Original Article
Middle East Journal of Cancer; October 2022; 13(4): 624-640

The Application of Bi-clustering and Bayesian Network for Gene Sets Network Construction in Breast Cancer Microarray Data

Ahmad Sohrabi*, MSc, Neda Saraygord-Afshari, PhD, Masoud Roudbari*♦, PhD**

*Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran
**Department of Medical Biotechnology, Faculty of Allied Medical Sciences, Iran University of Medical Sciences, Tehran, Iran

Please cite this article as: Sohrabi A, Saraygord-Afshari N, Roudbari M. The application of bi-clustering and Bayesian network for gene sets network construction in breast cancer microarray data. Middle East J Cancer. 2022;13(4):624-40. doi: 10.30476/mejc.2022.89998.1557.

Abstract

Background: Breast cancer is one of the most prevalent types of cancer in Iranian women and the second cause of death in women worldwide. Gene mutations are the key determinants of the disease; therefore, the genetic study of this disease is of paramount importance. One of the genetic evaluation methods of this disease is microarray technology, which allows the examination of the simultaneous expression of thousands of genes. Clustering is the method for analyzing high-dimension data, which we used in the present research for collecting similar genes in separated clusters.

Method: A descriptive and inferential statistical analysis was carried out to evaluate unsupervised learning models of gene expression analysis and five bi-clustering methods (including PLAID (PL), Fabia, Bimax, Cheng & Church (CC), and Xmotif) were compared. For this purpose, we obtained the microarray gene expression data for lapatinib-resistant breast cancer cell lines from previously published research. The enrichment efficacy of the clusters was evaluated with gene ontology, and the results of these five models were compared with the Jaccard index, variance stability, least-square error, and goodness of fit indices. Furthermore, the results of the best model were assessed for building a genes sets network with Bayesian networks.

Results: After preprocessing, clustering was performed on the data with the dimension (4710 × 18) of the genes. Four models, except for CC, successfully found bi-clusters in the data set. The data evaluation revealed that the results of the models were almost the same, but the PL model performed better than the others, finding 11 bi-clusters; this model was used to build the network of gene sets.

Conclusion: According to the results, the PL method was suitable for clustering the data. Accordingly, it could be recommended for data analysis. In addition, the gene sets network formed on gene expression data was incompetent.

Keywords: Breast cancer, Bi-clustering, Cluster analysis, Microarray data, Gene expression, Neoplasms, Bayesian network

♦Corresponding Author:
Masoud Roudbari, PhD
Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran
Email: roudbari.m@Iums.ac.ir

Middle East Journal of Cancer

Received: February 02, 2021; Accepted: January 31, 2022

Introduction

One of the most critical objectives of clinical researchers is finding strategies for early diagnosis and more effective treatments for diseases. The advancements in various fields of science, including clinical sciences, can be influenced by the emergence of new perspectives in other sciences.[1] One of these interdisciplinary sciences, which can be used to develop many others, is statistics.[1]

The completion of The Human Genome Decoding Project has provided numerous data for studying disease genetics. Completing this project is the start point for many other studies. Since all cells of every organism are made through multiple divisions of a primary cell (germ cells), all the cells of any organism have similar genes. Therefore, the reason behind the diversity of cells in tissues and organs is not genetic material but genes' different expressions under different conditions.[1] The events and environmental signals can lead to cell differentiation in various tissues during the development and formation of the embryo from the germ cell. However, under disease conditions, we can observe that the gene expression pattern is altered on many occasions, which puts health in danger. The aforementioned factors necessitate evaluating the effect of genome content in different conditions, as well as the amount and manner of expression while encountering stressful stimuli, along with the factors affecting the amount of gene expression.[2]

Identifying those genes with similar gene expression patterns is one of the basic principles of analyzing gene expression data, which is performed in different ways. Since the structure and function of each cell are determined based on the pattern of that cell gene expression, it is expected that information on the genetic origin of cellular manifestations be obtained by comparing the gene expression patterns in different conditions.[3]

Undoubtedly, cancers are one of the most important causes of mortality today; that said, following heart disease and accidents, they are the third leading cause of death. Breast cancer is one of the most prevalent types of cancer in women. According to studies, gene mutations are the determinants of the disease, and different types of mutations in different conditions and stimuli can lead to different types of cancers, including breast cancer.[3-4]

Modern life-style, new habits, like smoking, aging populations, as well as many other factors are considered as the influential factors increasing the rate of cancer prevalence. Several types of cancer can be prevented or even treated, if diagnosed early. Cancer is a general term for a complex set of diseases, and carcinogenesis, in which normal cells are transformed into abnormal cancerous cells. These abnormalities are biologically complex and can be presented in different stages. The noteworthy point about mutations leading to cancer is that the diversity of these mutations is not only very high in different types of tumors, but also in various samples of a particular tumor. This high diversity emphasizes the need to use advanced methods of genetic statistics in the genetic study of the disease.[4-8]

Unlike the genome, which is the same in all the cells of an organism (except in the abnormal cells and tissues cells that underwent mutation

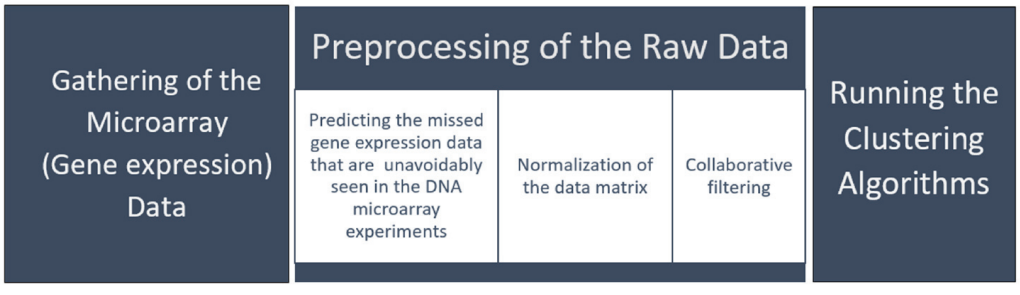

Figure 1. A brief overview of the data analysis processing steps is summarized in three main stages.

and may result in disease state), gene expression (i.e., production of protein products) is not identical in all the cells of an organism. In fact, gene expression is a dynamic process and varies between different cells, tissues, and organs. Even in identical cells, gene expression is influenced by various environmental signals. Environmental signals may lead to either a transient or permanent alteration in gene expression status. Therefore, gene expression can reflect momentary and lasting changes in the biological state of cells, tissues, organs, and organisms. Gene expression can be examined at different levels: the production rate of messenger RNA (mRNA) or the production of protein levels produced by genes. Data on gene expression provide helpful information on biological networks and contribute to understanding the cellular processes.[9,10] Currently, the microarray is one of the tools that allow observing and studying the simultaneous

