Original Article *Iran J Ped Hematol Oncol. 2020, Vol 11, No 2, 70-77*

Gene Identification from Microarray Data for Diagnosis of Acute Myeloid and Lymphoblastic Leukemia Using a Sparse Gene Selection Method

Razieh Sheikhpour PhD[1], Roohallah Fazli PhD[2], Sanaz Mehrabani MD[3,*]

1Department of Computer Engineering, Faculty of Engineering, Ardakan University, P.O. Box 184, Ardakan, Iran
2Department of Electrical Engineering, Faculty of Engineering, Ardakan University, P.O. Box 184, Ardakan, Iran
3Non-Communicable Pediatric Diseases Research Center, Health Research Institute, Babol University of Medical Sciences, Babol, Iran
*Corresponding author: Dr Sanaz Mehrabani, Non-Communicable Pediatric Diseases Research Center, Health Research Institute, Babol University of Medical Sciences, Babol, Iran. Email: Mehrabanisanaz@gmail.com. ORCID ID: 0000-0002-2062-0448

Received: 10 October 2020 **Accepted:** 24 December 2020

Abstract

Background: Microarray experiments can simultaneously determine the expression of thousands of genes. Identification of potential genes from microarray data for diagnosis of cancer is important. This study aimed to identify genes for the diagnosis of acute myeloid and lymphoblastic leukemia using a sparse feature selection method.

Materials and Methods: In this descriptive study, the expression of 7129 genes of 25 patients with acute myeloid leukemia (AML), and 47 patients with lymphoblastic leukemia (ALL) achieved by the microarray technology were used in this study. Then, the important genes were identified using a sparse feature selection method to diagnose AML and ALL tissues based on the machine learning methods such as support vector machine (SVM), Gaussian kernel density estimation based classifier (GKDEC), k-nearest neighbor (KNN), and linear discriminant classifier (LDC).

Results: Diagnosis of ALL and AML was done with the accuracy of 100% using 8 genes of microarray data selected by the sparse feature selection method, GKDEC, and LDC. Moreover, the KNN classifier using 6 genes and the SVM classifier using 7 genes diagnosed AML and ALL with the accuracy of 91.18% and 94.12%, respectively. The gene with the description "Paired-box protein PAX2 (PAX2) gene, exon 11 and complete CDs" was determined as the most important gene in the diagnosis of ALL and AML.

Conclusion: The experimental results of the current study showed that AML and ALL can be diagnosed with high accuracy using sparse feature selection and machine learning methods. It seems that the investigation of the expression of selected genes in this study can be helpful in the diagnosis of ALL and AML.

Keywords: Acute myeloid leukemia, Acute lymphoblastic leukemia, Gene, Identification, Microarray

Introduction

Leukemia is the blood cell cancer which is the most common cancer in children younger than 15 years (1,2). The cause of leukemia in children is still generally undiscovered. Few risk factors such as genetic susceptibility, infection, and ionizing radiation have been recognized, but they seem to describe only a small fraction of the cases (1). Acute leukemia includes a heterogeneous group of diseases determined by rapid and uncontrolled clonal expansion of progenitor cells of the hematopoietic system (2). Acute leukemia is categorized into myeloid and lymphoid, based on the immunologic markers determining their lineage commitment (3). Acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) are the frequent types of leukemia among children (1). ALL is the cancer of the lymphoid line of blood cells which is the most common childhood cancer (1,4,5). AML is the cancer of the myeloid line of blood cells that occurs due to blast accumulation and uncontrolled proliferation factors (4,5). In ALL and AML, the abnormal cells are rapidly reproduced in the bone marrow and blood which lead to disrupting the function of normal blood cells (4). Pale skin color,

