The Analysis of Colon Cancer Gene Expression Profiles and the Extraction of Informative Genes

Article *in* Journal of Computational and Theoretical Nanoscience · May 2013

DOI: 10.1166/jctn.2013.2812

Bin Yu
Qingdao University of Science and Technology
83 PUBLICATIONS **2,269** CITATIONS

Yan Zhang
Qingdao University of Science and Technology
50 PUBLICATIONS **771** CITATIONS

Journal of
Computational and Theoretical Nanoscience
Vol. 10, 1097–1103, 2013

The Analysis of Colon Cancer Gene Expression Profiles and the Extraction of Informative Genes

Bin Yu[1,*] and Yan Zhang[2]

[1] *College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao, 266061, China*
[2] *College of Electromechanical Engineering, Qingdao University of Science and Technology, Qingdao, 266061, China*

Based on the analysis of colon cancer gene expression profiles data, this paper proposes the new method of selecting its informative genes. First, considering the characteristics of the high dimensionality and small samples of gene expression profiles data, we use Bhattacharyya distance to reduce dimensionality of the data. Second, we establish two-class classification model for colon cancer based on GA-SVM and use the genetic algorithm (GA) to generate informative genes subset in order to utilize support vector machine (SVM) as classifier to get classification accuracy, which is 95.62%. Finally, multi-scale wavelet threshold denoising method is established to reduce the noise of the data in colon cancer gene expression profiles for getting higher classification accuracy. The results show that the extracted informative genes {R87126, L35545, R36977, M36634, H08393} can identify colon cancer with the accuracy of 98.06%.

Keywords: Gene Expression Profiles, Tumor Classification, Informative Genes, Genetic Algorithm, Support Vector Machine.

1. INTRODUCTION

With the development of the large-scale gene expression profiles technology, gene expression data analysis and modeling has become an important topic in bioinformatics area. Each cancer has its own characteristic gene expression profiles. Based on gene expression profiles, accurate identification of tumor subtypes at the molecular level is of great significance for the diagnosis and treatment of cancer.[1] Finding a set of genes tags, i.e., informative genes, which decide the sample class from thousands of genes measured by DNA chip, is the key to indicate the tumor type correctly in order to give a reliable diagnosis and simplify the experimental analysis, and also provide a shortcut for anti-cancer drug development.

Establishing the classification and prediction model of tumor to predict the tumor on the basis of the analysis of tumor gene expression profiles, has become the hot spot of bioinformatics these years. Now people have carried out a certain degree of exploration about recognition of tumor subtype and selection of information gene. Since Golub et al.[2] initiated cancer classification area based on gene expression profiles in 1999, a large number of researchers raised many methods of cancer classification based on gene expression profiles. At the same year, Alon et al.[3] made cluster analysis of gene expression profiles of colon cancer, using T-test method to select genes, obtained a correspondence between some expression profiles and the tumor. Similar to pattern classification, tumor sample classification has two classification methods of supervision and no-supervision. No-supervision classification methods mainly include clustering analysis,[4–5] self-organizing maps (SOM),[6–7] nonnegative matrix factorization (NMF),[8–9] principle component analysis (PCA)[10] and independent component analysis (ICA).[11–13] The advantage is that new tumor subtype may be found to provide research direction for biomedical experts and the disadvantage is that pre-existing sample category information is not used. Supervision classification methods mainly include k-nearest neighbor (k-NN),[14–16] support vector machines (SVMs),[17–19] rough set (RS),[20] artificial neural networks (ANN)[21] and multi-layer perceptron (MLP).[7] The advantage is that sample classification knowledge can be extracted according to pre-known sample category information and these methods have been applied in tumor classification area successfully.

Because the cost of gene chips and other reasons, the number of samples of the obtained data set was less than the number of genes (the dimensionality of the sample), which makes classifying the tumor using machine learning methods more difficult. There are a lot of noise genes among many genes and their presence will reduce the

[*] Author to whom correspondence should be addressed.

1546-1955/2013/10/1097/007 doi:10.1166/jctn.2013.2812

accuracy of classification. Therefore, how to select a small number of informative genes from tens of thousands of genes effectively in order to reduce the dimensionality of the sample has become a critical step in establishing prediction model of classification of tumors.

