Computational Complexity of Bi-clustering

by

Saeed Hassanpour Ghady

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2007

Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Saeed Hassanpour Ghady

Abstract

Bi-clustering, i.e. simultaneously clustering the rows and columns of matrices based on their entries, covers a large variety of techniques in data mining. The goal of all bi-clustering techniques is finding the partitions of the rows and the columns in which sub-rows and sub-columns show a similar behavior. Currently existing algorithms for bi-clustering problems are either heuristic, or try to solve approximations of the original problems. There is no efficient algorithm for exact bi-clustering problems.

The computational complexity of bi-clustering problems depends on the exact problem formulation, and particularly on the merit function used to evaluate the quality of a given bi-clustering partition. The computational complexity of most of the common bi-clustering problems is unknown. In this thesis, we present a formal definition for the homogeneous cover problem. This problem has many applications from bio-informatics to targeted marketing. We analyze its computational complexity and show that the problem is *NP-hard*.

Acknowledgments

I would like to express my gratitude to my supervisor, Prof. Shai Ben-David, for his support, encouragement and many productive discussions.

I would also like to thank professors Pascal Poupart and Alex López-Ortiz for reading my thesis and their constructive comments. I wish to thank Sharon Wulff for her helpful discussions, and Navid Hassanpour for proofreading this thesis.

Dedication

To my father, mother, brother, and sister.

Contents

List of Figures

Chapter 1

Introduction

A matrix is a common structure to present data in science. In many applications in different fields from astronomy to business management, matrices are widely used to show data in more structural and meaningful forms. In many situations in scientific analysis and data mining, we are interested in finding blocks that their rows and columns show a similar behavior. Generally, this problem is called bi-clustering or co-clustering, or two-mode clustering [25]. Different from clustering methods which are applied on either the rows or the columns of a matrix, bi-clustering methods perform clustering in the two dimensions simultaneously. Depending on the definition of the similarity, bi-clustering covers a large variety of problems.

Bi-clustering has many applications in different fields. For instance, in biology the expression level of a gene is represented by a real number that is the logarithm of the relative abundance of the mRNA of the gene. Several techniques such as DNA Chips, measure the expression level of genes within different experimental conditions [19]. Usually, gene expression data is arranged in a data matrix, where each gene corresponds to one row and each condition to one column. Relationship between certain genes and certain conditions is important because it can reveal the causes of some diseases like cancer. Clustering methods on rows and columns in gene expression data matrix are not useful, because many activation patterns are common to a group of genes only under specific experimental conditions. However by using bi-clustering methods we can find submatrices, that is sub-groups of genes and conditions, where the genes exhibit highly correlated activities for every condition (Figure 1.1) [1, 10, 13, 15, 20].

Bi-clustring is a common and useful approach in data mining [2, 3, 24]. For instance, in targeted marketing, online sellers want to offer interesting offers to each individual

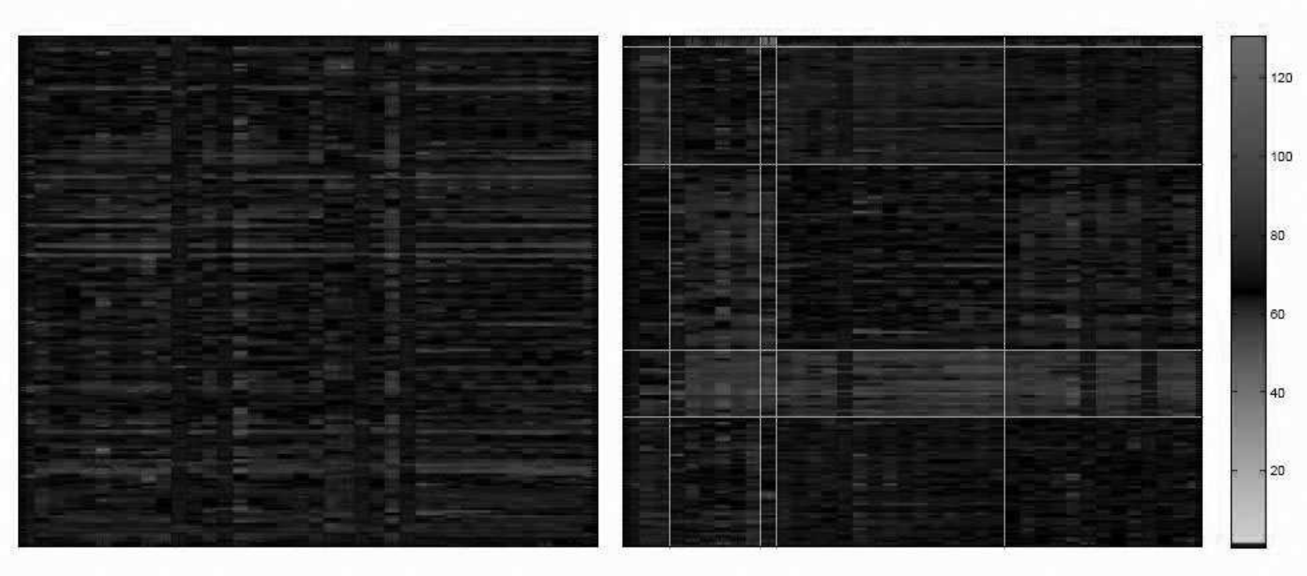


Figure 1.1: A sample of bi-clustering in computational biology. The left figure shows a gene expression data matrix, and the right figure shows the bi-clustering result on it.

costumer based on her record. For instance, in a website which sells movies, each costumer can rank her favorite movies. These rankings can be considered as the entries of a matrix, which its rows are web-site's costumers and its columns are all available movies in the website. A group of costumers that rank a group of movies similarly, usually have the same movie interests, and potentially are interested in similar movies. Finding similar groups of costumers and movies in this matrix is a typical example of bi-clustering.

Regarding the numerous applications of bi-clustering in various fields, efficiency of the bi-clustering algorithms is a critical issue. Currently existing algorithms for bi-clustering problems are either heuristic and customized for specific applications [4, 5, 6, 7, 9, 12, 21, 22, 23, 26], or try to solve approximations of the original problems [14, 16, 18]. The computation complexity of the most common bi-clustering problems is unknown. In this thesis, we analyze the computation complexity of one of the most common bi-clustering problems, the homogeneous cover problem.

