Oncogenes and Subtypes of Diffuse Large B-Cell Lymphoma Discoveries from Microarray Database

Ching-Hao Lai[1] Jun-Dong Chang[2] Meng-Hsiun Tsai[3*]

[1]Institute of Computer Science, National Chung Hsing University, 250, Kuo-Kuang Road, Taichung, Taiwan 402, R.O.C. (e-mail: harddrive1kimo@yahoo.com.tw)

[2]Graduate School of Computer Science and Information Technology, National Taichung Institute of Technology, 129 Sec. 3, San-min Road, Taichung, Taiwan 404, R.O.C. (e-mail: s18943102@ntit.edu.tw)

[3]Department of Management Information Systems, National Chung Hsing University, 250, Kuo-Kuang Road, Taichung, Taiwan 402, R.O.C. (e-mail: mht@nchu.edu.tw)

Abstract

This paper presents an effective analysis scheme for Diffuse Large B-Cell Lymphoma (DLBCL) microarray datasets. Analysis of variable (ANOVA) is a well known statistics tools. It is useful to get the oncogenes to distinguish the normal and cancerous tissues. But, it can not further obtain the sub-types of cancerous tissues effectively. Hierarchical clustering is a well known analysis method for data mining. Therefore, it is also useful and fit to classify oncogenes to obtain some sub-types. ANOVA and hierarchical clustering both are employed to help us analyze B-cell Lymphoma datasets. In our analysis results, ANOVA can obtain 11 oncogenes of DLBCL from Stanford DLBCL microarray database successfully and accurately. Then, the 11 oncogenes are used for hierarchical clustering to identify the sub-types of cancerous tissues. In our hierarchical clustering analysis, we use 20 GC B-like DLBCL and 15 Activated B-like DLBCL actual samples used for analyzing. The analysis result shows that the hierarchical clustering can distinguish GC B-like DLBCL and Activated B-like DLBCL samples successfully.

Keywords: Microarray, Analysis of Variance (ANOVA), hierarchical clustering, Diffuse Large B-Cell Lymphoma (DLBCL), data mining.

1. Introduction

In recent years, cancer has become the most general deathful disease for people. Hence, how to predicate and diagnose cancer is a very important work for saving many people's lives. Microarray data analysis is a useful and accuracy tool for doctors to diagnose cancer by gene expression [1][5-9]. Because of most microarray data include too many genes and few of these genes are useful for diagnosing cancer. These useful genes that must have high difference between the normal and cancerous cases are called oncogene. In order to obtain these genes from a huge gene data, some analyses must bhe performed to achieve it. In the other hand, some cancer can be classified into more than one type. Hence, it is also important to further distinguish the sub-types of all oncogenes. In this paper, we use analysis of variable (ANOVA) [2] to get the useful oncogenes, and then employee hierarchical clustering to get the sub-types.

In this paper, we download the Stanford Diffuse Large B-Cell Lymphoma (DLBCL) microarray database (it is downloaded from the website: [12]) to be used in analyses. This database includes 18432 genes and 38 samples. The samples are too few but the genes are too many to analyze in this database. It is a challenge for us to obtain the oncogenes and sub-types, but in our analyses, we have conquered this problem to get better analysis results [6][9].

Ash A. Alizadeh et al. [9] analyzed gene expression patterns to distinguish the normal and cancerous tissues for DLBCL microarray database. But they did not further classify GC B-like DLBCL and Activated B-like DLBCL. In this paper, a useful analysis scheme is proposed to further distinguish the sub-types of cancerous tissues for DLBCL.

