Implementation of plaid model biclustering method on microarray of carcinoma and adenoma tumor gene expression data

Gianinna Ardaneswari, Alhadi Bustamam, and Devvi Sarwinda
Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, Indonesia 16424
E-mail: {ninna, alhadi, devvi}@sci.ui.ac.id

Abstract. A Tumor is an abnormal growth of cells that serves no purpose. Carcinoma is a tumor that grows from the top of the cell membrane and the organ adenoma is a benign tumor of the gland-like cells or epithelial tissue. In the field of molecular biology, the development of microarray technology is used in the data store of disease genetic expression. For each of microarray gene, an amount of information is stored for each trait or condition. In gene expression data clustering can be done with a bicluster algorithm, thats clustering method which not only the objects to be clustered, but also the properties or condition of the object. This research proposed Plaid Model Biclustering as one of biclustering method. In this study, we discuss the implementation of Plaid Model Biclustering Method on microarray of Carcinoma and Adenoma tumor gene expression data. From the experimental results, we found three biclusters are formed by Carcinoma gene expression data and four biclusters are formed by Adenoma gene expression data.

1. Introduction
Bioinformatics is a branch of science that studies the application of computational techniques in managing and analyzing biological data. Molecular Biology research evolves through the development of the technologies used for carrying them out. It is not possible to research on a large number of genes using traditional methods. DNA Microarray is one such technology which enables the researchers to investigate and address issues which were once thought to be non traceable. One can analyze the expression of many genes in a single reaction quickly and in an efficient manner. DNA Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of the human body. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome.

DNA Microarray or gene-chip was developed in the 1990s to determine the transcriptional thousands of genes in one (or several) experiments on a small membrane or glass slide. There are a number of ways to make a DNA microarray. One method starts with PCR amplification of gene sequences that are known from an organism.

The development of microarray technology has provided an opportunity for researchers to explore the availability of genes an organism associated with genes we are studying. Microarray data provided are processed with the help of computerized measurement expression profiling, or by looking at the proximity of a pair of genes through stringensi degrees. From here generated a database that can predict the linkage of a gene with other genes. These results are usually displayed in the form of clusters which is the result of grouping genes that have similar expression patterns.
One of the most topics that used in bioinformatics is about clustering and also biclustering for two-dimensional data. In 2014, Das et al. proposed two-phase method for finding a bicluster. In the first phase a modified k-means algorithm is applied. In the second phase the residue score of Cheng and Church is applied to each of the columns of the clusters [1]. Biclustering is a very useful data mining technique which identifies coherent patterns from microarray gene expression data. A bicluster of a gene expression dataset is a subset of genes which exhibit similar expression patterns along a subset of conditions. Biclustering is a powerful analytical tool for the biologist and has generated considerable interest over the past few decades.

The aim of biclustering is to identify subset pairs (each pair consisting of a subset of genes and a subset of conditions) by clustering both the rows and columns of an expression matrix. Hence, biclustering algorithms must guarantee that the output biclusters are meaningful. This is usually done by accompanying statistical model or a heuristic scoring method that defines which of the many possible sub matrices represent a significant biological behavior. The biclustering problem is to find a set of significant biclusters in a matrix. Usually, gene expression data is arranged in a data matrix, where each gene corresponds to one row and each condition to one column [1].

Clustering is a search method for hidden patterns that may exist in datasets. It is a process of grouping data objects into disjointed clusters so that the data in each cluster are similar, yet different to the other clusters. K-means is one of the most famous and typical clustering algorithms and applied in many application areas. K-means has the advantages of fast convergence and ease of implementation, but it has poor performance in some applications with large dataset.

Poor performance as the computational problem can be solved by parallel computing simultaneously. Parallel processing is the processing of program instructions by dividing them among multiple processors with the objective of running a program in less time. Parallel computing is using multiple independent processors simultaneously or utilize GPU computing to solve a problem. In this issue typically used for processing very large data and high computational demands.

Therefore, by the development of bioinformatics analysis, related to increasing the number of data that to be analyzed, encourage researchers to propose methods for better results and efficient. In this study, we proposed a two-phase method. Two-phase in this method consists of two phases, the first phase is the implementation of parallel algorithms k-means on gene expression data to generate $k$ clusters. The second phase is the implementation of bicluster algorithm, we applied plaid model biclustering algorithm. The application of the two-phase method on these gene expression data aims to find good-quality bicluster effectively.

