Received: 1/16/2004 **Accepted: 8/1/2005**

Original Article

Application of Artificial Neural Networks in Cancer Classification and Diagnosis Prediction of a Subtype of Lymphoma Based on Gene Expression Profile

L. Ziaei MS, A. R. Mehri PhD**, M. Salehi PhD****

ABSTRACT

Background: Diffuse Large B-cell Lymphoma (DLBCL) is the most common subtype of non-Hodgkin's Lymphoma. DLBCL patients have different survivals after diagnosis. 40% of patients respond well to current therapy and have prolonged survival, whereas the remainders survive less than 5 years. In this study, we have applied artificial neural network to classify patients with DLBCL on the basis of their gene expression profiles. Finally, we have attempted to extract a number of genes that their differential expression were significant in DLBCL subtypes.

Methods: We studied 40 patients and 4026 genes. In this study, genes were ranked based on their signal to noise (S/N) ratios. After selecting a suitable threshold, some of them whose ratios were less than the threshold were removed. Then we used PCA for more reducing and Perceptron neural network for classification of these patients. We extracted some appropriate genes based on their prediction ability.

Results: We considered various targets for patients classifying. Thus patients were classified based on their 5 years survival with accuracy of 93%, in regard to Alizadeh et al study results with accuracy of 100%, and regarding with their International Prognosis Index (IPI) with accuracy of 89%.

Conclusion: Combination of PCA and S/N ratio is an effective method for the reduction of the dimension and neural network is a robust tool for classification of patients according to their gene expression profile.

Keywords: classification, gene expression, DLBCL, neural network, Perceptron

JRMS 2006; 11(1): 13-17

Tumors are currently diagnosed by pathology and immunohistochemistry techniques based on their morphology and protein expression, respectively. But the histological appearance of tumor cannot reveal the underlying genetic disorders or biological process that contributes to malignant process. Therefore mentioned methods are inaccurate for cancer classification [1].

The recent advent of microarray technology has allowed the simultaneous monitoring of thousands of genes, which has motivated the development in cancer classification, using gene expression. The gene expression data is very different from the data produced by all other previous methods. At first, it has very high dimensionality, usually contains thousands of genes. Secondly, the available data size is very small, some have sizes below 100. Thirdly, most genes are irrelevant to cancer distinction. It is obvious that traditional existing classification methods were not designed to handle this kind of data efficiently and effectively [2].

*Student of Biomedical Engineering, Department of Biomedical Physics and Engineering, Medical School, Isfahan University of Medical Sciences, Isfahan, Iran.
**Assistant Professor, Department of Biomedical Physics and Engineering, Medical School, Isfahan University of Medical Sciences, Isfahan, Iran.
*** Assistant Professor, Department of Genetic and Molecular Biology, Medical School, Isfahan University of Medical Sciences, Isfahan, Iran.
Correspondence to: Lia Ziaei, Department Medical Physics and Engineering, Medical School, Isfahan University of Medical Sciences, Isfahan, Iran. E-mail: lia57us@yahoo.com

Some researchers proposed that gene selection prior to cancer classification could improve its accuracy. Performing gene selection helps to reduce data size, thus improving the running time [3]. Two main categories of data selection methods have been used to analyze gene expression data: Clustering and Classification.

Clustering is the organization of a collection of unlabeled patterns (data vectors) into clusters, based on their similarity. So that the patterns within the same cluster are more similar to each other than those belonging to different clusters. Clustering is important for exploratory data analysis, since it will often reveal interesting structures of data, thereby allowes formulation of useful hypotheses to test [4].

Clustering techniques are divided into groups of: 1- methods in which the number of classes are preidentified, such as K-means and SOM (Self Organization Map) methods, and 2- methods in which the number of classes aren't preidentified, such as hierarchical method [5]. In contrast with clustering, classification techniques start with a collection of labeled (preclassified) expression patterns. The goal is to train a classification model that would be able to classify a new expression pattern. Classification has also been extensively used to distinguish (classify) different samples [5].

Artificial neural network is a robust tool recently used as either clustering or classification. Supervised models are used for classification and unsupervised models are used for clustering [6].

Diffuse Large B-cell Lymphoma (DLBCL), the most common subtype of non-Hodgkin's Lymphoma, is clinically heterogeneous. 40% of patients have higher overall survivals than the others [7]. Alizadeh et al (2000) showed that there is a diversity in gene expression among tumors of DLBCL patients, apparently reflecting the variation in tumor proliferation rate, host response and differentiation state of the tumor [7]. O'Neill and Song (2003) classified these DLBCL patients using 2 layers neural network with accuracy of 100% [8]. Lossos et al (2004) studied 36 genes whose expression had been reported to predict survival in diffuse large B_cell lymphoma of 66 patients. They showed that measurement of the expression of 6 genes is sufficient to predict overall survival in diffuse large B_cell lymphoma [9].

