2007－AL－113 (6)
2007／7／3

Hierarchical Clustering and Leaf Ordering Scheme for DNA Chip Expression Data

Sang-Soo Yeo, Kouichi Sakurai
Department of Computer Science & Communication Engineering,
Kyushu University, Fukuoka, Japan
ssyeo@itslab.csce.kyushu-u.ac.jp, sakurai@csce.kyushu-u.ac.jp

Abstract

Generally the result data from DNA chip experiments have lots of gene expression information. Scientists want to get perspective insight or want to find intuitive fact from that data. Hierarchical clustering is the most widely used method for analysis of gene expression data. In this paper, we address leaf-ordering, which is a post-processing for the dendrograms created by hierarchical clustering and we present a new approach for leaf-ordering scheme. And we show the comparison results for our approach and the existing approach.

1. Introduction

DNA chip experiment technology encourages producing huge sets of gene expression data in these days. DNA chip technology can be applied to functional genomics, genetic network, and association analysis and then it seems to be one of important methods for medical research fields such as genetic disease findings, new medicine development.

For accelerating these advantages, it is very important to process and to visualize gene expression data from DNA chip experiments. In other words, a large set of gene expression data could be meaningful information for scientists under the condition that it should be processed by computational methods and be presented by intuitive graphical representation.

DNA expression data clustering problem was evolved for this purpose. The key point of the problem is divide dataset into a number of clusters; gathering genes with similar expression pattern together and separating genes with different expression pattern. Clustering enables to predict functions of undiscovered genes by looking their involved clusters and it helps to find a gene set regulating a specific disease by comparing gene expression patterns of health persons and of patients.

However, DNA expression data clustering problem is known as NP problem in the view of

computational complexity class and there are a few heuristic methods until now. Usually these algorithms adopted graph theory, neural network, and/or statistical analysis to their main algorithms. The most famous one of them is hierarchical clustering method and many biologists use this method for their research. Eisen introduced this method into DNA chip data analysis domain and he also implemented it as a free software tool, which is used by many researchers now.

Output of hierarchical clustering is a dendrogram that is a kind of binary tree. This output dendrogram's leaf nodes respectively can be matched to each gene expression spot and its edges reflect similarities (or distances) of each pair of gene subgroups. In this method, an order of leaf nodes doesn't have any meanings. Eisen also introduced leaf-ordering scheme and he made it meaningful to consider an order of dendrogram's terminal nodes.

In this paper, our main concern is on how to represent leaf's order more efficiently and more intuitively for biological and medical scientists. In section 2, we introduce a formal definition of the leaf-ordering problem and some notations. In section 3, we mention the existing leaf-ordering methods and its drawback. And then we propose a new approach for leaf-ordering. This paper is a full paper version of our abstract that submitted to GIW 2002.

2. Definitions and Notations

The followings are notations for describing hierarchical clustering and leaf-ordering;

$D = \{d_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n\}$: m×n matrix of DNA chip experiment result data.

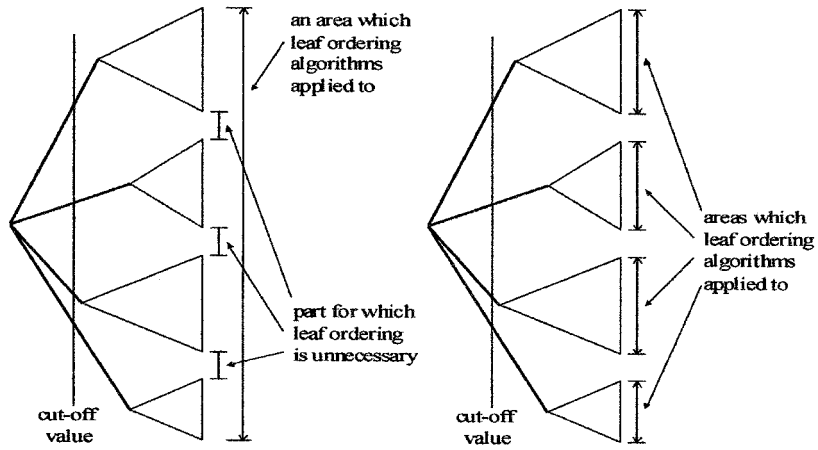
d_{ij}: the expression level of j-th experiment of i-th gene.

d_i: the row vector that include expression levels of all experiments of i-th gene.

T: binary tree that was produced by hierarchical clustering scheme. The number of T's leaf nodes is m. Edge's length of each parent's node means the similarity of the two children nodes.

