Application of Plaid Algorithm to Identifying Patterns in Breast Cancer Gene Expression Data

Hamid Alavi Majd[1], Ahmad Reza Baghestani[1], Seyyed Mohammad Tabatabaei[2], Soodeh Shahsavari[1*], Mostafa Rezaei Tavirani[3], Mohsen Hamidpour[4]

Abstract-numerous studies have correlated variation in gene expression between individuals to phenotypic diversity in breast tumors**.** The main goal of this study was to conduct plaid algorithm to biclustering of breast cancer gene expression data with the aim of identifying tumors subgroups with similar clinical features. The real dataset that had been used in this research is the one which was used in Breast cancer (docetaxel resistance) article in 2005 that was included in CGED. Gene expression profiling was done with on data matrix containing 44 patients and 2453 genes. Plaid algorithm was used to recognize gene expression patterns, after that, percent of significant genes in each bicluster was calculated with FDR. Biclustering algorithm has discovered 265 co-expressed genes which was divided to 6 subgroups with similar expression levels. 175 number of these genes was identified significant by FDR and expression levels were different in responder and non-responder. Randomization test and GO ontology did confirm the results of biclustering algorithm. The increasing clinical use of genomic profiling demands identification of more effective methods to segregate patients into prognostic and treatment groups. We have shown that biclustering can be used to select optimal gene sets for determining the prognosis of specific strata of patients.

Index Terms- Biclustering, Plaid Algorithm, Gene Expression, Breast Cancer

1 INTRODUCTION

In biology, the cell is the basic structure of any organism. All cells of an organism have the same genes that could be at different expression levels across numerous conditions [1]. Scientists have concluded that different conditions could affect it, in terms of whether a particular gene is expressed and how it could be expressed. The organism's health may be compromised due to the different expressions present. Therefore, it is crucial to evaluate the levels of genome when exposed to tense factors [2]. By comparing gene expression patterns, tissues types, different disease and time points, researchers are able to inference about genes or special cellular conditions [3]. Cancer, also known as a malignant tumor, is a disease caused by failed tissue growth regulation, involving abnormal cell growth with the potential to invade or spread to other parts of the body [4]. Therefore, a normal cell can transform into a cancer cell when the genes that regulate cell growth and differentiation are altered [5]. In 2012, about 14.1 million new cases of cancer occurred globally. It caused about 8.2 million deaths or 14.6% of all human deaths [6].The most common types of cancer in females are breast cancer, colorectal cancer, lung cancer, and cervical cancer [7]. Breast cancer is one of the most frequent cancers worldwide and the most frequent affecting women [8]. Excluding skin cancers, breast cancer is the most common cancer diagnosed among women in the United States, accounting for nearly 1 in 3 cancers. After lung cancer, it is also the second leading cause of cancer death among women [6].Breast tumors, based on many studies, are classified by grade and using certain molecular biomarkers. Also, numerous studies have correlated variation in gene expression between individuals to phenotypic diversity in breast tumors [9]. So, performing research in gene expression profiling of this disease can be useful to discover similar clinical characteristics and help researchers in diagnostics. In recent years, DNA microarray technology has provided monitoring of thousands of gene expression simultaneously when cells are under different conditions and various processes. This technology has a key role in accelerating and increasing the efficiency of gene expression studies [10]. The development of this technique has led to the availability of gene expression matrix with rows containing thousands of gene and columns containing hundreds of conditions [11]. Clustering is one of the most important techniques used for detecting pattern recognition [10]. But, traditional clustering methods will have some issues to discover similar patterns in gene expression data [12]. In order to overcome these constraints and for the purpose of finding the appropriate gene expression patterns, biclustering methods have proposed which computational framework is more flexible [13]. A bicluster is a subset of genes that has similar expression patterns over a subset of conditions; so, biclustering methods have determined homogeneous submatrices [14]. The first biclustering algorithm, the so-called Block Clustering, was developed by Hartigan [15]. Cheng and

Church proposed the first biclustering algorithm for the analysis of high-dimensional gene expression data [14]. Since then, many different biclustering algorithms have been developed. It is the opinion of the authors that biclustering may be able to identify clinical significant gene expression modules that stratify breast cancers according to inter-tumors heterogeneity [16]. In this study a biclustering technique was used, plaid algorithm [17], to group breast tumors from 44 patients into subgroups which were conditionally dependent on expression profiles of specific gene subsets. The main goal of this study was to conduct biclustering of breast cancer gene expression data, to identify tumors subgroups with similar clinical features.

