CLUSTERING DNA MICROARRAY DATA BY STOCHASTIC ALGORITHM

Ho Sun Shon[1], Sunshin Kim[2], Ling Wang[1], Keun Ho Ryu[1]

[1]Database/Bioinformatics Laboratory Chungbuk National University
[2]Division of Bio-Medical Informatics, Center for Genome Science, Korea National Institute of Health
{shon0621, khryu}@dblab.chungbuk.ac.kr, sskim04@hotmail.com

ABSTRACT ... Recently, due to molecular biology and engineering technology, DNA microarray makes people watch thousands of genes and the state of variation from the tissue samples of living body. With DNA Microarray, it is possible to construct a genetic group that has similar expression patterns and grasp the progress and variation of gene. This paper practices Cluster Analysis which purposes the discovery of biological subgroup or class by using gene expression information. Hence, the purpose of this paper is to predict a new class which is unknown, open leukaemia data are used for the experiment, and MCL (Markov CLustering) algorithm is applied as an analysis method. The MCL algorithm is based on probability and graph flow theory. MCL simulates random walks on a graph using Markov matrices to determine the transition probabilities among nodes of the graph. If you look at closely to the method, first, MCL algorithm should be applied after getting the distance by using Euclidean distance, then inflation and diagonal factors which are tuning modulus should be tuned, and finally the threshold using the average of each column should be gotten to distinguish one class from another class. Our method has improved the accuracy through using the threshold, namely the average of each column. Our experimental result shows about 70% of accuracy in average compared to the class that is known before. Also, for the comparison evaluation to other algorithm, the proposed method compared to and analyzed SOM (Self-Organizing Map) clustering algorithm which is divided into neural network and hierarchical clustering. The method shows the better result when compared to hierarchical clustering. In further study, it should be studied whether there will be a similar result when the parameter of inflation gotten from our experiment is applied to other gene expression data. We are also trying to make a systematic method to improve the accuracy by regulating the factors mentioned above.

KEY WORDS: Clustering, Microarray, MCL Algorithm, SOM, Hierarchical Clustering

1. INTRODUCTION

Clustering of gene expression data is used to analyze the result of microarray study. A cluster analysis in microarray data is a process to bind genes or samples that basically have similar information or expression forms. There is a higher similarity between samples that belong to the same cluster and a lower similarity between samples that belong to different cluster. The development of clustering algorithm of microarray experimental data will make a great contribution to analysis of functional genomics and genetic networks. That is, genes that are related to functions have a similar pattern, so if it is possible to find a gene which has a similar expression pattern, we may be able to predict a function of new gene from the gene that its function has been known. In this paper, a clustering which purposes to discover new biological subgroup or class by using gene expression information was done. For this experiment, the data used is a matrix which consists of 7129 of genes and 72 of samples. The analysis method applied to the data is MCL (Markov CLustering) algorithm [1] [2]. The MCL algorithm is based on probability and graph flow theory. MCL simulates random walks on a graph using Markov matrices to determine the transition probabilities among nodes of the graph. [2] [9]. MCL algorithm has been applied to a lot of biological data and made good results so far. This paper analyzes the result from applying to microarray data. Also, in order to do a comparative evaluation, we analyzed it by using a SOM clustering algorithm which is differed from hierarchical clustering in neural network.

2. RELETED WORKS

There are some studies on clustering algorithm using microarray data as followed. Hartuv [3] proposed a clustering algorithm on the basis of graph theory. Ben-Dor [4] also proposed CAST algorithm by using a graph. Kohonen [5] developed SOM (Self-Organizing Maps) algorithm [6]. Eisen and others [8] applied Hierarchical

Clustering algorithm which is widely used and analyzed in statistics of DNA microarray data clustering. The software; Cluster and Treeview composed by this study are being used by a lot of people [8].

3. CLUSTERING GENE EXPRESSION DATA

3.1 Markov Clustering Algorithm (MCL)

The MCL algorithm is based on probability and graph flow theory. MCL simulates random walks on a graph using Markov matrices to determine the transition probabilities among nodes of the graph. Therefore, the random walk in the graph that visits a dense cluster will likely stay in the cluster until many of its nodes have been visited [2] [9].