We can define the leaf-ordering problem as below;

We can put the root node to left side of dendrogram T and we can put leaf nodes to right side of T in a single vertical line. At this time we call the order of leaf nodes as leaf-ordering. When we draw T, we should decide which child node is put the above position for internal nodes of T. So there are $(m-1)$ internal nodes in T and the total number of possible leaf-orderings is 2^{m-1}. The objective function of this problem is to minimize the total summation of distance of all adjacent leaf node pairs. Distance is the opposite concept of similarity of each adjacent pair of leaf nodes. And we call the leaf-ordering with



(a) Overall leaf ordering approach (most of algorithms)

(b) Leaf ordering approach for each cluster (proposed approach)

Figure 1. An example for overall leaf-ordering and cluster leaf-ordering

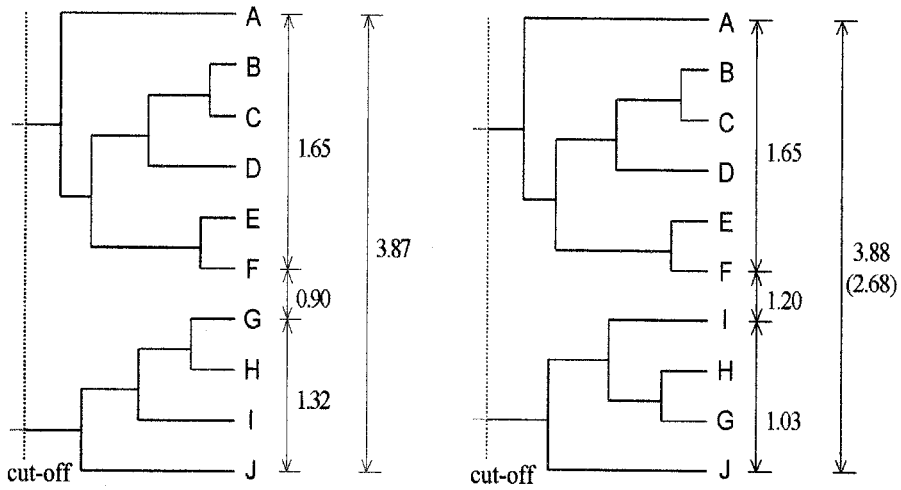
minimum value as the computationally optimal leaf-ordering and we call the algorithms that finding the best leaf-ordering for biological and medical research as leaf-ordering algorithms (or schemes).

3. Overall Leaf-Ordering Approach vs. Leaf-Ordering for Each Cluster Approach

Leaf-ordering problem could be a difficult problem in the view of computational complexity, because m, the number of genes, is usually very big[1]. So the exhaustive searching for finding the optimum is a difficult problem. Until now, there are several leaf-ordering schemes such as Eisen's algorithms[2,3], Alon's algorithm[4], Bar-Joseph's algorithms[5,6], Yeo's algorithm[7].

We think all of them have the same approach such that they try to find the best leaf-ordering for overall leaf nodes at once. We call this approach as overall leaf-ordering approach. And this approach is the contrary concept for our approach that will be introduced in the next section. Overall leaf-ordering approach may cause the reduction of effectiveness of leaf-ordering because all of algorithms would include pairs of genes located at each adjacent cluster's boundary for reordering. This is an unnecessary reordering.

The main problem of overall leaf-ordering approach is that it doesn't consider the fact that most of biological researchers try to analysis genes in each clusters, but not to inter-clusters. In overall leaf-ordering algorithms, they would apply their algorithmic procedure to all leaf nodes at once and this makes each cluster's internal leaf-ordering more inaccurate by considering adjacent leaf nodes out of each cluster.



(a) Overall leaf-ordering

(b) Leaf-ordering for each cluster

Figure 2. An example for overall leaf-ordering and cluster leaf-ordering

Now we propose a new approach, and its key point is that we should apply an algorithm to each cluster to improve each cluster's optimal reordering. This approach can get rid of the problem of overall leaf-ordering approach and it can be used for any leaf-ordering algorithms. Our approach is simply adopted to any algorithms as below;

1) Produce a dendrogram T using Hierarchical Clustering
2) Select a cut-off value within similarity boundary [-1,1]
3) Process a selected algorithm for each cluster respectively

To compare overall leaf-ordering approach and leaf-ordering for each cluster approach, we need to describe how to make clusters in T. After hierarchical clustering we would set up cut-off value for the output dendrogram T to make clusters. Cut-off value can be selected within similarity values' boundary, [-1,1]. We finally get many sub trees with this cut-off value, and these sub trees would be clusters. As the cut-off value is closer to 1, the number of clusters is bigger.