This paper analyzes colon cancer gene expression profiles data and studies the problem of the selection of classification informative genes. First, considering the characteristics of the high dimensionality and small samples of gene expression profiles data, we use Bhattacharyya distance[22] to reduce dimensionality of the data and support vector machine being as classifier, then established two-class classification model based on GA-SVM by using genetic algorithm (GA) to generate informative genes subsets. The results show that 10 informative genes are needed to get the accuracy of 95.62%. Finally, after reducing the noise of the data in colon cancer gene expression profile, only 5 informative genes are needed to get the accuracy of 98.06% to identify colon cancer.

2. MATERIALS AND METHODS

2.1. Dataset

The experimental data of this paper comes from colon cancer gene expression profiles data sets published by Alon.[3] The data is difficult to analyze compared with other gene chip data. It contains 40 colon tissue samples and 22 normal tissue samples, with each sample containing 2000 gene expression data. The data set is available at http://www.molbio.princeton.edu/colondata.

2.2. Methods

2.2.1. Data Preprocessing

The expression level of some genes in the sample is very close to each other in gene expression profiles and these genes will increase the computational complexity of the search of genetic information instead of providing useful information for the discrimination of sample types. Therefore, we need to remove irrelevant genes.

For the classification of tumor informative gene, we adopt pattern classification methods based on statistical learning, as Figure 1. Classification model mainly contains two basic processes: classification training process and classification test process. Assign the normal samples and tumor samples to the training set and test set at close to 2:1. The training set has 40 samples including 26 colon cancer tissue samples and 14 normal tissue samples and the test set has 22 samples including 14 colon cancer tissue samples and 8 normal tissue samples.

Considering that the data has high dimensionality and small sample characteristics, we need to filter the data and reduce the dimensionality first. The expression level of thousands of measured genes is very different. Only a few genes and samples have a strong correlation,while most genes are not associated with the type of sample, which do not contribute to classification, so these genes should be filtered from the data. Suppose the distribution of the two-class samples is subject to normal distribution. The i gene adopts Bhattacharyya distance as the separability criterion of the two categories, which is called Bhattacharyya feature score criterion (BFSC).

The Bhattacharyya distance between colon cancer samples and normal tissue samples of the i gene can be defined as follows:[22–23]

$$B(i) = \frac{1}{4}\frac{(\mu_+(i) - \mu_-(i))^2}{\sigma_+^2(i) + \sigma_-^2(i)} + \frac{1}{2}\ln\left(\frac{\sigma_+^2(i) + \sigma_-^2(i)}{2\sigma_+(i)\sigma_-(i)}\right) \quad (1)$$

$\mu_+(i)$ and $\sigma_+^2(i)$ refer to the mean and variance of the gene in colon cancer samples (positive samples). $\mu\_(i)$ and $\sigma_-^2(i)$ refer to the mean and variance of the gene in the normal tissue samples (negative sample).

The greater the distance is, the stronger the relationship between gene and colon cancer is. After calculating Bhattacharyya distance, we descend the order of them,

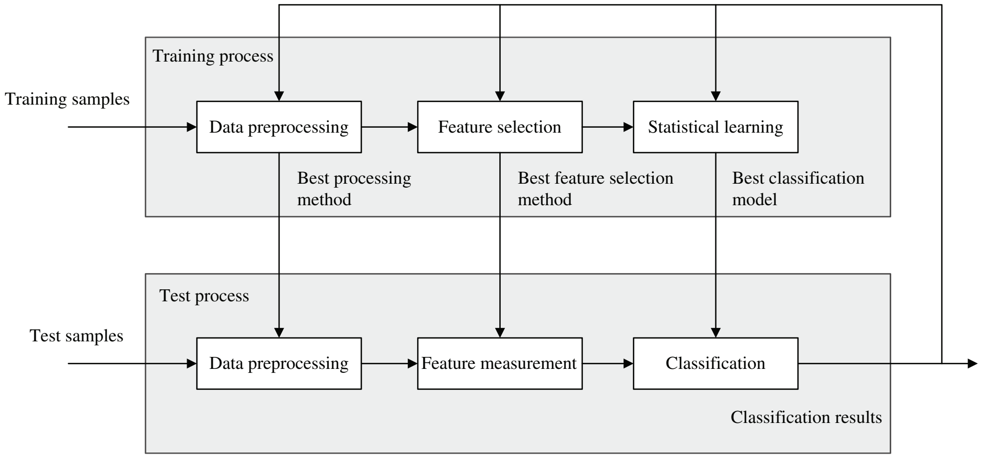

Fig. 1. The pattern classification methods based on statistical learning.

with many quantities of small numerical value genes and few quantities of big numerical value genes. For reliability, we take the first 600 gene expression profiles data as the next selection and classification of the subset of the genes, which reduces the data dimensionality.

