Genomics & Informatics Vol.1(1) 20-24, September 2003

Global Optimization of Clusters in Gene Expression Data of DNA Microarrays by Deterministic Annealing

Kwon Moo Lee[1], Tae Su Chung[2] and Ju Han Kim[3]*

[1]Bioinformatics Project, IT R&D Center, Samsung SDS, Seongnam, Korea
[2]Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea
[3]SNUBiomedical Informatics, Seoul National University College of Medicine, Seoul, Korea

Abstract

The analysis of DNA microarry data is one of the most important things for functional genomics research. The matrix representation of microarray data and its successive 'optimal' incisional hyperplanes is a useful platform for developing optimization algorithms to determine the optimal partitioning of pairwise proximity matrix representing completely connected and weighted graph. We developed Deterministic Annealing (DA) approach to determine the successive optimal binary partitioning. DA algorithm demonstrated good performance with the ability to find the 'globally optimal' binary partitions. In addition, the objects that have not been clustered at small non-zero temperature, are considered to be very sensitive to even small randomness, and can be used to estimate the reliability of the clustering.

Keywords: cluster analysis, DNA microarray, gene expression, global optimization, annealing, clustering quality

Introduction

The cluster analysis is one of the most prominent methods for analyzing DNA microarray data. It explores the internal structure of complex data by organizing them into meaningful groups. Genes of a similar expression pattern may share similar function, clustering gene expression profiles can be used for tentative assignment of functional annotation of the unknown genes based on the functional annotations of the known genes (Eisen *et al.*, 1998). Cluster analysis is also useful in classifying tissues on the basis of gene expression. Cancerous and normal tissues can be distinguished with genes of subtle differences in gene expression (Alon *et al.*, 1999). Golub *et al.* suggested molecular classification of cancer using cluster analysis, which categorizes leukemia into acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without prior knowledge about their gene expression patterns (Golub *et al.*, 1999). They also developed leukemia class predictor by supervised learning method based on gene expression data. Since there is the correlation between the motif existence in promoter region and gene expression profile, it has been shown that co-regulated genes can be identified by cluster analysis (Tavazoie and Church, 1998; Tavazoie *et al.*, 1999). Holmes and Bruno considered gene expression data and promoter sequences simultaneously to find co-regulated genes by cluster analysis (Holmes and Bruno, 2000).

The matrix representation of microarray data and its successive 'optimal' incisional hyperplanes by Kim *et al.* (Kim *et al.*, 2001) constitutes a useful platform for developing optimization algorithm to determine the optimal partitioning of pairwise proximity matrix representing completely connected and weighted graph. In this method, the clustering and the clustering analysis is formulated as an optimization problem whose global optimum is found by various global optimizers. Previously, MITree algorithm (Kim *et al.*, 2001) and evolution strategy (Lee *et al.*, 2001) are applied to this cluster analysis. Both of them demonstrated good performances with the ability to find the 'globally optimal' successive binary partitions. MITree-K is a K-partitioning extension of the same geometric principle (Kim *et al.*, 2002).

In this paper, we applied a heuristic global optimization method, Deterministic Annealing (DA), to the same clustering method. In DA approach, the cost function is locally minimized subject to a constraint on a given randomness (Shannon entropy), controlled by 'temperature' that is gradually lowered. As the temperature goes slowly down to zero, we obtain the best binary partitioning by the analogy of statistical physics in

*Corresponding author: E-mail juhan@snu.ac.kr
Tel +82-02-740-8320, Fax +82-02-747-4830

Abbreviations: ALL, acute lymphoblastic leukemia; AML, acute myelold leukemia; DA, deterministic anealing; SOM, self-organizing map; MII, matrix Incision Index; SA, simulated anealing.

Accepted 22 March 2003

annealing process. Even though global optimization approach produces the high quality clustering results, in general, it requires much computational cost. But, since the speed of DA algorithm depends on that of the local optimizer, if we employ high performance local optimization technique, the computation cost can be substantially reduced. It was also demonstrated that error-prone objects can be identified by monitoring the annealing process, which can be applied to increase the quality of clustering analysis.