Figure 2 shows an example of this drawback of overall leaf-ordering approach. We assume that figure 2 (a) and (b) have the same status at first time, and then a cut-off value was set up for dividing it into two clusters, A~F cluster and G~J clusters. Table 1 is this example's similarity table and Table 2 is this example's distance table. You can use these tables for applying any algorithms to this example case. We

used Bar-Joseph's optimal algorithm for this example. Figure 2 (a) shows the result that we applied that algorithm to all genes at once and figure 2 (b) shows the result that we applied it to each clusters. As you can see, (a)'s total distance is 3.87 and smaller than (b)'s. However, in the view of cluster level, the distance summation of (a)'s the second cluster (gene G~J) is 1.32 and larger than (b)'s. In other words, we need to consider only genes in the cluster for leaf-ordering. And we think that it is more reasonable to figure 2 (b)'s total distance is 2.68.

Table 1. Similarity table for figure 1 example

	A	B	C	D	E	F	G	H	I	J
A	1.00	0.40	0.20	0.10	0.06	0.04	-0.05	-0.07	-0.09	-0.09
B		1.00	0.90	0.65	0.50	0.30	0.07	0.03	0.00	-0.04
C			1.00	0.70	0.45	0.28	0.09	0.02	-0.02	0.00
D				1.00	0.50	0.32	0.10	-0.01	0.00	-0.10
E					1.00	0.85	-0.01	-0.04	0.01	-0.03
F						1.00	0.10	-0.10	-0.20	-0.21
G							1.00	0.80	0.64	0.57
H								1.00	0.60	0.41
I									1.00	0.28
J										1.00

Table 2. Distance table for figure 1 example

	A	B	C	D	E	F	G	H	I	J
A	0.00	0.60	0.80	0.90	0.94	0.96	1.05	1.07	1.09	1.09
B		0.00	0.10	0.35	0.50	0.70	0.93	0.97	1.00	1.04
C			0.00	0.30	0.55	0.72	0.91	0.98	1.02	1.00
D				0.00	0.50	0.68	0.90	1.01	1.00	1.10
E					0.00	0.15	1.01	1.04	0.99	1.03
F						0.00	0.90	1.10	1.20	1.21
G							0.00	0.20	0.36	0.43
H								0.00	0.40	0.59
I									0.00	0.72
J										0.00

gene distance = (1 – gene similarity)

4. Experiments and Results

We used various dataset for testing these two approaches. The three random dataset was created by random number generator and normalized by Cluster software[3]. Another three dataset are Spellman's yeast data[8] and the last one is Alon's cell life cycle experiment data[4]. We used Bar-Joseph's optimal leaf-ordering algorithms for the two approaches.

Table 3 shows experimental results of overall leaf-ordering approach and each cluster approach under Bar-Joseph's algorithm. Especially figure 3 shows a snap shot of experiment for Spellman800 dataset. We have set cut-off values for making 16-19 clusters through all experiments. We found that total distances are always smaller using our proposed approach than using overall leaf-ordering approach in Table 3. So we sure that leaf-ordering for each cluster approach is more reasonable approach.

Table 3. Comparison of total distance of all clusters for overall leaf-ordering approach and leaf-ordering for each cluster approach

	Bar-Joseph's Scheme			
Dataset	cut-off value	number of clusters	Overall Leaf-Ordering	Leaf-Ordering for Each Cluster
Random300_20	0.20	17	112.65	111.82
Random500_20	0.20	19	174.52	173.53
Random800_20	0.20	19	271.29	269.52
Spellman800	0.40	17	270.16	269.40
Spellman523	0.25	18	186.49	185.68
Spellman1033	0.30	16	479.12	478.62
Alon2000	0.35	19	660.80	660.45

5. Conclusion

In this paper, we proposed a new approach for leaf-ordering algorithm and we presented the experimental results for comparing the existing approach and our approach. The main problem of overall leaf-ordering approach is that it doesn't consider the fact that most of biological researchers try to analysis genes in each clusters. However, in our approach, leaf-ordering can be processed in each class independently. This will offer a better perspective view for scientists in the field of DNA chip expression analysis.