2.2.2. Genetic Algorithm and Support Vector Machine

Genetic algorithm, which is first proposed by Professor John H. Holland in the university of Michigan, United States, is random adaptive global search algorithm.[24] Produce a set of string to form the initial population by random and the task of genetic manipulation is imposing a certain operation on individuals of groups according to their degree of adaptation to the environment to achieve the evolutionary process of survival of the fittest. Those chromosomes with higher fitness value are more likely to produce offspring. The offspring are the product of parents and they are represented by a combination of genes from each parent. This process is called "hybrid." If the solutions of the new generation can be sufficiently close or equal to expected value, then the search is ended. Otherwise, the new generation will repeat breeding program of their parents and evolve from generation to generation until a terminating condition is satisfied.

For the parameters of genetic algorithm, this paper uses dynamic parameters. Changes in parameters are associated with the evolution algebra. Early in the population, there is high crossover rate, low mutation rate. Late in the population, there is low crossover rate, high mutation rate. We reduce the possibility of early populations by this way. After several tests the parameters are found as follows:

$$\begin{aligned} P_c &= 1 - \frac{\text{generation}}{\text{Generation} \times 8} \\ P_m &= \frac{\text{generation}}{\text{Generation} \times 16} \end{aligned} \tag{2}$$

At this point, the algorithm convergence can be very good.

Generation value of 10 represents the total evolutionary generations, generation indicates the current generation number, so the range of crossover probability P_c value changes from 0.875 to 1 and the range of mutation probability P_m value changes from 0 to 0.0625.

Support vector machine is a new machine learning methods raised by Vapnik et al.[25] according to statistics theory. According to the principle of structural risk minimization proposed by Vapnik, its greatest feature is to try to improve the generalization ability of learning. It has shown unique advantage in solving the problem of the small sample, nonlinear and high dimensionality pattern recognition. The basic idea of applying support vector machine for classification can be briefly described as: through certain nonlinear relation, map input space samples to a feature space and structure optimization classification hypeplane in the feature space in order to make two-class samples (They can be generalized to multi-class samples) can be divided in this feature space. Feature mapping only relate to input vector with low dimensionalityal and dot product of feature space (the dot product of feature mapping can be replaced by a kernel function), so as to avoid "dimensionality disaster" and solve high dimensionality characteristics problem.

For a given category sample set, assume a hyperplanes $u \cdot x + b = 0$ to separate tumor samples and normal samples. u, x are vectors, $u \cdot x$ is the inner product operation and $2/\|u\|$ is the nearest point distance interval of two kinds. SVM classifier can build such hyperplanes, correctly separate out most of the data points and make the spacing distance between any kind and hyperplanes the largest. The best hyperplanes meet the constraints.

$$\text{Maximize} \quad L(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i y_i \alpha_j y_j K(x_i, x_j) \tag{3}$$

$$\text{subject to} \quad \sum_{i=1}^{m} \alpha_i y_i = 0, \quad \alpha_i \in [0, C], \;\; i = 1, \ldots, m \tag{4}$$

The α_i is Lagrange multiplier, $K(x_i, x_j)$ is the kernel function, C is a regularization parameter. For test sample x, We adopt the decision function $f(x) = \operatorname{sgn}(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x_j) + b)$ and use its plus-minus to determine the category of the sample.

The concrete layout of the kernel function has major effect on the classification performance of SVM. Several typical kernel functions are

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^p \tag{5}$$

$$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2) \tag{6}$$

Equation (5) is the polynomial kernel function of degree p which will revert to the linear kernel function when $p = 1$. Equation (6) is the radial basis function (RBF) kernel with one parameter γ.

SVM is suitable to this tumor data set with the characteristics of high dimensionality and small sample.[17–18] We adopted LIBSVM software package developed by Chang et al.[26] as the classifier.