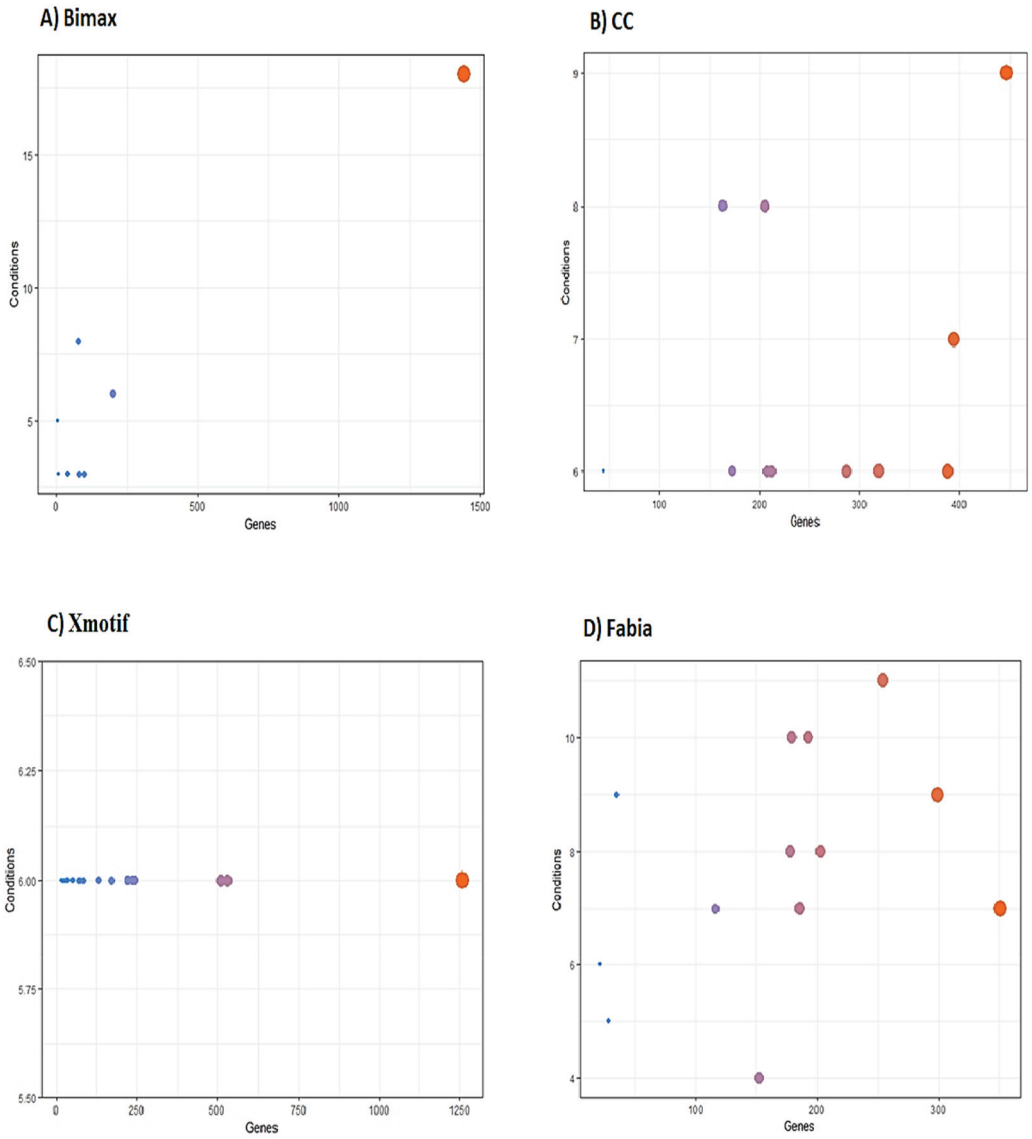

Figure 2. Scatter plots (x-axis: genes, y-axis: experimental conditions) of the four bi-clustering methods (A: Bimax, B: CC, C: Xmotif, and D: Fabia) used in this study. These plots depict the size and number of the gene bi-clusters detected by these bi-clustering methods. The color and size of the spots represent the number of the genes located in each cluster. As can be seen, the number of the genes located in the clusters varies clearly. Moreover, some of the bi-clusters produced by a particular method are obviously different with the others. As an example, notice the big orange spot in the Bimax model.

expression of thousands of genes. Considering the difference in the expression levels defined for genes, microarrays also exist in different types of biochips and protein chips. Using these tools, the cell expression pattern can be investigated when the cell is exposed to various conditions, such as disease, injury, and any other tensions. These tools are so powerful thereby it can be acknowledged that the emergence of different array methods has played an essential role in the development of gene expression studies.[11] One of the main goals of such studies is to investigate the interactive effect of gene expression on each other and also how genes are expressed in healthy cells and those affected by stressful conditions. The diversity of cancer-causing mutations is very high in different types of tumors and in different samples of a particular tumor. Due to this broad spectrum, there are also differences in the pathological features and the response to treatment in various tumors. However, in addition to diversity, cancer cells have common features, and discovering the genetic origin of common features of different tumors can have a remarkable positive impact on cancer diagnostic-treatment methods. The most important applications of microarray technology are the study of gene expression (genome),[12] using comparative genome hybridization,[13] and identifying single-nucleotide polymorphism.

These studies are used for different purposes, such as determining the genotypic structure of individuals and diseases and different conditions, measuring the probability of some diseases, estimating mutations in germ and somatic cells in investigating the cancer genetic origin, and

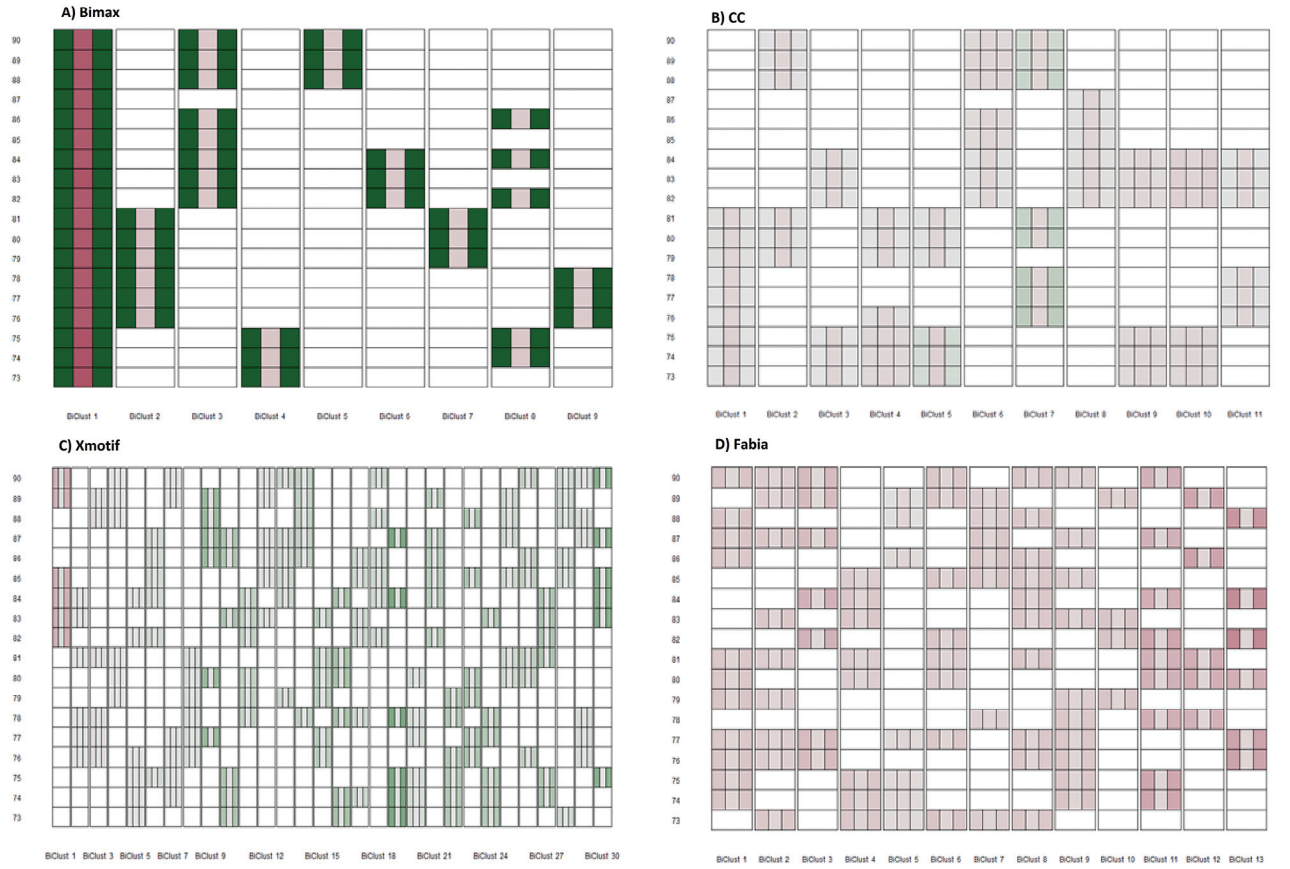

Figure 3. Bi-cluster visualization. Each heatmap shows the membership graph for all the extracted bi-clusters generated by the four bi-clustering methods (A: Bimax, B: CC, C: Xmotif, and D: Fabia) used in this study. The horizontal axis (columns) represents the biclusters, and the vertical axis (rows) represents the experimental conditions. The colored-filled cells are representers f the biclusters conditions. As can be seen, not all the conditions are covered by a single bi-cluster; instead, most of the bi-clusters include some subsets of the experimental condition.

genetic linkage analysis.[14]

Several genome projects have been defined to decode the hidden genetic code in the DNA sequences of various living organisms; the largest and most important one is the human genome project to recognize the entire human genome.