2 METHODS

2.1 Data Resource

The real dataset that was used in this research is related to Breast cancer (docetaxel resistance) article in 2005 that was included in CGED [18]. First, all the data were obtained using an advanced version of quantitative RT–PCR, that produces data of better quality than those based on hybridization techniques. Second, tissue samples were obtained mainly from a single hospital. This eliminated deterioration of clinical data by a difference in medical practice, commonly found in data collected from multiple hospitals .Forty-four (44) breast tumor tissues (22 responders and 22 non-responders) were sampled through biopsy. Numbers of assayed genes were 2453. This gene expression profiling of breast cancer samples were designed to develop a method for prediction of patients' response to docetaxel. Individuals that had more than 50% response to treatment were defined as responders. The data in this database are quite unique both in analytical and clinical aspects. First, all the data were obtained using an advanced version of quantitative RT–PCR that produces data of better quality than those based on hybridization-based techniques. Second, tissue samples were obtained mainly from a single hospital. This eliminated deterioration of clinical data by a difference in medical practice, commonly found in data collected from multiple hospitals [19].

2.2 Biclustering Algorithm

Distribution parameter identification is one of the biclustering algorithms, in which it is assumed that the data structures follow a statistical model and tries to fit its parameters to the data by minimizing a certain criterion through an iterative approach. Plaid models, Spectral biclustering and Rich Probabilistic Models are some examples of this kind of biclustering. Among them, the Plaid model [17] is arguably one of the most flexible biclustering models up to now. This model was proposed by Lazzeroni and Owen and modified by Turner et al. It defines the expression levels as a sum of layers, constructed as biclusters. This model assumes that the level of matrix entries is sum of the uniform backgrounds and k biclusters. So the expression matrix with I genes (Rows) and J conditions (Columns) is represented as

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk}$$

Where μ_0 is a general matrix background and $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$, and μ_k is the added background in bicluster k, and α and β are column specific additive constants in bicluster k. Also, $\rho_{ik} \in \{0,1\}$ and $\kappa_{jk} \in \{0,1\}$ are gene-bicluster membership and condition-bicluster membership indicator variables. The general biclustering problem is now formulated as finding parameters values, so that the resulting matrix would fit the original data as much as possible. Formally, the problem is about minimizing $\sum_{ij}\left[Y_{ij} - \sum \theta_{ijk}\rho_{ik}\kappa_{jk}\right]^2$.

2.3 Evaluation

Validation of the biclustering algorithm is difficult. In this study, evaluation of discovered biclusters was performed in two ways:

1. A test statistics was conducted for evaluation of significance of biclusters. For this purpose, a hypothesis test was conducted to demonstrate that known biclusters were not identified just by chance. This was performed by constructing 1000 permutations random sample of the gene expression dataset with size 20*5 and then compared the number of known biclusters identified by the method, with the number of discovered biclusters in the random set. Number of corrected bicluster was computed and then the mont-carlo p-value was calculated. If p-values smaller than 0.05, then the hypothesis of discovering bicluster by chance is rejected.

2. The result of the different biclustering techniques in microarray data is groups of genes, strongly co-expressed with each other. These genes are expected to have the same functions. Gene ontology

biological process could be the function that measures these similarities and cover three domain cellular component, molecular function and biological process. GO Enrichment Validation is a hypergeometric test for GO enrichment. This statistical test is significant if the genes in the biclusters are annotated with GO terms, and are not specified by chance [20].

3 RESULT

Gene expression profiling was done with a data matrix containing 44 patients and 2453 genes. First, the data set was normalized with median approach and then missing values were imputed using the K Nearest neighborhood method. Information about discovered biclusters is shown in Table 1. In this table, the first column contains the label of each bicluster. The second and third columns report the number of genes and conditions respectively and the last columns contains the mean square residue (MSR) of the biclusters. MSR is a measure that shows how each bicluster are homogenies and so small values of that are better.

Table 1: Information about the bicluster result

Label	Genes	Conditions	MSR
A	187	2	3.21
B	37	5	3.59
C	23	8	4.46
D	15	6	2.24
E	4	12	1.14
F	3	8	2.40

The results of the method show that MSR are proper. All biclusters which were in gene expression levels of responders, demontrated that 265 co-expressed genes are divided to 6 subgroups with similar expression levels. Figure 1 shows the biclusters in the bubble plot and each circle consists of co-expressed genes. As shown in this graph, there is very small overlap between discovered biclusters and the almost independent biclusters. The evaluation of the significance of the discovered biclusters was performed by F test. The result is shown in Table 2, which shows statistical significance.

