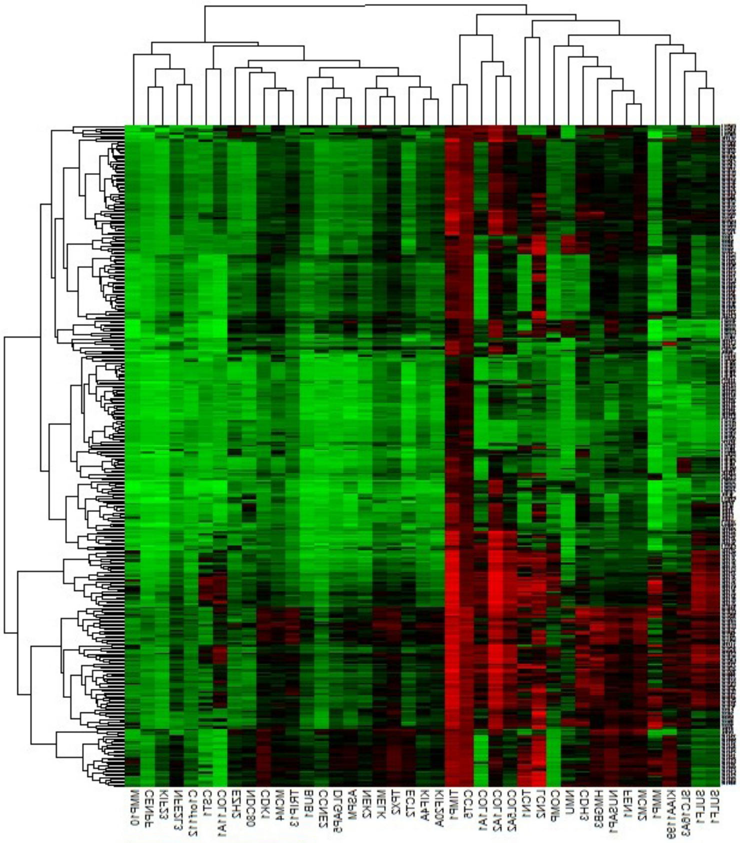

Figure 3. Heat map of 41 core gene profile.

We further applied IPA to systematically analyze these 41 genes. IPA Network Analysis constructs three networks among 41 genes (**Supplementary File 5**). The most significant network (Figure 4) consists of 22 genes. It is noteworthy that the most significant feature is mapped to *TIMP1*, a gene that is expressed in almost all carcinomas.

IPA Functional Analysis maps these genes to a set of diseases and disorders, determined by the P values of the Fisher exact test P values (**Supplementary File 6**). The top 10 categories are displayed in Figure 5. The most significant function and disease category is cancer, which includes a number of cancer-related functional annotations. Cancer is also one of the top 5 significant disease categories, along with organismal injury and abnormalities, reproductive system disease, connective tissue disorders, and tumor morphology, which include 40, 36, 32, 15, and 11 of these 41 genes, respectively. Therefore, the majority of the identified genes are known to be related to cancer. IPA Canonical Pathway Analysis identifies metabolic and cell signaling pathways where these 41 genes are enriched. The top 10 significant pathways, ordered by their Fisher exact test negative log P values, are displayed in Figure 6.