Gene expression data extracted from DNA microarrays are usually represented as a matrix of gene expression levels under different experimental or biological conditions. The rows and columns of this matrix represent genes and the experimental conditions, respectively. Its entries indicate gene expression. Since we can access large amounts of data using microarrays, accurate and robust methods and tools are needed to analyze these data. Clustering is an analytical method that helps researchers interpret gene expression data by finding groups of genes with similar expression patterns.[11] The objective of gene clustering is to find different gene subsets with very similar naturally occurring attributes based on some common information; hence, it is expected that while a high level of similarity could be found within the members of a particular cluster, a high level of difference could also be found between the members of two separate clusters.[15] Although typical clustering methods (such as hierarchical clustering and k-means) are useful for analyzing microarray data, they have the following limitations:[16]

1. Classical clustering methods assume that sets of related genes behave similarly in a particular experimental condition in which measurement occurs. This assumption makes sense once the data set is in a limited condition of a simple experiment. However, for larger data, which includes hundreds of heterogeneous conditions and a large number of experiments, this assumption is not valid, and using classical clustering methods does not make sense.[11]
2. In clustering, heterogeneous populations are often divided into several homogeneous subpopulations, which have no commonalities with each other; in fact, each gene can only belong to one of the clusters. However, biological facts show that certain genes may be involved in several biological activities and cannot be allocated to one cluster.[17]
3. There may be genes that are not active in any experimental conditions in the researcher's data; nonetheless, all genes are classified into clusters in classical clustering methods.[18]

The concept of bi-clustering was introduced to overcome limitations and shortcomings and to find appropriate patterns.[19] Bi-clustering is a method that simultaneously clusters both genes (rows) and conditions (columns). Thus, this method generates clusters representing subsets

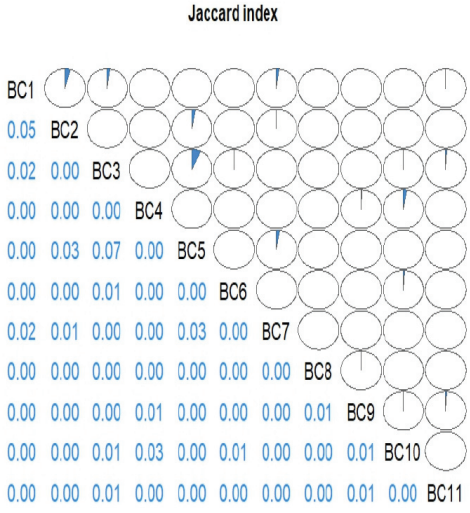

Figure 4. This figure depicts the box and whiskers graph in the Plaid and Fabia models.

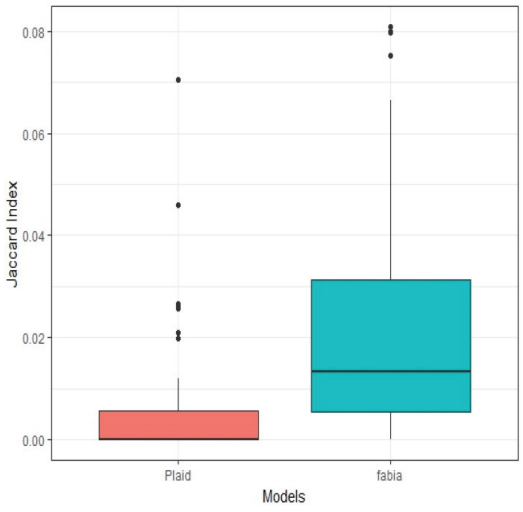

Figure 5. The graph shows the Jaccard index in the Plaid and Fabia models. The highest overlap using the Jaccard index in the Fabia model is 8%, and in the Plaid model is 7%.

Table 1. Different experimental conditions in the study of Komurov, K. et al.[32]

The sample code	Cell type	The received Lapatinib dose	Cell type	The sample code
GSM940882	Resistant to lapatinib	0	Sensitive to lapatinib	GSM940873
GSM940883		0		GSM940874
GSM940884		0		GSM940875
GSM940885		0.1		GSM940876
GSM940886		0.1		GSM940877
GSM940887		0.1		GSM940878
GSM940888		1		GSM940879
GSM940889		1		GSM940880
GSM940890		1		GSM940881

of genes related to subsets of conditions. In contrast, classical clustering methods can be applied for two purposes, namely genes' or conditions' clustering. Hence, when using the bi-clustering method, instead of searching for gene expression data, we should detect a subset of genes with similar expression patterns in a subset of experimental or biological conditions; that is, there are homogeneous submatrices in the data matrix, which are definable in a set of conditions, called bi-cluster.[20,21] The bi-clustering methods could be effective provided that the data set is in the form of a matrix with real values, such that the set values of aij represents the relationship between its i^{th} row and j^{th} column, and the goal is to identify the subsets of rows with similar behavior in a subset of columns. Bi-clustering was proposed by Hartigan, in which, using an algorithm, the original data matrix is decomposed into a set of submatrices that are the bi-clusters, and the variance index is employed to evaluate the quality of each bi-cluster.[22] Cheng and Church (CC) were the first who applied the concept of bi-clustering to gene expression events. They defined bi-clustering with a greedy approach, as a subset of rows and columns, with a high similarity score and used the mean square error as this score. This algorithm is applied through two stages of deletion and addition and will continue its function until there are no rows and columns eligible for deletion or addition.[23]

Lazzeroni and Owen proposed the PLAID (PL) model, which is a statistical modeling method for gene expression data analysis. The basic idea for expressing a matrix here is to form layers based on bi-clusters. This model assumes that the level of matrix entries is the sum of a constant and k bi-clusters. The iterative exploratory search method has been utilized to estimate the parameters of this model. The characteristics of the PL model, which solves the problems in the CC algorithm, have introduced it as a suitable

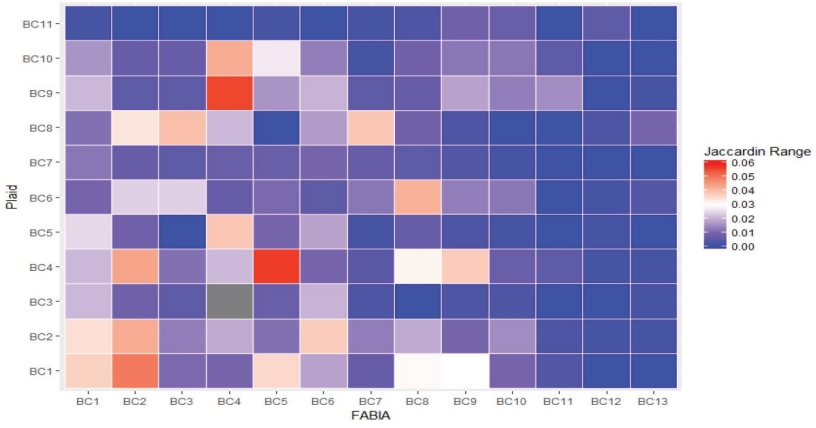

Figure 6. This graph represents the rate of communality between clusters Plaid and Fabia models.