Results and Discussion

Fisher's iris data

The Fisher's iris data set has been widely used for evaluating the performance of cluster algorithms (Fisher, 1936). The iris data set consists of four measurements (petal/sepal length and petal/sepal width) of 50 *Iris Setosa*, 50 *Iris Versicolor*, and 50 *Iris Virginica*. In this test, the similarity measures between the iris flowers are the square values of the Pearson' s correlation coefficient. And, temperature cooling parameter α is equal to 0.95. As seen in Fig. 1, the first binary partitioning separated all *Setosa* from the entire data set without any error. The next partitioning of 50 *Versicolor*' s and 50 *Virginica*' s gave rise to six errors. Three *Versicolor*' s (objects 9, 12, 40) were clustered to *Virginica* and three *Virginica*' s (objects 66, 77, 81) to *Versicolor*. The overall accuracy of clustering was 96% (144/150). The repeated results of our algorithm were the same.

The Fig. 2 shows how x_i' s are reduced to 0 or 1 as the temperature decreases to zero. We see that x_9, x_{12}, x_{40}, and x_{59} are reduced to 0 or 1 slowly relative to other x_i' s and these variables correspond to the objects misclassified as above. Because the misclassified objects might be near the cluster boundaries, they are very susceptible to randomness. Hence, even at small temperature, they are not bounded and easily agitated from 0 or 1. This concept is very helpful to find significant clusters which are robust to any possible errors or randomness. In addition, we can reduce computation time substantially using this concept. Because the slowly relaxing variables may not contribute to clustering quality and require great time-consumption to be relaxed to 0 or 1 as Fig. 2, we can terminate the algorithm before all the variables are reduced to 0 or 1. In fact, we can use the following stopping criteria

1) terminate the algorithm if pre-assigned number of iteration is reached. And the point whose value remains the region $\{0 < x < 1\}$ is regarded as an outlier, or
2) terminate the algorithm if all x_i' s are in the neighborhood $\{0 \leq x < \varepsilon\}$ of 0 or neighborhood $\{1 - \varepsilon < x \leq 1\}$ of 1, where ε is a small positive parameter (e.g. $\varepsilon = 0.1$) when pre-assigned number of iteration is reached.

Golub's leukemia data

We also tested our algorithm using Golub' s leukemia data set from 38 human acute leukemia cells, which was used for training class predictor of acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) (Golub *et al.*, 1999). The ALL group consists of T-cell ALL' s (T-ALL) and B-cell ALL' s (B-ALL). In the experiment, RNA extracted from each leukemia sample was hybridized to high density microarrays containing 6,817 human genes. After scanning the microarrays and image processing, the expression levels of each gene for each leukemia cell were quantified. It has been shown that leukemia class prediction was

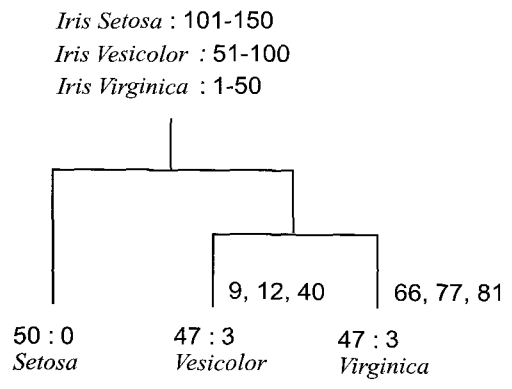

Fig. 1. Clustering result of Fisher' s iris flowers by DA approach. The lists of numbers in the terminal leaves are those of misclassified objects.

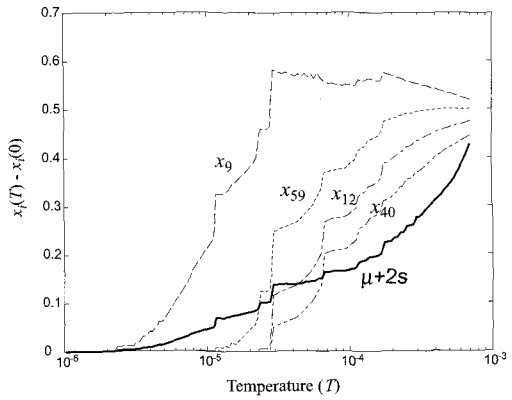

Fig. 2. Behaviors of some x_i' s susceptible to randomness (μ is the mean of all x_i' s and σ is standard deviation).