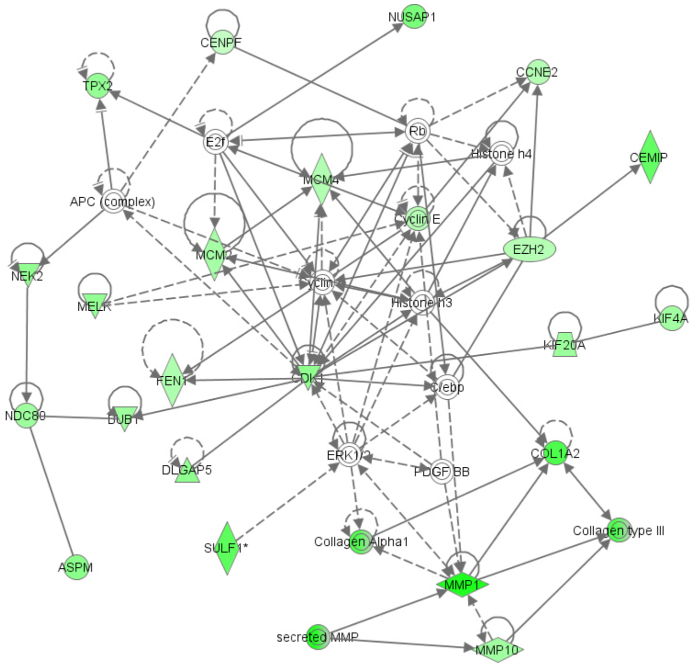

Figure 4. Network from 3 IPA networks. The solid line represents a relationship between the genes such that they are considered a member of the same.

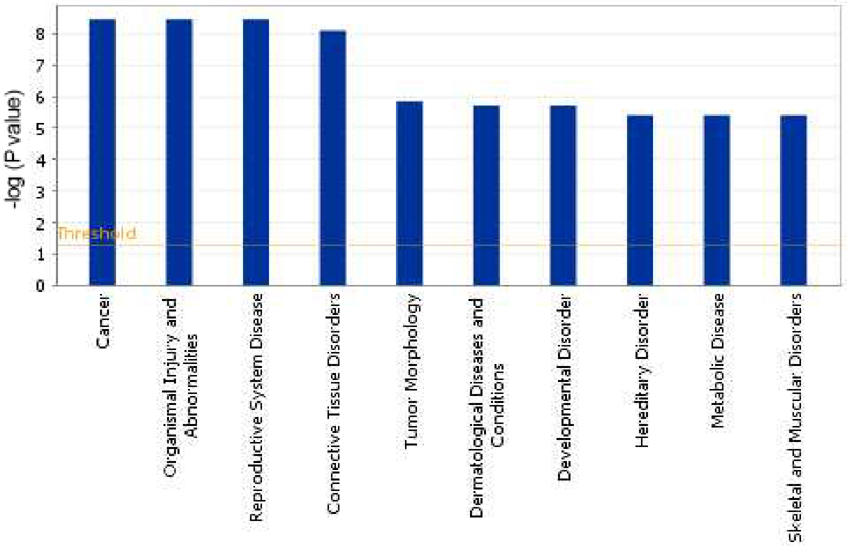

Figure 5. Functional analysis by IPA.

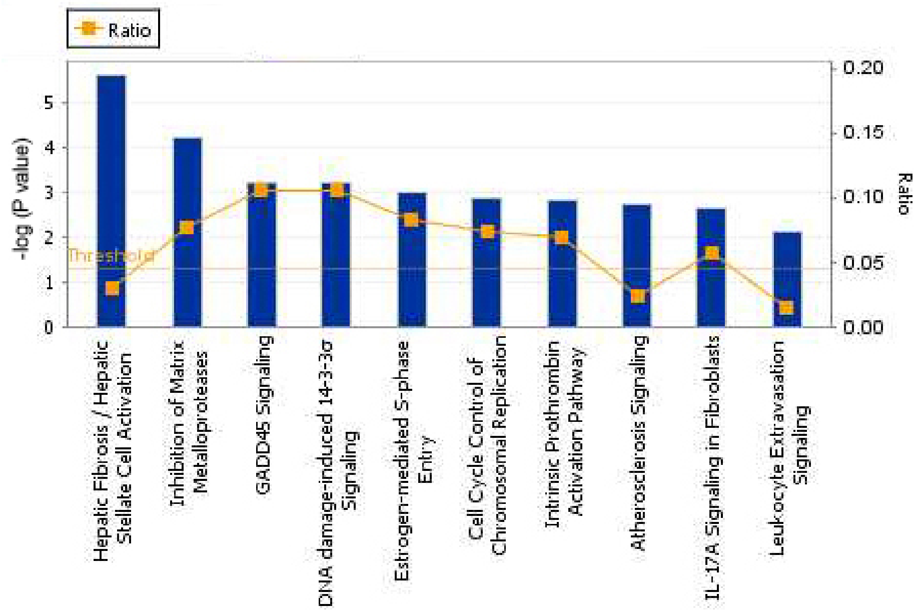

Figure 6. Disease categories and pathway analysis by IPA.