Table 2. Dimensions of the extracted bi- clusters by the XM model

Number of columns	Number of rows	Cluster number	Number of columns	Number of rows	Cluster number
6	31	16	6	1259	1
6	35	17	6	531	2
6	25	18	6	511	3
6	28	19	6	236	4
6	32	20	6	245	5
6	18	21	6	170	6
6	30	22	6	131	7
6	24	23	6	222	8
6	20	24	6	74	9
6	19	25	6	69	10
6	21	26	6	84	11
6	15	27	6	48	12
6	17	28	6	50	13
6	16	29	6	54	14
6	18	30	6	38	15

model in bi-clustering.[24] This method also has certain limitations; firstly, in this model, the distributions are assumed to be normal, and secondly, there are restrictions and conditions on the membership parameters proposed in the model, so they can only accept zero and one value.

Murali and Kasif (MK) introduced the Xmotif (XM) algorithm for finding the pattern of the protected gene expression. This algorithm searches the rows with constant values in a subset of columns. An XM is a subset of genes simultaneously protected in a subset of columns; hence, XM is known as a bi-cluster. MK hypothesized that the expression matrix consisted of several MK bi-clusters and found that its largest bi-cluster contains the largest number of protected genes.[25]

Prelić et al. introduced the simple Bimax (BM) algorithm, which detects bi-clusters in a binary gene expression data matrix.[26] In this algorithm, the data must be converted to binary, which has a great effect on running BM. In addition, the algorithm is sensitive to errors in the data, and if there is an error, it will not be able to detect the optimal bi-clusters.

The Fabia (FA) model was proposed by Hochreiter et al., in which the $X_{n\times m}$ matrix is modeled as the sum of the K bi-clusters plus an additive error. The factor analysis model is then used for goodness of fit in the data set. This model is additive (i.e., can be written as the arithmetic sum of predictor variables' and individual effects.), and in estimating the parameters, assumes that all effects are in a normal distribution; meanwhile, gene expression data, even after the logarithmic transformation, have long tails and do not have a

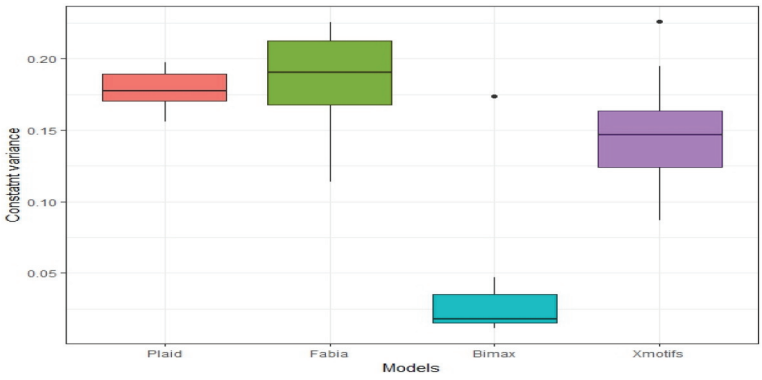

Figure 7. This figure shows the variance stability values of the box and whiskers clusters for all the models.

Table 3. Dimensions of the extracted bi-clusters by Fabia, Bimax, and Plaid models

Cluster number	1	2	3	4	5	6	7	8	9	10	11	12	13
Fabia model													
Number of rows	254	299	186	203	351	178	116	193	179	152	35	28	21
Number of columns	11	9	7	8	7	8	7	10	10	4	9	5	6
Bimαx model													
Number of rows	1442	200	78	97	79	41	37	5	6				
Number of columns	18	6	8	3	3	3	3	5	3				
Plaid model													
Number of rows	447	388	207	394	287	205	163	172	211	319	44		
Number of columns	9	6	6	7	6	6	8	6	6	6	6		

normal distribution.[27]

Abdalla et al. applied the bi-clustering method to compare the expression data between malignant (cancerous) and benign (non-cancerous) tissues of breast cancer patients. They concluded that using this method will differentiate the gene expression rate identified in the early stages of tumor growth in both groups; accordingly, breast cancer could be detected in the same early stages of the disease.[28]

The present study aimed to present the application of different algorithms in clustering breast cancer gene expression data and determine the most biologically efficient algorithm in these clusters or the highest efficient MSE criteria for the above-mentioned gene expression. Furthermore, we sought to determine the best pattern in the bi-clustering of breast cancer gene expression data and present the application of Bayesian network in the formation of networks between the genes placed in a particular cluster in each of the chosen algorithms. A Bayesian network is a probabilistic graph-based statistical model conducive to depicting complex stochastic processes, including both variables and conditional dependencies.[29,30] Moreover, gene expression-based clustering of the genes associated with a particular subtype of a neoplasm or a particular condition is of noticeable medical value; for example, it could contribute to developing specialized treatment or diagnostic tools for patients suffering from similar subtypes of the disease or treated with similar therapeutic regimens.[31]

Materials and Methods

This research is a descriptive and inferential statistical analysis conducted in 2017-2018 in the School of Public Health, Iran University of Medical Sciences. Gene expression data obtained from the study by Komurov et al.[32] were used to conduct this paper. These data are related to breast cancer patients in two groups of sensitive and resistant to lapatinib (LAPT) cases in three

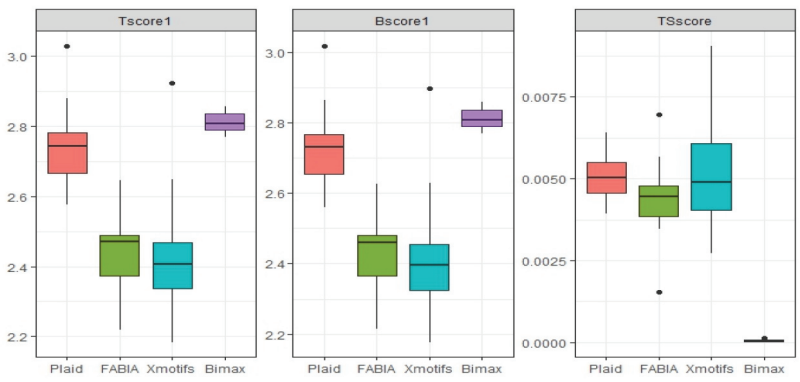

Figure 8. These diagrams show the goodness of fit indices for the models that were fit for the present study.

Table 4. Comparison of different clustering models

Model	Percentage of clusters with significant expression	The significance of all expression, at the level of 5%	Percentage of significant clusters in terms of sample membership	MSE
Cheng & Church	100.0	79.65	63.6	2.73
Bimax	77.8	83.27	66.7	3.14
Fabia	69.2	71.32	0.00	2.48
Xemotif	30.0	88.30	6.7	2.63

MSE: Mean square error

different doses (0 for those who did not receive LAPT, 0.1, and 1). Herein, there was no sampling; we utilized all the gene expression data provided by Komurov et al. (i.e., a set of gene expression data in dimensions of 48803 × 18 (18 is the number of conditions) obtained from Affymetrix technology in the corresponding research). Table 1 represents different experimental conditions.

Primarily, the data were downloaded and retrieved in R software. Afterward, the Bioconductor database was used to develop R software in the field of biological analysis. Data preprocessing, clustering, and cluster validation were then performed. Finally, we formed a Bayesian network.[33] In the validation section, in addition to R software, we used Gene Ontology (GO). Data preprocessing of gene expression was carried out through the following steps:

In microarray analysis and also in the image analysis phase, some color intensities are usually lost. Hence, where we see a missing value (for any reason), these values must be added to the analysis by an appropriate method. The K Nearest Neighbor (KNN) method is among the methods that are used for microarray missing values. In this method, k is considered to be the entry with the shortest distance to the missing data, and the mean of this k entry is used to estimate the missing value of putting embed.[34] The implementation of this method was performed herein using the Impute package available in the Bioconductor database to estimate the missing values.