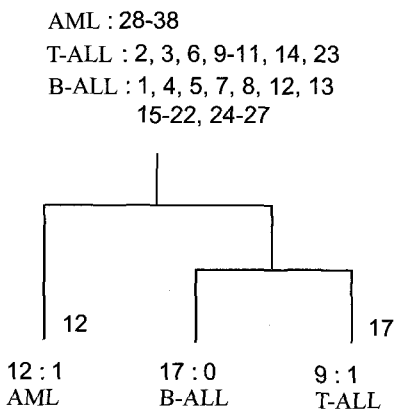

Fig. 3. Clustering result of Golub' s leukemia data by DA approach. The lists of numbers in the terminal leaves are those of misclassified objects.

feasible by this gene expression monitoring without additional biological knowledge. Golub *et al.* selected 50 genes highly correlated with ALL-AML class distinction. The Fig. 3 demonstrates the results of our cluster algorithm applied to leukemia data using the 50 genes. The first binary partitioning distinguished ALL from AML with 1 error, and the second distinction of ALL between TALL and B-ALL gave rise to 1 error. Two B-ALL samples were misclassified into AML and T-ALL, respectively. The overall accuracy of the result was 94.7% (36/38). The repeated results of our algorithm were the same.

Discussion

Due to the recent development of high-throughput experimental techniques in functional genomics, we are facing the flood of large-scale gene expression data. The needs of analysis of such biological data shed light on the importance of cluster analysis, basic methodology to reveal internal structure of complex data. Because cluster analysis is the first phase of mining useful information, the reliability of clustering results is responsible for the quality of information extracted in the following stages. The direct maximization of the fig. of merit (clustering index) which is main difference with conventional cluster algorithms can be good strategy to increase clustering quality. We have shown that deterministic annealing was effectively applied to find globally optimal clusters. Previously, we applied two different approaches MITree algorithm and evolution strategy to this problem and have the same results (Kim *et al.*, 2001; Lee *et al.*, 2001). We also find possibility that we can identify the clustered objects that are susceptible to errors or randomness and irrelevant to increase clustering quality. This concept has good possibility to be applied to find relevant and robust clusters and reduce computation time substantially. In addition, our approach does not needs any prior assumption about data structure and considers global perspective of data in clustering. We can also apply this algorithm with different clustering indices and similarity measures, which is not possible in algorithm dependent methods such as K-means and SOM. Even though the computational cost is quite high in this preliminary study with iris and leukemia data sets, we have shown that our approach is promising for the further investigation with a variety of data sets, similarity measures and local optimization algorithms.

Methods

Data representation

A set of DNA microarray experiments data can be viewed as a completely connected and weighted graph of genes or arrays (i.e., cell lines) with similarity measures. The vertices of the graph correspond to the objects to be clustered and the edges represent the similarity between the objects. The similarity measures can be stored in pairwise proximity matrix. This representation makes the algorithm independent of similarity measures, while the other algorithms like K-Means and SOM (Self-Organizing Maps) mingle with the given similarity measure.

Clustering as an optimization problem

The cluster analysis proposed in this paper is done by successive binary partitioning, which produces top-down hierarchical tree. Binary partitioning decomposes the graph into two parts with optimized fig. of merit (clustering index) called MII (Matrix Incision Index) suggested by Kim *et al.* (Kim *et al.*, 2001). The MII includes homogeneity and separation of binary partition. Homogeneity indicates that each object within the same cluster should have high similarity. On the other hand, separation means that the objects between clusters should have low similarity. These requirements are incorporated into MII as follows:

$$\mathrm{MII} = \frac{bm/(n+m) + cn/(n+m)}{a} \quad (1)$$

where m and n are the numbers of objects in groups 1 and 2, respectively, a is the average link strength between groups 1 and 2, b and c are within-group average link strength of group 1 and 2, respectively. The numerator of MII corresponds to homogeneity, the denominator to separation. Since homogeneity/ separation should be as high/low as possible in binary partitioning, we should maximize the MII. The index is defined directly from the similarity matrix without prior information regarding the structure of data set. After defining the MII , we should find

how to get to its global maximum. Because there is no general and rigorous mathematical method for the global optimization problem, a feasible way is to use heuristic methods to adopt randomness to escape local optima, one of which is deterministic annealing used in this paper. The advantage of global optimization of clustering is to increase clustering quality because the clustering index is directly optimized differing from other algorithms such as K-means and SOM. In addition, we do not have any prior assumption about data structure. One possible disadvantage is that this global optimization algorithm might be relatively slower than other local optimization algorithms.

Deterministic annealing to find globally optimized clusters

We describe a heuristic algorithm creating a hierarchical tree from complex data by iterative optimal binary partitioning, based on DA. The best binary partitioning can be found by the global maximization of MII introduced in the previous section as a clustering index (Eq. 1). Motivated by the annealing process of heating and slowly cooling materials to obtain the most stable structure, Kirkpatirck *et al.* have proposed this stochastic optimization method called Simulated Annealing (SA) in an attempt to find global optimum (Kirkpatrick *et al.*, 1983). Using SA, the optimal solution is randomly searched over the cost function landscape with higher probability for low cost function. As the temperature, the parameter representing randomness in searching decreases gradually, the solution is converged to global minimum.

SA was applied to cluster analysis of temporal gene expression profiles and finding the optimal number of clusters (Lukashin *et al.*, 2001). A deterministic version of SA called Deterministic Annealing requires much less computational cost because stochastic simulation in SA is replaced by the corresponding expectation values. The randomness is incorporated into cost function to be minimized, and the function is locally minimized in the course of lowering temperature. Rose *et al.* applied DA to the central clustering analysis which requires clustering centroids as K-means algorithm (Rose *et al.*, 1990; Rose, 1998) and it was used in analyzing tumor and normal colon cancer tissues probed by oligonucleotide arrays (Alon *et al.*, 1999).

To specify data partition, we assign 0 or 1 to each object. Although it is essentially a discrete (combinatorial) optimization to find the best partitioning, the problem is transformed into a continuous version, if the clustering assignments denoted by x_i are in the range [0, 1]. Here, x_i' s denote the average of the many discrete states in stochastic simulation of the annealing process. Using these average clustering assignments, we can represent MII as follows.

$$\mathrm{MII} = \frac{(n_0 L_0^{av} + n_1 L_1^{av})/N}{L_{bet}^{av}} \tag{2}$$

where

$$L_0^{av} = \frac{\sum_{i<j} L_{ij}(1-x_i)(1-x_j)}{\sum_{i<j}(1-x_i)(1-x_j)} \tag{3}$$

$$L_1^{av} = \frac{\sum_{i<j} L_{ij} x_i x_j}{\sum_{i<j} x_i x_j} \tag{4}$$

$$L_{bet}^{av} = \frac{\sum_{i<j} L_{ij}\{x_i(1-x_j)+x_j(1-x_i)\}}{\sum_{i<j}\{x_i(1-x_j)+x_j(1-x_i)\}} \tag{5}$$

$$n_0 = \sum_i (1-x_i) \tag{6}$$

$$n_1 = \sum_i x_i \tag{7}$$

Here, L_{ij} is similarity between object i and j, and N is the number of objects to be partitioned.

Instead of optimizing MII alone, we define free energy, F which includes randomness of clustering as follows, and carry out local optimization at given temperature, T.

$$F = -\mathrm{MII} - TS \tag{8}$$

where

$$S = -\sum_i \{x_i \log x_i + (1-x_i)\log(1-x_i)\} \tag{9}$$

Here, S is the Shannon entropy which measures the level of randomness, and T controls randomness of clustering assignment during the annealing. We carry out local optimization at each stage of temperature cooling process base on the optimization result from previous temperature. As the temperature goes slowly down, we obtain the best binary partitioning by the analogy of statistical physics in annealing process. This DA approach was successfully applied to Traveling Salesman Problem (TSP) by Hopfield and Tank in a similar manner (Hopfield and Tank, 1985).

