Health Promotion Perspectives, 2013, 3(1), 129-134
doi: 10.5681/hpp.2013.015
http://journals.tbzmed.ac.ir/HPP

A Comparison of Selective Classification Methods in DNA Microarray Data of Cancer: Some Recommendations for Application in Health Promotion

Tohid Jafari Koshki [1], *Ebrahim Hajizadeh [1], Mehrdad Karimi [2]

[1] *Department of Biostatistics, School of Medical Sciences, Tarbiat Modares University, Tehran, Iran*
[2] *Department of Biostatistics and Epidemiology, School of Health, Tehran University of Medical Sciences, Tehran, Iran*

ARTICLE INFO

Article type:
Original Article

Article history:
Received: Oct 22 2012
Accepted: Jan 28 2013
e-published: Jun 30 2013

Keywords:
Breast Cancer,
Classification,
DNA Microarray Analysis,
Marker-gene

***Corresponding Author:**
Ebrahim Hajizadeh
Tel: +98 21 82880;
e-mail:
hajizadeh@modares.ac.ir

ABSTRACT

Background: The aim of this study was to apply a new method for selecting a few genes, out of thousands, as plausible markers of a disease.

Methods: Hierarchical clustering technique was used along with Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers to select marker-genes of three types of breast cancer. In this method, at each step, one subject is left out and the algorithm iteratively selects some clusters of genes from the remainder of subjects and selects a representative gene from each cluster. Then, classifiers are constructed based on these genes and the accuracy of each classifier to predict the class of left-out subject is recorded. The classifier with higher precision is considered superior.

Results: Combining classification techniques with clustering method resulted in fewer genes with high degree of statistical precision. Although all classifiers selected a few genes from pre-determined highly ranked genes, the precision did not decrease. SVM precision was 100% with 22 genes instead of 50 genes while the NB resulted in higher precision of 97.95% in this case. When 20 highly ranked genes selected to be fed to the algorithm, same precision was obtained using 6 and 5 genes with SVM and NB classifiers respectively.

Conclusion: Using hybrid method could be effective in choosing fewer number of plausible marker genes so that the classification precision of these markers is increased. In addition, this method enables detecting new plausible markers that their association to disease under study is not biologically proved.

Citation: Jafari Koshki T, Hajizadeh E, Karimi M. A Comparison of Selective Classification Methods in DNA Microarray Data of Cancer: Some Recommendations for Application in Health Promotion. Health Promot Perspect 2013; 3(1): 130-135

Introduction

Cancer ranks first health threatening issue and breast cancer is the leading malignancy among Iranian women [1, 2]. According to American Cancer Society report, estimated number of deaths due to breast cancer in 2012 ranks second after lung cancer among US women while estimated new cases of breast cancer in 2012 is above the rest kinds of cancer [3].Trend of cancer was increasing from 1994 to 1999. This trend was decreasing from 1999 to 2006 with 2% per year attributed to the reduction in the use of Menopausal

Hormone Therapy (MHT) [4]. According to reports published in 2005 to 2007, 12.15% of females born today, would experience breast cancer in their life; i.e. one out of 8 [3].

Early detection of breast cancer could result in mortality reduction and improve patients' prognosis. Although, mammography and other screening methods could acceptably detect breast cancer in early stages, but they would be partially effective and even ineffective in specific age groups [5].Cancer is considered a heterogeneous disease in terms of many aspects including its cellularity, different genetic alternation, and diverse clinical behavior that could be, in turn, due to heterogeneity of malignant cells and the patients' baseline factors. Now, cancer is classified based on clinical and histomorphological features that they could only partially reflect this heterogeneity and this will lead to lower possibility of perfect diagnosis, prognosis and prescription of relevant treatments. Beyond the diagnosis, determining the type of cancer is crucial to adopt relevant therapies. Because anti-cancer agents do not differentiate between normal and cancerous cells and sometimes lead to disastrous toxicity and an inconsistent efficacy. On the other hand, development of innovative drugs that selectively target cancer cells is of great interest and this adds to the importance of diagnosis and differentiating molecular events related to incidence and development of cancer [6]. Since this phenomenon is under influence of several genes, biological techniques that enable studying hundreds or thousands of molecular factors simultaneously are of high interest. These techniques could untangle the complexities existing between diseases and cell physiology.

US Food and Drug Administration (FDA) encourages using new technologies such as microarrays that could improve the assessment of medical products to promote the public health. Using this state-of-the-art technology in various clinical and diagnostic disciplines would have an important impact on public health, epidemiology, and other fields. Hamelin et al. studied virulence genes in environmental *E. coli* isolates using microarrays that were impossible before [7]. Ramaswamy et al. conducted a microarray analysis to identify primary and metastatic adenocarcinoma [8]. By the advent of such a technology, it would be possible to diagnose diseases in earlier stages and accurately define the type of a multi-class disease and use the best treatment as well as comparing the efficacy of various available treatments. This, certainly, could promote public health and prevent the growing burden, of all kind, of diseases on a society by screening and timely and accurately detection of these health threats.

