Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2015; 1(1-2): 16-21.

Original Article

Predicting the categories of colon cancer using microarray data and nearest shrunken centroid

Mehri Khoshhali[1], Azam Moslemi[1], Massoud Saidijam[2], Jalal Poorolajal[3], Hossein Mahjub[3*]

[1] Department of Biostatistics & Epidemiology, Hamadan University of Medical Sciences, Hamadan, Iran
[2] Department of Genetics and Molecular Medicine and Research Center for Molecular Medicine, Hamadan University of Medical Sciences, Hamadan, Iran
[3] Department of Biostatistics & Epidemiology and Research Center for Health Sciences, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran

ARTICLE INFO

Received 07.04.2014
Revised 21.07.2014
Accepted 07.09.2014
Published 20.2.2015

Available online at:
http://jbe.tums.ac.ir

Key words:
colon cancer,
gene expression,
microarray,
prediction,
nearest shrunken centroid

ABSTRACT

Background & Aim: It is very helpful to classify and predict the clinical category of a sample based on its gene expression profile. This study was conducted to predict tissues of colorectal adenoma, adenocarcinoma, and paired normal in colon based on microarray data using nearest shrunken centroid method.

Methods & Materials: In this study, the colon cancer dataset were used including, 18 adenocarcinoma, 4 colorectal adenoma, and 22 paired normal colon samples with 2360 common gene expression measurements. In order to predict categories of colon cancer was used nearest shrunken centroid method. R software was used for data analysis.

Results: Based on our findings, performance of nearest shrunken centroid method was successful to reduce 2360 genes to a set of eleven genes containing rig, BIGH3, GLI3, Homo sapiens guanylin, p78, 54KDa, XBP-1, CO-029, desmin, MLC-2, and HMG-1. This method predicted three classes. It predicted two classes-colorectal adenoma and adenocarcinoma with error of zero and normal class with error of 4.5%.

Conclusion: Nearest shrunken centroid method succeeded to reduce several 1000 genes to 11 genes that were able to characterize colon tissue samples to one of the three classes of adenocarcinoma, colorectal adenoma and normal with 97.7% accuracy.

Introduction

Cancers of colon and rectum are not common in developing countries, however are the second most prevalent malignancy in wealthy communities. More than 940,000 cases happen annually worldwide, and about 500,000 die from it each year (1). A major reason is a diet rich in fat, refined carbohydrates and animal protein, combined with low physical activity. Genetic susceptibility appears to be involved in less than five percent of cases. Epidemiological studies suggest that risk can be reduced by decreasing meat consumption, particularly processed meat, and increasing the intake of vegetables and fruit (1).

The recent development of microarray technologies to investigate gene expression in model organisms, cell lines, and human tissues has become an important part of biological research over the last several years. Microarrays allow the study of expression levels of thousands genes simultaneously in a given cell type (2).

Different statistical methods applied for analysis of microarray data such as selection of

* Corresponding Author: Hossein Mahjub, Postal Address: Department of Biostatistics and Epidemiology, School of Public Health, Hamadan University of Medical Sciences, P.O. Box 65715-4171, Hamadan, Iran Email: mahjub@umsha.ac.ir

Please cite this article in press as: Khoshhali M, Moslemi A, Saidijam M, et al. Predicting the categories of colon cancer using microarray data and nearest shrunken centroid. *J Biostat Epidemiol. 2015; 1(1-2): 16-21*

genes, clustering of genes with similar expression profiles and classification (3).

Recent advances in microarray technology have enabled researchers to reproduce gene expression in order to predict and classify specific subgroups of high- and low-risk patients (4). The problem of classification using microarray is challenging because there are a large number of variables (genes) but a much smaller number of samples. Hence, detecting relevant genes that distinguish samples is also greatly favorable in practice (5, 6).

There is a distinction between classification and clustering. If the classes are pre-existing, then classification analysis is more appropriate than clustering analysis (3). Many classification methods have been used for microarray data including support vector machines (7) linear discriminant analysis (8), random forests (9), neural networks (10), generalized models, nearest centroid (11), and nearest shrunken centroid (5).

Tibshirani et al. proposed the nearest shrunken centroid method for class prediction in DNA-microarray studies. In this method, shrunken centroids are used as prototypes for each class and are determined subsets of genes that best characterize each class (5). The nearest shrunken centroid could make the classifier more accurate by reducing the effect of noisy genes, and it performs automatic gene selection (12). Furthermore, this method is desired because it is easy to implement and interpret, and it could apply when there are more than two classes (5, 6).