The Markov matrix is, in general, represented as in equation (1) each element in the matrix represents the probability of random walk in the graph.

$$M = \begin{pmatrix} a_{11}\, a_{12}\, a_{13} \cdots a_{1m} \\ a_{21}\, a_{22}\, a_{23} \cdots a_{2m} \\ \vdots \quad \vdots \quad \vdots \quad \quad \vdots \\ a_{n1}\, a_{n2}\, a_{n3} \cdots a_{nm} \end{pmatrix} \qquad (1)$$

It used Markov matrix which applies mathematical concept of random work in the graph, that is, the matrix used a probability value. For the process of computing probability of random work through graph, two operators that transform probability sets are used. Two operators, inflation factor and diagonal item, can be optimized through simulation. First, we made genes from each sample into a matrix and transformed by using Euclidean distance of equation (2).

$$Euclidean\,Dis(S)_{ij} = \sqrt{\sum_{k=1}^{n} (M_{ki} - M_{kj})^2} \qquad (2)$$

Then, it applies MCL algorithm to the new matrix, 72×72 , which was found by Euclidean distance. Here we did a simulation to consider two factors of inflation of matrix above and diagonal term, and seek for the optimum factor. The diagonal terms have less effect than the inflation factor for clustering nodes in the graph [2] [9]. The equation (3) below was used to find an inflation factor.

$$Inflation(M)_{ij} = (M_{ij})^r / \sum_{i=1}^{k} (M_{ij})^r \qquad (3)$$

Finally, our method has improved the accuracy through using the threshold, namely the average of each column. If the value of Euclidean distance is bigger than threshold, Markov matrix will be made and if it is smaller than threshold or the same, an inverse number will be taken to the value of Euclidean distance. Also, in this experiment, diagonal terms were determined when the clustering was done the best through the simulation.

3.2 Hierarchical Clustering

Hierarchical clustering is a method that composes a tree which makes genes with similar expression-patterned be neighbours. This method visualizes clustering result into dendrogram which is in a shape of tree and enables to catch the whole expression pattern. In this paper, we used a method that defines a distance between two clusters as an average distance of all individuals that belong to each cluster and binds clusters with a big similarity by using Average Linkage.

3.3 SOM Algorithm

Self-Organizing Maps is a kind of neural networks learning method developed by Kohonen. It is an algorithm that lets reference vectors decided before seeking for the final vector by learning according to the input when the input value in the form of vector is given. The purpose of SOM shows all of the points of high dimension space in the target space of low dimension by keeping the distance and adjacency relation maximally.

When it does clustering with SOM, it needs to fix the number of reference clusters [5] [6].

4. EXPERIMENT AND EVALUATION

Experimental data is leukaemia data and the numbers of genes are 7129 and samples are 72. These data consist of two classes: ALL and AML. Therefore, the experiment has been performed by using R-language to know how well whole samples differentiate these classes [7]. First, genes between samples are made as a matrix by using Euclidean distance and transformed. The data are row data in Figure 1; rows mean genes and columns mean samples. If we show this matrix by using Euclidean distance of formula 1, it will be the same Figure 2.

−214	−139	−76	−135	−106		15	−318	−32	−124	−135
−153	−73	−49	−114	−125		−114	−192	−49	−79	−186
−58	−1	−307	265	−76		2	−95	49	−37	−70
88	283	309	12	168		193	312	230	330	337
−295	−264	−376	−419	−230		−51	−139	−367	−188	−407
−558	−400	−650	−585	−284	⋯	−155	−344	−508	−423	−566
199	−330	33	158	4		29	324	−349	−31	−141
−176	−168	−367	−253	−122		−105	−237	−194	−223	−315
252	101	206	49	70		42	105	34	−82	206
		⋯						⋯		
185	169		240	156		173	225	36	348	209
511	837	1199	835	649		492	737	592	938	634
−125	−36	33	218	57		54	63	57	−15	−58
389	442	168	174	504		277	472	215	433	375
−37	−17	52	−110	−26	⋯	−13	33	−22	−2	−23
793	782	1138	627	250		279	737	588	1170	2315
329	295	777	170	314		51	227	361	284	250
36	11	41	−50	14		6	−9	−26	39	−12
191	76	228	126	56		2484	371	133	298	790
−37	−14	−41	−91	−25		−2	−31	−32	−3	−10

Figure 1. M_{ij} = Raw data (7129 x 72)