DNA Microarray first was developed and used by Schenaet al. in the early 1990 [9]. It is an analytical tool that makes quick and precise genomic assessment possible. In this technology, gene expression levels in healthy and diseased tissues are evaluated and differences are determined. Hereby, genes with different levels of expression are identified and considered as potential markers of the disease to be studied in future studies.

In 1998, Eisenet al. clustered genes to analyze the gene expression patterns and identify genes with similar expression level [10]. Golub et al. used DNA microarray to diagnose and predict the class of cancerous tissues with Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) [11].

DNA microarray is a silicon chip on which thousands of single stranded oligonucleotides so called targets are spotted. These sequences are attached to the chip in two ways: Delivery or Contact Print in which targets are products of molecular techniques that are transferred on the chip. The second procedure is Synthesis or Photolithography. In this method, targets are constructed on the chip directly. Every chip is partitioned to distinct rows and columns called cells. Thousands of oligonucleotides of a gene are synthesized in a cell using a robot; each cell pertaining to a gene. This way, each cell contains only one type of oligonucleotides.

Then, mRNA from both healthy and diseased tissues is extracted. Two types of florescent tags is attached to the end of these two types of single stranded mRNAs such that they turn into red or green under laser radiation. These mRNA samples are washed and combined in equal amounts and poured on the chip. Therefore, single stranded mRNAs would bind to complementary single stranded oligonucleotides spotted on the chip. After a while, the chip is washed and put under the laser radiation and red or green lights are produced by tags. In each cell, there exist some red and some green tags. Thus, spectrum of colors would appear on the chip depending on the relative amount of these tags. Finally, produced spectrum is scanned and turned into digits using computers. These digits pertaining to relative density of tags indicate the relative expression level of genes in healthy and diseased tissues [6, 9].

Breast cancer is one of major diseases with regard its prevalence, incidence, and cost-burden issues [3]. As a health problem, it could be screened in various ways such as mammography [3, 12]. However, evaluating novel screening procedures to update or replace previous methods by more ones that are reliable is appealing. Hence, we used this disease as a case study due to its importance and availability of relevant data to clarify microarray application in health related areas. However, microarray analysis application scope is much wider and could be applied to various health problems in a same manner.

Statistical and machine learning methods are widely used in DNA microarray data analyses. The aim of this study was to combine these techniques to select fewer genes as markers of a disease via an algorithm. Breast cancer data was used to assess the performance of the algorithm according to its precision in predicting the type of cancer of a new patient.

Materials and Methods

This cross-sectional study was done based on data prepared by Farmer et al. by expression level of 22,215 genes from 49 patients with previously known type of cancer: basal, luminal, and apocrine[13]. Biopsies were taken from patients enrolled in a clinical trial before any treatment. The patients' characteristics and the trial duration were not mentioned. We used all data on gene expression levels obtained from these patients. This data is freely available on NCBI-GEO database with accession number GSE1561. The data were normalized before usage.

Statistical Modeling

Clustering is a famous statistical technique to discover similar groups in a dataset [14]. We used hierarchical clustering method to cluster the most similar genes in a group.

Among different distance criteria, we used average linkage distance:

$d(p, q)=1-r(G_p, G_q)$

wherer (G_p, G_q) represents Pearson correlation coefficient between genes p and q. Distance between two clusters each containing a number of genes is calculated as average of distance between pairwise genes.

Bayes learning is a useful machine learning technique often called Naïve Bayes. A Bayesian technique is as follows. Consider x representing feature (genes here) and y is the target function with range V. A set of learning samples is prepared and the machine is expected to classify new sample $x=(G_1, G_2, \dots ,G_n)$ and predict its label. Bayes classifier assigns the label with maximum possibility to this sample [15]. This maximum probability is called ν_{MAP} and calculated as:

$$\nu_{MAP} = \arg\max_{\nu_j \in V} P(\nu_j | G_1, G_2, \dots, G_n)$$

Cortes et al. introduced a class of hyperplanes with sign function to allocate a new sample to one of spaces created by optimal plane [16]. Consider observations $\{x_i, y_i\}$, $i=1, 2, \dots ,L$ that x_i indicates variable values of