In recent years, nearest shrunken centroid was used to diagnosis of multiple cancer types based on gene expression (5, 6, 11, 13).

In this study, we used the colon cancer data, which was previously analyzed by Notterman et al. (9). They performed clustering method to classify colon samples (14).

Since in colon cancer data, classes are already specified, we have applied nearest shrunken centroid to classify and predict categories for colon tissue samples based on gene expression data. This method also identified subsets of genes, which are the best description of each class.

Methods

In this study, we used gene expression data of colon cancer (14). These data consists of two dataset. Dataset of adenocarcinoma consists of eighteen adenocarcinoma samples and paired normal tissue with 7457 genes. The colorectal adenoma dataset consists of four colorectal adenoma samples and paired normal tissue with 7087 genes. Data are available at: http://genomics-pubs.princeton.edu/oncology/.

A computer program was written to distinguish the common genes into both data sets. Accordingly, 2360 genes for 44 samples that containing eighteen adenoma carcinoma, four colorectal adenoma and twenty two their paired normal samples in colon were obtained.

Nearest shrunken centroid method was applied to classify and predict the categories of colon tissue samples based on common gene expression data.

Assuming there are *n* samples, and each sample including *p* gene expressions. Also each sample belongs to one class and the size sample in class *k* is n_k.

In nearest shrunken centroid method, the class centroid $\overline{\mathrm{X}}_{\mathrm{ik}}$ for gene *i* and class *k* is compared to the overall centroid $\overline{\mathrm{X}}_{\mathrm{i}}$ by

$$\overline{\mathrm{X}}_{\mathrm{ik}} = \overline{\mathrm{X}}_{\mathrm{i}} + \mathrm{m_k}(\mathrm{s_i} + \mathrm{s_0})\mathrm{d_{ik}} \quad (1)$$

where s_i is the pooled within-class standard deviation of gene *i* and s_o is an offset to guard against genes with low expression levels. In Eq. (1), $\mathrm{m_k}$ is $\sqrt{\frac{1}{\mathrm{n_k}} - \frac{1}{\mathrm{n}}}$ and d_{ik} is shrinkage of soft threshold. Each $\mathrm{d_{ik}}$ is reduced by a value of threshold (Δ) in absolute value, until it reaches zero. Genes with $\mathrm{d_{ik}} = 0$ for all classes do not contribute to the classification and it can be removed. Thus, the number of genes that will be used in predicting the classes varies as the threshold parameter Δ changes. When Δ becomes larger, fewer genes participate to the classification and only the most significant genes as discriminating features use among different classes. Value of threshold Δ is determined by cross-validation (5, 6).

To predict the class for a new sample with expression levels $\mathrm{x}^* = (\mathrm{x}_1^*, \mathrm{x}_2^*, \dots, \mathrm{x}_\mathrm{p}^*)$ for genes that survived at the given threshold. The discriminant score that is defined for class k, is

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - x'_{ik})}{(s_i + s_0)^2} - 2\log\pi_k$$

The first part of the equation is the standardized squared distance of x* to the k^th^ shrunken centroid. The second part is a correction based on the class prior probability π_k, where their sum is equal one. A sample belongs to the class that has the lowest score among all classes.

Also, like Gaussian linear discriminant analysis, it is possible to calculate estimates of the class probabilities using the discriminant score as:

$$\hat{p}_k(x^*) = \frac{e^{-\frac{1}{2}\delta_k(x^*)}}{\sum_{i=1}^{k} e^{-\frac{1}{2}\delta_i(x^*)}}$$

Thus, a sample belongs to the class that has a higher probability (5, 6).

We performed nearest shrunken centroid method to classify and predict the categories in the colon data. Also, cross-validation was used to assess the misclassification error (5, 6).

For data analysis, we used the package Pamr in R (version 12.2.2, R Development Core Team, 2011) environment that can be downloaded from: http://cran.r-project.org/web/packages/PAM.

Results

Figure 1 shows misclassification error and number of genes for thresholds ranging from 0 to 14. The values of different threshold will result different misclassification error and various number of genes.