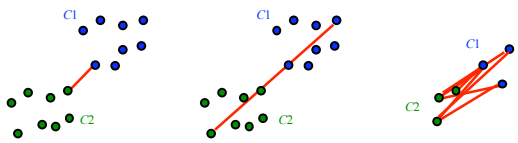
Generally speaking, the clustering algorithms include the hierarchical clustering and the nonhierarchical clustering [4]. The hierarchical clustering supports a nested partition order which is called dendrogram. The hierarchical clustering can be agglomerative or divisive. The agglomerative analysis looks all data as different clusters. These clusters are merged to product a nested order. The divisive analysis looks all data as the same cluster and splits it. The hierarchical clustering not only can present data practically but also can reduce the affection of the initialization and local minima. In addition, the number of the clusters doesn't need to assign beforehand. The user can decide a correct threshold to

let the number of the clusters is correct and the clusters all are accurate. There still are several disadvantages in the hierarchical clustering. In the merging or dividing phase, it only considers the local neighbors and doesn't know the sharp and the size of the clusters. In addition, because the hierarchical analysis is static, when the data has been clustered already, it can't be clustered any more. The agglomerative hierarchical clustering is used in this paper. The general distance methods between each member are Euclidean, Seuclidean, Mahalanobis, Cityblock, Minkowski, Cosine and Correlation distance; the general linkage methods between each cluster are single, complete and average linkage. The single linkage is the smallest distance of all components between two clusters. It can be obtained by equation (1) and shown as Fig. 1(a). The complete linkage is the largest distance of all components between two clusters. It can be obtained by equation (2) and shown as Fig. 1(b). The average linkage is the average distance of all components between two clusters. It can be obtained by equation (3) and shown as Fig. 1(c).

$$SD(C1, C2) = \mathrm{Min}(d(i, j)), i \in C1, j \in C2. \quad (1)$$

$$CD(C1, C2) = \mathrm{Max}(d(i, j)), i \in C1, j \in C2. \quad (2)$$

$$AD(C1, C2) = \mathrm{Avg}(d(i, j)), i \in C1, j \in C2. \quad (3)$$



(a) (b) (c)

Fig. 1. The single, complete and average linkage between $C1$ and $C2$.

2. The analysis scheme

In Taiwan, the microarray chips are very expensive hence it is very difficult to obtain the microarray data. Nowadays, some research organizations provide and publicize some microarray database on their websites. In this paper, our analyses focus on a public Diffuse Large B-Cell Lymphoma (DLBCL) database [12][5]. This database includes 38 microarray samples and 18432 genes. The microarray samples can be classified in normal, GC B-like DLBCL, and Activated B-like DLBCL classes. In this paper, we do not only get the oncogenes of DLBCL but also discover the subtypes of cancerous tissues.

In the DLBCL microarray database, the number of DLBCL oncogenes may be little, but the number of genes in whole database is very great (18432 genes). So it is difficult to use a standard and fixed method to perform analyzing. In this paper, a linear regression analysis and analysis of variance (ANOVA) is used to find oncogenes. First, all data are performed with using linear regression analysis to test the difference of each gene between normal and cancerous samples. If the genes have greater distance between the two kinds of samples, they will be remained to analyze with using ANOVA. The linear regression analysis is a data selection strategy in our analyses, it is very useful to reduce the number of genes and remain fewer and important genes. In our linear regression analysis, the coordination (x, y) of certain gene are defined as:

$$x = CH2I - CH2B; \quad y = CH1I - CH1B. \quad (4)$$

where $CH1I$ is the total strength of certain gene in cancerous tissues, $CH1B$ is the background value of cancerous tissues, $CH2I$ is the total strength of certain gene in normal tissues and $CH2B$ is the background value of cancerous tissues [7][10]; x and y are the values on x-axis and y-axis for the certain gene in the linear regression analysis. When all data are analyzed, we can obtain a regression line to represent the trend of most data. The outliers have greater distance between the regression line that must be analyzed to obtain the oncogenes. Our database includes 38 samples, so 38 linear regressions are performed to get 38 regression lines. In each linear regression, 18432 residuals can be obtained for 18432 genes, and these genes will be sorted by their residuals. The genes that have greater residuals are more possible to be outliers, and they are needed to be analyzed in this paper. In equation (5), S is the slope of certain gene in regression distribution. It can be used to present the different expression of a gene between normal and cancerous tissues. When S of the certain gene is greater or fewer than 1, the gene possibly has more different expression between normal and cancerous tissues.