0	84238.56	88486.18	59252.05	73384.9		94746.69	101833.4	79537.14	76635.59	96466.56
84238.56	0	84387.97	80073.6	78625.07		98521.84	93605.26	87139.42	75879.26	94001.11
88486.18	84387.97	0	82472.39	94609.8	...	111021	114629.3	92532.81	94613.08	113824.6
59252.05	80073.6	82472.39	0	60754.52		83123.32	95270.14	70046.57	71920.86	94915.14
73384.9	78625.07	94609.8	60754.52	0		79995.79	90409.3	75443.54	82968.78	96019.62
		...						...		
94746.69	98521.84	111021	83123.32	79995.79		0	98261.85	80665.26	90737.26	97952.79
101833.4	93605.26	114629.3	95270.14	90409.3		98261.85	0	85488.21	85221.71	89968.41
79537.14	87139.42	92532.81	70046.57	75443.54	...	80665.26	85488.21	0	67921.84	89268.49
76635.59	75879.26	94613.08	71920.86	82968.78		90737.26	85221.71	67921.84	0	78707.62
96466.56	94001.11	113824.6	94915.14	96019.62		97952.79	89968.41	89268.49	78707.62	0

Figure 2. Matrix (72x72) using Euclidean distance

We used the expression to get an inflation factor of equation (3) and diagonal terms are the optimum values gotten from the simulation. We also used a threshold according to the following condition (4), and diagonal terms got the optimum values through the simulation. So, the matrix of Figure 3 can be gotten.

$$IF\ S_{ij} > Threshold \Rightarrow NS_{ij} = S_{ij}$$
$$IF\ S_{ij} \le Threshold \Rightarrow NS_{ij} = 1/S_{ij} \quad (4)$$
$$(IF\ i = j),\ then\ Diagonal\ term = 0.0005$$

Where, S_{ij} = Euclidian distance

5.00E-04	1.19E-05	1.13E-05	1.69E-05	1.36E-05		6e-05	0.000	0e+00	1.257	4e-05	1.30	7e-05	1.03	9.00E-05
1.19E-05	5.00E-04	1.19E-05	1.25E-05	1.27E-05		3e-05	1.06	6e-05	1.147	6e-05	1.317	3e-05	1.06	7.00E-05
1.13E-05	1.19E-05	5.00E-04	1.21E-05	1.06E-05	...	0e+00	0.000	0e+00	1.08	8e-05	1.05	6e-05	0.000	0.00E+00
1.69E-05	1.25E-05	1.21E-05	5.00E-04	1.65E-05		2e-05	1.04	7e-05	1.427	2e-05	1.39	7e-05	1.05	3.00E-05
1.36E-05	1.27E-05	1.06E-05	1.65E-05	5.00E-04		6e-05	1.10	1e-05	1.325	5e-05	1.20	3e-05	1.041	4.00E-05
			...						...					
1.06E-05	1.02E-05	0.00E+00	1.20E-05	1.25E-05		5.00E-04		1.02E-05	1.24E-05		1.10E-05		1.02E-05	
0.00E+00	1.07E-05	0.00E+00	1.06E-05	1.11E-05		1.02E-05		5.00E-04	1.17E-05		1.17E-05		1.11E-05	
1.26E-05	1.15E-05	1.08E-05	1.43E-05	1.33E-05	...	1.24E-05		1.17E-05	5.00E-04		1.47E-05		1.12E-05	
1.30E-05	1.32E-05	1.06E-05	1.39E-05	1.21E-05		1.10E-05		1.17E-05	1.47E-05		5.00E-04		1.27E-05	
1.04E-05	1.08E-05	0.00E+00	1.05E-05	1.04E-05		1.02E-05		1.11E-05	1.12E-05		1.27E-05		5.00E-04	

Figure 3. NS_{ij} = Matrix Using Threshold

The simulation that controls an inflation factor by applying MCL algorithm has been practiced repeatedly. When inflation factor was 1.215 and diagonal term was 0.0005, the accuracy was the highest. Figure 4 shows the result of our experiment.