If the threshold is equal to 6.390, then 33 genes were selected with overall error 9% and threshold 6.882 gives overall error 9% with 17 genes. Also, with the value of the threshold 7.865, four genes were selected with overall error 3%.

The value of the threshold was selected to minimize misclassification error in each three classes. Therefore, the threshold value of 7.373 was chosen and eleven genes survive at this value using nearest shrunken centroid method. This threshold value is specified with an arrow in figure 1.

Table 1 shows the list of selected genes and their standardized centroids for each class. Zero value in each class expresses that the mean expression gene value in that class is equal to the overall mean for that gene. Based on the nearest shrunken centroid method, the genes that their standardized centroids were zero for all classes then they were eliminated from classifier.

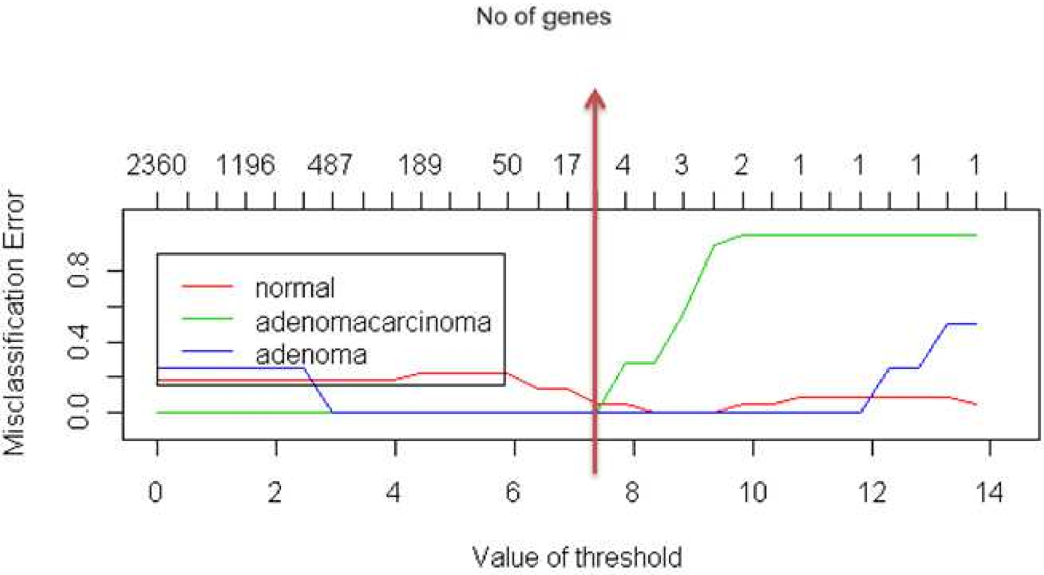

Figure 1. Misclassification error and number of genes for values of different threshold. Red arrow specifies the threshold value that was specified in this study

Table 1. List of the genes that survive at the threshold 7.373 from the nearest shrunken centroid classifier with standardized centroids for each class

Gene name	Description	Standardized centroid class			Reference number
M32405	Human homologue of rat insulinoma gene (rig) exons 4-Jan	0.0000	0.0000	3.2809	(15)
M77349	Human transforming growth factor-beta induced gene product (BIGH3) mRNA, complete cds	0.0000	0.0000	1.4829	(24)
M57609	Human DNA-binding protein (GLI3) mRNA, complete cds	0.0000	0.0000	−0.3686	(17)
M97496	(Homo sapiens guanylin) mRNA, complete cds	0.3286	−0.0551	0	(25)
M33882	Human (p78) protein mRNA, complete cds	0.0000	0.0000	−0.2116	(26)
U02493	Human (54 kDa) protein mRNA, complete cds	0.0000	0.0000	0.0979	(18)
M31627	Human X box binding protein-1 (XBP-1) mRNA, complete cds	0.0000	0.0000	0.093	(19)
M35252	Human (CO-029)	0.0000	0.0000	0.0869	(20)
M63391	Human (desmin) gene, complete cds	0.0649	0.0000	0.0000	(27)
J02854	Human 20-kDa myosin light chain (MLC-2) mRNA, complete cds	0.0356	0.0000	0.0000	(28)
D63874	Human mRNA for (HMG-1), complete cds	0.0000	0.0166	0.0000	(21)

Table 2. A cross-tabulation of the true versus predicted values, from a nearest shrunken centroid fit