enlarged lymph nodes, easy bleeding or bruising, feeling tired, fever, spleen, and liver are symptoms of ALL, while reddish dots on the skin, feeling tired, bleeding gum, and shortness of breath are symptoms of AML (4). Diagnosis of AML from ALL is important with regard to prognosis and treatment. One of the most accurate and important ways to diagnose AML and ALL is to use people's DNA and their genetic information. With the use of DNA microarray technology, it is possible to take a genome-wide method to diagnose AML and ALL and measure the expression level of thousands of genes simultaneously (6–9). Microarray gene expression data are widely utilized for the discovery of cancer biomarkers or gene signatures and diagnosis of cancer. In microarray data, gene numbers are significantly larger than the sample number, which leads to the curse of dimensionality phenomenon and challenges the classification process (7–9). Most genes in microarray data are redundant, and a few relevant genes may be useful for cancer diagnosis and appropriate therapeutic selection in clinical management. Therefore, an important step in analyzing microarray data is to decrease the number of genes and select appropriate genes for the classification of cancer which leads to the decreased processing time of classification and misclassification rate (8,10,11). In gene selection, a number of relevant genes which has been widely utilized in microarray data are selected (12). Gene selection methods can be split into the filter, wrapper, and embedded methods. Filter methods rank the genes based on their certain characteristics independent of the classifiers. These methods are fast and simply applied to microarray data sets that have thousands of genes. Wrapper methods use some criteria to choose a number of genes that have the best performance for a specified classifier. Wrappers usually have good performance but the computational cost of these methods is high. Embedded methods carry out gene selection in the training process and are usually specific to a classifier (11,13). Classical gene selection methods ignore the correlation among genes and evaluate the importance of each gene individually. To solve the problem, sparse feature selection was presented to consider the correlated information among different features in the dimension reduction process (14). This study aimed to select the relevant genes in the diagnosis of ALL and AML using the leukemia microarray gene expression data. For this purpose, a sparse feature selection method based on l2,1-norm minimization on regularization was applied to consider the correlated information among different genes in the gene selection process. The feature selection method also preserved the geometry structure of all leukemia gene expression data. Then, some classifiers such as k-nearest neighbor (KNN), support vector machine (SVM), Gaussian kernel density estimation based classifier (GKDEC), and linear discriminant classifier (LDC) were applied to the selected genes to diagnose ALL and AML.

Materials and Methods

Data set

In this descriptive study, the microarray gene expression data collected from the bone marrow of patients with leukemia cancer provided by Golub et al. (15) was used. The data set included 72 samples of leukemia that were classified into 23 AML and 49 ALL samples. Each sample in this data was indicated by the expression of 7129 genes. To evaluate the efficiency and performance of machine learning methods, the data should be divided into two training and test sets. The training set was applied to construct the model and the test set to evaluate the model. The gene expression data of leukemia utilized in this study were previously split into training and test sets. The training and test sets include 38 leukemia patients (25 ALL and 13 AML) and 34 leukemia patients (24 ALL and 10 AML), respectively.

Sparse feature selection method

In this study, a sparse feature selection method based on $l_{2,1}$-norms and graph Laplacian was applied to identify important genes in the diagnosis of AML and ALL. The sparse method considers the correlation among various genes and preserves the geometry structure of the data. The objective function of the sparse method used in this study is as Eq. (1) (16–18):

$$\arg\min_{W,b} Tr(W^T X L X^T W) + \mu \|X^T W + 1_n b^T - Y\|_F^2 + \lambda \|W\|_{2,1}$$

where L denotes the graph Laplacian, $b \in R^c$ is the bias term and $1_n \in R$ is a column vector in which all n its elements are 1 and n is the number of training data. X and Y are the training data and their labels, respectively. μ and λ indicate regularization parameters. The $l_{2\text{-}1}$-norm regularization in Eq. (1) ensures that the most sparse genes were selected and the correlation among genes was considered in the gene selection process. For the computation of graph Laplacian, a graph S was created with n nodes which node i specifies sample x_i. In the graph, close samples were connected to each other. The weight matrix of the graph was defined as Eq. (2):

$$S_{ij} = \begin{cases} 1 & \text{if } x_i \in KNN(x_j) \text{ or } x_j \in KNN(x_i) \\ 0 & otherwise \end{cases}$$

The graph Laplacian is calculated through $L = D - S$, where D is computed as $D_{ii} = \sum_j S_{ij}$.