Through the comparison of different classification results with different kernel functions and different parameters, we select the linear kernel function which can classify the colon cancer data ideally and the linear support vector machine as classifier, defining fitness function for SVM classifier sample classification accuracy, using training set to leave-one-out cross-validation (LOOCV) procedure, identifying samples from a set of test. The calculation process of the proposed GA-SVM two-class classification model as shown in Figure 2. The simulation experiment computing environment: MATLAB programming, Pentium(R) Dual-Core CPU E5300@ 2.60 GHZ, 2.59 GHZ and 2.00 GB RAM.

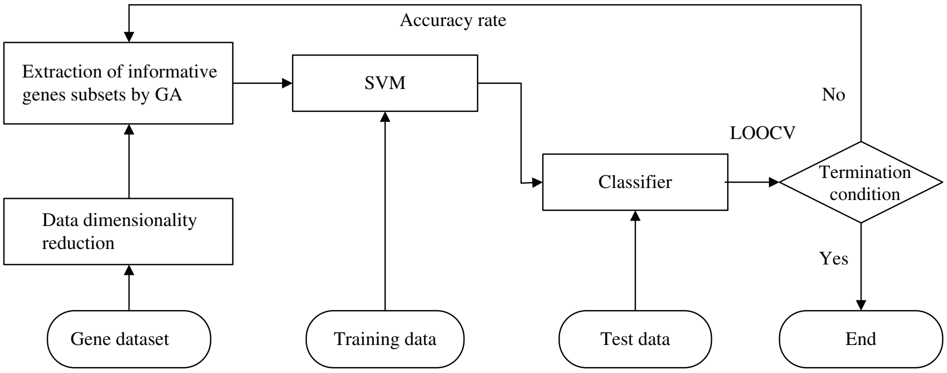

Fig. 2. Flow chart of GA-SVM classification model.

2.2.3. Multi-Scale Wavelet Thresholding

Inevitably gene expression profiles contain noise, some noise intensity is even bigger, deviating when extracting informative genes from gene expression profiles containing noise.[2] Because the randomness and discreteness of noise is similar to the data of gene mutations and genetic changes, some noise makes gene expression level improve. It is easier to search the gene which has close relationship with colon cancer genes. This kind of noise is favorable to the extraction of informative genes. Some noise appear in genetic sequences with relatively high level of gene expression. Because of the interference of the noise, the gene, which was not extracted as informative genes, improve the level of gene expression and cause search error. This kind of noise is obviously unfavorable. Some noise has no effect on expression level of informative genes, because the noise may appear in small amount in the genetic sequence whose expression level is very low originally. Even if there is noise, they won't be searched as informative genes.

Noise in gene expression profiles can be caused by the error of system and the informative genes. But compared to the noise caused by systematic errors, the noise caused by informative genes is more obvious. First, obtain the corresponding value of noise of each gene in different patients, the samples which have larger noise values can be considered causing by informative genes, while the samples which have smaller noise values can be considered causing by systematic errors. And then filter the data. When the noise value is greater than 0.5, then consider the noise is caused by informative genes. If the number of genes with a noise value which is more than 0.5 is more than 20 in the patient's genes, then retain the data, otherwise filter out the corresponding data.

In this paper, we adopted multi-scale wavelet threshold denoising method.[27–29] Through many experiments, we used the important Daubechies (dbN) wavelet series as mother wavelet and selected db10 as the optimal wavelet basis after analyzing all data of the test dataset as well as reconstruct wavelet from five different scale levels.

Process wavelet coefficients by employing penalty threshold, Birge-Massart threshold and hard threshold respectively. Experiments show maximum root mean square (RMS) and minimum error of penalty threshold function and the best effect of reducing noise at the scale level 4. We employed the method of wavelet thresholding for noise reduction to get the 389 gene expression data.

Table I. A brief description of selected genes based on GA-SVM algorithm.