$$S = \frac{CH1I - CH1B}{CH2I - CH2B}. \quad (5)$$

After our linear regression analysis, the genes will be further analyzed in an ANOVA. In our analyses, The ANOVA is used to measure and test the difference of the 200 genes between normal and cancerous tissues. An ANOVA is a statistical test method also called F test, it is similar to t test. T test is used to test the difference between two small populations, but an ANOVA can be used to test more than two small populations [11]. In our analyses, we want to get the oncogenes which are different from normal, GC B-like DLBCL and Activated B-like DLBCL samples. Hence, an ANOVA is more adaptive for our analyses and the input data of the ANOVA are the genes obtained by the linear regression. The ANOVA is useful to help us obtain fewer and more accuracy genes to be the DLBCL oncogenes of our analyses.

In our analyses, the samples of the database can be classified into normal, GC B-like DLBCL and

Activated B-like DLBCL samples. After the ANOVA, we can obtain the DLBCL oncogenes of cancerous tissues, but the ANOVA can not present these oncogenes are belonged to GC B-like DLBCL or Activated B-like DLBCL. In order to further discover the unknown subtype of cancerous tissues subtypes from the oncogenes of our analyses, a hierarchical clustering is employed to achieve this work. In a hierachical clustering, all distance and linkage parameters are performed in some individual clustering procedures, the results of each clustering procedure with different distance and linkage parameter will be analyzed and compared to choose the best parameter. The results of our whole analyses are shown in Section 3.

3. Data Analyses and Results

After our linear regression analysis, we obtain 18432 residuals in each sample of the microarray database. In other words, each gene can get 38 residuals from 38 linear regressions analyzing for all samples. Because of we want to obtain the genes with greater residuals. In order to measure the summary residuals of all genes in 38 linear regressions, each gene must compute its total residual by 38 residual. After this computing, all genes are sorted by their total residuals. Some genes with their 38 residuals and total residuals are shown in Table I, and the top 10 genes are shown in Table II. In our analyses, we pick the genes that have 100 smallest and biggest total residuals to be the input data of the ANOVA. In order to show the analysis results more conveniently, we define three symbols to represent the normal, GC B-like DLBCL, and Activated B-like DLBCL classes, and they are defined as N, G, and A. In most cases, the confidence usually is set to 99% (P-value is 0.001), but in our analyses, we also perform the case with setting the confidence to 99.9% (P-value is 0.0001). After the analyses, we find the results of the two cases are the same, and we find 11 most significant difference genes in N vs. G and N vs. A. In G vs. A, we list their number and name in Table III. According to the results, we find the gene 20, 21, 22 and 24 are the same gene. In the past DLBCL microarray analysis researches, the gene CD44 was discovered to be the oncogene and also can distinguish G and A. Excepting the gene CD44, our analyses discover 7 new oncogenes for DLBCL. They all can powerfully identify the subtypes of DLBCL.

We can find the normal tissues and the cancerous tissues which can be identified in the previous experimental results but the classification for the subtype of cancerous tissues is still difficult. A hierarchical clustering for the subtype classification is presented in this section.

In this analysis, the dataset includes 20 GC B-like DLBCL and 15 Activated B-like DLBCL actual samples used for analyzing. The 11 specific genes are used to be the variables in our hierarchical clustering. The analysis results show that to use Seuclidean and average linkage can get the best clustering result and the result is shown in Fig. 2. In Fig. 2, the x-axis represents the 35 actual samples and the y-axis is the linkage distance of each sample in hierarchical clustering. The samples 1 to 20 are GC B-link DLBCL samples, and the samples 21 to 35 are Activated B-like DLBCL samples. Fig. 2 shows that 14 GC B-like DLBCL and 4 Activated B-like DLBCL samples can be classified in our hierarchical clustering analysis successfully.

Table I. The 38 residuals and the total residuals of several genes.

Sample number / Gene number	1	2	…	37	38	Total Residual
1	157.23	45.63	…	89.81	20.13	1029.17
2.	-2.59	39.62	…	-111.49	2.39	614.26
…	…	…	…	…	…	…
18431	-364.96	2644.09	…	-1198.56	-1831.91	12569.17
18432	180.68	-178.97	…	-153.63	3020.63	1477.82

Table II. The top 10 genes in linear regression analysis.

