Available online at **www.ijmrhs.com**

Special Issue 9S: Medical Science and Healthcare: Current Scenario and Future Development

ISSN No: 2319-5886

International Journal of Medical Research & Health Sciences, 2016, 5, 9S:500-506

Use of LDA combined with PLS for classification of lung cancer gene expression data

Keyghobad Ghadiri[1], Mansour Rezaei[2], Seyyed Mohammad Tabatabaei[3], Meisam Shahsavari[4] and Soodeh Shahsavari[5]*

[1]Nosocomial Infection Research Center, Kermanshah University of Medical Sciences, Kermanshah, Iran
[2]Biostatistics and Epidemiology Department, Faculty of Public Health, Kermanshah University of Medical Sciences, Kermanshah, Iran
[3]Medical Informatics Department, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[4]Nursing Department, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[5]Health Information Technology Department, Faculty of Paramedical Sciences, Kermanshah University of Medical Sciences, Kermanshah, Iran
**Corresponding Author: soodeh_shahsavari@yahoo.com*

ABSTRACT

Reliable and precise classification is essential for successful diagnosis and treatment of cancer. Thus, improvements in cancer classification are increasingly sought. Linear discriminant analysis (LDA) is the most effective method of cancer classification in high-dimensional prediction, but there are drawbacks to tumor classification by a formal method such as LDA. We propose a method for lung cancer gene microarray classification that combines a feature reduction approach, partial least squares (PLS), and discriminate method, LDA, for improving classification performance. The real dataset used related to lung cancer gene expression. After bioinformatics data pre-processing, data reduction and feature selection were carried out using PLS and then LDA was used for classification. The results were validated using the accuracy index and gene ontology analysis. Of the total of more than 50,000 genes, 214 genes were shown to have relevance. The classification accuracy of this method was 94.5% and gene ontology analysis results were good. It can be said that the LDA classifier combined with PLS is powerful method. This method can identify gene relationships warranting further biological investigation.

Keywords: lung cancer, gene expression data, linear discriminant analysis, partial least squares, classification

INTRODUCTION

Cancer is a term that covers a complex set of diseases [1]. Carcinogenesis is the transformation of a normal cell into a cancer cell and is a multifaceted process with many stages [2]. In molecular and cellular biology, cancers are rare diseases that caused by the same molecular defects that occur in cellular activity and cause similar changes in cell genes. It is a disease caused by abnormal gene expression [3]. Lung cancer is a leading causes of cancer death worldwide [4–7]. Its high incidence and poor prognosis make it the seventh most common cause of death; it will be responsible for 3% of mortalities by 2030 [8]. Lung cancer deaths have increased dramatically in recent years; therefore, healthcare policymakers who determine research and treatment priorities based on death rate as an indicator of burden of disease should pay special attention to this underreported data and develop strategies against this form of cancer [9]. Despite extensive clinical research, the five-year survival rate of non-small cell lung cancer (NSCLC), the major histologic subtype, has improved only slightly (from 14% to 18%) [10]. Recently, targeted treatment based on molecular distortion has significantly improved results in subsets of patients with NSCLC [11, 12]. A diagnosis of cancer and how to treat it has a major effect on the day-to-day life of patients and disruption of

normal life is a serious consequence of treatment. It has economic consequences for the patient and the patient's family and negatively affects their psychological status and quality of life [13]. Lung cancer is a disease in which the uncontrolled growth in certain cells in the lungs is formed a tumor. These abnormal cells cannot function as normal cells and is capable of both invading surrounding normal tissue and spreading throughout the body via the circulatory or lymphatic systems [14 , 15]. All cells of an organism have the same genes, but different conditions affect the way particular gene is expressed and how it could be expressed. It is crucial to evaluate the levels of genome in these situations [16]. DNA microarray technology has allowed the monitoring of thousands of gene expressions simultaneously under different conditions and processes. This technology has accelerated and increased the efficacy of gene expression studies [17].