1.1 Outline

The structure of this thesis is as follows: in Chapter 2, we investigate the common problems in the bi-clustering literature and known facts about their computational complexity.

In addition, we present the existing algorithms that try to solve these bi-clustering problems. In Chapter 3, we present a special version of the homogeneous cover problem with some restrictions on the number of row and column clusters, and show that it is *NP-hard*. Then, we present the general form of the homogeneous cover problem, and prove that it is *NP-hard* too. In Chapter 4, we conclude the thesis and present future work.

Chapter 2

Background

In this chapter, we present the background from the bi-clustering literature.

2.1 Bi-clustering Problems

There are various problems in the field of bi-clustering. The common idea in all of these problems is searching for the blocks such that their columns and rows have similar patterns. Following is the list of common problems in the bi-clustering literature.

Problem 1 *For a $m \times n$ matrix of real numbers, $\mathcal{A}$, and two integers k and l ($1 \leq k \leq m$, $1 \leq l \leq n$), divide the rows into k subsets: $r_1, r_2, \ldots, r_k$ and divide the columns into l subsets: $c_1, c_2, \ldots, c_l$, such that for the resulting $k \times l$ blocks, $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{k\times l}$, $\sum_{p=1}^{k\times l} \sum_{i,j\in\mathcal{B}_p} (\mathcal{A}(i,j) - average(\mathcal{B}_p))^2$ is minimized (For a matrix $\mathcal{M}$, $average(\mathcal{M})$ is the average of all $\mathcal{M}$'s elements.) [8].*

The objective function of problem 1 is close to k-means problem. By minimizing the objective function in this problem, in each block, deviation of the elements from their average is minimized, and the resulting blocks are approximately constant.

Problem 2 (Ben-David 07) *For a $m\times n$ matrix of ones and zeros, $\mathcal{A}$, and two integers k and l ($1 \leq k \leq m$, $1 \leq l \leq n$), divide the rows into k subsets: $r_1, r_2, \ldots, r_k$ and divide the columns into l subsets: $c_1, c_2, \ldots, c_l$, such that for the resulting $k \times l$ blocks, $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{k\times l}$, $\sum_{p=1}^{k\times l} \sum_{i,j\in\mathcal{B}_p} |\mathcal{A}(i,j) - Majority(\mathcal{B}_p)|$ is minimized, where*

$$Majority(i,j) = \begin{cases} 1 & \textit{if more than half of the elements of } A(i,j) \textit{ are one}; \\ 0 & \textit{otherwise.} \end{cases}$$

The objective function in problem 2 is minimizing the minority in each block. In this problem, the goal is building approximately constant blocks by minimizing the sum of the errors in the all blocks.

In both problems the resulting blocks are approximately constant, therefore, we can summarize a large matrix to a $m \times n$ matrix that reveals the similarities and patterns in the original matrix. These problems have many applications in different fields of science such as data mining and computational biology. The applications in targeted marketing and gene expression data, were mentioned in introduction, can be considered as one of these problems.

2.2 Computational Complexity of Bi-clustering Problems

Problem 1 is presented in 1972 by Hartigan. Since that time this problem has been appeared in many fields and applications. However, there is no known result about the computational complexity of this problem. In bi-clustering literature, there are several heuristics and approximation algorithms to solve this problem, without any guarantee for their results. Recently, Ben-David and Wulff proved that problem 2 is *NP-hard.*

Chapter 3

The Homogeneous Cover Problem in Bi-clustering

An important problem in bi-clustering is the homogeneous cover problem. This problem has many applications in different fields from computational biology to data mining. All applications in genetics and targeted marketing which were mentioned in the introduction can be tailored to this problem. Therefore, the computational complexity of the homogeneous cover problem is an important concern in many areas. In this chapter, We investigate the computational complexity of the homogeneous cover problem.

3.1 Computational Complexity of the Homogeneous Cover Problem.

To present the homogeneous cover problem, first, we provide some related definitions.

Definition 3.1.1 (Homogeneous matrix) *A matrix, $\mathcal{M}$, is homogeneous if and only if all of its elements are equal.*

Definition 3.1.2 (Area of a matrix) *For a matrix, $\mathcal{M}$, area of the matrix, $|\mathcal{M}|$, is the number of the elements in $\mathcal{M}$.*

Problem 3 (The homogeneous cover problem) *For a $m \times n$ matrix of ones and zeros ($m, n > 2$), $\mathcal{A}$, and an integers k ($2 \leq k \leq m, n$), divide the rows into k subsets: $r_1, r_2, \ldots, r_k$ and the columns into k subsets: $c_1, c_2, \ldots, c_k$, such that $\sum_{1 \leq i,j \leq k} U(i,j)$ is*

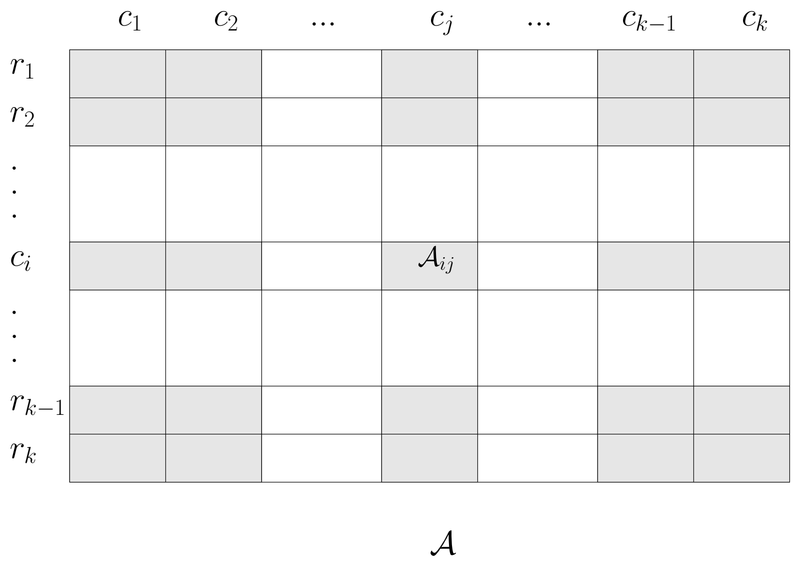


Figure 3.1: The homogeneous cover problem, the rows and the columns of the matrix are divided into k groups.