The used data had experienced initial normalizations, such as background correction. After reviewing the data, it was found that the gene expression involved different wide ranges. These values were below 100 or ranged from 2000 to 50000. Thus, for the final and comparable data normalization (data are inherent with a high variance and noise), the data structure was checked, and logarithmic transformations of the data were performed on the basis of 2. Following this preprocessing, only 4710 genes remained in the data matrix, and the 4710 × 18 gene expression matrix was used for other statistical analyses; number 18 indicates the number of the samples in this data set. This procedure contributes to the elimination of technical variations and systematic experimental biases, which is essential to detect biological variations accurately.[35]

In microarray analysis, the expression of thousands of genes is measured simultaneously; hence, a high number of variables will result in a prolonged data processing time and may lead to

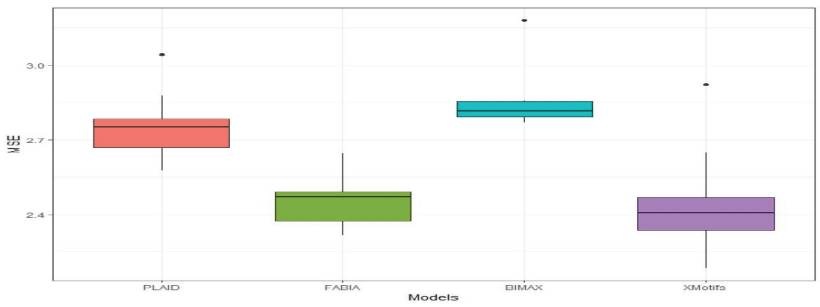

Figure 9. This figure represents the mean square error indices for all the models used in this study.
MSE: Mean square error

many unwanted errors in the results of these processes. Therefore, before analysis, filters are applied to reduce the data dimension in such data. The filtering process itself will be through two approaches, namely specific and general approaches. In the specific one, the filtering and deletion of genes are done based on a specific purpose, according to the conditions of the data. On the other hand, in the general approach, the existing conditions of the data are not used, and all the conditions are considered to be the same. Therefore, certain indicators, such as the index of dispersion, are examined instead of the conditions; the cut-off point is adjusted; accordingly, the genes are filtered, and other analyses are then performed on the remaining data.

As mentioned, in the process of data preparation, a deliberate method for filtering was applied. As indicated in table 1, based on the experimental conditions, we considered three main criteria, namely receiving lapatinib regimens (case and control groups), the dosages used for lapatinib treatment, and being sensitive or resistant to lapatinib. An independent samples t-test and analysis of variance were performed. Regarding the level of significance of the test, a *P*-value of 0.3 was considered, according to which the genes were filtered. The sum of these three filters was considered as the final filtering criterion. Ultimately, the genes whose *P*-values were less than 0.3 in one of the three mentioned tests remained for further analysis, and the rest of the genes were filtered out. Employing these methods, further genes remained for analyses, and a limited number of genes were filtered. This filtering schedule reduced the number of genes from 48803 to 4710.

Five algorithms were used in this study for clustering: CC algorithm 23 with greedy search method and sensitivity to noise, PL method 24 with comprehensive search method and non-noise sensitivity, BM method with Divide, and solution search method that is relatively noise sensitive,[36] FA method with extensive search method and non-noise sensitivity,[27] and XM with greedy search method and non-noise-sensitivity.[25] Subsequently, the results of these five algorithms will be evaluated and compared in terms of biological and statistical aspects. Eventually, the algorithm with the best performance on data will be selected, and its results will be entered into the next stage to implement Bayesian networks (Figure 1).

Like classical clustering methods, the bi-clustering method depends on similarities between the genes or conditions. An appropriate criterion for evaluating a bi-clustering algorithm is identifying the type of bi-clusters that the algorithm can find. Bi-clusters are divided into four general categories as follows:

1. bi-cluster with constant values;
2. bi-cluster with constant values on rows or columns;
3. bi-cluster with logical values;
4. bi-cluster with logical circulation.

Available biological data is used to evaluate and validate the results of bi-clustering to determine whether the resulting clusters contain genes with the same function. Thus, after uploading the data about each bi-cluster in the database and recording their initial information, the database will perform the necessary processing on the gene set of that bi-cluster and store the results. These processes take place in three areas of the biological process, molecular function, and cellular position, and there will be charts, tables, and related indicators for each area. Therefore, by examining and comparing these indicators, for each bi-cluster, the set of genes data algorithm will be obtained. Afterward, the best algorithm and its bi-cluster will be selected by examining these data. According to the studies by Tanay et al. and Padilha and Campello, each of the algorithms is useful in some cases, and considering the analyzed data, each of the algorithms can be more efficient. Therefore, they recommended that several methods be used in this field and that the results be evaluated and validated.[20,26]

After selecting the best clustering algorithm and extracting the gene expression matrix, the Bayesian bi-clusters are applied to these data, and the possible interactions and potential relationships between the genes will be

investigated. Moreover, drawing the relevant graphs makes it possible to present, describe, and analyze these networks. To this end, the Bnlearn and Bayesian Network packages available in the official R database will be used.[37]

In this study, we used internet data. The names and profiles of none of the members in these data set were mentioned, and all were specified using a code. In addition, in this research and the extracted articles from it, the data source will be cited; therefore, the present study did not require special ethical considerations.

To compare the clustering models used in this study, we utilized various indicators, such as the Jaccard index, which shows the commonalities of models and clusters, variance stability, goodness of fit, mean square error, and biological index. Moreover, to compare the groups, independent samples t-test, and analysis of variance were applied.

Results

Given the fact that there were no missing data, the data were explicitly filtered. According to table 2, the groups were compared. Moreover, the genes were filtered based on the cut-off point of 0.3 as the significance level, and the combination of these three filters was considered as the final filter. For this purpose, three tests in the mentioned groups were performed on each of the genes and the genes with a *P*-value of at least less than 0.3 were considered; otherwise, they were filtered, after which 4710 genes remained out of the 48803 primary ones. It must be mentioned that at this step, we did not look for effective genes; instead, we tried to eliminate those genes with no expression variations. Hence, our approach was conservative. It is expected that taking the level of $P < 0.3$ for the filtering process will result in an unwanted increase in the number of true-positive data. Moreover, it can be shown that filtering multiple comparison corrections would not be an appropriate strategy.

Since we used data that had undergone initial normalizations, such as background correction, after reviewing the data, it was found that the gene expression involved a wide range, including values below 100, or was in the range of 2000 to 50,000. Therefore, for the final normalization and making data to be comparable, after examining the data structure, logarithmic transformations of the data were performed on the basis of 2. Following this preprocessing, only 4710 genes remained in the data matrix, and the 4710×18 gene expression matrices were used for other statistical analyses, and number 18 indicated the number of the samples in this dataset.

Plaid model results

After data preprocessing, the model PL was implemented for the data set and the ultimate output model was extracted with 11 bi-clusters from the data matrix. In this model, the clusters also overlapped in some cases, with the highest number of gene overlaps belonging to gene 4 and the highest number of condition overlaps belonging to condition 6. Table 3 demonstrated the information on other clusters, with 11 extracted clusters.

According to our observations, the color and size of the points were affected by the clusters' number of genes and the clusters included a wide range of genes and conditions (Figure 2).