In cancer treatment or therapy, the classification of normal and abnormal patterns of cells is one of most important processes in the diagnosis of cancer. Modern cancer diagnoses are achieved using an expert classifier system [18]. Cancer classification based on microarray can be used to detect subtypes of cancers and produce therapies. Many studies have developed methods for cancer classification [19–21]. These methods include principal component analysis (PCA) [22,23], *k*-nearest neighbor (*k*-NN) [24], hierarchical clustering analysis (HCA) [25], support vector machine (SVM) [26], Bayesian [28], partial least squares (PLS), discriminant analysis (DA) [28], and ensemble methods [29]. Reliable and precise classification is essential for successful diagnosis and treatment of cancer; thus, improvements in cancer classification are increasingly sought [30,31]. The conceptually simple approach of linear discriminant analysis (LDA) and its variants [32,33], remain among the most effective procedures in the domain of high-dimensional prediction. Drawbacks exist that are associated with tumor classification by LDA. One property of microarray data is that the number of genes, *p*, exceeds the number of tissue samples (patients) [34]. Except for a few classification methods using all genes [30], classification is generally performed using a selection of significant genes for constructing accurate prediction models. A small number of genes is usually strongly significant for a disease and most are not used for cancer classification. These extra genes can produce noise that decreases classification accuracy [35]. As a dimension reduction technique, PLS has been used in gene expression data analysis even where the number of genes exceeds the number of samples [12]. We propose a method for lung cancer gene microarray classification that combines a feature reduction approach, PLS, and discriminate method, LDA, for improving classification performance.

MATERIALS AND METHODS

Data Resources

The real dataset that was used in this research was obtained from a lung cancer gene expression study in 2010 that included GEO datasets [36]. The genetic mechanisms of carcinogenesis in non-smokers is unclear, but semaphorin have been suggested to play a role in lung tumor suppression. This report is a comprehensive analysis of the molecular signature of nonsmoking female lung cancer patients in Taiwan with a focus on the semaphorin gene family. Sixty pairs of tumor and adjacent normal lung tissue specimens were analyzed using Affymetrix U133+2.0 expression arrays.

Dimension reduction

Feature selection can be employed to improve classification accuracy or aid model explanation by establishing a subset of discriminating features within a class. Factors such as experimental noise, choice of technique and threshold selection can adversely affect the set of features selected. The high dimensionality and multicollinearity inherent in gene expression data can exacerbate discrepancies between the set of features retrieved [37]. Selecting an optimal number of features to use for classification is a complicated task. Given these issues, after preprocessing of bioinformatics data, the PLS method was used for data reduction and features selection and the genes identified were tested using FDR. In the PLS algorithm, class labels can be used for the dependent y vector. In the two-class case, the values of the dependent variable are usually assigned a value of 1 for one class and 0 or -1 for the other class. Feature selection methods less prone to the effect of bias of multicollinear data include those based on variable influence on projection (VIP) values derived from PLS-DA. The VIP value is the weighted sum of squares of the PLS weights (*w*) which takes into account the variance of each PLS dimension. The VIP score of a predictor is a summary of the importance of the projections to finding latent variables. [38]. Analysis of this section was carried out using the ropls R package.

It is evident that good classification and prediction requires good predictors. Even after feature selection, the number of genes retained is often large. LDA was used for the purpose of final feature selection and classification. A permutation test evaluated whether the specific classification of the individuals between groups is significantly better than random classification in any two arbitrary groups [39]. MASS and SMA R package were used for analysis in this section.

RESULTS

The data comprised 120 pairs of cancer and adjacent normal lung tissue specimens from nonsmoking female lung cancer patients admitted to National Taiwan University Hospital and Taichung Veterans General Hospital. As shown by Lu [36], the mean ± SD of the age of patients used for the microarray experiments was 61 ± 10 years. Most tumors were adenocarcinoma (93%) and 78% were in stage I or II. Discriminant analysis was used for classification and feature selection of lung cancer gene expression data. The important gene relationships warranting further biological investigation were identified using the PLS. Figure 1 shows that this method is properly fitted to the data and showed an R2 of about 80%. R2Y and Q2Y of the model were compared with the corresponding values obtained after random permutation of the y response in the top left figure. The top right graph is an inertia bar plot that suggests that the orthogonal components captured most of the inertia. The bottom left graph showed no important outlier in this graph. The bottom right plot of the x-score shows the number of components and the cumulative R2X, R2Y and Q2Y. These indices show that PLS was the proper approach to implement in this dataset.