maximized. Where $\mathcal{A}_{i,j}$ is a submatrix which its rows belong to r_i and its columns belong to c_j, and $U(i,j)$ and $Majority(i,j)$ are defined as below (Figure 3.1):

$$U(i,j) = \begin{cases} |\mathcal{A}_{i,j}| & \text{if } \forall a \in \mathcal{A}_{i,j} : a = Majority(i,j); \\ 0 & \text{otherwise.} \end{cases}$$

$$Majority(i,j) = \begin{cases} 1 & \text{if more than half of the elements of } \mathcal{A}_{i,j} \text{ are one}; \\ 0 & \text{otherwise.} \end{cases}$$

There is no efficient algorithm to solve the homogeneous cover problem. Here we show that an efficient algorithm for this problem does not exist, unless P=NP.

Theorem 3.1.3 *The homogeneous cover problem is* NP-hard.

Proof We prove this theorem by induction. First we show the *NP-hardness* of the homogeneous cover problem when $k = 2$, as the basis for our induction. To do so, we prove that the largest homogeneous rectangle on a restricted domain is *NP-hard*, then we present α structure for a matrix and show that the largest homogenous rectangle on a restricted domain problem on a matrix with α structure is reducible to the homogeneous cover problem with $k = 2$.

Claim 3.1.4 *The homogeneous cover problem with two clusters on rows and columns is* NP-hard.

To prove this claim, first, we present some definitions and investigate related problems and their computational complexity.

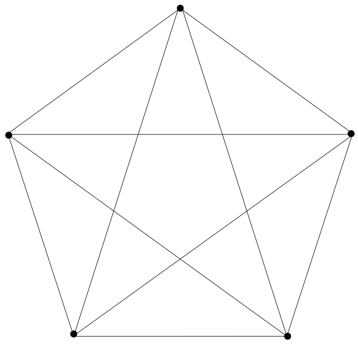


Figure 3.2: A clique of size 5.

Definition 3.1.5 (Clique) *A clique in an undirected graph, G, is a set of vertices, V, such that for every two vertices in V, there exists an edge connecting them (Figure 3.2).*

Problem 4 (Clique problem) *Given a graph $G = (V, E)$ and a positive integer K, does G contain a clique with at least size K?*

Fact 3.1.6 *The Clique Problem is* NP-complete *[11].*

Definition 3.1.7 (Bipartite graph) *A bipartite graph is a graph whose vertices can be divided into two disjoint sets, V_1 and V_2, such that every edge connects a vertex in V_1 and a vertex in V_2; i.e. there is no edge between two vertices in the same set (Figure 3.3).*

Definition 3.1.8 (Bi-clique) *Let $G = (V, E)$ be a graph with vertex set V and edge set E. A pair of two disjoint subsets A and B of V is called a bi-clique if $\{a, b\} \in E$ for all $a \in A$ and $b \in B$ (Figure 3.4).*

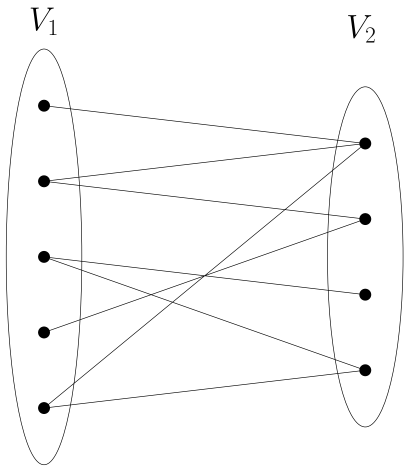


Figure 3.3: A bipartite graph with two disjoint sets, V_1 and V_2

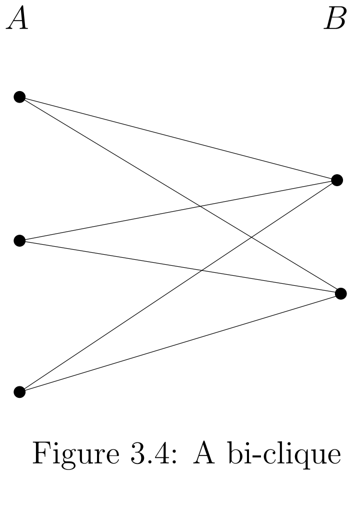


Figure 3.4: A bi-clique

Problem 5 (The maximum edge bi-clique problem) *Given a bipartite graph $G = (V_1 \cup V_2, E)$ and a positive integer K, does G contain a bi-clique with at least K edges?*

Peeters in [17] has reduced the maximum edge bi-clique problem to the clique problem.

Claim 3.1.9 *The maximum edge bi-clique problem is* NP-complete *[17].*

Problem 6 (The largest homogeneous rectangle on a restricted domain) *For a real number, $C \leq 1$, and for a $m \times n$ matrix of 1's and 0's, $\mathcal{A}$, such that there exists a homogeneous rectangle $\mathcal{B} \subseteq \mathcal{A}$ and $|\mathcal{B}| = C|\mathcal{A}|$, find the largest homogeneous rectangle in $\mathcal{A}$.*

Claim 3.1.10 *The largest homogeneous rectangle on a restricted domain problem is* NP-hard.