We can re-interpret this problem from the viewpoint of constraint optimization. Here, we should minimized the object function, -MII subject to the constraint, S=0, which requires every x_i should be 0 or 1. To deal with such a constraint optimization problem, we define Lagrangian function which is the same as free energy, Eq. 8 with undetermined Lagrangian multiplier, T. We can find

extrema of constraint object function and Lagrangian multiplier, T by the condition, $\partial F/\partial x_i = \partial F/\partial T = 0$. If minimal point of -MII is located at the corner of hypercube defined by $\{x_i \mid 0 < x_i < 1\}$, the critical temperature is positive. On the other hand, the minimum point located at the middle of hypercube results in negative critical temperature, differing from what we see in statistical physics. The annealing can be seen to find such critical points of Lagrangian function as it scans given range of T.

To find globally optimized binary partitioning, we implemented the DA approach as follows.

(1) Initialize: random initialization of x_i, temperature initialization ($T \leftarrow T_0$), Set cooling coefficient ($0< \alpha <1$)
(2) Local optimization of F at given T
(3) If all x_i' s are 0 or 1
    - Termination
else
    - Temperature decreases as $T \leftarrow T_0$, Go to (2)

In this paper, we used the standard multidimensional local optimization technique which incorporates one-dimensional line search method (Bazaraa *et al.*, 1993). The improving feasible direction of the line search is determines by calculating gradient of cost function in the feasible region $\{x_i \mid 0 < x_i < 1\}$. At the boundary of the feasible region, the generation of the improving feasible direction is done by solving a sub-problem of linear programming.

Acknowledgments

This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health and Welfare, Republic of Korea (01-PJ10-PG6-01GM01-0004).

References

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96, 6745-6750.

Bazaraa, M.S., Sherali, H.D., and Shetty, C.M. (1993). Nonlinear Programing Theory and Algorithms, 2nd eds., (Wiley).

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863-14868.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems, Annals of Eugenics 7, Part II, 179-188.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Caasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L. , Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531-537.

Holmes, I. and Bruno, W.J. (2000). Finding regulatory elements using joint likelihoods for sequence and expression profile data. Intelligent Systems for Molecular Biology, 202-210.

Hopfield, J.J. and Tank, D.W. (1985). Neural computation of decision in optimization problems. Biol. Cybern. 52, 141-152.

Kim, J.H., Ohno-Machado, L., and Kohane, I.S. (2001). Unsupervised learning from complex data: The Matrix incision tree algorithm. Pacific Symposium on Biocomputing, 30-41.

Kim, J.H., Ohno-Machado, L., and Kohane, I.S. (2002) Visualization and evaluation of clustering structures for gene expression data analysis. J Biomed Inform 35, 25-36.

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimization by simulated annealing. Science 220, 671-680.

Lee, K., Kim, J.H., Chung, T.S., Moon, B.S., Lee, H., and Kohane, I.S. (2001). Evolution strategy applied to global optimization of clusters in gene expression data of DNA microarrays. Proc. IEEE Cong. on Evol. Comp. 845-850.

Lukashin, A.V. and Fuchs, R. (2001). Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. Bioinformatics 17, 405-414.

Rose, K., Gurewitz, E., and Fox, G. (1990). Statistical mechanics and phase transition in clustering. Phys. Rev. Lett. 65, 945-948.

Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Proc. IEEE 86, 2210-2239.

Tavazoie, S. and Church, G.M. (1998). Quantitative wholegenome analysis of DNA-protein interactions by *in vivo* methylase protection in *E. coli*. Nature Biotechnol. 16, 566-71.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. (1999). Systematic determination of genetic network architecture. Nature Genetics 22, 281-285.