Informative gene number	Gene ID	Gene description
49	T61661	PROFILIN I (HUMAN).
227	M64445	Human GM-CSF receptor mRNA, complete cds.
237	T50334	14-3-3-LIKE PROTEIN GF14 OMEGA (Arabidopsis thaliana).
493	R87126	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus).
869	L35545	Homo sapiens endothelial cell protein C/APC receptor (EPCR) mRNA, complete cds.
1042	R36977	P03001 TRANSCRIPTION FACTOR IIIA.
1048	X01060	Human mRNA for transferrin receptor.
1635	M36634	Human vasoactive intestinal peptide (VIP) mRNA, complete cds.
1771	J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds.
1772	H08393	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens).

3. RESULTS AND DISCUSSION

In this paper, we first reduce sample data dimensionality by using BFSC and take the first 600 gene expression

Table II. Classification accuracy of different classification function.

Classification function	Classification accuracy (%)
Linear kernel ($p = 1$)	95.62
Polynomial kernel ($p = 2$)	91.57
RBF kernel	93.36

Table III. A brief description of selected genes based on GA-SVM algorithm.

Informative gene number	Gene ID	Gene description
493	R87126	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus).
869	L35545	Homo sapiens endothelial cell protein C/APC receptor (EPCR) mRNA, complete cds.
1042	R36977	P03001 TRANSCRIPTION FACTOR IIIA.
1635	M36634	Human vasoactive intestinal peptide (VIP) mRNA, complete cds.
1772	H08393	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens).

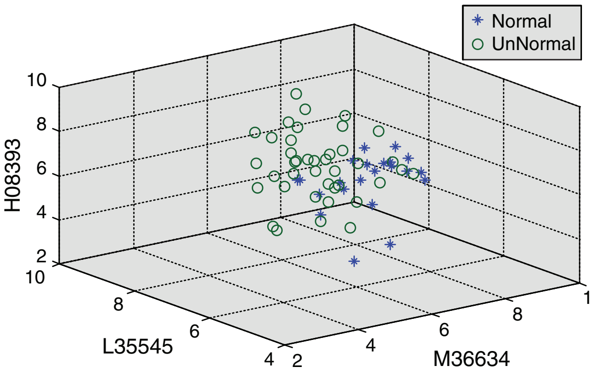

Fig. 5. Three-dimensional scatter plot composed by three genes of colon cancer data set {M36634, L35545, H08393}.

data as the next selection and classification of the subset of the genes. Use variable 0–1 showing whether we take the gene as informative genes. A genetic algorithm is used to experiment on 600 genes samples, run 5 generations, 10 times in the colon cancer training set and get the statistics of various genes appearing frequency in subsets. Because genetic algorithm is random method, we can't guarantee every time operation can get the same gene subset. Therefore, we operate algorithm 10 times and each time we can select the gene subsets from 600 genes. With the data of these genes subsets, we structure classification model and examines the classification ability of GA-SVM model in the test set. In the 22 test samples, one of the samples can't be correctly classified by the model. After many experiments, we found that informative genes only need 10 genes at least, as shown in Table I. Support vector machine being as classifier, three different types of kernel functions (linear, polynomial, and radial basis) are included in the algorithm, we get classification accuracy through calculation as shown in Table II.

From Table I, when we classify colon cancer gene by using GA-SVM algorithm for classification and the linear SVM as classifier, the extracted informative gene codes is {T61661, M64445, T50334, R87126, L35545, R36977, X01060, M36634, J05032, H08393}. It can achieve a higher prediction precision and accuracy rate is 95.62%.

Due to the noise existing in the gene expression profiles, in this paper, we use multi-scale wavelet threshold denoising method, adopting a commonly used Daubechies (dbN) wavelet series as mother wavelet. Through the analysis of the data of test set, we choose db10 as the optimal wavelet basis and reconstruct wavelet from five different scale levels. Then we get 389 gene expression profiles after noise reduction, we input the microarray data into

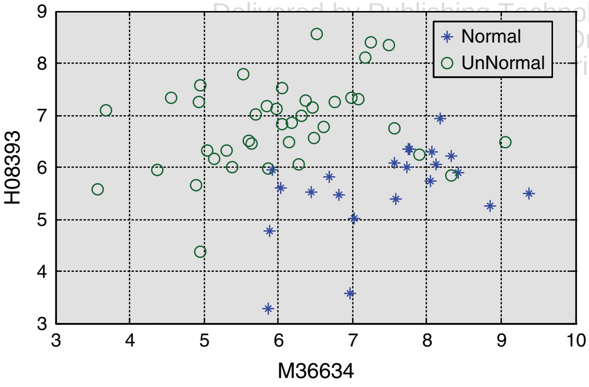

Fig. 3. Two-dimensional scatter plot composed by two genes of colon cancer data set {M36634, H08393}.