Predicted	Normal	Adenocarcinoma	Adenoma	Class error (%)
Normal	21	00	1	4.5
Adenocarcinoma	00	18	0	0.0
Adenoma	00	00	4	0.0
Overall error percent				2.3

The nonzero value of standardized class centroids for each gene determines that the gene contribute to the nearest-centroid computation. The selected genes for normal class are Homo sapiens guanylin, desmin, and MLC-2. For adenocarcinoma class the selected genes are Homo sapiens guanylin and HMG-1. Furthermore, for colorectal adenoma class the genes are rig, BIGH3, GLI3, p78, 54kDa, XBP-1, and CO-029. If the absolute value of standardized class centroids of each gene is high, it indicates the gene is important for classification. The most important selected genes for classification are rig, BIGH3, GLI3, and Homo sapiens guanylin, respectively.

Based on table 1, Homo sapiens guanylin gene was more highly expressed in normal tissue than in the adenocarcinoma.

As was noted, this dataset was used to predict the three classes using nearest shrunken centroid approach as it is shown in table 2. The table shows that the used method correctly predicted the two classes of colorectal adenoma and adenocarcinoma with error of zero. The error rate for classification of normal tissue was obtained 4.5% in which one normal tissue was incorrectly classified as adenoma tissue. Also the total error was 2.3%.

Discussion

In this study, nearest shrunken centroid method was successful in reducing 2360 genes to a set of eleven genes containing rig, BIGH3, GLI3, Homo sapiens guanylin, p78, 54KDa, XBP-1, CO-029, desmin, MLC-2, and HMG-1 for predicting three classes with total error of 2.3%.

Based on the previous biological studies, seven genes of 11 selected genes were concerning cancer tumors. Shiga et al. showed that rig was activated in different human tumors such as esophageal cancers and colon cancers (14). Also, Skonier et al. isolated BIGH3 from a human lung adenocarcinoma cell line (15). Ruppert et al. expressed that GLI gene amplification in primary human tumors was important in the neoplastic process (16). Ma et al. expressed that gene 54 kDa is post-transcriptionally regulated protein in human breast tumors leading to reduced expression in estrogen receptor (ER)− tumors and the expression of an amino terminal altered isoform in a subset of ER+ tumors (17). Hsiao et al. demonstrated XBP-1 significantly correlates with LMP1 expression in Nasopharyngeal carcinoma tumor biopsies (18). Sela et al. showed that the CO-029 antigen expressed on gastric, colon, rectal, and pancreatic carcinomas, but not on most normal tissues (19). Also, Xiang et al. found new information on *HMG-1* mRNA expression in the human gastrointestinal cancer and suggested a correlation between *FM1* mRNA expression to the differentiation and the stage of human gastrointestinal adenocarcinoma (20).

The mentioned present method was also applied performed on these seven genes for predicting the classes of colorectal adenoma and normal with an error rate of zero percent. Also, for classification of adenocarcinoma the error rate was 22%. The total error rate for these genes was achieved equal to 9.1%.

Notterman et al. applied clustering algorithm for this data set of colon cancer and successfully distinguished normal tissue, adenocarcinoma and adenoma. In their work, two genes rig and XBP-1 set in adenoma cluster and gene HMG-1 set in adenocarcinoma cluster. Also they demonstrated that gene Homo sapiens guanylin was more highly expressed in normal tissue than in the adenocarcinoma (13). Also, finding of this study were the same for these genes.

Jaeger et al. applied three methods including Correlation, Clustering and Masked out Clustering for classification adenocarcinoma dataset. They selected seven genes containing S17, S24, PPH alpha, HKSP, Gaba, CLPP, and ASB. Also, they predicted classes of adenocarcinoma and normal with an error rate of zero percent (21). These genes were dissimilar with those genes that we obtained in this study. The difference may be the result of using different methods of gene selection.

Park et al. proposed hierarchical nearest shrunken centroid classifier when there are no pre-defined classes. This Algorithm uses fewer genes in prediction than the regular nearest shrunken centroid classifier. They show that this method performs as well as the nearest shrunken centroid method, but provides additional useful information (22).

The neural network approach as a form of discriminant analysis is used to classify cancers based on their gene expressions. It is more complex than nearest shrunken centroid. When there are many genes and few samples, it is desired that used simpler statistical methods will perform as well as or better than neural networks (5).

