Complexity Study on Two Clustering Problems
(Extended Abstract)*

Louxin Zhang and Song Zhu

1 Department of Mathematics & KRDL, National University of Singapore
Singapore 117543
matzlx@nus.edu.sg
2 Bioinformatics Lab, Kent Ridge Digital Labs
Singapore 119613
zhusong@krdl.org.sg

Abstract. The complexity issues of two clustering problems are studied. We prove that the Smooth Clustering and Biclustering problems are NP-hard; we also propose an 0.5 approximation algorithm and 0.8 inapproximability for a simplified clustering problem.

1 Introduction

By allowing biologists to observe thousands of genes being on or off under certain conditions, cDNA microarray technology is becoming a powerful and versatile tool for studying many important attributes of genes. It has been used for gene functional assignments; it has been used for study of gene regulation networks of a living cell; it has also been used for cancer classification and diagnosis. Because a large number of genes are often involved in rich cDNA microarray experiments, clustering is necessary for gene expression analysis as it would partition genes into different classes each containing genes expressing in similar patterns under certain conditions (such as tissues, environments) (see [1,5,8,7,2,4]).

Recently, the authors proposed a novel clustering approach for overcoming data errors such as data missing [9] and expression inconsistency across different experiments [3] in the stage of clustering. The approach is based on the so-called smooth score [10]. Gene expression data generated from cDNA microarray experiments are usually given as matrices, where each entry is a gene expression value under a condition. We assume the rows of such a matrix correspond to the genes and the columns to the conditions. The smooth score is not defined as a pairwise dissimilarity measure like Euclidean distance; instead, it measures the deviation of the expression level of a gene from the average expression level of all the involved genes under a condition. For a cluster $I$ of genes obtained by considering conditions in $J$, its smooth score is formally defined in Equation (1).

We formulate the Smooth Clustering problem as, given a set of conditions,

* Supported in part by a NSTB grant.

© Springer-Verlag Berlin Heidelberg 2001
finding a largest cluster of genes with its smooth score below a threshold under the given conditions. We would also like to find a largest smooth ‘bicluster’ with its smooth score below a threshold, grouping genes and conditions simultaneously as proposed in [4], which is called the Smooth BiClustering problem. The authors proposed efficient greedy algorithms for the Smooth Clustering and BiClustering problems in [10]. These algorithms were shown to perform well in finding co-regulation patterns in a test with a yeast data set.

In this paper, we study the computational complexity of these two clustering problems. The Smooth Clustering and Biclustering problems are introduced in Section 2 and are proved to be NP-hard in Section 3. In Section 3, a variant of the Smooth Clustering problem is also proved to be NP-hard. In Section 4, we focus on a special case of the Smooth Clustering Problem where the input gene expression matrix has only one column. A polynomial-time $\frac{1}{2}$-approximation algorithm is presented for the special case. However, unless NP = P, there is no polynomial-time algorithm of approximation ratio better than $\frac{1}{2}$, even for the special case. Finally, we conclude the paper with an open problem in Section 5.

2 Clustering Problems

The gene expression data from cDNA microarray experiments is usually presented as a matrix. Each entry represents the relative abundance of the mRNA of a gene under a specific condition. Here, we assume that each row corresponds to a gene and a column to a condition. The logarithm transformation is often applied to gene expression values for converting multiplicative changes of the relative abundance into additive increments.

Let $A = (a_{ij})$ be a gene expression matrix with gene set $X$ and condition set $Y$. Any subsets $I \subseteq X$ and $J \subseteq Y$ specify a submatrix $A(I,J)$. We associate it with the following smooth score

$$s(I,J) = \max_{j \in J} \max_{i \in I} |a_{ij} - \frac{1}{|I|} \sum_{k \in I} a_{kj}|,$$

where $\frac{1}{|I|} \sum_{k \in I} a_{kj}$ denotes the average expression level of genes in $I$ under condition $j$. The smooth score $s(I,J)$ is actually a refinement of $L_\infty$-distance $d_\infty(\cdot,\cdot)$, a popular metric in functional analysis. Recall that, for any two $n$-dimensional vectors $x = (x_i)$ and $y = (y_i)$, $d_\infty(x,y) = \max_i |x_i - y_i|$. If a gene expression level is considered as a function with condition as variable, clustering aims to classify genes into groups each containing genes with expression functions in similar shapes. Thus, we propose the smooth score for gene expression analysis. If $A(I,J)$ has the smooth score $s(I,J)$, then, for any rows $v$ and $v'$ in $A(I,J)$, $d_\infty(v,v') \leq 2s(I,J)$.

Given a small number $\epsilon > 0$, $A(I,J)$ is an $\epsilon$-smooth cluster if $s(I,J) \leq \epsilon$. We formulate clustering as the following problem.