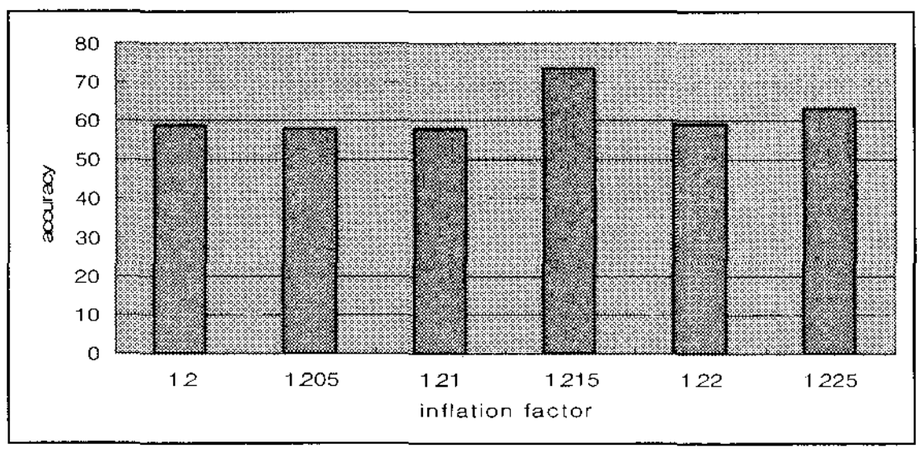

Figure 4. Inflation Experiment results with simulation

We experimented and applied SOM algorithm and hierarchical algorithm to our data by using Cluster & Treeview tool that enables to use gene expression data [8]. As a result, SOM algorithm made 8 nodes and there found a cluster of sample. From these results, a few of the best classified clusters are subsets in a class of the two classes (ALL and AML).

Table 1. Nodes of SOM result

Node NO	Sample
Node0	5,13,15,20,21,24,31,32,34,35,36,37,38
Node2	16,19,44,52,68
Node3	9,11,14,17,18,26,30,40,41,47
Node4	2,10,25,39,45,55
Node5	1,3,23,28,46,67,66
Node6	6,22,49,56,71
Node7	33
Node8	4,7,8,12,27,29,50,51,53,54,57,59, 60,61,62,63,64,65,69,70,72

Table 1 is a result by using SOM and there are 9 nodes made of the sample. But there is no sample for node 1, so in fact there are 8 nodes made.

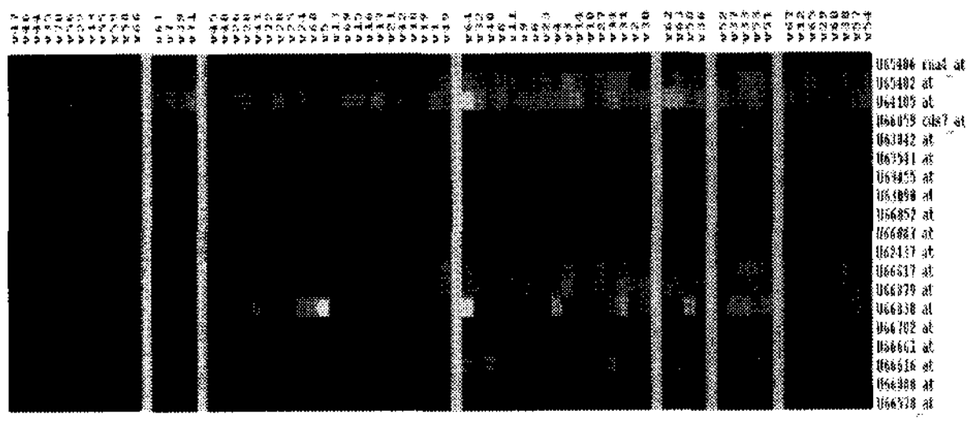

Figure 5. Result of Hierarchical Clustering

Figure 5 is the result that used hierarchical clustering. As a result, there are 7 classes divided.

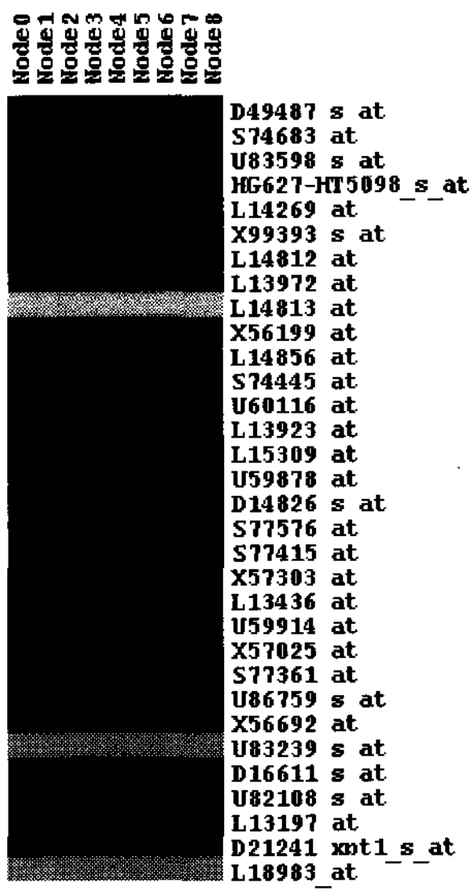

Figure 6. Result of SOM

Figure 6 is the result which SOM was used and there are 9 nodes made for sample.