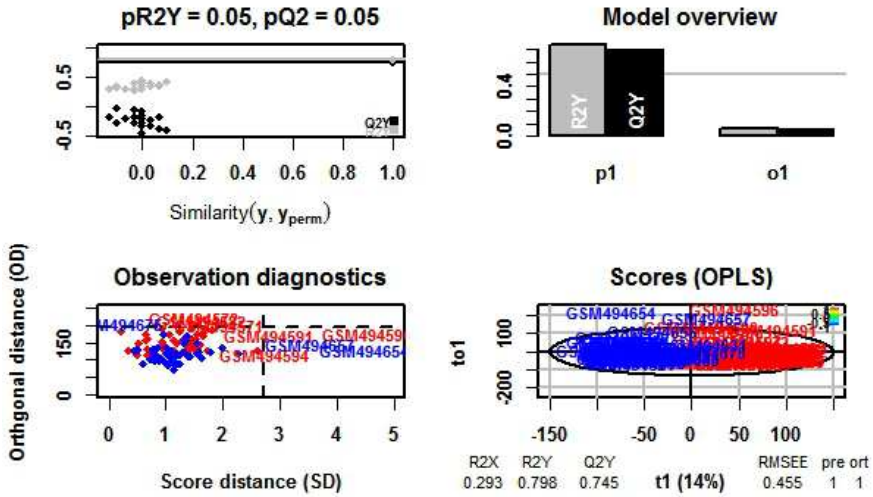

Figure 1. Score plot of OPLS-DA model of status of disease

Of the more than 50,000 genes, 238 show important changes in mRNA levels and have a significant VIP index. The rest of the genes were dropped from further analysis. Figure 2 shows the results.

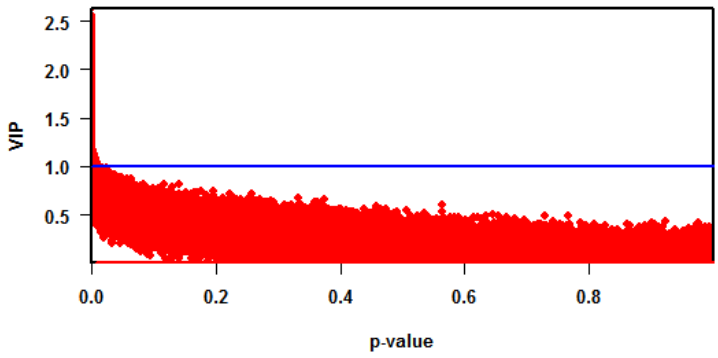

Figure 2. Level of VIP significance from one-predictive OPLS models

FDR is an important aspect of feature selection, especially in microarray data. The t-statistic filter can readily compute an expected FDR based on the p-value. Of the 214 genes selected for their contribution to classification, PCA analysis of the identified gene expression data showed that the first three components contained 92.7% of the variance. This indicates that the data can be summarized in just three gene expression features that explain most of the total variability observed. The linear discriminant function was obtained for proper classification of genes within groups (Figure 3). The accuracy of this method was 94.5% and can be said that the LDA classifier was powerful. Each class of genes is a group that are strongly co-expressed. These genes are expected to have the same function.

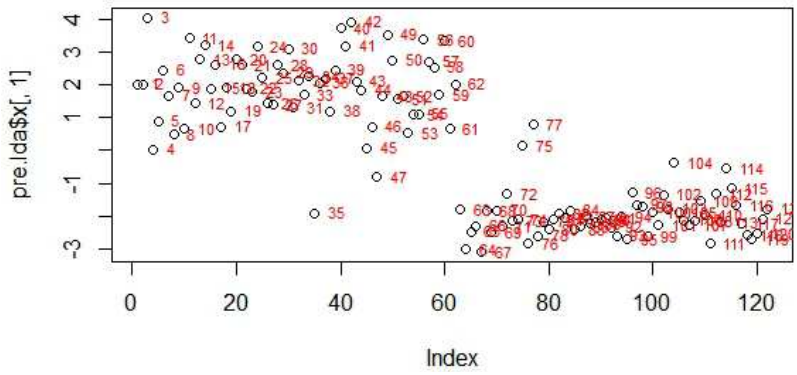

Figure 3: Linear discriminant function for classification of genes within groups

Biological process ontology is a function that measures these similarities and covers three domains: cellular component, molecular function and biological process. GO enrichment validation is a hypergeometric test for GO enrichment. This statistical test is significant if the genes in the biclusters are annotated with GO terms and are not specified by chance [40]. Table 1 shows the significant GO terms for the set of genes discovered by discriminant classification along with their p-values. The web tool David was used to evaluate the enrichment analysis of the discovered clusters [41, 42].