Figure 3 shows a bi-cluster membership graph. Each row of this graph represents one of the conditions and each column represents a bi-cluster, with each colored cell illustrating one's membership in the cluster. Each colored cell are divided into three parts; the color of the two outer parts represents the mean of all the genes in the cluster and the color of the inner part shows the mean of all the genes for that person. According to this graph, the first cluster (bi-cluster 1) identified LAPT-sensitive specimens and the second cluster identified specimens receiving a dose of 1 unit. In certain cases, such as in bi-cluster 3, which identifies resistant specimens, one person (code 87) is not identified and other clusters are interpreted similarly.

BM model results

After data preprocessing, the BM model was implemented for the data set and the final output model was extracted with nine bi-clusters from the data matrix. The clusters had no overlapping

in any cases (this model could not find overlapping clusters). Table 2 depicts the information on other clusters with nine extracted clusters.

According to figure 1, the row and column dimensions (genes and conditions) of the bi-clusters produced by this model are introduced, whose color and size of the points are affected by the number of the genes in each cluster. As can be seen, clusters contain a wide range of genes and conditions, and cluster number 1 is very different from other clusters and includes all the conditions with 1442 genes.

Figure 2 illustrates a bi-cluster membership diagram which identified that cluster 2 and cluster 4 had spotted the LAPT-resistant specimens that received LAPT and the resistant ones that did not receive LAPT, respectively. In some cases, such as in cluster 3, which identified sensitive specimens, one person (code 87) was not identified. Validation of the BM model using the results of Gene Ontology interpretation showed that in two clusters out of the found clusters, there was no significant expression, and in the other clusters, about 83% of the total expressions were significant at the level of 0.05.

Fabia model results

Following data preprocessing, the FA model was implemented for the data set and the final output model was extracted with 13 bi-clusters from the data matrix, in which the clusters had also overlapped in several cases. The highest number of gene and condition overlaps were 8 and 8, respectively. Information on other clusters, with 13 extracted clusters, can be seen in table 2.

Figure 1 exhibits the row and column (genes and conditions) dimensions of the bi-clusters produced by this model, in which the color and size of the points are affected by the number of the genes in each cluster. As can be seen, the clusters contain a wide range of genes and conditions (from 21 to 351).

The bi-cluster membership diagram is presented in figure 2, which shows that all the clusters identified LAPT-sensitive and LAPT-resistant groups' specimens. We observed no correlations between the two doses of the received medication and resistance or sensitivity to medication in the results. Validation of the FA model using the results of Gene Ontology interpretation showed that in four clusters, there was no significant expression, and in the other clusters, about 71% of the total expressions were significant at the level of 0.05.

Xmotif model results

After data preprocessing, the XM model was implemented for the data set and the final output model was extracted with 30 bi-clusters from the data matrix. The clusters had no overlapped cases (this model could not find overlapping clusters). Table 3 represents the information on other clusters with 30 extracted clusters.

The row and column dimension (genes and conditions) of the bi-clusters produced by this model is presented in figure1, in which the color and size of the points present the number of genes in each cluster. As shown, the clusters contain a wide range of genes and conditions.

Figure 2 shows a bi-cluster membership diagram that indicates the first cluster identified LAPT-sensitive specimens. The second cluster identified specimens receiving 1 unit dose; however, in some cases, like in cluster 6, which identified resistant specimens, one person (code 87) was not identified. The other clusters were interpreted likewise.

Validation of the XM model using the results of Gene Ontology (GO) interpretation revealed that in 21 clusters out of the 30 found ones, there was no significant expression, and in the other clusters, about 88% of the total expressions were significant at the level of 0.05.

Cheng and Church model results

The Cheng and Church (CC) model failed to find any bi-clusters in the data. As mentioned earlier, only the PL and FA models could spot the overlapping bi-clusters. Figure 4 demonstrates the commonalities between the clusters of the PL model and figure 5 shows the overlap of each model. According to this graph, the highest overlap using the Jaccard index in the FA model was 8%, in the PL model, it was 7%, and in general, the overlap of the PL model was less than FA.

The examination of the graph of commonalities between the models implied that the

commonalities of BM and XM were 2%, which was the lowest and those between PL and FA models were 10%, which was the highest.

Since the highest commonalities among the models belonged to FA and PL, the inter-cluster commonalities of these two models were examined; according to figure 6, the highest rate of commonalities was about 6%, which was related to cluster 5 of the FA and the cluster 4 of the PL model, as well as the cluster 4 of the FA and cluster 9 of the PL model.

Using the variance stability index, which shows the variance of the gene expression values in each bi-cluster, figure 6 shows the box and whiskers diagram of this index for all the clusters extracted from all the bi-clusters of the model; this indicated that the internal variance of the models was optimal and the highest variance was 0.25, which was related to the FA model (Figure 7).

Using the goodness of fit indices, which include three indices of T-score, B-score, and T-score; the row effects, column effects, and combined row and column effects were examined respectively based on the indices. If their values are close to zero, the first two indices indicate that the model has weak row or column effects; meanwhile, in the third index, being close to zero, means the model has both optimal rows and columns effects. According to figure 8, the four models are suitable and do not significantly differ from each other.

The low value of the mean square error (MSE) for each model indicates better cluster performance and more remarkable internal similarity of clusters. Figure 9 depicts the box and whiskers diagram of this index for each model. According to the results, the MSE indices of FA and XM models was 0.3 unit less than those of PL and BM models.

Table 4 exhibits the comparative results of biological and statistical indices of different clustering models.

Given the results of different indices (significance of clusters in terms of membership graph, ontology chart, and model goodness of fit indices), it seems that the PL model generally has a better function on the data compared with the other models; therefore, this model was chosen and its results were examined.

Discussion

The results obtained in the present study indicated that all the methods, except for CC, were able to find bi-clusters on the data. Following the examination of the statistical indices, it was found that the performances of these four models on the data were almost the same. We also used the Gene Ontology (GO) database for examining the biological relevance of the genes in each cluster. GO is a primary bioinformatics resource for the attributes of the genes, especially biological functions and their products.[38,39] After reviewing the results of the membership diagram and GO, we observed that the performance of the PL model was much better than that of the others; thus, it was selected as a model with better performance in finding bi-clusters on the data set. The results were used to form the gene sets network by the Bayesian networks, which showed that the Bayesian network results could not be interpreted due to the extraction of many genes (about 2837 genes in 11 clusters in the PL model).

The first study on bi-clustering in Iran was conducted by Jahromi et al. in 2005. They utilized the PL model for the bi-clustering of leukemia data. They also compared their results with hierarchical cumulative clustering and demonstrated the efficiency of the bi-clustering method using the PL model. On account of the novelty of the array data at the time, their research did not include the concepts and principles of bi-clustering. Moreover, they did not use external validation methods for evaluating the biological significance of bi-clustering results.[40] In the present study, we investigated the concepts of bi-clustering in terms of type, the structure of obtained bi-clusters, and the algorithm used to identify these clusters. In addition, the biological significance of the results obtained from this method was studied using the GO database.