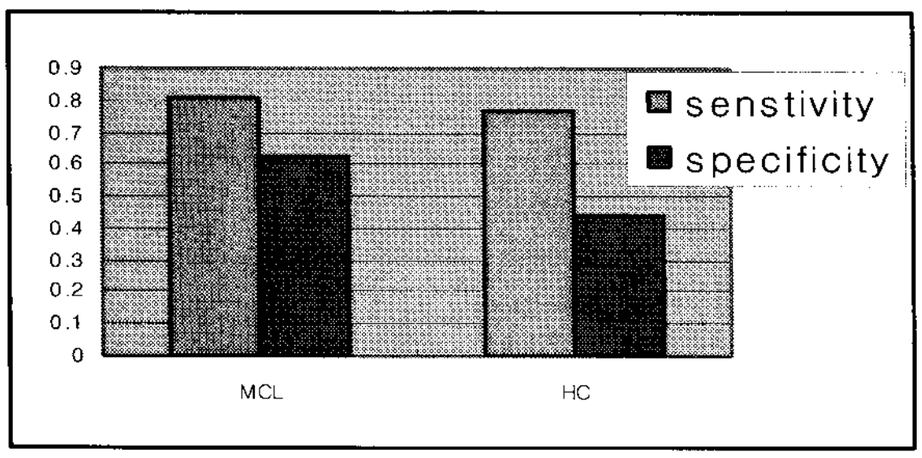

Figure 7. Accuracy of MCL Algorithm and HC

It is impossible to discern and compare which cluster each sample belongs to by SOM algorithm, because the node for the sample is created in the Cluster & Treeview tool. But it is possible to compare between MCL algorithm and hierarchical algorithm. Figure 7 shows the result that compared two analysis methods. And we also confirm MCL algorithm is well clustered best.

5. CONCLUTION

Because of generalization of microarray data experiment and rapid development of study with genes, microarray data are being produced continuously. Clustering algorithm takes the lead to achieve significant information from this mass information. In this paper, microarray data that consisted of 72 samples and 7129 genes are tested by using MCL algorithm that is based on graph theory. In order to classify Class well, it simulated inflation factors and diagonal terms. That's how it could find the factor that has the highest accuracy. Our experimental result shows about 70% of accuracy in average compared to the class that is known before. We performed the experiments of the SOM algorithm and the hierarchical clustering by using Cluster & Treeview tool, and had a comparative analysis to MCL algorithm. As a result of comparing MCL algorithm through the simulation of stochastic flow and the hierarchical clustering method, it was clear that MCL algorithm has the highest accuracy. In further study, it should be studied whether there will be a similar result when the parameter of inflation parameter gotten from our experiment is applied to other gene expression data. We are also trying to make a systematic method to improve the accuracy by regulating the factors mentioned above.

Acknowledgements

This work was supported by the RRC program of MOCIE and ITEP.

References

[1] Ho Sun Shon, Sunshin Kim, Chung Sei Rhee, Keun ho Ryu, "Clustering DNA Microarray Data by MCL Algorithm, ISMB, 2007

[2] Stijn Marinus van Dongen, GRAPH CLUSTERING by FLOW SIMULATION, 1969.

[3] E. Hartuv et al., An Algorithm for Clustering cDNAs for Gene Expression Analysis, RECOM B 99, 1999, pp.188- 197.

[4] A. Ben-Dor , R. Shamir , Z. Yakhini, Clustering Gene Expression Patterns , Journal of Computational Biology, 1999, pp. 281- 297.

[5] T. Kohonen, Self-Organizing Maps, Springer Verlag, New York, 1997.

[6] P.Tamayo, D. Slonim, J. Mesirov , Q. Zhu, S. Kitareewan , E. Dmitrovsky , E. S. Lander and T . R. Golub , Interpreting patterns of gene expression with self-organizing maps: Methods and application to Hematopoeitic differentiation, PNAS, vol96, 1999, pp. 2907-2912

[7] http://www.r-project.org/

[8] EisenLab, http://rana.lbl.gov/EisenSoftware.htm

[9] Sunshin Kim, Clustering Methods for Finding Orthologs among Multiple Species, 2007