Table1: Gene Ontology and Enrichment Analysis for Discovered Genes

class	Ontology	Number of GO term	Percent of significant pvalue
tumor	Biological process	1309	72.0
	Molecular function	164	75.0
	Cellular component	420	95.0
normal	Biological process	1654	64.0
	Molecular function	683	33.0
	Cellular component	2850	97.3

DISCUSSION

The increasing clinical use of genomic profiling demands identification of more effective methods to segregate patients into prognostic and treatment groups [43]. The problem of reliable classification is important in many scientific areas. Linear discriminant analysis is the most effective method of cancer classification in high-dimensional prediction but has drawbacks for implementation with gene expression data. There are some common issues with this method in general. First critical issue with all of genes is not important for a lung cancer disease and so they are oblivious to errors in the data and affected on validity of results. The PLS method was implemented to avert this problem and to select important features. Second issue with at least some of the genes is correlated and Discriminant classification features should be independent; thus, PCA analysis was implemented for solve this problem. Discriminant analysis was performed to component that achieved by PCA.

To reduce error and guarantee the selected genes, after implementation of PLS, the identified genes were tested with FDR. A total of 214 genes were selected for final classification. Discriminant analysis based on the Fisher criterion were proposed for classification. The results show that the proposed classification method is appropriate. Insight into the role of a gene in a biological process may be gained by studying the biological functions of the top predictor genes. For lung cancer data, the findings are often consistent with current knowledge of the biological roles of the top predictor genes. Unfortunately, the exact biological role of some of these genes is not known.

This method can help identify gene relationships that warrant further biological investigation. Hofmann et al., [44] in study that was performed in 2006 were shown that, Out of 59,620 examined genes or ESTs, only 0.6% (n=344) were expressed significantly differentially between lung and tumor, which could be an indication that only few genes change their expression during the processes of tumorigenesis and metastases. The classification according to biological process gives insights into molecular changes occurring in tumor development and progression. In the present work, the largest part of the highly expressed genes in the tumor tissues was involved in the processes of cell growth. Nancy et al. [45] used the SVM method for feature selection in lung cancer microarray data with 1000 genes. The number of identifying features with different kernels was at least 246 genes and the results were similar to those of the present study. Wang and Gotoh [46] present a method for classification of cancer based on gene expression profiles using single genes. The principal advantage of the single-gene models is that the predication

procedures and results are understood with ease, because our models are based on rules.In the lung cancer dataset, when α=0.90, 56 genes are identified. One might doubt the utility of this model is so complexity of cancerous pathogenesis that it should be connected with many genes. Shi et al. [47] performed feature selection on lung cancer microarray data and the number of identified features was at least 2961 genes. Their study used a simple linear model and ignored complicated relationships in the data, which was likely the reason that the number of genes identified were high and different those found by Nancy et al. and the present study. Li and Xiong [48] introduced classification method combining fisher's linear discriminant analysis and feature selection based on a stepwise optimization process for tumor classification. The results demonstrated that this method achieved high classification accuracy. Measuring the microarray gene expression data of the cancer and applying an efficient models and algorithms for analysis them may lead to improvements in diagnostics and therapy decisions.

Acknowledgement
This study was sponsored by Nosocomial Infection Research Center, Kermanshah University of Medical Sciences, Kermanshah, Iran.

Conflict of interest
There is no conflict of interest in this study.