SVM

SVM is known as a classification method which applies a nonlinear mapping to turn the microarray data space into a higher dimension. This classifier searches a linear optimal separating hyperplane in the new dimension which separates the sample of ALL from AML (19).

KNN

KNN is based on learning by analogy which searches the gene space for k close samples to the new sample. In fact, this method calculates the distance of the new sample to training samples and searches the gene space to find k leukemia samples in the training set which are closest to the new sample. This method needs a distance criterion such as Euclidean distance or Manhattan distance to find similarities between samples (19).

GKDEC

GKDEC is a non-parametric classification method. In kernel estimator, kernel bandwidth and kernel function affect the probability density estimation. A popular kernel function is the Gaussian kernel that extensively utilized in GKDEC (20).

LDC

LDC assumes that AML samples are linearly separable from ALL samples. This method estimates the parameters of the linear discriminant directly from the leukemia microarray data.

Ethical Consideration

Current study was approved by Ethical committee of Shahid Sadoughi University of Medical Sciences (number: IR. SSU. MEDICINE.REC.1399.224).

Results

In this study, different experiments were conducted on the leukemia microarray data set to diagnose AML and ALL. For this purpose, the rank of each gene in leukemia microarray data set was calculated using the method defined in Eq. (1). Then, the genes with the highest ranks were selected and the classifiers such as KNN, SVM, LDC, and GKDEC were applied on the selected genes to diagnose AML and ALL. For evaluation of the performance of the classifiers on the identified genes, accuracy, sensitivity, and specificity were used as the evaluation measures. Accuracy is a measure that refers to the percentage of correctly predicted leukemia cancer samples. Sensitivity and specificity were utilized to identify the percentage of correctly predicted AML and ALL cancer samples, respectively. The leukemia microarray data set was originally split into training and test sets with 38 and 34 samples, respectively. Gene Ranking and

model construction were carried out on the training data, and the evaluation of the models on the selected genes was done on the test data. Table I show the performance of different classifiers on the different number of genes selected by the sparse feature selection method. The bold indicates the best Performance. It is clear from Table I, GKDEC and LDC using 8 selected genes diagnosed AML and ALL with the accuracy of 100%. These methods were able to correctly diagnose all ALL and AML samples. Moreover, the accuracy of the KNN classifier using 6 genes and the SVM classifier using 7 genes in diagnosing of AML and ALL was 91.18% and 94.12%, respectively. The result of Table I show that all classification models constructed on the small number of selected genes have high performance in the diagnosis of AML and ALL, which indicates the ability of the sparse feature selection method presented in Eq.(1) in identification of the most relevant genes. This is because the method considers the correlation among different genes in the gene selection process and keeps the geometry structure of microarray data in the construction of graph Laplacian. In table II, 10 top-ranked genes of the microarray leukemia data set identified by the sparse feature selection method are shown.

Table I: Performance of different classifiers on different number of genes

Method	Number of genes	Accuracy	Sensitivity	Specificity
KNN	5	88.24	100	88.33
SVM		88.24	100	88.33
GKDEC		97.06	90	100
LDC		94.12	100	91.67
KNN	6	91.18	100	87.5
SVM		91.18	100	87.5
GKDEC		97.06	90	100
LDC		94.12	100	91.67
KNN	7	91.18	100	87.5
SVM		94.12	100	91.67
GKDEC		97.06	100	95.83
LDC		94.12	100	91.67
KNN	8	91.18	100	87.5
SVM		94.12	100	91.67
GKDEC		**100**	**100**	**100**
LDC		**100**	**100**	**100**
KNN	9	91.18	100	87.5
SVM		94.12	100	91.67
GKDEC		97.06	100	95.83
LDC		100	100	100
KNN	10	91.18	100	87.5
SVM		94.12	100	91.67
GKDEC		100	100	100
LDC		100	100	100