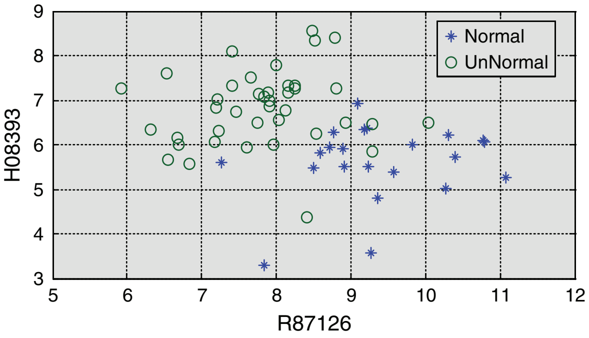

Fig. 4. Two-dimensional scatter plot composed by two genes of colon cancer data set {R87126, H08393}.

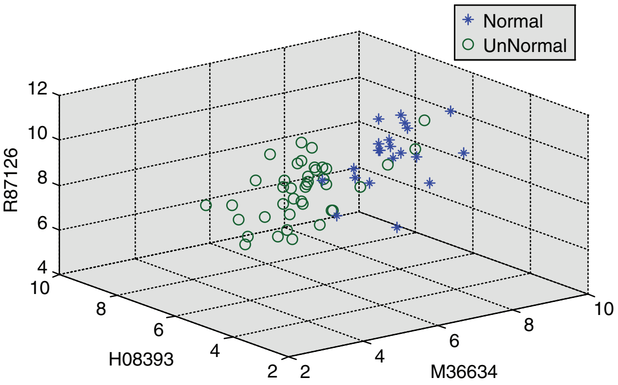

Fig. 6. Three-dimensional scatter plot composed by three genes of colon cancer data set {M36634, H08393, R87126}.

Table IV. The comparison of classification results by using different classification methods.

No.	Methods of gene selection or feature extraction	Classifier	Sample set	The quantity of informative genes	Accuracy (%)	Reference
1	Primary gene selection by BFSC, then refined selection by dynamic genetic algorithm	SVM (linear kernel)	Colon	5	98.06	This paper
2	Feature score criterion	SVM (linear kernel)	Colon	—	90.30	[17]
3	Primary gene selection by Relief-F, then refined selection by HykGene	Naïve Bayes	Colon	3	91.90	[33]
4	Genetic algorithms	SVM (polynomial kernel)	Colon	12	93.55	[34]
5	Revised locally linear embedding	SVM (RBF kernel)	Colon	—	91.00	[35]
6	Multi-objective evolutionary algorithms	The discrimination of gene expression difference	Colon	7	97.00	[36]
7	RFSC and factor analysis	SVM (RBF kernel)	Colon	—	91.94	[37]
8	Partial least square	Logistic discrimination	Colon	—	93.50	[38]

GA-SVM tumor classification model. After many experiments, we finally get five informative genes and at this time the classification accuracy is 98.06%. The results of informative genes are as shown in Table III.

From the results we can see that when the informative genes reduced 5 after denoising, the accuracy rate increased by 2.44%, which shows that the noise existing in the colon cancer gene expression profiles has a detrimental effect on determining informative gene. Figures 3 and 4 are two-dimensional scatter plots composed by the subset of the genes {M36634, H08393} and {R87126, H08393} selected from informative genes at random. Figures 5 and 6 are three-dimensional scatter plots composed by the subset of the genes {M36634, L35545, H08393} and {M36634, H08393, R87126} selected from informative genes at random. Apparently, the demarcation between tumor samples and normal samples of Figures 4 and 5 is more blurred, which shows that the corresponding information can not do high-precision distinguish on cancer well. But the demarcation between the two-class samples of Figures 3 and 6 is clearer, which shows the corresponding information can do high-precision distinguish on cancer well. For those high-resolution gene scatters, we consider the corresponding subset of the genes can be used as a potential subset of informative genes.