REFERENCES

[1] Johnson C, Warmoes MO, Shen X, Locasale JW. Epigenetics and cancer metabolism. Cancer Letters. 2013; 28;356(2 Pt A):309-14.
[2] Bishak YK, Payahoo L, Osatdrahimi A, Nourazarian A. Mechanisms of Cadmium Carcinogenicity in the Gastrointestinal Tract. Asian Pacific J Cancer Prev. 2015;16(1):9–21.
[3] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet. Nature Publishing Group; 2000;24(3):227–35.
[4] Karim-Kos HE, de Vries E, Soerjomataram I, Lemmens V, Siesling S, Coebergh JW: Recent trends of cancer in Europe: A combined approach of incidence, survival and mortality for 17 cancer sites since the 1990s. Eur J Cancer. 2008; 44: 1345-1389.
[5] Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA: Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc. 2008; 83:584-594.
[6] Tyczynski JE, Bray F, Aareleid T, Dalmas M, Kurtinaitis J, Plesko I, Pompe-Kirn V, Stengrevics A, Parkin DM: Lung cancer mortality patterns in selected Central, Eastern and Southern European countries. Int J Cancer. 2004; 109: 598-610.
[7] Janssen-Heijnen ML, Coebergh JW: The changing epidemiology of lung cancer in Europe. Lung Cancer. 2003; 41: 245-58.
[8] Injuries violence and disabilities biennial report 2004-2005. Switzerland: WHO; 2006. World Health Organization.
[9] Vahedi M, Pourhoseingholi MA, Baghestani AR, Abadi A, Sobhi S, Fazeli Z., Bayesian Analysis of Lung Cancer Mortality in the Presence of Misclassification, Iran J Cancer Prev 2013; 1(Suppl.):1 -5.
[10] Howlader N, Noone AM, Krapcho M, Neyman N, Aminou R, Altekruse SF, et al. (eds). SEER cancer statistics review, 1975–2009 (Vintage 2009 Populations). Bethesda, MD: National Cancer Institute [cited 2012 Aug 13].
[11] Keedy VL, Temin S, Somerfield MR, Beasley MB, Johnson DH, McShane LM, et al. American Society of Clinical Oncology provisional clinical opinion: epidermal growth factor receptor (EGFR) mutation testing for patients with advanced non-small-cell lung cancer considering first-line EGFR tyrosine kinase inhibitor therapy. J Clin Oncol 2011;29:2121–7.
[12] Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in nonsmall-cell lung cancer. Nature 2007;448:561 –6.
[13] Javad Shahidi , Ali Taghizadeh-Kermani2, Mahmood Reza Gohari, Mohammad Reza Ghavamnasiri, Fahimeh Khoshroo, Leila Pourali5, S. Robin Cohen, Changes in Daily Activities of Cancer Patients after Diagnosis: How Do Canadian and Iranian Patients Perceive the Change, Iranian Journal of Cancer Prevention, 2014; 1:28-34.
[14] Sterner-Kock A, Thorey Is, Koli K, Wempe F, Otte J, Bangsow T, Kuhlmeier K, Kirchner T, Jin S, Keski-Oja J, Von Melchner H. Disruption of the gene encoding the latent transforming growth factor-beta binding protein 4 (LTBP-4) causes abnormal lung development, cardiomyopathy, and colorectal cancer. Genes Dev 2002; 16: 2264-2273.
[15] The Cell: A Molecular Approach. 2nd edition. Cooper GM. Sunderland (MA): Sinauer Associates; 2000.
[16] K. L. Jae, "Analysis issues for gene expression array data," Clinical Chemistry 2001; 47(8): 1350–1352.
[17] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically signifiant biclusters in gene expression data," Bioinformatics, vol. 18, no. 1, pp. S136–S144, 2002.