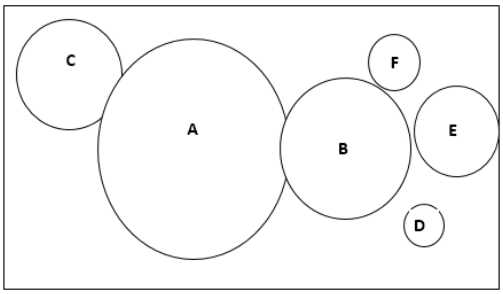

Figure1: bubble plot for genes identified by biclustering using plaid algorithm

Table 2: Test for significantly of discovered biclusters

	F stat	Pvalue
Row Effect	25.51	2.59e-21
Column Effect	2.87	1.02e-02
Tukey Test	0.019	8.91e-01

3.1 Biological Significant

Table 3 shows the significant GO terms for the set of genes, discovered by each biclusters result along with their p-value. The web tool David was used to evaluate the enrichment analysis of discovered biclusters [21]. For each bicluster, we first denoted numbers of GO term and then calculated percent of significance of the GO terms.

3.2 Molecular Features

Fold change is often used in analysis of gene expression data in microarray experiments, for measuring change in the expression level of a gene. The definition of the fold-change in this study for each gene was defined as mean of expression levels of the gene in responders to non-responders. This measure was evaluated for significance by t-test statistics and FDR method. Figure 2 shows percent of significant gene that discovered in every bicluster. Fold-change is measured to select the differentially expressed genes as the representatives of these Biclusters.

3.3 Randomization test

In this study, for each permuted sample, we calculated number of biclusters and number of genes by Plaid model and by random selection. If number of genes in random sample was at least at 80 percentile fold of number of the genes discovered by Plaid model, we

accepted that model was identified biclusters by chance. In this study number of identified biclusters by chance was 15 and p-value=0.015, so hypothesis of selected by chance was rejected.

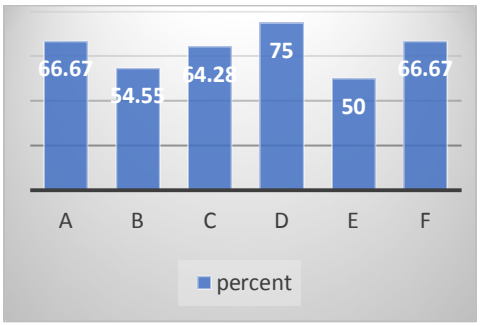

Figure 2: percent of significant gene in each biclusters with fold change method

4 Discussion

The increasing clinical use of genomic profiling demands identification of more effective methods to segregate patients into prognostic and treatment groups. We have shown that biclustering can be used to select optimal gene sets for determining the prognosis of specific strata of patients.Gene expression patterns associated with docetaxel sensitivity and resistance are highly complex. In the past, investigators use single gene biomarker to assess sensitivity and resistance and so they did not carry out any correlation between commonly measured predictive and prognostic markers. There is a little information about the usefulness of gene expression array in human breast cancer. The aim of this study was to discover patterns of many genes that could be used as a predictive test in patients as well as exclude genes with low and differentially expressed breast cancer. This study have shown that the biclustering analysis of data reveal correlated structures in the responders to docetaxel. There are 6 biclusters to be discovered. These results suggest that the patterns of gene expression are likely to involve many genes in pathways and support the idea that the patterns of expression levels of many genes could be successful in distinguishing between sensitive and resistant tumors. Randomization method and gene ontology analysis, have shown that the biclusters results are proper. To identify molecular features, differential expression levels in each bicluster were examined between responders and non-responders by Fold Change. The results show that at least 50% of the genes in responders and non-responders in the biclusters were significant. As shown here, using the Plaid algorithm, it can be concluded that good results are obtainable using the biclustering search.