We can find from the biomedical literature that gene M36634 (vasoactive intestinal peptide, VIP) are closely related to colon cancer. VIP can promote the value-added of colon cancer cell lines Lovo and HT29 and others, the growth of colon cancer induced by AOM, and increase activity of ornithine decarboxylase of cancer cells and ODC_Mrna.[30–31] At the same time, R36977 has something to do with the formation of tumor and the regulation of cell cycle.[32] Gene L35545 is a gene related to the long arm APC of chromosome 5. H08393 is gene markers related to colon cancer and was applied for the United States' patent in 2005, the patent NO. 20050165556. From these we can see the obtained informative genes have large reference value.

It is because test methods of tumor classification based on the gene expression profiles is expected to be used in medicine clinical diagnosis that tumor classification problems are widely researched. Colon cancer samples use different methods of gene selection or feature extraction and the comparison of classification experimental results are listed in Table IV. Now these are very good experimental results for tumor classification problem. By comparison, our experimental result has obvious superiority in quantity and classification performance of informative genes. It also showed that algorithms we designed have strong competitive advantage over other gene selection methods. Its biggest advantage is that it can be able to find the optimal informative genes subset and this not only creates the conditions for the development process of analyaing tumors from the angle of metabolic pathways but also prepares conditions for developing tumor function classification gene chip.

4. CONCLUSION

That the emergence of tens of thousands of gene expression of DNA microarray technology can be measured by an experiment simultaneously provides a new research method for oncology research and holds much attention to medical pharmaceutical and clinical application area. So the method of tumor molecular diagnosis based on the gene expression profiles has gradually formed one of the important research areas in bioinformatics. But because of the characteristics of high dimensionality, much noise, the small sample set and the phenomenon of common expression among genes, it has brought great difficulties in mining biological medical knowledge contained in gene expression profiles deep and accurately and make the selection of informative genes or feature extraction become the most difficult problem for cancer classification.

In this paper, we analyze the gene expression profiles data of colon cancer and study the extraction problem of informative genes. We do primary gene selection by BFSC and then refined selection by dynamic genetic algorithm. Support vector machine being as classifier, we build two-class classification model of colon cancer based on GA-SVM. The experimental results show that only 10 informative genes are needed to get the classification accuracy of 95.62%. Due to the noise existing in the gene expression profiles, we use multi-scale wavelet threshold denoising method to reduce the noise and informative genes reduced 5 after denoising. Only 5 informative genes are needed to get the classification accuracy of 98.06%, which shows that the noise existing in the colon cancer gene expression profiles has a detrimental effect on determining informative gene.

The present study is just to find genes related to cancer as much as possible and take these informative genes as a diagnosis of cancer subtype. But the appearance and development of tumor is a very complicated system as life system, so we can not be able to find the occurrence and development of tumor mechanism only by researching the genes related to the single tumor and must study tumor by the systematic and integral approach. Because tumor classification model based on the gene expression profiles is valuable in application and it is hopeful to apply to medicine clinical practice and medical pharmaceutical and other areas, we need to further analyze the expression regulation mechanism of informative genes, verify prediction model, mine the common law of multiple tumors classification model and discuss the technological possibility of developing functional classification gene chip according to informative genes of tumor, its success will bring great convenience for the early clinical diagnosis of tumor.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (No. 30871341), and Scientific Research Foundation for the Excellent Youth Scholars of Shandong Province of China (No. BS2012DX009).