[18] Akay, M. F. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications 2009; 36, 3240–3247.
[19] Fang Z, Yang J, Li XY, Luo QM, Liu L. Knowledge guided analysis of microarray data. J Biomed Inform 2006;36:401-11.
[20] Wong H-S, Wang H-Q. Constructing the gene regulation-level representation of microarray data for cancer classification. J Biomed Inform 2008; 41:95-105.
[21] Wang H-Q, Wong H-S, Zhu HL, Yip TTC, A neural network-based biomarker association information extraction approach for cancer classification. J Biomed Inform 2009;42:654-66.
[22] Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics 2001;17:763-74.
[23] Liu JJ, Cai WS, Shao XG. Cancer classification based on microarray gene expression data using a principal component accumulation method. Sci China 2011;54:802-11.
[24] Li LP, Weinberg CR, Thomas AD,Pedersen LG. Gene selsction for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001;17:1131-42.
[25] Bushel PR, Hamadeh HK, Bennett L, Green J, Ableson A, Misener S, et al., Computional selection of distinct class and subclass specific gene expression signatures, J Biomed Inform 2002;35:160-70.
[26] Pan F, Wang BY, Hu X, Perrizo W. Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis. J Biomed Inform 2004;37:240-8.
[27] Blanco R, Inza M, Merino H, Quiroga J, Larranaga P. Feature selection in Bayesian classification for the prognosis of survival of cirrhotic patients treated with TIPS. J Biomed Inform 2005;38:376-88.
[28] Lutz U, Lutz RW, Lutz WK, Metabolic profiling of glucuronides in human urine by LC-MS/MS and partial least square discriminant analysis for classification and prediction of gender. Anal chem 2006;78:4564-71.
[29] Peng YH, A novel ensemble machine learning for robust microarray data classification. Comput Biol med 2006;36:553-73.
[30] Lancashire LJ, Lemetre C, Ball GR: An introduction to artificial neural networks in bioinformatics – application to complex microarray and mass spectrometry datasets in cancer studies. Brief Bioinform. 2009, 10: 315-329.
[31] Liao JG, Chin KV: Logistic regression for disease classification using microarray data: model selection in a large p and small n case. Bioinformatics. 2007, 23: 1945-1951.
[32] Tusher VG, Tibshirani R and Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001, 98:5116–5121.
[33] Guo Y, Hastie T and Tibshirani R: Regularized linear discriminant analysis and its application in microarrays. Biostatistics 2005, 8:86–100.
[34] Schäfer J and Strimmer K: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 2005, 4.
[35] Chandra, B., & Gupta, M. (2011). An efficient statistical feature selection approach for classification of gene expression data. Journal of Biomedical Informatics, 44, 529–535.
[36] Lu TP, Tsai MH, Lee JM, Hsu CP et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. Cancer Epidemiol Biomarkers Prev 2010 Oct;19(10):2590-7.
[37] Raamsdonk L, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh M, Berden J, Brindle K, Kell D, Rowland J, et al.: A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nature Biotechnology 2001, 19: 45–50.
[38] Bylesjö M, Rantalainen M, Cloarec O, Nicholson J, Holmes E, Trygg J: OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J Chemometrics 2006, 20: 341–351.
[39] Moore, D. S., G. McCabe, W. Duckworth, and S. Sclove (2003): Bootstrap Methods and Permutation Tests.
[40] Al-Akwa FM, Kadah YM (2009). An Automatic Gene Ontology Software Tool for Bicluster and Cluster Comparisons, IEEE ,163-7.
[41] Sherman B.T., Tan Q., Guo Y., Bour S., Liu D., Stephens R., Baseler M.W., Lane H.C., Lempicki R.A. (2007), DAVID knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics,8,426.
[42] Alavi Majd, H., Shahsavari, S., Baghestani, A. R., Tabatabaei, S. M., Khadem Bashi, N., Rezaei Tavirani, M., & Hamidpour, M. Evaluation of Plaid Models in Biclustering of Gene Expression Data 2016. *Scientifica*.
[43] Alavi Majd H, Baghestani AR, Tabatabaei S.M, Shahsavari S, Rezaei Tavirani M, Hamidpour M, Application of Plaid Algorithm to Identifying Patterns in Breast Cancer Gene Expression Data, International Journal of Scientific & Engineering Research 2015, Volume 6, Issue 9.
[44] Hofmann, Hans-Stefan, et al. "Identification and classification of differentially expressed genes in non-small cell lung cancer by expression profiling on a global human 59.620-element oligonucleotide array." Oncology reports 16.3 (2006): 587-596.
[45] Nancy S.G, Balamurugan A, A Comparative Study of Feature Selection Methods for Cancer Classification using Gene Expression Dataset, Journal of Computer Applications (JCA) 2013, Volume VI, Issue 3.

[46] Wang X1, Gotoh O. Cancer classification using single genes. Genome Inform 2009;23(1):179-88.
[47] Shi W, Liu K, Xu S, Zhang J, Yu L, Xu K, Zhang T, Gene expression analysis of lung cancer, European Review for Medical and Pharmacological Sciences 2014; 1 8: 21 7-228.
[48] D. Wuju, Li, and Xiong Momiao. "Tclass: tumor classification system based on gene expression profile." Bioinformatics 2002; 18(2): 325-326.