Table3 Gene Ontology and Enrichment Analysis for Discovered Biclusters

Bicluster	Ontology	Number of GO term	Percent of significant pvalue
A	BF	97	80.41
	MF	27	92.59
	CC	17	64.71
B	BF	511	86.11
	MF	76	88.16
	CC	93	82.79
C	BF	60	75
	MF	10	60
	CC	11	81.82
D	BF	4	50
	MF	1	100
	CC	3	100
E	BF	118	81.36
	MF	20	90
	CC	35	88.57
F	BF	5	62.12
	MF	2	50
	CC	3	66.67

BF: Biological process
MF: Molecular function
CC: Molecular function

ACKNOWLEDGMENT

The authors would like to thank the Academic & Research Affairs of Para Medical Faculty of Shahid Beheshti University of Medical Sciences for financial Support.

AUTHORS

Correspondence Authors: Soodeh Shahsavari, Biostatistics Department, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
Email: Soodeh_shahsavari@yahoo.com

1. Biostatistics Department, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

2.Medical Informatics Department, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

3. Proteomics Department, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

4. Hematology Department, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

REFERENCES

1. Crick F. (1970) Central Dogma of Molecular Biology. Nature, 227, 561-563.
2. Jae K. L. (2001), Analysis Issues for Gene Expression Array Data. Clinical Chemistry, 47, 1350-1352.
3. Divina F, Aguilar-Ruiz J. (2006) Biclustering of Expression Data with Evolutionary Computation, IEEE Transactions on Knowledge & Data Engineering, 18(5): 590-602.
4. "Cancer Fact sheet N°297". World Health Organization. February 2014. Retrieved 10 June 2014.
5. Cooper GM. The Cell: A Molecular Approach. 2nd edition, Sunderland (MA): Sinauer Associates; 2000.
6. Rebecca Siegel, Deepa Naishadham, and Ahmedin Jemal (2012), Cancer statistics: 2012, CA: A Cancer Journal for Clinicians, 62(1): 10–29.
7. Jemal A, Siegel R, Ward E, Hao Y, Xu J Murray6 T, and J Thun M (2008), Cancer Statistics: 2008, CA: A Cancer Journal for Clinicians, 58(2): 71–96.
8. WHO, Cancer, Accessed 30.01.2011, [http://www.WHO.int/ ediacentre/factsheets/fs297/en].
9. Zhaoqi Liu, Xiang-Sun Zhang, Shihua Zhang (2014), Breast tumor subgroups reveal diverse clinical prognostic power, Scientific Reports, 4:4002.
10. Tanay A., Sharan R., Shamir R (2008), Discovering Statistically Significant Biclusters in Gene Expression Data. Bioinformatics, 18, 136-144.
11. Liu X., Wang L (2007), Computing the Maximum Similarity Biclusters of Gene Expression Data, Bioinformatics, 23, 50-56.
12. Madeira S, Oliveira A (2004), Biclustering Algorithms for Biological Data Analysis: A Survey, IEEE/ACM, Transactions of Computional Biology and Bioinformatics,1(11):24-25.
13. Chea Gan X., Wee A., Liew C., Yan H (2008), Discovering Biclusters in Gene Expression Data Based on High-dimentional Linear Geometrics. BMC Bioinformatics, 9, 209.
14. Cheng Y., Church G.M. (2000), Biclustering of Gene Expression Data. Intelligent Systems in Molecular Biology, 93-103.
15. Hartigan J.A. (1972), Direct Clustering of a Data Matrix. American statistical association (JASA), 67, 123-129.
16. Gupta R., Rao N., Kumar V (2011), Discovery of Error-Tolerant Biclusters from Noisy Gene Expression Data. BMC Bioinformatics, 12.
17. Lazzeroni L., Owen A (2002), Plaid Models for Gene Expression Data. Citeseer, 61-86.
18. Kato K.,Yamashita R., Matoba R., Monden M., Noguchi S., Takagi T, Nakai K (2005), Cancer Gene Expression Database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues. Nucleic Acids Research, 33, 533-536.
19. Kyoko Iwao-Koizumi, Ryo Matoba, Noriko Ueno, Seung Jin Kim, Akiko Ando, Yasuo Miyoshi, Eisaku Maeda, Shinzaburo Noguchi, and Kikuya Kato (2005), Prediction of Docetaxel Response in Human Breast Cancer by Gene Expression Profiling, JOURNAL OF CLINICAL ONCOLOGY, 23(3).
20. Al-Akwa FM, Kadah YM (2009). An Automatic Gene Ontology Software Tool for Bicluster and Cluster Comparisons, IEEE ,163-7.
21. Sherman B.T., Tan Q., Guo Y., Bour S., Liu D., Stephens R., Baseler M.W., Lane H.C., Lempicki R.A. (2007), DAVID knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics, 8, 426.