Table II: Top-ranked genes of microarray leukemia data set identified by the sparse feature selection method

Rank	Accession	Description
1	U45255_s_at	Paired-box protein PAX2 (PAX2) gene, exon 11 and complete cds
2	D14134_at	RECA Replication protein A (E coli RecA homolog, RAD51 homolog)
3	L02950_at	CRYM Crystallin Mu
4	HG3725-HT3981_s_at	Insulin-Like Leydig Hormone
5	X89430_at	Methyl CpG binding protein 2
6	M11973_cds1_at	Gamma-B-crystallin gene (gamma 1-2)
7	U11878_at	Interleukin-8 receptor type B (IL8RB) mRNA, splice variant IL8RB10, partial cds
8	Y10812_at	Fructose-1,6-bisphosphatase
9	U94333_at	Clq/MBL/SPA receptor C1qR(p) mRNA
10	U09411_at	ZNF132 Zinc finger protein 132 (clone pHZ-12)

Discussion

In the current study, the classification of microarray data of leukemia patients into ALL and AML was carried out using KNN, SVM, GKDE, and LD classifiers. The classifiers diagnosed ALL and AML using a small number of genes identified by the sparse feature selection. GKDE and LD classifiers diagnosed AML and ALL with the accuracy of 100% using 8 top-ranked genes identified by the sparse feature selection method. Moreover, KNN and SVM classifiers achieved the accuracy of 91.18% and 94.12%, using 6 genes 7 genes to diagnose AML and ALL, respectively.

Alshamlan et al. (21) proposed the genetic bee colony (GBC) algorithm which combines the genetic algorithm (GA) and artificial bee colony (ABC) algorithm. They employed the mRMR method on leukemia microarray data set to select top relevant genes. The accuracy of the SVM classifier on 50, 100, and 150 relevant genes selected by the mRMR method was 91.66%, 97.22%, and 100%, respectively. Alshamlan et al. also carried out the comparison of the performance of the GBC algorithm with ABC and mRMR-ABC algorithms. The mean accuracy of the SVM on 5 genes of leukemia microarray data set selected by GBC, mRMR-ABC, and ABC algorithms were 96.43%, 92.82%, and 91.89%, respectively.

A method was presented by Bolón-Canedo et al. (22) which distributed the data by features and carried out a merging procedure for updating the feature subset based on the improvement of accuracy. They achieved the classification accuracy of 91.18%, 97.06%, 94.12%, and 94.12% using C4.5, SVM, KNN, and naïve Bayes classifiers, respectively on the leukemia microarray data set.

Aziz et al. (7) modeled the leukemia data using the independent component analysis (ICA) method and selected the relevant genes using the fuzzy backward feature elimination (FBFE) method. The classification accuracy of SVM and NB classifiers with ICA feature vector was 88.23% and 86.21%, respectively. Moreover, Aziz et al. (7) achieved the accuracy of 94.2% and 95.12% for SVM and NB classifiers, respectively using the FBFE method on the independent component feature vector extracted by ICA. FBFE eliminated the irrelevant genes from the independent components and selects 35 genes for SVM and 30 genes for NB.

Apolloni et al.(9) presented a hybrid feature selection method called BDE-X_{Rank} which combines an FS method based on a binary differential evolution (BDE) algorithm with a filter feature selection method. SVM, KNN, NB, and C4.5 classifiers were used on the leukemia data set to evaluate the performance of the BDE-X_{Rank} method in the identification of

the most predictive genes for the classification of ALL and AML. Classification accuracies of 82.4%, 97.1%, 91.2%, and 91.2% were obtained by SVM, KNN, NB, and C4.5 classifiers, respectively constructed on the genes identified by BDE-XRank.