Proof Consider a graph $G = (V, E)$, we construct a bipartite graph, $H = (V_1 \cup V_2, E')$, from G like this: $(k = \frac{1}{2}|V|)$

$V_1 = V$
$V_2 = E \cup \{e_1, \ldots, e_{\frac{1}{2}k^2 - k}\}$
$E' = \{\{v, e\} : v \in V; e \in E; v \notin e\} \cup \{\{v, e_i\} : v \in V; i = 1, \ldots, \frac{1}{2}k^2 - k\}$

Peeters [17] proves that H has a bi-clique with at least $k^3 - \frac{3}{2}k^2$ edges if and only if G has a clique of size k. In addition, adjacency matrix of H contains a homogeneous (all-1) submatrix of size $2k \times \frac{1}{2}k^2 - k$ which is equal to $C|H|$, for a $C \in \mathbb{R}, C < 1$. And because this all-1 rectangle's area is larger than $\frac{|\mathcal{H}|}{2}$, the largest homogeneous rectangle in $\mathcal{H}$ is an all-1 rectangle. If there is a polynomial algorithm for the largest homogeneous rectangle on a restricted domain problem, by applying it on H, we can decide if there is a clique of size k in the graph G or not. However, the clique problem is *NP-complete* and this is a contradiction. So, there is no such an algorithm for the largest homogeneous rectangle on a restricted domain problem, and it is *NP-hard.*

Proof of Claim 3.1.4 We reduce the largest homogeneous rectangle on a restricted domain problem to the homogeneous cover problem with $k = 2$.

For any $m \times n$ matrix of ones and zeros, $\mathcal{A}$, which contains a homogeneous rectangle $\mathcal{H} \subseteq \mathcal{A}$ and $|\mathcal{H}| = C|\mathcal{A}|$ $(C \in \mathbb{R}, C \leq 1)$, we construct a new matrix $\mathcal{A}'$, with the following structure. We call it an α structure.

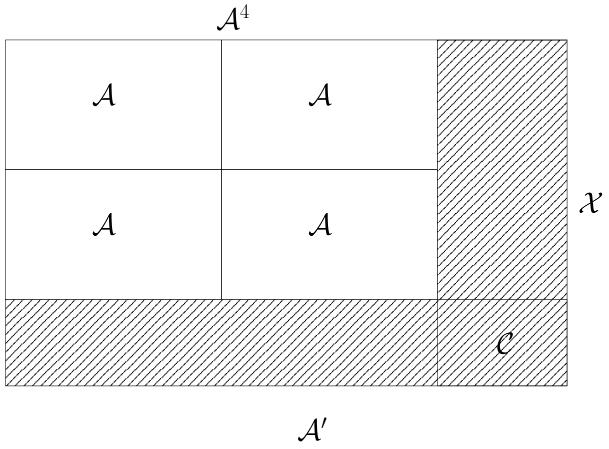


Figure 3.5: α structure.

Definition 3.1.11 **(α structure)** *A' has the α structure iff A' consists of two disjoint parts A^4 and $\mathcal{X}$ (Figure 3.5), with the following properties:*

1. *$\mathcal{A}^4$ is a $2m \times 2n$ matrix and consists of four copies of $\mathcal{A}$.*
2. *$|\mathcal{H}| > \frac{|\mathcal{A}|+2|\mathcal{X}|}{4}$.*
3. *No two rows and two columns of $\mathcal{X}$ are identical.*
4. *For every row or column which does not have any intersection with $\mathcal{A}^4$, half of the elements are ones and half of the elements are zeroes.*
5. *In a row (column) of $\mathcal{A}'$ which does not have any intersection with $\mathcal{A}^4$, each two elements on two similar columns (rows) of two adjacent $\mathcal{A}$'s are different.*
6. *In submatrix $\mathcal{C}$, none of the rows and columns are all-1 or all-0.*

We can always construct a matrix $\mathcal{A}'$ with α structure, from $\mathcal{A}$. One way to construct $\mathcal{A}'$ from $\mathcal{A}$ is:

First, four copies of $\mathcal{A}$ are combined together to make a $2m \times 2n$ matrix, $\mathcal{A}^4$. Then we add $\lceil \log m \rceil$ columns to $\mathcal{A}^4$ and put the binary presentations of 0 to $m-1$ on their first m rows, and put the binary presentations of the 0 to $m-1$ in the reverse order on the next m rows. Then we add $\lceil \log n \rceil$ rows to the matrix which was just created, put the binary presentations of 0 to $n-1$ on their first n columns, and put the binary presentations of 0 to $n-1$ in the reverse order on the next n columns. For the remaining $\lceil \log n \rceil \times \lceil \log m \rceil$ submatrix, we divide it into four quarters, and assign 0's to two quarters on the main

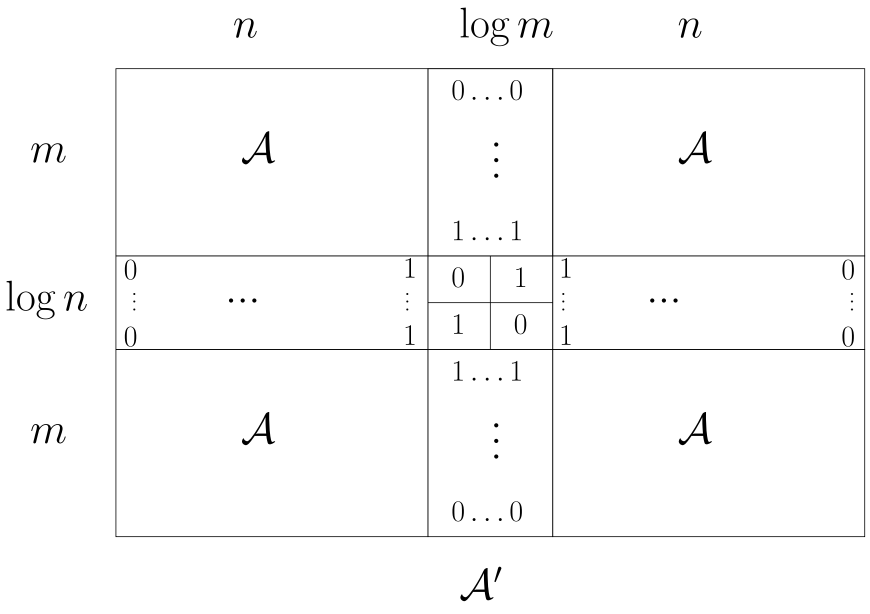


Figure 3.6: Construction of $\mathcal{A}'$ from $\mathcal{A}$.