*i*th subject and y_i the class of *i*th subject. Observations are classified into two groups with a hyperplane. Among all possible separating hyperplanes, we find the plane **w.x**+b=0 such that has maximum distance from both groups. Therefore, the decision function to classify a new observation would be based on sign function below:

$$f(x) = \text{sign}\left(\sum_{i=1}^{L} y_i \alpha_i . (\mathbf{x}.\mathbf{x}_i) + b\right)$$

where b is calculated by:

$$\alpha_i . [y_i ((\mathbf{x}_i . \mathbf{w}) + b) - 1] = 0;\ i = 1, \ldots, L$$

The Algorithm to select markers

We followed following steps to find best markers based on their precision in determining the class of new samples:

First, relief-F filtering technique, introduced by Kira and Rendel [17], was performed on the data. This would sort genes upon their relevance to represent the data. Then an arbitrary number of genes with highest ranks are selected. Now, the average linkage hierarchical clustering is done and dendrogram is plotted. A line is put on the upper part of dendrogram and is moved downward for clusters to be created. At first, we have only one cluster containing all selected genes. Moving the line would divide this cluster to more and more sub-clusters. At each step, one gene is extracted from each sub-cluster. This gene, which is the representative gene of pertaining cluster, would be the gene with minimum summed squared distance from other genes in that sub-cluster. When these representative genes are selected, one sample is excluded from subjects and the classifier is constructed according to these genes. Then, the class of the left-out gene is specified by the classifier. Now, the left-out subject is returned to the sample, another subject is excluded from sample, and the process is repeated until all subjects are excluded and returned to sample one by one. When this process is completed for this step, the rate of accurately classified subjects is considered as the classifier's performance [18]. Aforementioned process is repeated at each step of the line movement and the prediction precision is recorded for the respective number of clusters or genes.

At the final step, the number of sub-clusters or genes equals the number of genes selected at the earliest stage.

This algorithm was implemented in Waikato Environment for Knowledge Analysis (WEKA) [19, 20] and MATLAB 7.0.

Results

Table 1 shows the prediction precision for 50 selected genes with highest ranks.

Table 1: Number and precision of candidate markers based on 50 highly ranked genes as the input of the algorithm

NCBI accession number for selected markers	(Number, Precision(%)) for candidate markers	Precision (%) for all selected genes	Classification method
209604_s_at, 214431_at, 205009_at, 209459_s_at, 205186_at, 214053_at, 204580_at, 204822_at, 204667_at, 207039_at, 203574_at, 203780_at, 205376_at, 203453_at, 203636_at, 209900_s_at, 208484_at, 203256_at, 214079_at, 205030_at, 217528_at, 214451_at	22, 100	100	SVM
209603_at, 205186_at, 205819_at, 204667_at, 205376_at, 203636_at, 209900_s_at, 208484_at, 203256_at, 214079_at, 205030_at, 217528_at	12, 97.95	95.91	NB

Table 2: Number and precision of candidate markers based on 20 highly ranked genes as the input of the algorithm

NCBI accession number for selected markers	(Number, Precision(%)) for candidate markers	Precision (%) for all selected genes	Classification method
209603_at, 204667_at, 204580_at, 218963_s_at, 214079_at, 205030_at	6, 97.95	97.95	SVM
209604_s_at, 204580_at, 218963_s_at, 214079_at, 205030_at	5, 95.91	95.91	NB

The result indicates that for SVM classifier, one could construct a function that classifies a new patient with higher accuracy using 22 genes instead of 50 genes. The result for NB classifier is even more interesting. Using all selected 50 genes, the precision is 95.91% but 100% for 22 selected genes by the algorithm. The algorithm was implemented for 20 highly ranked genes separately. The results are shown in Table 2. This table suggests that we could gain same precision with smaller number of genes.

Discussion

Traditional techniques usually fall short of perfect diagnosis, prediction, adopting the best treatment and evaluating the prescribed treatments as well. Therefore, researchers endeavor to make improvements in each area. In this regard, DNA microarray analysis has been center of attention in recent years and many studies have been done to improve this technology. Using this technology helps the researchers to circumvent previous limitations of the traditional laboratory tools. Hamelin et al. used microarrays to study the virulence genes in environmental *E. coli* isolates that had technological limitations before [7]. Ramaswamy et al. distinguished primary and metastatic adenocarcinoma by using a microarray analysis [8]. They reported a challenging result contradicting the previous notion that metastases arise from rare cells within a primarytumor that have the ability to metastasize. Microarrays could be utilized to detect a disease accurately and quickly and help to decrease the cost of diseases and improve the quality of life in the society. By developments in this area, microarray would be able to serve as a cheap and quick screening tool before too long.