A hybrid method based on relief and convolutional neural network (CNN) was presented by Kiliçarslan et al. (23) for the diagnosis of ALL and AML on leukemia gene expression data. They applied the relief method which is a dimension reduction algorithm on the leukemia data to select the relevant genes. Then, a convolutional neural network with Softmax function was used on the genes selected by the relief method to diagnose ALL and AML. They achieved an accuracy of 99.86% in the diagnosis of ALL and AML.

ALL and AML were diagnosed using the leukemia gene expression data with an accuracy of 94.85% by Arunkumar and Ramakrishnan (24). They used the CFS method for the selection of the relevant genes. Then, the genes selected by CFS were employed to calculate the final minimal reduct set utilizing a customized fuzzy triangular norm operator based on the fuzzy rough quick reduct (FRQR) algorithm.

Potharaju and Sreedevi (25) presented a distributed feature selection (DFS) method utilizing symmetrical uncertainty (SU) and multilayer perceptron (MLP) by distributing across the multiple clusters. They evaluated the DFS method using ridor, simple cart (SC), KNN, and SVM classifiers and compared the DFS method with some classical methods such as IG, gain ratio (GR), and chi-squared attribute evaluator (Chi). They obtained the classification accuracy of 93.05%, 94.44%, 95.83%, and 98.61% using ridor, SC, KNN, and SVM classifiers, respectively which was better than IG, GR, and Chi methods.

Karimi and Farrokhnia (26) presented a method based on the combination of dimension reduction and gene selection techniques for the microarray data set. This method used the GA to select the relevant genes and combined it with linear discriminant analysis (LDA). In this method, some relevant genes of the leukemia data set were selected using GA, and LDA was performed on the selected genes instead of the whole data set. Karimi and Farrokhnia identified 22 relevant genes and achieved an accuracy of 94.21% on leukemia data set in the diagnosis of ALL and AML.

A hybrid feature selection method was proposed by Santhakumar and Logeswari (27) for the classification of ALL and AML which was based on the combination of Ant Colony Optimization (ACO) and Ant Lion Optimization (ALO) algorithm. The accuracy of 95.45%, 93.94%, and 90.91% was achieved using the ant lion mutated ant colony optimizer feature selection, ant colony optimizer feature selection, and ant lion mutated feature selection, respectively.

A gene selection method was presented by Kyun Park et al. for microarray data (28). This method combined an unsupervised gene selection method with a supervised one to identify the top-ranked genes. Kyun Park et al. (28) achieved an accuracy of 100% in the classification of ALL and AML using 13 top-ranked genes of leukemia data.

Conclusion

The results of this study indicated that sparse feature selection and machine learning methods can be applied for diagnosis of AML and ALL with high accuracy. Moreover, the results showed that the sparse feature selection based on $l_{2,1}$-norm identifies the most relevant genes of microarray data for diagnosis of AML and ALL. This is because the sparse method considers the useful information among different genes and preserves the geometry structures of microarray data in the gene selection process. Therefore, it seems that investigating the expression of

the genes identified by the sparse feature selection method can be used in the diagnosis of ALL and AML.

Conflict of interest

The authors declare no conflict of interest.