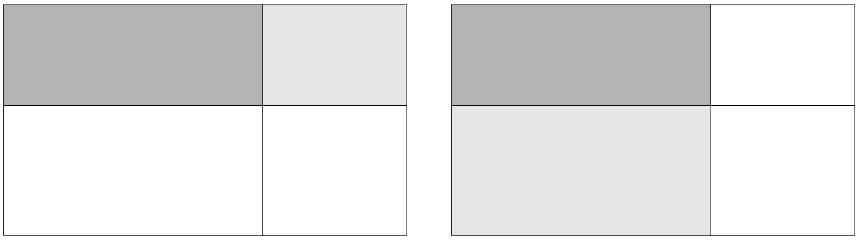
diagonal and assign 1's to the two remaining quarters. Because of property 2, $|\mathcal{X}|$ is exponentially less than $|\mathcal{A}|$. Therefore, initially by choosing a large enough C, where $|\mathcal{H}| = C|A|$, we can always satisfy property 2. It is easy to check that this matrix has the other properties of α structure, too. Figure 3.6 shows this construction.

Remark From property 2, we know that there is a homogeneous submatrix $\mathcal{H}$ in $\mathcal{A}$ such that $|\mathcal{H}| > \frac{|\mathcal{A}|+2|\mathcal{X}|}{4}$. By combining the four copies of $\mathcal{H}$'s from four $\mathcal{A}$'s, we can have a homogeneous submatrix of size $|\mathcal{A}| + 2|\mathcal{X}|$. Therefore, in the optimal partition of the homogeneous cover problem with $k = 2$ on $\mathcal{A}'$, the sum of the homogeneous areas is larger than $|\mathcal{A}| + 2|\mathcal{X}|$.

Lemma 3.1.12 *The optimal partition of the homogeneous cover problem with $k = 2$, on a matrix with α structure, $\mathcal{A}'$, has only one homogeneous block.*

Proof Assume that the optimal partition has more than one homogeneous block. This means there should be at least two homogeneous blocks in the optimal partition. Because the optimal partition has only four blocks, these homogeneous blocks can be in horizontal (have common rows, Figure 3.7(a)) or vertical (have common columns, Figure 3.7(b)) or diagonal positions (have no common row or column, Figure 3.8).

If they are in horizontal positions, because we already know that sum of the homogeneous areas is larger than $|\mathcal{A}| + 2|\mathcal{X}|$, there are at least 2 complete rows of $\mathcal{A}'$ in



(a) Two homogeneous blocks in horizontal positions.

(b) Two homogeneous blocks in vertical positions.

Figure 3.7: Horizontal and vertical positions

that horizontal area. However, from property 3, no two rows in $\mathcal{X}$ are identical. Therefore, no two rows in $\mathcal{A}' = \mathcal{A} + \mathcal{X}$ are identical and making two horizontal homogeneous blocks is not possible. By a similar argument we can show that the vertical positions are impossible too.

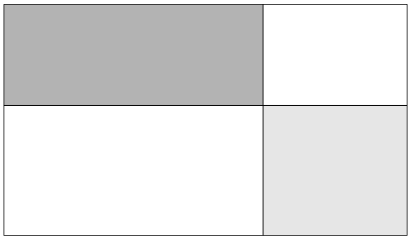


Figure 3.8: Two homogeneous blocks in diagonal positions.

In the case that two homogeneous blocks are in diagonal positions, if one of the blocks' width is larger than $\mathbf{Width}(A')/2$, this block cannot contain a column from $\mathcal{X}$, because only half of the element of each column from $\mathcal{X}$ are ones or zeros (property 4). Therefore the other block covers all the columns from $\mathcal{X}$. This block does not have any intersection with submatrix $\mathcal{C}$, because none of $\mathcal{C}$'s rows and columns are homogenous (property 6). Therefore, all of the block's rows are coming form out side of $\mathcal{C}$. If this block has more than one row, it contains at least two rows from $\mathcal{X}$. However, no two rows of $\mathcal{X}$ are identical (property 3), and the block cannot be homogeneous. If it has only one row, the wide block has to have more than two columns (because we know that the sum of the areas of the homogeneous blocks in the optimal partition is larger than $|\mathcal{A}| + 2|\mathcal{X}|$). Therefore, the wide block covers at least three columns, and only one element in each column is outside of the wide block. Because the wide block is homogeneous, two of these three columns have to be identical, which is against property 3. With a similar

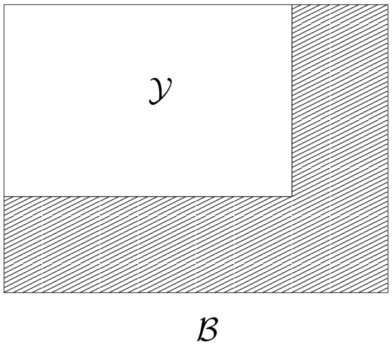


Figure 3.9: $\mathcal{B}$, when all elements of $\mathcal{Y}$ are from one $\mathcal{A}$.

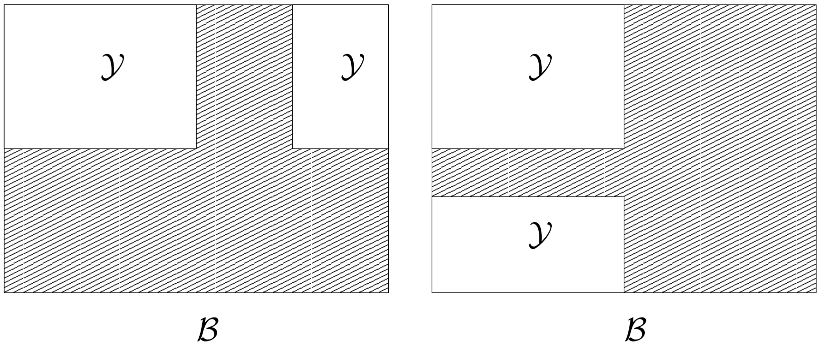
argument we can show that the length of the blocks cannot be larger than **Length**$(A')/2$ in diagonal positions.

In the case that dimensions of two diagonal blocks are equal to half of the dimensions of $\mathcal{A}'$, because dimensions of $\mathcal{A}^4$ are larger than 4 (because $m, n > 2$), to satisfy property 3 at least three rows or columns have to be added to $\mathcal{A}^4$. If three or more columns are added to $\mathcal{A}^4$, because half of the elements of each column are 0's and other half are 1's (property 4), and two diagonal blocks are homogeneous, there are only two possible orders for the added columns' elements, which are exactly reverse of each other in terms of the order of zeros and ones. Therefore, at least two columns from these three added columns are identical. However, it is against property 3. By a similar argument, we can show that we have a contradiction in the case that three or more rows are added to $\mathcal{A}^4$, and the two homogeneous blocks are in diagonal positions. Hence, diagonal positions for two homogeneous blocks are impossible.