CONCLUSION

In this study, a novel hybrid approach was introduced and 41 core cancer genes were identified from a microarray dataset covering 9 cancer types. The combined use of two mechanisms implemented in these methods makes the feature selection more reliable and robust. Using far fewer genes, our approach is able to offer better (or the same) accuracy compared with conventional approaches.

All selected genes are either known cancer genes or probably related to cancers. The analysis has clearly shown that the selected genes play important roles in cancers and form important networks through their interactions with each other. According to drug targets, we are currently collaborating with a pharmacologist to develop anticancer drugs.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the PhD Start-up Fund of Natural Science Foundation of Harbin Normal University of Heilongjiang Province of China (#XKB201313).

Supplementary material

REFERENCES

Breiman L (2001). Random Forests. *Mach. Learn*. 45: 5-32.

Chang CC and Lin CJ (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intel. Syst. Technol*. 2: 21-27.

Cheng Q and Cheng J (2009). Sparsity optimization method for multivariate feature screening for gene expression analysis. *J. Comput. Biol*. 16: 1241-1252.

Chuang LY and Yang CH (2009). Tabu Search and Binary Particle Swarm Optimization for Feature Selection Using Microarray Data. *J. Comput. Biol*. 12:▪1689-1703.

Cruz JA and Wishart DS (2007). Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*. 2:59-77.

Dasarathy B (1991). Nearest Neighbor Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, Los Alamitos, CA, USA.

Freedland SJ, Pantuck AJ, Paik SH, Zisman A, et al. (2003). Heterogeneity of molecular targets on clonal cancer lines derived from a novel hormone-refractory prostate cancer tumor system. *Prostate* 55: 299-307.

Gao S, Xu S, Fang Y and Fang J (2013). Prediction of core cancer genes using multi-task classification framework. *J. Theor. Biol*. 317: 62-70.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249-264.

Jimenez MA, Collado RM, Ramirez BM, Arce C, et al. (2009). Biological pathway analysis by ArrayUnlock and ingenuity pathway analysis. *BMC Proc*. 3 (Suppl 4): S6.

Khalil IG and Hill C (2005). Systems biology for cancer. *Curr. Opin. Oncol*. 17: 44-48.

Kononenko I (1994). Estimating attributes: analysis and extension of RELIEF. In: ECML-94 Proceedings of the European conference on machine learning on Machine Learning: 1994, Catania, Italy, 171-182.

Kreeger PK and Lauffenburger DA (2010). Cancer systems biology: a network modeling perspective. *Carcinogenesis* 31: 2-8.

Li J, Liu H and Wong L (2003). Mean-entropy discretized features are effective for classifying high-dimensional biomedical data. BIOKDD03, Washington, DC.

Li T, Zhang C and Ogihara M (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* 20: 2429-2437.

Lisboa PJ and Taktak AF (2006). The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Netw*. 19: 408-415.

Liu H, Li J and Wong L (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform.*13: 51-60.

Rhodes DR, Yu J, Shanker K, Deshpande N, et al. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. U. S. A*. 101: 9309-9314.

Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, et al. (2007). Oncomine 3.0: Genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 9: 166-180.

Robnik SM and Kononenko I (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn*. 53: 23-69.

Segal E, Friedman N, Koller D and Regev A (2004). A module map showing conditional activity of expression modules in cancer. *Nat. Genet*. 36: 1090-1098.

Stratton MR, Campbell PJ and Futreal PA (2009). The cancer genome. *Nature* 458: 719-724.

Wang Y, Makedon FS, Ford JC and Pearlman J (2005). HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics* 21: 1530-1537.