References

1. Filippini T, Heck JE, Malagoli C, Giovane C Del, Vinceti M. A review and meta-analysis of outdoor air pollution and risk of childhood leukemia. J Environ Sci Heal - Part C Environ Carcinog Ecotoxicol Rev2015;33(1):36–66.
2. Shukla S, Chhikara A, Bundela T, Sharma S, Chandra J. Clinical, morphological and immunophenotypical findings in acute leukemia: A study from a tertiary care hospital. Iran J Pediatr Hematol Oncol 2020;10(3):136–143.
3. De Leeuw DC, Van Den Ancker W, Denkers F, De Menezes RX, Westers TM, Ossenkoppele GJ, et al. MicroRNA profiling can classify acute leukemias of ambiguous lineage as either acute myeloid leukemia or acute lymphoid leukemia. Clin Cancer Res 2013;19(8):2187–2196.
4. Masilamani V, Devanesan S, AlSalhi MS, AlQahtany FS, Farhat KH. Fluorescence spectral detection of acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML): A novel photodiagnosis strategy. Photodiagnosis Photodyn Ther 2020;29: 101634-101637.
5. Rasool M, Farooq S, Malik A, Shaukat A, Manan A, Asif M, et al. Assessment of circulating biochemical markers and antioxidative status in acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) patients. Saudi J Biol Sci 2015;22(1):106–111.
6. Hong J-HH, Cho S-BB. Gene boosting for cancer classification based on gene expression profiles. Pattern Recognit 2009;42(9):1761–1767.
7. Aziz R, Verma CK, Srivastava N. A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data. Genomics Data 2016;8:4–15.
8. Elyasigomari V, Lee DA, Screen HRC, Shaheed MH. Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. J Biomed Inform2017;67:11–20.
9. Apolloni J, Leguizamón G, Alba E. Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl Soft Comput J 2016;38:922–932.
10. Chen Y, Zhang Z, Zheng J, Ma Y, Xue Y. Gene selection for tumor classification using neighborhood rough sets and entropy measures. J Biomed Inform 2017;67:59–68.
11. Sheikhpour RR, Sarram MAMA, Chahooki MAZMAZ, Sheikhpour RR. A kernelized non-parametric classifier based on feature ranking in anisotropic Gaussian kernel. Neurocomputing 2017;267:545–555.
12. Peng Y, Li W, Liu Y. A Hybrid Approach for Biomarker Discovery from Microarray Gene Expression Data for Cancer Classification. Cancer Inform 2006;2:301–311.
13. Nie F, Huang H, Cai X, Ding CH. Efficient and robust feature selection via joint ℓ2, 1-norms minimization. In: Advances in neural information processing systems 2010; 1813–1821.
14. Sheikhpour R, Sarram MA, Gharaghani S, Chahooki MAZ. A robust graph-based semi-supervised sparse feature selection method. Inf Sci 2020;531:13–30.
15. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–527.
16. Ma Z, Nie F, Yang Y, Uijlings JRR, Sebe N, Member S, et al. Discriminating joint feature analysis for multimedia data understanding. IEEE Trans Multimed 2012;14(6):1662–1672.

17. Ma Z, Yang Y, Nie F, Uijlings J, Sebe N. Exploiting the entire feature space with sparsity for automatic image annotation. Proc 19th ACM Int Conf Multimed - MM '11. 2011;283.
18. Shi C, Ruan Q, An G. Sparse feature selection based on graph Laplacian for web image annotation. Image Vis Comput 2014;32(3):189–201.
19. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.
20. Alpaydin E. Introduction to Machine Learning. Second Edi. MIT press; 2014.
21. Alshamlan HM, Badr GH, Alohali YA. Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Comput Biol Chem 2015;56:49–60.
22. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. Distributed feature selection: An application to microarray data classification. Appl Soft Comput J 2015;30:136–50.
23. Kilicarslan S, Adem K, Celik M. Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network. Med Hypotheses 2020;137:109577-1095779.
24. Arunkumar C, Ramakrishnan S. Prediction of cancer using customised fuzzy rough machine learning approaches. Healthc Technol Lett 2019;6(1):13–18.
25. Potharaju SP, Sreedevi M. Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance. Clin Epidemiol Glob Heal 2019;7(2):171–176.
26. Karimi S, Farrokhnia M. Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: Combining data dimension reduction and variable selection technique. Chemom Intell Lab Syst 2014;139:6–14.
27. Santhakumar D, Logeswari S. Hybrid ant lion mutated ant colony optimizer technique for Leukemia prediction using microarray gene data. J Ambient Intell Humaniz Comput 2021;12:2965–2973.
28. Park DK, Jung EY, Lee SH, Lim JS. A composite gene selection for DNA microarray data analysis. Multimed Tools Appl 2015;74(20):9031–9041.