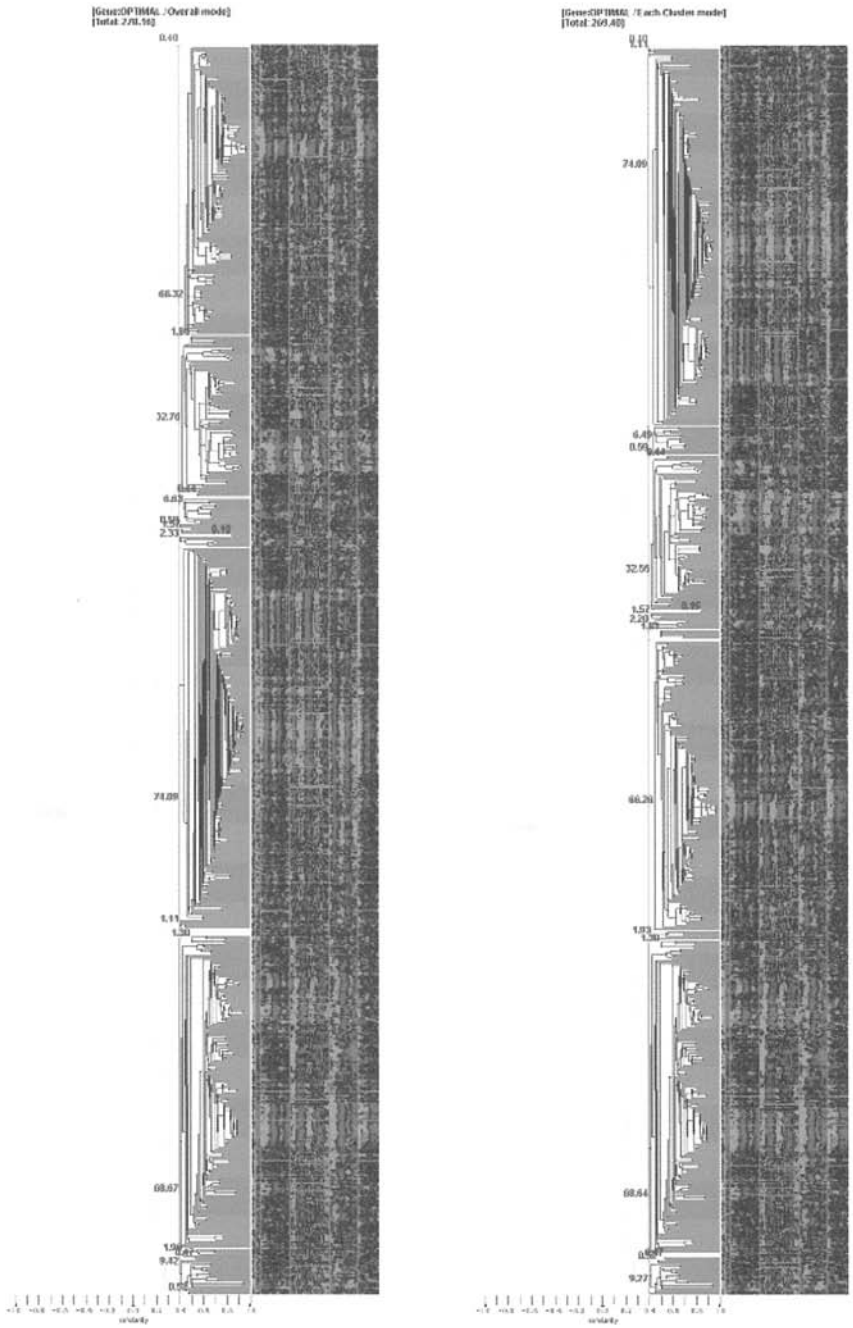


Figure 3. Experimental results for Spellman800 dataset

References

[1] R. Shamir and R. Sharan, "Algorithmic approaches to clustering gene expression data", Current Topics in Computational Biology, MIT Press.

[3] M. Eisen et al., "Cluster analysis and display of genome-wide expression patterns", Proc. of Natl. Acad. Sci., 95:14863-14867, 1998.

[4] M. Eisen, "Cluster and TreeView Manual", Eisen Lab. Homepage (http://rana.lbl.gov/ manuals/Cluster TreeView.pdf)

[5] U. Alon et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", Proc. Natl. Acad. Sci., 96:6745-6750, 1999.

[6] Z. Bar-Joseph et al., "Fast optimal leaf ordering for hierarchical clustering", Proceedings of ISMB 2001. pp.s22-s29.

[7] Z. Bar-Joseph, Therese Biedl, et al., "Optimal Arrangement of Leaves in the Tree Representing Hierarchical Clustering of Gene Expression Data", Bioinformatics Research Group Homepage of University of Waterloo (http://monod.uwaterloo.ca/ supplements/01expr/art.pdf)

[8] S.-S Yeo, J.W. Rhee, S.K. Kim, "Development of Leaf Ordering Algorithm for Hierarchical Clustering for DNA Microarray Data", Proceedings of 2002 KISS Conference, Vol. A-29(1), pp.706-708, April 2002.

[9] P.T. Spellman et al. "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization", Molecular Biology of the Cell 9:3273-97.1998.