Gene Number	Gene name
1	Ferritin light chain
2	beta-2-microglobulin
3	B-actin,1314-1660
4	B-actin,1314-1660
5	immunoglobulin kappa light chain
6	invariant chain=Ia-associated invariant gamma-chain
7	cathepsin B
8	Fibronectin 1

9	B-actin,1099-1372
10	manic fringe

Table III. The significant different genes between GC B-link DLBCL and Activated B-like DLBCL.

Gene number	Gene name
20	Mig=Humig=chemokine targeting T cells
21	Mig=Humig=chemokine targeting T cells
23	Mig=Humig=chemokine targeting T cells
24	Mig=Humig=chemokine targeting T cells
31	MHC Class I=HLA-A2
50	cathepsin B
64	KIAA0279
93	PARP = poly (ADP-ribose) polymerase
109	Immunoglobulin mu
110	BLC=BCA-1=B lymphocyte chemoattractant BLC=CXC chemokine
194	CD44=Pgp-1=extracellular matrix receptor-III=Hyaluronate receptor

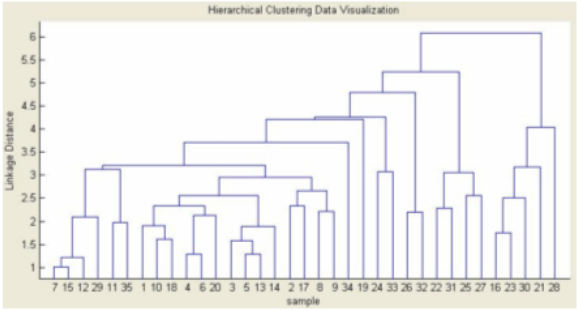


Fig. 2. The best clustering result with Seuclidean and average linkage.

4. Conclusions

In this paper, the linear regression analysis is employed to reduce the number of genes in each sample. The analysis results confirm that linear regression analysis is very helpful to remove most unimportant genes. In the other hand, the ANOVA is implemented to get 11 oncogenes, and 7 of them are the new discovery for previous researches. Finally, the hierarchical clustering is used to find the subtypes of DLBCL successfully. It is very useful for distinguishing GC B-like DLBCL and Activated B-like DLBCL. Besides, the DLBCL subtypes findings are less to present in other researches.

5. References

[1] M. Jean, "DNA Arrays Reveal Cancer in Its Many Forms," Science, Vol. 289, No. 5485, 2000, pp. 1670-1672

[2] R. V. Hogg and J. Ledolter, Engineering Statistics, MacMillan Publishing Company, 1987.

[3] Seber, G.A.F., Multivariate Observations, Wiley, New York, 1984.

[4] H. Spath, Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples, translated by J. Goldschmidt, Halsted Press, New York, pp. 226, 1985.

[5] F. Azuaje, "Interpretation of Genome Expression Patterns: Computational Challenges and Opportunities," *IEEE Eng. Med. Biol. Mag.*, pp. 119.

[6] F. Azuaje, "An Unsupervised Neural Network Approach To Discovering Gene Expression Patterns In B-Cell Lymphoma," *Online Journal of Bioinformatics*, pp. 26-41, 2001.

[7] M. B. Eisen and P. O. Brown, "DNA Arrays for Analysis of Gene Expression," *Meth. Enzymol.*, vol. 303, pp. 179-205, 1999.

[8] Q. John, "Computitional Analysis of Microarray Data," *Nature*, Vol. 2, pp. 418, 2001.

[9] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani,G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage,R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Bird, D.Botstein, P. O. Brown, and L. M. Staudt, "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling," *Nature*, vol. 403, pp. 503-511, 2000.

[10] http://rana.lbl.gov/EisenSoftware.htm.

[11] Patten, Mildred L, Understanding research methods: An Overview of the essentials (3rd ed.), Los Angeles: Pyrczak Publishing, 2002.

[12] http://llmpp.nih.gov/lymphoma/data.shtml.