Conclusion

In this study, nearest shrunken centroid approach succeeded that reduces several thousand genes to eleven genes that was able to characterize colon tissue samples to one of three classes adenoma carcinoma, colorectal adenoma and normal with a high accuracy.

Acknowledgments

This article was a part of research work in the Hamadan University of Medical Sciences. We would like to thank Research Center for Health Sciences and Faculty of Health for their valuable collaboration with this study.

References

1. World Health Organization. Global cancer rates could increase by 50% to 15 million by 2020. Geneva, Switzerland: WHO; 2003.
2. Lee JW, Lee JB, Park M, Song SH. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis 2005; 48(4): 869-85.
3. Hou J, Aerts J, den HB, van IW, den BM, Riegman P, et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS One 2010; 5(4): e10312.
4. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002; 99(10): 6567-72.
5. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science 2003; 18(1): 104-17.
6. Wang L, Zhu J, Zou H. Hybrid huberized support vector machines for microarray classification and gene selection. Bioinformatics 2008; 24(3): 412-9.
7. Xu P, Brock GN, Parrish RS. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. Computational Statistics & Data Analysis 2009; 53(5): 1674-87.

8. Stokowy T. Classification of DNA microarray data with random forests. Advances in Intelligent and Soft Computing 2010; 69, 2010: 305-8.
9. Soares C, Montgomery L, Rouse K, Gilbert JE. Automating microarray classification using general regression neural networks. Proceedings of 4th International Conference on Machine Learning and Applications; 2008 Dec 11-13; San Diego, CA: IEEE Computer Society; 2008. p. 508-13.
10. Dabney AR. Classification of microarrays to nearest centroids. Bioinformatics 2005; 21(22): 4148-54.
11. Suarez-Farinas M, Shah KR, Haider AS, Krueger JG, Lowes MA. Personalized medicine in psoriasis: developing a genomic classifier to predict histological response to Alefacept. BMC Dermatol 2010; 10: 1.
12. Wang S, Zhu J. Improved centroids estimation for the nearest shrunken centroid classifier. Bioinformatics 2007; 23(8): 972-9.
13. Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res 2001; 61(7): 3124-30.
14. Shiga K, Yamamoto H, Okamoto H. Isolation and characterization of the human homologue of rig and its pseudogenes: the functional gene has features characteristic of housekeeping genes. Proc Natl Acad Sci U S A 1990; 87(9): 3594-8.
15. Skonier J, Neubauer M, Madisen L, Bennett K, Plowman GD, Purchio AF. cDNA cloning and sequence analysis of beta ig-h3, a novel gene induced in a human adenocarcinoma cell line after treatment with transforming growth factor-beta. DNA Cell Biol 1992; 11(7): 511-22.
16. Ruppert JM, Vogelstein B, Arheden K, Kinzler KW. GLI3 encodes a 190-kilodalton protein with multiple regions of GLI similarity. Mol Cell Biol 1990; 10(10): 5408-15.
17. Ma S, Song X, Huang J. Supervised group Lasso with applications to microarray data analysis. BMC Bioinformatics 2007; 8: 60.
18. Hsiao JR, Chang KC, Chen CW, Wu SY, Su IJ, Hsu MC, et al. Endoplasmic reticulum stress triggers XBP-1-mediated up-regulation of an EBV oncoprotein in nasopharyngeal carcinoma. Cancer Res 2009; 69(10): 4461-7.
19. Szala S, Kasai Y, Steplewski Z, Rodeck U, Koprowski H, Linnenbach AJ. Molecular cloning of cDNA for the human tumor-associated antigen CO-029 and identification of related transmembrane antigens. Proc Natl Acad Sci USA 1990; 87(17): 6833-7.
20. Xiang YY, Wang DY, Tanaka M, Suzuki M, Kiyokawa E, Igarashi H, et al. Expression of high-mobility group-1 mRNA in human gastrointestinal adenocarcinoma and corresponding non-cancerous mucosa. Int J Cancer 1997; 74(1): 1-6.
21. Jaeger J, Sengupta R, Ruzzo WL. Improved gene selection for classification of microarrays. Pac Symp Biocomput 2003; 53-64.
22. Park MY, Hastie T. Hierarchical classification using shrunken centroids. Stanford. CA: Department of Statistics, Stanford University; 2005 [Online]. [cited 2005]; Available from: URL: http://www-statstanfordedu/~hastie/Papers/hpampdf.