Therefore, in the optimal partition for the homogeneous cover problem, with $k = 2$, having two homogeneous blocks in any positions is impossible. Because the number of the homogeneous blocks in the optimal partition is greater than or equal to one, we conclude that the number of the homogeneous blocks in the optimal partition is one.

Lemma 3.1.13 *If $\mathcal{B}$ is the largest homogeneous block in a matrix with α structure, $\mathcal{A}'$, then either $\mathcal{B} \subseteq \mathcal{A}^4$ or $|\mathcal{B} \cap \mathcal{X}| > |\mathcal{B} \cap \mathcal{A}^4|$.*

Proof Assume that $\mathcal{B} \nsubseteq A^4$. Let $\mathcal{Y} = \mathcal{B} - \mathcal{X}$. If all elements of $\mathcal{Y}$ are coming form one of $\mathcal{A}$'s in $\mathcal{A}'$ (Figure 3.9), then $|B| \leq |\mathcal{A}| + |\mathcal{X}|$. However, we already know that $|B| > |\mathcal{A}| + 2|\mathcal{X}|$, and this is a contradiction.

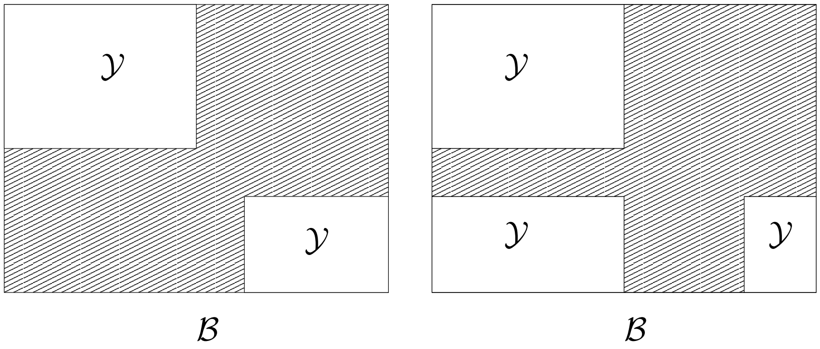


(a) Horizontal positions.

(b) Vertical positions.

Figure 3.10: Two possible cases, when all elements of $\mathcal{Y}$ are from two $\mathcal{A}$'s.

If all elements of $\mathcal{Y}$ are from two $\mathcal{A}$'s in the horizontal positions (Figure 3.10(a)), then $|\mathcal{B} \cap \mathcal{X}| > |\mathcal{B} \cap \mathcal{A}^4|$, because otherwise we can build a bigger homogeneous matrix only by duplicating $\mathcal{Y}$. A similar argument shows the similar result for the vertical positions (Figure 3.10(b)).



(a) All elements of $\mathcal{Y}$ are from two diagonal $\mathcal{A}$'s.

(b) All elements of $\mathcal{Y}$ are from three $\mathcal{A}$'s.

Figure 3.11: Two impossible cases for $\mathcal{B}$.

Because of the symmetric structure of $\mathcal{A}'$, $\mathcal{Y}$ does not consist of the elements of only two $\mathcal{A}$'s in the diagonal positions (Figure 3.11(a)), or three $\mathcal{A}$'s (Figure 3.11(b)). If the elements of $\mathcal{B}$ are from four $\mathcal{A}$'s (Figure 3.12), $|\mathcal{B} \cap \mathcal{X}| > |\mathcal{B} \cap \mathcal{A}^4|$. From property 5, in this case we know that there are no more than two symmetric parts in $\mathcal{Y}$ and duplicating $\mathcal{Y}$ guarantees a larger homogeneous area than $\mathcal{B}$, and this is a contradiction.

Therefore, in all possible cases for the largest homogeneous block in $\mathcal{A}'$, $\mathcal{B}$, either

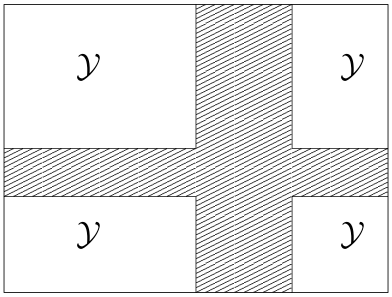


Figure 3.12: $\mathcal{B}$, when all elements of $\mathcal{Y}$ are from four $\mathcal{A}$'s.

$\mathcal{B} \subseteq \mathcal{A}^4$ or $|\mathcal{B} \cap \mathcal{X}| > |\mathcal{B} \cap \mathcal{A}^4|$.

Till now we proved that the optimal partitions of the homogeneous cover problem on a matrix with α structure, $\mathcal{A}'$, has one homogeneous block and the largest homogeneous block in $\mathcal{A}'$ is either a subset of $\mathcal{A}^4$ or its intersection with $\mathcal{X}$ is larger than its intersection with $\mathcal{A}^4$. We show that the largest homogeneous block in $\mathcal{A}'$ is one of the blocks of the optimal partition of the homogeneous cover problem with $k = 2$.

Claim 3.1.14 *One of the blocks in the optimal partition of the homogeneous cover problem with $k = 2$ on a matrix with α structure, $\mathcal{A}'$, is the largest homogeneous submatrix in A'.*

Proof The only homogeneous block in the optimal partition of the homogeneous cover problem with $k = 2$ is the largest homogeneous block in $\mathcal{A}'$. Otherwise, we can make a better partition by dividing rows and columns to have the largest homogeneous submatrix in $\mathcal{A}'$ as one of the blocks in the homogeneous cover problem's partition, and have a partition with a larger homogeneous block than the optimal partition.

From claim 3.1.14, we know that there is only one homogeneous block in the optimal partition of the homogeneous cover problem with $k = 2$ on a matrix with α structure, $\mathcal{A}'$, and it is either a subset of $\mathcal{A}^4$ or its intersection with $\mathcal{X}$ is larger than its intersection with $\mathcal{A}^4$. Now we show that the latter case is impossible.