2. Material and Methods

2.1. K-Means Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume $k$ clusters) fixed a priori. The main idea is to define $k$ centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate $k$ new centroids as barycenters of the clusters resulting from the previous step. After we have these $k$ new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the $k$ centroids change their location step by step until no more changes are done. In other words centroids do not move any more [7].

2.2. Parallel K-Means Algorithm

K-means algorithm is one of the most famous unsupervised clustering algorithms. K-means has the advantages of fast convergence and ease of implementation, but it has poor performance in some
applications with large dataset.

Meanwhile, poor performance as the computational problem can be solved by parallel computing simultaneously. Parallel computing is a type of computational method that the execution of processes are carried out simultaneously. Large problems can often be divided into smaller ones, which can be solved at the same time. So, we can apply parallel computing in this k-means algorithm.

2.3. Plaid Model Biclustering

In the field of molecular biology, the development of microarray technology is used in the data store of disease genetic expression. Biclustering has become a popular technique for the study of gene expression data. For each of microarray gene, an amount of information is stored for each trait or condition. In gene expression data clustering can be done with a bicluster algorithm, that is clustering method which not only the objects to be clustered, but also the properties or condition of the object.

Lazzeroni and Owen proposed plaid models. In this approach, the genes-condition matrix is represented as a superposition layers, corresponding to bicluster [2].

\[ Y_{ij} = \sum_{k=0}^{k} \theta_{ijk} \rho_{jk} \kappa_{jk} \]  

(1)
Figure 2. Clustering & Biclustering Analysis

Where $Y_{ij}$ refers to the expression level of gene $i$ under sample $j$ in the input matrix, $k$ is the number of biclusters, $\theta_{ij0}$ describes the background layer and $\theta_{ijk}$ represents four different types of models, depending on the types of biclusters. Each $\rho_{jk} \in \{0, 1\}$ is 1 if gene $i$ is in the $k$th bicluster, zero otherwise. Similarly, each $\kappa_{jk} \in \{0, 1\}$ is 1 if sample $j$ is in the $k$th bicluster, zero otherwise. Using this equation, a bicluster is assumed to be the sum of a bicluster background level plus row-specific and column-specific constants.

In order to find $k$ biclusters in the data, Lazzeroni and Owen proposed a greedy algorithm that adds one layer at a time. The process seeks for a plaid model minimizing the sum of squared errors when approximating the data matrix to the model [4]. In the eq. (1) allows a gene to be more than one bicluster or in none at all.

In order to find $k$ biclusters in the data, Lazzeroni and Owen proposed a greedy algorithm that adds one layer at a time. The process seeks for a plaid model minimizing the sum of squared errors when approximating the data matrix to the model. For this purpose, an iterative approach is adopted with each cycle updating $h$ values, $q$ values and $j$ values in turns. Assuming that residual data becomes more and more unstructured noise as each layer is removed from the data, authors propose a simple rule for stopping the process, in which only a small number of extra layers can be extracted once the data have been reduced to noise.

2.4. Proposed Method
Biclustering is one of data mining techniques are very useful to identify a coherent pattern of microarray gene expression data. A bicluster of gene expression dataset is a subset of genes that showed similar expression patterns along the subset of the conditions.

By the development of bioinformatics analysis, related to increasing the number of data that to be analyzed, encourage researchers to propose methods for better results and efficient. In this study, we proposed a two-phase method. Two-phase in this method consists of two phases, the first phase is the implementation of parallel algorithms k-means on gene expression data to generate $k$ clusters. The second phase is the implementation of bicluster algorithm, we applied plaid model biclustering algorithm. The application of the two-phase method on these gene expression data aims to find good-quality bicluster effectively.
3. Results
In this study, experiments are conducted on the Carcinoma and Adenoma microarray database that can be accessed via http://genomics-pubs.princeton.edu/.

- Carcinoma Data: There are 7457 genes and 36 samples.
- Adenoma Data: There are 7086 genes and 8 samples.