References

1. E. S. Lander and R. A. Weinberg, *Science* 287, 1777 (**2000**).
2. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Dowing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, *Science* 286, 531 (**1999**).
3. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, *Proc. Natl. Acad. Sci. USA* 96, 6750 (**1999**).
4. H. P. Zhang, C. Y. Yu, B. Singer, and M. M. Xiong, *PNAS* 98, 6730 (**2001**).
5. K. Simek, K. Fujarewicz, A. Swierniak, M. Kimmel, B. Jarzab, M. Wiench, and J. Rzeszowska, *Eng. Appl. Artificial Intelligenee* 17, 417 (**2004**).
6. P. Toronen, M. Kolehmainen, G. Wong, and E. Castren, *FEBS Lett.* 451, 42 (**1999**).
7. S. B. Cho and H. H. Won, Machine learning in DNA mieroarray analysis for cancer classification, *Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics* 189 (**2003**).
8. J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, *Proc. Natl. Acad. Sci. USA* 101, 4164 (**2004**).
9. C. H. Zheng, D. S. Huang, L. Zhang, and X. Z. Kong, *IEEE Trans. Information Technol. Biomed.* 13, 599 (**2009**).
10. J. F. Pinto da Costa, H. Alonso, and L. Roque, *IEEE/ACM Trans. Computat. Biol. Bioinformatics* 8, 246 (**2011**).
11. X. W. Zhang, Y. L. Yap, D. Wei, F. Chen, and A. Danehin, *Eur. J. Human Genetics* 13, 1303 (**2005**).
12. D. S. Huang and C. H. Zheng, *Bioinformatics* 22, 1855 (**2006**).
13. H. Henry and X. L. Li, *BMC Bioinformatics* 12, s7 (**2011**).
14. L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, *Bioinformatics* 17, 1131 (**2001**).
15. S. Dudoit, J. Fridlyand, and T. P. Speed, *J. American Stat. Assoc.* 97, 77 (**2002**).
16. J. W. Lee, J. B. Lee, M. Park, and S. H. Song, *Comput. Stat. Data Anal.* 48, 869 (**2005**).
17. T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, *Bioinformatics* 16, 906 (**2000**).
18. M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr., and D. Haussler, *Proc. Natl. Acad. Sci. USA* 97, 262 (**2000**).
19. I. Guyon, J. Wston, and S. Barnhill, *Machine Learn.* 46, 389 (**2002**).
20. H. Midelfart, J. Komorowski, K. Norsett, F. Yadetie, A. K. Sandvik, and A. Lægreid, *Fundamenta Informaticae* 53, 155 (**2002**).
21. J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Atonescu, C. Peterson, and P. S. Meltzer, *Nat. Med.* 7, 673 (**2001**).
22. O. R. Duda, P. E. Hart, and G. D. Stork, Pattern Classification, 2nd edn., John Wiley & Sons, New York (**2001**), p. 46.
23. S. Theodoridis and K. Koutroumbas, Patter Recognition, 2nd edn., Academic Press, New York (**2003**), p. 177.
24. J. H. Holland, *Sci. American* 44 (**1992**).
25. V. N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag New York Inc., New York (**1995**).
26. C. C. Chang and C. J. Lin, LIBSVM: A library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, (**2001**).
27. S. G. Chang, B. Yu, and M. Vetterli, *IEEE Trans. Image Processing* 9, 1532 (**2000**).
28. F. Luisier, T. Blu, and M. Unser, *IEEE Trans. Image Process* 16, 593 (**2007**).
29. S. G. Chang, B. Yu, and M. Vetterli, *IEEE Trans. Image Process* 9, 1522 (**2000**).
30. D. Yu, P. K. Seitz, P. Selvanayagam, S. Rajaraman, C. M. Townsend Jr., and C. W. Copper, *Endocrinology* 131, 1188 (**1992**).
31. M. Tatsuta, H. Iishi, M Baba, R. Yamamoto, H. Uehara, and A. Nakaizumi, *Cancer Lett.* 93, 219 (**1995**).
32. H. Y. Chuang, H. K. Tsai, and Y. F. Tsai, *J. Information Sci. Eng.* 19, 953 (**2003**).
33. Y. Wang, F. Makedon, J. C. Ford, and J. D. Pearlman, *Bioinformatics* 21, 1530 (**2005**).
34. S. Peng, Q. Xu, X. B. Ling, X. Peng, W. Du, and L. Chen, *FEBS Letter* 555, 358 (**2003**).
35. S. Chao, and H. C. Li, *Proc. APBC* 211 (**2005**).
36. D. Kalyanmoy and A. R. Reddy, *BioSystems* 72, 111 (**2003**).
37. S. L. Wang, J. Wang, H. W. Chen, and W. S. Tang, The Classification of Tumor Using Gene Expression Profile Based on Support Vector Machines and Factor Analysis, Intelligent Systems Design and Applications, IEEE Computer Society Press, Jinan, China (**2006**), Vol. 2, p. 471.
38. D. V. Nguyen and D. M. Rocke, *Bioinformatics* 18, 39 (**2002**).

Received: 14 March 2012. Accepted: 29 April 2012.