Claim 3.1.15 *In the optimal partition of the homogeneous cover problem with $k = 2$ on a matrix with α structure, $\mathcal{A}'$, there is only one homogeneous block, $\mathcal{B}$, where $\mathcal{B} \subseteq \mathcal{A}^4$.*

Proof If $|\mathcal{B} \cap \mathcal{X}| > |\mathcal{B} \cap \mathcal{A}^4|$, the sum of the areas of the homogeneous blocks in $\mathcal{A}'$, is equal to the size of $\mathcal{B}$ and it is less than $|2\mathcal{X}|$. However, we know that the sum of the

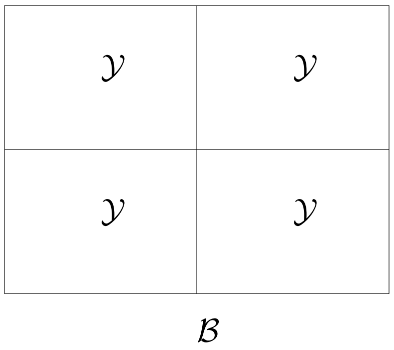


Figure 3.13: $\mathcal{B}$ consists of four similar blocks from four copies of $\mathcal{A}$'s.

areas of the homogeneous blocks in the optimal partition on a matrix with α structure is larger than $|\mathcal{A}| + 2|\mathcal{X}|$ which is a contradiction.

Claim 3.1.16 *In the optimal partition of the largest homogeneous cover problem with $k = 2$ on a matrix with α structure, $\mathcal{A}'$, there is only one homogeneous block, $\mathcal{B}$, which consists of four similar blocks from four copies of $\mathcal{A}$ in $\mathcal{A}'$ (Figure 3.13).*

Proof From claim 3.1.15, we know that in the optimal partition of the largest homogeneous cover problem with $k = 2$ on a matrix with α structure, $\mathcal{A}'$, one homogeneous block, $\mathcal{B}$, exists and $\mathcal{B} \subseteq \mathcal{A}^4$. If claim 3.1.16 is not true, then $\mathcal{B}$ consists of different blocks with either different or similar sizes form different $\mathcal{A}$'s. In the first case by adding four copies of the largest block from different $\mathcal{A}$'s, we can build a homogeneous submatrix larger than $\mathcal{B}$, and in the second case by duplicating $\mathcal{B}$, we can make a matrix twice larger than $\mathcal{B}$ and have contradictions in both cases.

Claim 3.1.17 *In the optimal partition of the homogeneous cover problem with $k = 2$ on a matrix with α structure, $\mathcal{A}'$, there is only one homogeneous block, $\mathcal{B}$, which consists of four copies of the largest homogeneous submatrix in $\mathcal{A}$.*

Proof From claim 3.1.16, we know that in the optimal partition of the largest homogeneous cover problem with $k = 2$, the homogeneous block, $\mathcal{B}$, consists of four similar blocks from four $\mathcal{A}$'s. This repeated block has to be the largest homogeneous submatrix in $\mathcal{A}$. Otherwise, we can build a larger homogeneous submatrix by combining the four copies of the largest homogeneous submatrix in $\mathcal{A}$.

$$\begin{array}{c|c|c|} & n & n \\ \hline m & \mathcal{A} & 0 \\ \hline m & 0 & 1 \\ \hline \end{array}$$

$$\mathcal{A}'$$

Figure 3.14: Construction of $\mathcal{A}'$ from $\mathcal{A}$.

Consider the homogeneous cover problem with $k = 2$ on a matrix with α structure, $\mathcal{A}'$. From claim 3.1.17, the projection of the homogeneous block in the optimal partition of the homogeneous cover problem, gives us the largest homogeneous submatrix in a matrix which contains a homogeneous submatrix with a size greater than or equal to $|\mathcal{H}|$. Therefore if there is an efficient algorithm for the homogeneous cover problem with $k = 2$, by applying it on $\mathcal{A}'$, we can solve the problem of the largest homogeneous rectangle on a restricted domain for $\mathcal{A}$. Therefore, the largest homogeneous rectangle on a restricted domain problem, is reduced to the homogeneous cover problem with $k = 2$. From claim 3.1.10, the largest homogeneous rectangle on a restricted domain problem is *NP-hard*. Therefore, the homogeneous cover problem with $k = 2$ is at least as hard as the largest homogeneous rectangle on a restricted domain problem, and is *NP-hard*.

Till now, we proved that the homogeneous cover problem with $k = 2$ is *NP-hard*. This is the basis for our induction. Now we prove that if the homogeneous cover problem is *NP-hard* for $k = p$ $(2 \leq p)$, then the homogeneous cover problem with $k = p + 1$, is *NP-hard*.

Consider a $m \times n$ matrix of zeros and ones, $\mathcal{A}$. We construct a new $2m \times 2n$ matrix from $\mathcal{A}$. We add m rows of zero to the bottom of matrix $\mathcal{A}$, and n columns to the right side of the matrix which is just created. We put elements of the first m rows of the new added columns equal to zero, and the elements of the next m rows of the new added columns equal to one. We call the constructed matrix A'. From the construction, it is obvious that, A' consists of four disjoint $m \times n$ submatrices: $\mathcal{A}$, two all-zero matrices and one all-one matrix (Figure 3.14).

If we consider the optimal partition for the homogeneous cover problem with $k = p+1$, in the optimal partition, no row from the first m rows of A' can be grouped with any row form the last m rows of A'. Because in this case just by putting the last m rows of A' in one row cluster, we can have a better partition, and this is a contradiction.

If the entire last m rows of A' are grouped to one row cluster, it means that the first m rows of A' are divided to p groups. If the last m rows of A' are divided into more than one row clusters, it means that the first m rows of A' are dividable into p' ($p' < p$) row clusters and each row cluster consists of similar rows. Because all the rows in each of these x' row clusters are similar, any optimal row cluster for the homogeneous cover problem with $k = p$ ($p > p'$), is achievable from these p' row clusters (Just by dividing several clusters into two clusters, to make enough row clusters.).

