International Journal of Bioelectromagnetism
Vol. 9 No. 2 2007

Clustering Approach using MCL Algorithm for Analyzing Microarray Data

Ho Sun Shon, Sunshin Kim, Chung Sei Rhee, Keun Ho Ryu
Database/Bioinformatics Laboratory Chungbuk National University
{shon0621, sskim04, khryu}@dblab.chungbuk.ac.kr, csrhee@chungbuk.ac.kr

Abstract

Clustering gene expression data is used to analyze the result of microarray study. This method is often useful in understanding how a class of genes performs together during a biological process. In this work, we have tried to cluster open leukemia data using MCL (Markov CLustering) algorithm, which is a clustering method for graphs based on simulation of stochastic flow. We have tuned to set the proper factors of inflation, diagonal terms of matrix, and threshold. Our experimental result shows about 70 percent of accuracy in average compared to the class that is known before.

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.

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 is MCL (Markov Clustering) algorithm. MCL algorithm is an algorithm that clusters nodes in the graph through the simulation with probability flow within the graph and is fast and excellent at extension [1] [2]. 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. We analyzed the result by applying microarray data in this paper. Although the result was not compared and evaluated with other algorithms, it analyzed self-organizing map (SOM) which is differentiated from neural network and hierarchical clustering.

2. Related Works

There are some studies on clustering algorithm using microarray data as followed. Hartuv proposed a clustering algorithm on the basis of graph theory and algorithm [3]. Ben-Dor also proposed CAST algorithm by using a graph and embodied a program [4]. Tamayo developed SOM (Self-Organizing Maps) algorithm and implemented [5]. Eisen and others 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 [7].

3. Clustering gene expression data

3.1 Markov CLustering algorithm (MCL)

MCL algorithm was developed by Stijn van Dongen from Netherlands. It is a kind of algorithm that clusters node in the graph through simulation by computing probability and it is also fast and excellent in extension. 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, genes between samples are made as a matrix by using Euclidean distance and transformed. Then, it applies MCL algorithm to the new matrix, 72×72 ,

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD), by the RRC program of MOCIE and ITEM, and by the Korea Science and Engineering Foundation (KOSEF)

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] [8]. Finally, our method has improved the accuracy through using the threshold, namely the average of each column.

3.2 Hierarchical Clustering

Hierarchical clustering is a method that composes a tree which makes genes with similar expression-patterned be neighbors. 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. When it does clustering with SOM, it needs to fix the number of reference clusters [5].

4. Experiments and Evaluation

Experimental data is leukemia 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 [6]. 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 1 shows the result of our experiment.

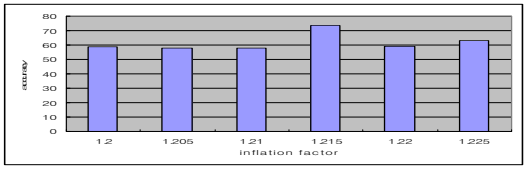

Figure 1. Our experiment result with Inflation Factors

And we experimented and applied SOM algorithm and hierarchical algorithm to our data by using Cluster and Treeview tool that enables to use gene expression data [7]. As a result, 8 of node were composed in SOM algorithm and a cluster was divided into 7. From these results, a few of the best classified clusters are subsets in a class of the two classer (ALL and AML).

5. Conclusions

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 R-language 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 experiment SOM algorithm and hierarchical clustering by using Cluster and Treeview tool, but it is impossible to compare between the clusters by MCL and those by each method mentioned above, because they are divided into several cluster different each other. 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.

6. 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 FFLOW 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] http://www.r-project.org/
[7] EisenLab, http://rana.lbl.gov/EisenSoftware.htm
[8] Sunshin Kim, Clustering Methods for Finding Orthologs among Multiple Species, 2007