From the experimental result, we found three biclusters are formed by Carcinoma gene expression data and four biclusters are formed by Adenoma gene expression data. We got sum of squares (SS) of the clusters is 66.9% for Carcinoma gene expression data and 71.5% for Adenoma gene expression data.

In Carcinoma gene expression data, the three biclusters shows that they have matrices size $11 \times 15$, $16 \times 2$, and $3 \times 12$, respectively.

In Adenoma gene expression data, the four biclusters shows that they have matrices size $4 \times 3$, $4 \times 3$, $4 \times 5$, and $25 \times 4$, respectively.

Furthermore, we can identify what genes that affect the most, according to samples or conditions. So, from the experimental results it can be used for further identification or diagnosis.

4. Conclusion
In this paper, the proposed method is introduced which is divided in two phase. To do bicluster microarray of carcinoma and adenoma gene expression data, we can use plaid model biclustering.
Algorithm 1 Two-Phase Algorithm

**Input:** a gene expression matrix

**Output:** bicluster

1: **procedure** TWO-PHASE (gene_expression_matrix)
2:     Phase 1 :
3:     Generate clusters using parallel k-means
4:     Initialize the centroid \( k = 1 : K \)
5:     **for** \( i = 1 \) to \( K \) **do**
6:         **for** \( j = 1 \) to \( J \) **do**
7:             read \( C_{ij} \)
8:         **end for**
9:     **end for**
10:     Create \( k \) clusters
11:     Recompute the centroid for each cluster until there is
12:     no change in the centroids
13:     Phase 2 :
14:     **for** each cluster **do**
15:         Apply plaid model biclustering algorithm
16:     **end for**
17: **end procedure**

Figure 4. Resume of Two-Phase Method
and K-means or parallel K-means algorithm as a tool to preprocessing data. Our results show that Carcinoma gene expression data produces three biclusters and Adenoma gene expression data produces four biclusters.

Acknowledgment
This research is supported by Hibah PITTA UI 2016 research grant.

References
[1] Das M, Borah B 2014 Biclustering of Gene Expression Data Using Two-Phase Method International Journal of Computer Applications 103(13) November 2014 pp 0975–8887
[2] Lazzeroni L, Owen A 2002 Plaid models for gene expression data Statistica Sinica 12 pp 61–86
[3] Notterman D A, Alon U, Sierk A J, Levine A J 2001 Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays. Cancer Research 61 pp 3124–3130
[4] Pontes B, Giraldes R, Aguilar-Ruiz S 2015 Bioclustering on expression data: A review Elseiver Journal of Biomedical Informatics 57 pp 163–180
[5] Turner H, Bailey T, Krzanowski W 2005 Improved biclustering of microarray data demonstrated through systematic performance tests *Elsevier Computational Statistics & Data Analysis* 48 pp 235–254

[6] Cheng Y and Church G 2000 Biclustering of Expression Data *International Conference on Intelligent Systems for Molecular Biology* pp 93–103

[7] MacQueen J B 1967 Some Methods for classification and Analysis of Multivariate Observations *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability Berkeley, University of California Press* I pp 281–297

[8] Pollard K S, Van der Laan M J 2002 Statistical inference for simultaneous clustering of gene expression data *Math. Biosci.* 176 pp 99–121

[9] Kanungo T, Mount D, Netanyahu N, *et al.* 2002 An efficient k-means clustering algorithm: analysis and implementation *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7) pp 881–892

[10] Madeira S C and Oliveira A L 2004 Biclustering Algorithms for Biological Data Analysis: A Survey *IEEE/ACM Transactions on Computational Biology and Bioinformatics* I pp 24–45

[11] Bustamam A, Ardaneswari G, Lestari, Dian, Tasman H 2014 Performance Evaluation of Fast Smith-Waterman Algorithm for Sequence Database Searches using CUDA GPU-Based Parallel Computing *Journal of Next Generation Information Technology* 5(2) pp 38–46

[12] Permata T S, Bustamam A 2015 Clustering protein-protein interaction network of TP53 tumor suppressor protein using Markov clustering algorithm *International Conference on Advanced Computer Science and Information Systems (ICACSIIS)* IEEE Proceeding pp 221–226