Wang et al. assessed the performance of this algorithm on three datasets [18]. They found that hybrid method could result in fewer marker-genes. Chittenden et al. used this method to classify coronary artery disease [21]. They evaluated transcriptional correlation in monocyte types from these patients and fewer genes with higher degree of significance. Same result was found on independent datasets. They also reported that selected marker-genes had roles in molecular functions and biological processes. The results from present study show that many of genes selected by the algorithm have a role or are related to breast cancer: GATA3, AR, NBR1, ESR1, ERBB4 and GHR to name a few [12]. ER expression level was known as the source of differentiation between luminaland basal types. Farmer et al. reported that ERBB4 was commoner in molecular apocrine than the other groups [13]. These facts support the appropriateness of the hybrid algorithm in cancer type differentiations. Nevertheless, functionality of some of genes identified as markers of cancer differentiation is not clear yet. This could illuminate a path for future studies to clarify their possible relation to breast cancer.

DNA microarray analysis enables us to study thousands of genes simultaneously, which leads us to patterns and markers that their roles are not known in diseases as well. In the bioinformatics context, statistical techniques are considered as the final chain of data mining process. In this manuscript, we tried to briefly introduce microarray technology and evaluate the performance of a hybrid approach combining two statistical techniques. The bur-

den of health problems, that could be prevented by effective diagnostic and screening in advance, imposes a great demand on new techniques to improve existing diagnostic and screening methods [3]. Microarray is a new approach that is widely used in treatment assessment, disease diagnosis and classification. We, here, used this technique to highlight its advantage in cancer classification. It could be applied to another health related issues from different viewpoints by designing a relevant study and gathering data on the subject.

Conclusion

We used some of well-known statistical techniques via an algorithm. However, other classification and clustering methods or other statistical and machine learning techniques could be used in this algorithm and results compared with methods used in this study.

Acknowledgement

We wish to thank Maryam Nabavi for her help on data preparation. We also appreciate Habib Zeighami, Ph.D., for his comments on manuscript. The authors declare that there is no conflict of interest.

References

1. Haskell CM, Bereck J (2001). Cancer treatment. 5th ed. Philadelphia: WB Saunder.
2. Mousavi SM, Montazeri A, Mohagheghi MA, Jarrahi AM, Harirchi I, Najafi M, et al. Breast cancer in Iran: An Epidemiological Review. *Breast J* 2007; 13(4): 383-91.
3. American Cancer Society. Cancer facts & figures 2012. Atlanta: American Cancer Society; 2012 .
4. American Cancer Society. Cancer facts & figures 2010. Atlanta: American Cancer Society; 2010.
5. Montazeri A, Vahdaninia M , Harirchi I, Mahmood Harirchi A, Sajadian A, Khaleghi F, et al. Breast cancer in Iran: need for greater women awareness of warning signs and effective screening methods. *Asia Pac Fam Med* 2008; 7(6).
6. Mocellin S(2007). Microarray technology and cancer gene profiling: Springer.
7. Hamelin K, Bruant G, El-Shaarawi A, Hill S, Edge TA, Bekal S, et al. A Virulence and Antimicrobial resistance DNA microarray detects a high frequency of virulence genes in *Escherichiacoli* isolates recreational waters. *Appl Environ Microbiol* 2006, 4200–4206.
8. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. *Nature Genetics* 2003; 33: 49-54.
9. Schena M(2003). Microarray analysis New Jersey: John-Wiley & Sons.
10. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proceedeings of National Academic Science* 1998; 95: 14863–8.
11. Golub TR, Slonim DK, Tamayo P, Huard , Gaasenbeek M, Mesirov JP, et al. Molecular classifcation of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 1999; 286: 531-7.
12. Knudsen S (2006). Cancer diagnostics with DNA microarrays. John Wiley & Sons, Inc.
13. Farmer P, Bonnefoi H, Becette, Becette V, Tubiana-Hulin M, Fumoleau P , et al. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* 2005; 24: 4660–71.
14. Rencher AC (2002). Methods of multivariate analysis. 2nd ed. Wiley & Sons, Inc.
15. Mitchell M (1997). Machine learning: McGrow-Hill.
16. Cortes C, Vapnik V. Support-Vector networks. *Mach Learn* 1995; 20: 273-97.
17. Kira K, Rendell L. A practical approach for feature selection. *In Proceedings of the Ninth International Conference on Machine Learning*, 1992. 249–56.
18. Wang Y, Makedon FS, C. Ford JC, Pearlman J. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics* 2005; 21(8): 1530–7.
19. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH(2009); The WEKA data mining software: an update; SIGKDD explorations, Volume 11.
20. Bouckaert RR, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, et al (2009). WEKA manual for version 3-6-1 Hamilton, New Zealand: University of Waikato.
21. Chittenden TW, Sherman JA, Xiong F, Hall AE, Lanahan AA, Taylor JM, et al. Transcriptional profiling in coronary artery disease. Indications for novel markers of coronary collateralization. *J Am Heart Assoc* 2006; 114: 1811-20.