The above-mentioned study [7], using gene expression profile, identified two molecularly distinct forms of DLBCL: Germinal center B-like DLBCL and Activated B-like DLBCL. In this study we are going to use gene expression data of these DLBCL patients [7] to differentiate between two forms of DLBCL, using supervised neural network.

Data and Methods

The data presented in first figure of the Alizadeh et al report (http://llmpp.nih.gov/lymphoma/data.shtml) [7] were used in this study. These data were from 42 patients and corresponded with 4026 genes expression levels. At first, data was divided into 2 groups with survival of less and more than 4 years.

The data analysis consisted of the following steps:

1. Initial cut
2. Principal component analysis
3. Artificial neural network (supervised regression model)
4. Extraction of relevant genes

1. Initial cut

Initial cut was performed using Signal to Noise (S/N) ratio as follows:

$S/N = (\mu_A - \mu_B) / (\mu_{A+}\mu_B)$

μ and σ are mean and standard deviation per class, respectively [10].

This ratio is just usable in two class problems.

The genes based on their S/N ratios, were ranked and after selecting suitable threshold, the genes whose S/N ratios were less than threshold (0.2) were removed.

2. Principal component analysis

To allow for supervised regression model with no over-training (i.e. low number of samples compared to the number of genes) we reduced the dimensionality of the samples using PCA.Thus each sample was represented by 10 features, which were the results of projection of gene expression using the 10 dominant PCA eigenvectors, which correspond with greatest 10 eigenvalues [5, 11].

3. *Artificial neural network*

For prediction, we employed an artificial neural network (ANN) classifier. Due to the limited amount of training data and the fact that one output was needed, we limited ourselves to linear Perceptron (LP) with 10 input nodes representing the PCA components described above. Using more than 10 components did not improve the classification of the samples. Classification steps are as bellow:

At first, the 40 labeled samples were randomly split into 3 equal groups. 2 groups [26] were used as trained and one group [14] was used as test. Since there were not enough samples available, we performed a leave one out cross validation on 26 training samples in which one sample is held, a predictor is trained on the remaining samples, the left sample is classified by this predictor, and the process is repeated iteratively. In this method, 26 networks were trained. In the final step, these 26 models were tested on 14 blinded test samples. Then we used average committee vote to classify these 14 test samples. It means that the outputs of 26 networks were averaged on each of the 14 test samples and this average forced to 0 or 1.

Results

At first, initial cut applied on data and the number of genes reduced from 4026 to 1547.

This method was run using various numbers of eigenvectors. The accuracy of using 14 eigenvectors was 93% (i.e. 1 out of 14 was misclassified). In the next experiment we changed sample labels and used the results of Alizadeh et al study. It means that patients were divided into 2 groups: germinal center and activated like B-cell lymphoma. Indeed, we altered the targets of the networks, using 10 eigenvectors with accuracy of classification as 100%.

Identifying significant genes was another aim of this study. These genes were relevant to some special biological processes. Furthermore, specifying significant genes and training the network using just these genes, increased classification accuracy, and helped to decrease complicated computation and running time.

Next, we removed PCA and used various most significant subsets of genes that were ranked according to S/N ratio. Using 14 most significant genes, accuracy of classification was 100% (Table 1), but in some permutation of training samples, the accuracy was 93%. However, in those permutations using PCA, error was 0%. In another experiment, 20 trained samples and 20 test samples were studied. In this situation, the error was 20%, without using PCA, and 0%, with using it. Thus we concluded that the PCA is a robust tool for reduction of dimensions.

In the final experiment, we considered another target for neural network and then classified patients based on their IPI (International Prognosis Index); data was divided into 2 groups with IPI, of less and more than 2. This method was run on samples, but our results demonstrated that the error wasn't 0%.

Discussion

The first generation of gene expression analysis methods has been successfully applied in a variety of clustering and classificating settings. Alizadeh et al (2000) used hierarchical clustering to divide patterns into two subgroups [7]. Michael C O'Neill and Li Song used two layers neural network for classification of DLBCL patients. Their classification accuracy was 100% and they were able to extract 34 significant genes (Table 2), but they did not claim that the gene sets extracted in their procedure were the "best" gene sets [8].

The 14 genes that were extracted in our study have no overlap with the genes extracted in O'Neill and Song study.

We examined 34 genes in our program and the error was 7% (one out of fourteen was misclassified). So, we have not any claim that our extracted genes are the best sets.

Table 1. Most significant genes extracted in this study.