Okada et al. used the bimodal algorithm for bi-clustering the gene expression data. They evaluated the efficiency of this method compared to the bi-clustering methods of ISA, OPSM, BM, Samba, CC, and XM and employed the

FuncAssociate web tool to validate the obtained bi-clusters. All the bi-clusters obtained by the bimodal method contained at least one significant enriched expression at the significance level of 0.001%. ISA, OPSM, and BM methods also showed a high percentage of enriched bi-clusters at different levels of significance (about 90% - 100% in the OPSM, about 72% - 99% in the BM, and about 80% - 91% in the ISA method).[41]

Padilha and Campello evaluated 17 clustering methods concerning different aspects (noise in data, number of clusters, cluster overlap, and cluster size) with two search approaches, in simulated data and real data. The models did not perform well in all aspects, but they introduced five models that generally performed well in all cases, namely COALESCE, LAS, Bibit, BM, and Samba. Subsequently, they argued that because the type of latent bi-cluster in real data is not known, the two models that can detect the most various types of clusters are the PL and CPB models. Hence, they recommended the models as the best ones considering their statistical performance. They further examined the performance of the ontology of the results of the models on the real data and concluded that the LAS, OPSM, and Samba models had the best *P*-values. They also concluded that the results of PL model were satisfactory, and the advantage of the PL model over other models is that all the clusters of this model had a significant expression.[26]

Like any other cancer, genetic mutations and failure of DNA repair systems, which are almost associated with the activation/overexpression of the proto-oncogenes or deactivation/downregulation of the tumor suppressor genes, are the main reason behind various types of breast cancer[42,43] (hormone receptor-positive (HR+), HER2-positive (HER+), and triple-negative (TN)). The followings are among the most important genetic mutations reported in breast neoplasms: a variety of genetic alterations, such as alteration in mucin-like cancer-associated antigen (MCA), tissue polypeptide-specific antigen (TPS), tissue polypeptide antigen (TPA), human epidermal growth factor receptor 2 (HER2), soluble HER2 (sHER2), carcinoembryonic antigen (CEA), cancer antigen 27-29 (CA 27-29 or BR 27-29), and cancer antigen 15-3 (CA 15-3). These genetic alterations are used as the remarkable theranostics biomarkers of this cancer.[44-48] Each subtype of breast cancer has its unique mutation profile, suggesting that clustering the gene alteration profiles can be conducive to improving our understanding about breast malignancies. Related gene clusters are valuable representers of the first-level biological networks.

Moreover, mining the genomic data and clustering the relevant genes could be helpful for understanding the molecular mechanisms of the therapeutic regimes. This application of genome mining is of pivotal clinical value to understand and overcome therapeutic barriers which occur during drug resistance and therapeutic failure leading to relapse or progress of the disease. Hence, in the present study, using the free available microarray data, we assessed five gene expression bi-clustering methods (PL, Fabia, Bimax, CC, and Xmotif) to detect relevant genetic networks involved the lapatinib resistance in breast cancer cell lines. Lapatinib is an effective, widely used, and orally active drug for treating $Her2^{+}$ breast malignancies and is a favorable choice for combination therapy.

Overall, we should mention that due to the limitation in the availability of the microarray data regarding the lapatinib resistance in breast cancer at the time of the study, we cannot easily over-generalize the results of the present study for direct clinical applications. The microarray dataset used herein was extracted from the cell lines cultivated in laboratory conditions and was not prepared from real patients. However, our study revealed the ability of the bi-clustering methods to extract valuable data for basic research (or trials) on lapatinib resistance in breast cancer.

Further research could be recommended on validation and finding standard methods to compare clustering models.

Bi-clustering algorithms are usually performed using parameter values suggested by the designer of these algorithms. Of course, these values may not always be the best choice and may lead to

the poor performance of the algorithm. It is recommended to use the general effect method, which has recently been added to bi-clustering methods, for examining the various parameters in each model to present the obtained clusters. Simulation studies on methods for finding suitable parameters for each data type are also recommended.

Further research on bi-clusters, whose genes have not been interpreted in any genes ontology at different levels of significance, is of great necessity. It means the genes within bi-clusters categorize the samples well in the membership chart; nonetheless, no laboratory study has been conducted on the genes inside this cluster.

The research also encountered certain limitations. The first limitation was the lack of sufficient gene expression data, which is related to the availability of relevant microarray data in Iran. To address this problem, we noted the high cost of the microarray technology and its analyzing methods as a key probable cause. Moreover, as mentioned earlier in the discussion section, the second limitation was the lack of real-patient gene-expression microarray data for lapatinib resistance when we conducted the present study. Most of the available research data were performed on the cancerous cell lines in laboratory conditions, which lowers the applicability of genome mining research for direct clinical use. The third limitation was the lack of sufficient knowledge of statistical research with life sciences, especially in the field of genetics, biotechnology, and bioinformatics and also the lack of sufficient knowledge of bio-scientists with statistics.

Finally, there are also some advantages and disadvantages over the algorithms utilized for bi-clustering methods; for example, the greedy search method used for the CC bi-clustering approach is fast. It always provides a locally optimal solution, but using this algorithm may be associated with wrong decisions because the method might take loose bi-clusters. Likewise, the comprehensive search method used for PL bi-clustering is costly and takes considerable computing time. The basics of the methodological discussions were not the aim of the present study; hence, readers are ″encouraged to use the following references in this regard.[20,49,50]

Conclusion

In this article, to introduce the most appropriate bi-clustering method, the ability of five relevant methods (PL, Fabia, Bimax, CC, and Xmotif) were examined and compared in order to find different gene subsets related to various types of breast cancer.

All the methods, except for CC, succeeded in finding bi-clusters on the data; the results were almost the same and further investigation showed that the performance of the PL model was much better than that of the others to construct bi-clusters. Furthermore, we found it to be more appropriate for mining gene expression data. However, we cannot guarantee an explicit generalization on the application of this method until the method is verified by dataset-based tests and simulation.

Acknowledgments

The authors would like to express their sincere gratitude to the Vice-Chancellor for Research and Technology of Iran University of Medical Sciences for funding this research.

Conflict of Interest

None declared.

References

1. Crick F. Central dogma of molecular biology. *Nature.* 1970;227(5258):561-3. doi: 10.1038/227561a0.
2. Lee JK. Analysis issues for gene expression array data. *Clin Chem.* 2001;47(8):1350-2. doi: 10.1093/clinchem/47.8.1350.
3. Divina F, Aguilar-Ruiz JS. Biclustering of expression data with evolutionary computation. *IEEE Trans Knowl Data Eng.* 2006;18:590-602. doi: 10.1109/tkde.2006.74.
4. Smith RA, Cokkinides V, von Eschenbach AC, Levin B, Cohen C, Runowicz CD, et al. American Cancer Society guidelines for the early detection of cancer. *CA Cancer J Clin.* 2002;52(1):8-22. doi: 10.3322/canjclin.52.1.8.
5. Society AC. Cancer News: American Cancer Society, 1947. [Access date: 7/30/2019]. Available from: https://www.cancer.org/about-us/who-we-are/our-