By a similar argument, we can show that the optimal partition for the homogeneous cover problem with $k = p + 1$, builds the optimal p column clusters on the first n columns. Therefore, the homogeneous cover problem with $k = p$ on $\mathcal{A}$, is reducible to the homogenous cover problem with $k = p + 1$, on A'.

Because the theorem 3.1.3 is true for the basis of the induction ($k = 2$), and all natural numbers are achievable from the basis of the induction by applying appropriate numbers of increments, the homogeneous cover problem is *NP-hard* for any fixed k ($k \in \mathbb{N}$).

We proved that the homogeneous cover problem is *NP-hard* in the general case, where the number of row and column clusters can be any fixed arbitrary integer. As most computer scientists believe that $P \neq NP$, it is probable that there is no efficient algorithm for the homogeneous cover problem.

Chapter 4

Conclusion

In this thesis, we have introduced a formal definition for the homogeneous cover problem. This problem has a variety of applications from computational biology to data mining. We have shown that the homogeneous cover problem can be reduced to the clique Problem on an undirected graph, and consequently, the homogeneous cover problem is *NP-hard*.

4.1 Future Work

Bi-clustering covers a large variety of problems. The computational complexity of bi-clustering problems depends on the exact problem formulation, and particularly on the merit function used to evaluate the quality of a given bi-clustering partition. The computational complexity of most of the common bi-clustering problems is unknown. We can continue our work by investigating the computational complexity of other problems in the field.

In this thesis, we investigated the computational complexity of the homogeneous cover problem, and showed that this problem is *NP-hard*, and there exists no efficient algorithm to solve it, unless P=NP. The next logical step is finding an algorithm that finds a result close to the problem's optimal partition in a reasonable time. Probably heuristic techniques and approximation algorithms will be useful for finding such a solution.

Bibliography

[1] S. Busygin, G. Jacobsen, and E. Kramer. Double conjugated clustering applied to leukemia microarray data. In *2nd SIAM ICDM, Workshop on clustering high dimensional data*, 2002.

[2] J. Carrasco, D. Fain, K. Lang, and L. Zhukov. Clustering of bipartite advertiser-keyword graph, 2003.

[3] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274, New York, NY, USA, 2001. ACM Press.

[4] Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretic co-clustering. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98, New York, NY, USA, 2003. ACM Press.

[5] Gad Getz, Erel Levine, and Eytan Domany. Coupled two-way clustering analysis of gene microarray data. *PNAS*, 97(22):12079–12084, 2000.

[6] Michelangelo Grigni and Fredrik Manne. On the complexity of the generalized block distribution. In *IRREGULAR '96: Proceedings of the Third International Workshop on Parallel Algorithms for Irregularly Structured Problems*, pages 319–326, London, UK, 1996. Springer-Verlag.

[7] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data, 2002.

[8] J. A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

[9] Nicholas A. Heard, Christopher C. Holmes, David A. Stephens, David J. Hand, and George Dimopoulos. Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. *PNAS*, 102(47):16939–16944, 2005.

[10] Je-Gun Joung, Dongho Shin, Rho Hyun Seong, and Byoung-Tak Zhang. Identification of regulatory modules by co-clustering latent variable models: stem cell differentiation. *Bioinformatics*, 22(16):2005–2011, 2006.

[11] Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller and James W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[12] Yuval Kluger, Ronen Basri, Joseph T. Chang, and Mark Gerstein. Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome Res.*, 13(4):703–716, 2003.

[13] Abraham Kupfer, S. J. Singer, Charles A. Janeway, and Susan L. Swain. Coclustering of CD4 (L3T4) Molecule with the T-Cell Receptor is Induced by Specific Direct Interaction of Helper T Cells and Antigen-Presenting Cells. *PNAS*, 84(16):5888–5892, 1987.

[14] Stefano Lonardi, Wojciech Szpankowski, and Qiaofeng Yang. Finding biclusters by random projections. *Theor. Comput. Sci.*, 368(3):217–230, 2006.

[15] Sara C. Madeira and Arlindo L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(1):24–45, 2004.

[16] Nina Mishra, Dana Ron, and Ram Swaminathan. On finding large conjunctive clusters. In *COLT*, volume 2777 of *Lecture Notes in Computer Science*, pages 448–462. Springer, 2003.

[17] René Peeters. The maximum edge biclique problem is NP-complete. *Discrete Appl. Math.*, 131(3):651–654, 2003.

[18] Cecilia M. Procopiuc, Michael Jones, Pankaj K. Agarwal, and T. M. Murali. A monte carlo algorithm for fast projective clustering. In *SIGMOD '02: Proceedings*

of the 2002 ACM SIGMOD international conference on Management of data, pages 418–427, New York, NY, USA, 2002. ACM Press.

[19] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–70, 1995.

[20] Karen Schluter and Detlev Drenckhahn. Co-Clustering of Denatured Hemoglobin with Band 3: Its Role in Binding of Autoantibodies against Band 3 to Abnormal and Aged Erythrocytes. *PNAS*, 83(16):6137–6141, 1986.

[21] Qizheng Sheng, Yves Moreau, and Bart De Moor. Biclustering microarray data by gibbs sampling. *Bioinformatics*, 19:196–205, 2003.

[22] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data, 2002.

[23] Chun Tang, Li Zhang, Murali Ramanathan, and Aidong Zhang. Interrelated two-way clustering: An unsupervised approach for gene expression data analysis. In *BIBE '01: Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering*, page 41, Washington, DC, USA, 2001. IEEE Computer Society.

[24] William-Chandra Tjhi and Lihui Chen. A partitioning based algorithm to fuzzy co-cluster documents and words. *Pattern Recogn. Lett.*, 27(3):151–159, 2006.

[25] Iven Van Mechelen, Hans-Hermann Bock, and Paul De Boeck. Two-mode clustering methods: A Structured Overview. *Statistical Methods in Medical Research*, 13(5):363–394, 2004.

[26] Haixun Wang, Wei Wang, Jiong Yang, and Philip S. Yu. Clustering by pattern similarity in large data sets. In *SIGMOD '02: Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pages 394–405, New York, NY, USA, 2002. ACM Press.