1	19289	(Unknown UG Hs.169565 ESTs, Moderately similar to ALU SUBFAMILY SB WARNING ENTRY [H.sapiens]; Clone=825217)
2	13394	(Unknown UG Hs.120716 ESTs; Clone=1334260)
3	19288	*Unknown; Clone=825199
4	13812	(Unknown UG Hs.224323 ESTs, Moderately similar to alternatively spliced product using exon 13A [H.sapiens]; Clone=1338448)
5	19274	(Unknown UG Hs.136345 ESTs; Clone=746300)
6	14422	*Unknown; Clone=1353015
7	14423	*Unknown; Clone=1353041
8	19365	*Unknown UG Hs.105261 EST; Clone=824088
9	16886	*JAW1=lymphoid-restricted membrane protein; Clone=417502
10	19321	*JAW1=lymphoid-restricted membrane protein; Clone=815539
11	14963	*Unknown UG Hs.124922 ESTs; Clone=1358244
12	18436	*Deoxycytidylate deaminase; Clone=1302032
13	20585	(Unknown UG Hs.208410 EST, Moderately similar to ALU SUBFAMILY SB WARNING ENTRY !!!!
14	14671	(Unknown UG Hs.169081 ets variant gene 6 (TEL oncogene); Clone=1355435)

Table 2. Most significant genes extracted in O'Neill and Song study [8].

1	14706	Unknown Hs.180836	18	17856	Interferon alfa/beta receptor-2
2	21367	Unknown Hs.134746	19	21653	Unknown Hs.1510936
3	13601	Similar to high mobility group	20	15656	Unknown
4	20397	FBPI FUSE binding protein I	21	14393	Unknown Hs.29205
5	17901	*Pre-pro-orphanin	22	16631	Adenosine kinase
6	13097	Unknown	23	13318	Unknown Hs.122428
7	14560	Unknown Hs.32533	24	18330	Topoisomerase II beta
8	13867	Unknown	25	14983	Unknown
9	15664	Unknown	26	17721	IdI inhibitor of DNA binding I
10	20490	Unknown Hs.122407	27	16850	PM5 protein=homology to collagenase
11	13650	Unknown	28	20481	Unknown Hs.37629
12	18252	Myosin_IC	29	17398	Receptor r_IBB ligand
13	16886	JAWI	30	14772	Unknown
14	18593	Receptor protein_tyrosin kinase	31	19280	BENE
15	20759	Unknown Hs.33053	32	21603	Unknown Hs.33431
16	17802	Thymosin beta_4	33	19258	Tre-2
17	17887	A-rafc-raf-I kinase	34	21091	Unknown Hs.199250

Lossos et al (2004) studied 36 genes from 66 patients with invariant accuracy analysis; genes were ranked based on their ability to predict the survival. In their study, Kaplan-Merier method was used for classification. They showed that measurement of the expression of six genes: LMO_2, BCL_6, $CCND_2$, $SCYA_3$, and BCL_2 is sufficient to predict overall survival in diffuse large B-cell lymphoma [9].

Our results clearly demonstrated that the combination of S/N ratio and PCA is a suitable method for reducting dimension and a simple neural network was a nearly perfect tool for this classification.

Acknowledgment

The authors would like to thank Alizadeh et al. for preparing Lymphoma database.

References

1. *Khan J, Wei JS, Ringner M, Saal LH, Landanyi M, Wetermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7(6):673-9.*
2. *Lossos IS, Morgensztern D. Non-Hodgkin's lymphoma in the microarray era. Clin Lymphoma. 2004;5(2):128-9.*
3. *Lu Y, Han J. Cancer Classification using gene expression data. Information System. 2003;28(4): 243-68.*
4. *Poulsen CB, Borup R, Nielsen FC, Borregaard N, Hansen M, Gronbaek K, et al. Microarray-based classification of diffuse large B-cell lymphoma. Eur J Haematol. 2005;74(6):453-65.*
5. *Caustion HC, Quachenbush J, Brazma A. Microarray Gene Expression Data Analysis: A Beginner's Guide. Oxford UK: Blackwell Publishing; 2003. p.230-1.*
6. *Zini G, d'Onofrio G. Neural network in hematopoietic malignancies. Clin Chim Acta. 2003;333(2):195-201.*
7. *Alizadeh A, Eisen MB, Davis RE, Ma C, Losses IS, Resenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403 (6769):503-11.*
8. *O' Neill MC, Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. BMC Bioinformatics. 2003;4:13.*
9. *Lossos IS, Czerwinski DK, Alizadeh A, Wechser MA, Tibshirani R, Botstein D, et al. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. N Engl J Med. 2004;350(18):1828-37.*
10. *Gloub TR, Slonim DK, Jamayo 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):531-7.*
11. *Orango CA, Jones DT, Thornton JM, editors. Bioinformatics genes, proteins & computers. Oxford UK: BIOS Scientific Publishers; 2003. p.190-4.*