history.html
6. Vlahovic TA, Wang YC, Kraut RE, Levine JM. Support matching and satisfaction in an online breast cancer support community. The conference on human factors in computing systems; 2014 April 26- May1;Toronto, Canada: ACM SIGCH; 1625-34 p. doi: 10.1145/2556288.2557108
7. Savad S, Mehdipour P, Miryounesi M, Shirkoohi R, Fereidooni F, Mansouri F, et al. Expression analysis of MiR-21, MiR-205, and MiR-342 in breast cancer in Iran. *Asian Pac J Cancer Prev.* 2012;13(3):873-7. doi: 10.7314/apjcp.2012.13.3.873.
8. Azizi F, Hatami H, Janghorbani M. Epidemiology and control of common diseases in Iran. 2nd ed. Tehran: Khosravi Pub; 2011. p. 542-557.
9. Beltrame F, Papadimitropoulos A, Porro I, Scaglione S, Schenone A, Torterolo L, et al. GEMMAóA Grid environment for microarray management and analysis in bone marrow stem cells experiments. *Future Gener Comput Syst.* 2007;23:382-90. doi: 10.1016/j.future.2006.07.008.
10. Knudsen S. Guide to analysis of DNA microarray data. 2nd ed. New York: John Wiley & Sons, Inc.; 2004. p. 23-110.
11. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. *Bioinform*. 2002;18:S136-S44. doi: 10.1093/bioinformatics/18.suppl_1.s136
12. Adomas A, Heller G, Olson Å, Osborne J, Karlsson M, Nahalkova J, et al. Comparative analysis of transcript abundance in Pinus sylvestris after challenge with a saprotrophic, pathogenic or mutualistic fungus. *Tree Physiol.* 2008;28:885-97. doi: 10.1093/treephys/28.6.885.
13. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*. 1999;23:41-6. doi: 10.1038/12640.
14. Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, Ghandour G, et al. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet*. 1999;22:164-7. doi: 10.1038/9674.
15. Fei X, Lu S, Pop HF, Liang LR. GFBA: A biclustering algorithm for discovering value-coherent biclusters. In: Mändoiu I, Zelikovsky A, editors. Bioinformatics research and applications. ISBRA 2007. Lecture Notes in Computer Science, vol 4463; 2007. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72031-7_1.
16. Gu J, Liu JS. Bayesian biclustering of gene expression data. *BMC Genomics.* 2008;9 Suppl 1(Suppl 1):S4. doi: 10.1186/1471-2164-9-S1-S4.
17. Yang MS. A survey of fuzzy clustering. *Math Comput Model.* 1993;18:1-16. doi: 10.1016/0895-7177(93)90202-A
18. Zhang Y, Wang H, Hu Z. A novel clustering and verification based microarray data bi-clustering method. In: Tan Y, Shi Y, Tan KC, editors. Advances in swarm intelligence. ICSI 2010. Lecture Notes in Computer Science, vol 6146; 2010. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-13498-2_80.
19. Gan X, Liew AW-C, Yan H. Discovering biclusters in gene expression data based on high-dimensional linear geometries. *BMC Bioinform.* 2008;9:1-15. doi: 10.1186/1471-2105-9-209.
20. Tanay A, Sharan R, Shamir R. Biclustering algorithms: A survey. Handbook of computational molecular biology. Chapman and Hall/CRC Financial Mathematics Series. 1st ed. USA: Taylor & Francis; 2005. 9,122-4 p.
21. Xie J, Ma A, Fennell A, Ma Q, Zhao J. It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data. *Brief Bioinform.* 2019;20(4):1449-64. doi: 10.1093/bib/bby014.
22. Hartigan JA. Direct clustering of a data matrix. *J Am Stat Assoc.* 1972;67:123-9. doi: 10.2307/2284710.
23. Cheng Y, Church GM. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol.* 2000;8:93-103.
24. Lazzeroni L, Owen A. Plaid models for gene expression data. *Statistica Sinica.* 2002;12(1):61-86.
25. Murali T, Kasif S. Extracting conserved gene expression motifs from gene expression data. *Pac Symp Biocomput.* 2003:77-88. doi: 10.1142/9789812776303_0008.
26. Padilha VA, Campello RJ. A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics.* 2017;18(1):55. doi: 10.1186/s12859-017-1487-1.
27. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, Kasim A, et al. FABIA: factor analysis for bicluster acquisition. *Bioinformatics*. 2010;26(12):1520-7. doi: 10.1093/bioinformatics/btq227.
28. Abdalla M, Tran-Thanh D, Moreno J, Iakovlev V, Nair R, Kanwar N, et al. Mapping genomic and transcriptomic alterations spatially in epithelial cells adjacent to human breast carcinoma. *Nat Commun.* 2017;8:1-11. doi: 10.1038/s41467-017-01357-y.
29. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol.* 2000;7:601-20. doi: 10.1089/106652700750050961.
30. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1988.p.77-93.
31. Ma S, Huang J, Shen S. Identification of cancer-associated gene clusters and genes via clustering penalization. *Stat Interface.* 2009;2(1):1-11. doi: 10.4310/sii.2009.v2.n1.a1.

32. Komurov K, Tseng JT, Muller M, Seviour EG, Moss TJ, Yang L, et al. The glucose-deprivation network counteracts lapatinib-induced toxicity in resistant ErbB2-positive breast cancer cells. *Mol Syst Biol.* 2012;8:596. doi: 10.1038/msb.2012.25.
33. Nielsen TD, Jensen FV. Bayesian networks and decision graphs: Information Science and Statistics. 2nd ed. New York: Springer-Verlag; 2007.p.35-74.
34. Parry R, Jones W, Stokes T, Phan J, Moffitt R, Fang H, et al. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *Pharmacogenomics J.* 2010;10:292-309. doi: 10.1038/tpj.2010.56.
35. Liu X, Li N, Liu S, Wang J, Zhang N, Zheng X, et al. Normalization methods for the analysis of unbalanced transcriptome data: a review. *Front Bioeng Biotechnol.* 2019;7:358. doi: 10.3389/fbioe.2019.00358.
36. Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinform.* 2006;22:1122-9. doi: 10.1093/ bioinformatics/btl060.
37. Dey DK, Ghosh S, Mallick BK. Bayesian modeling in bioinformatics. 1st ed. Boca Raton. Florida: CRC Press, 2010.p.221-270.
38. Oghabian A, Kilpinen S, Hautaniemi S, Czeizler E. Biclustering methods: biological relevance and application in gene expression analysis. *PLoS One*. 2014;9:e90801. doi: 10.1371/journal.pone.0090801.
39. Ranganathan S, Nakai K, Schonbach C. Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics: 1st ed. Amsterdam: *Elsevier Science,* 2019.p.135-217.
40. Naghizadeh Jahromi MM, Hajizadeh E, Kazmnejad A. cDNA microarray data normalization. *Iran J Biotechnol.* 2005;3(1):55-63.
41. Okada Y, Fujibuchi W, Horton P. A biclustering method for gene expression module discovery using a closed itemset enumeration algorithm. *IPSJ Digital Courier.* 2007;3:183-92. doi: 10.2197/ipsjdc.3.183.
42. Waks AG, Winer EP. Breast cancer treatment: A review. *JAMA*. 2019;321(3):288-300. doi: 10.1001/jama.2018.19323.
43. Sung H, DeSantis CE, Fedewa SA, Kantelhardt EJ, Jemal A. Breast cancer subtypes among Eastern-African-born black women and other black women in the United States. *Cancer.* 2019;125(19):3401-11. doi: 10.1002/cncr.32293.
44. Bui MM, Riben MW, Allison KH, Chlipala E, Colasacco C, Kahn AG, et al. Quantitative image analysis of human epidermal growth factor receptor 2 immunohistochemistry for breast cancer: Guideline from the College of American Pathologists. *Arch Pathol Lab Med.* 2019;143:1180-95. doi: 10.5858/arpa.2018-0378-cp.
45. Lebeau A, Denkert C, Sinn P, Schmidt M, Wöckel A. Update of the German S3 breast cancer guideline : What is new for pathologists? [Article in German] *Pathologe.* 2019;40(2):185-98. doi: 10.1007/s00292-019-0578-3.
46. Duffy MJ, McDermott EW, Crown J. Blood-based biomarkers in breast cancer: From proteins to circulating tumor cells to circulating tumor DNA. *Tumor Biol.* 2018;40(5):1010428318776169. doi: 10.1177/1010428318776169.
47. Cappelletti V, Appierto V, Tiberio P, Fina E, Callari M, Daidone MG. Circulating biomarkers for prediction of treatment response. *J Natl Cancer Inst.* 2015;2015:60-3. doi: 10.1093/jncimonographs/lgv006.
48. Saraygord-Afshari N, Naderi-Manesh H, Naderi M. Enhanced reproducibility of the human gel-based tear proteome maps in the presence of di-(2-hydroxyethyl) disulfide. *Biotechnol Appl Biochem*. 2014; 61(6):660-7. doi: org/10.1002/bab.1221.
49. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform.* 2004;1:24-45. doi: 10.1109/tcbb.2004.2.
50. Bozdag D, Kumar AS, Catalyurek UV. Comparative analysis of biclustering algorithms. 1st ACM International Conference on Bioinformatics and Computational Biology; 2010 August 2-4; Niagara Falls. New York. USA: Association for Computing Machinery; 2010. 265-74. doi: 10.1145/1854776.1854814.