TELKOMNIKA, Vol.17, No.4, August 2019, pp.2017~2024
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v17i4.12773 ■ 2017

Co-clustering algorithm for the identification of cancer subtypes from gene expression data

Logenthiran Machap*[1], Afnizanfaizal Abdullah[2], Zuraini Ali Shah[3]
[1,2]Synthetic Biology Research Group, School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia
[3]Artificial Intelligence and Bioinformatics Group, School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia
*Corresponding author, e-mail: logmac_87@yahoo.com[1], afnizanfaizal@utm.my[2], aszuraini@utm.my[3]

Abstract

Cancer has been classified as a heterogeneous genetic disease comprising various different subtypes based on gene expression data. Early stages of diagnosis and prognosis for cancer type have become an essential requirement in cancer informatics research because it is helpful for the clinical treatment of patients. Besides this, gene network interaction which is the significant in order to understand the cellular and progressive mechanisms of cancer has been barely considered in current research. Hence, applications of machine learning methods become an important area for researchers to explore in order to categorize cancer genes into high and low risk groups or subtypes. Presently co-clustering is an extensively used data mining technique for analyzing gene expression data. This paper presents an improved network assisted co-clustering for the identification of cancer subtypes (iNCIS) where it combines gene network information with gene expression data to obtain co-clusters. The effectiveness of iNCIS was evaluated on large-scale Breast Cancer (BRCA) and Glioblastoma Multiforme (GBM). This weighted co-clustering approach in iNCIS delivers a distinctive result to integrate gene network into the clustering procedure.

Keywords*: cancer subtype, clustering, co-clustering, gene expression, gene network*

1. Introduction

Gene expression is the process by which the genetic information in deoxyribonucleic acid (DNA) is transcribed into a Ribonucleic acid (RNA) then translate to the protein where the process called transcription and translation respectively [1]. Microarray technology is a fundamental tool used to measure the gene expression levels of thousands of genes simultaneously [2]. Hence, the gene expression data matrix that has been produced from this technology under various conditions, where the row represent gene and column represents sample [3, 4]. Many different methods have been applied by scientists, for instance, early stage screening to identify cancer types. Nevertheless, assessment on accurate disease outcome is still remaining interesting and challenging tasks for physicians and pharmaceutical fields. Therefore, biomedical and bioinformatics researchers are focusing on machine learning techniques [5]. From high dimensional datasets, it is able to discover patterns and relationships by applying these techniques, besides cancer type able to be predicted effectively in future [6].

Hence, various computational analysis was carried out which are divided into comparative and non-comparative analysis. Figure 1 shows the general taxonomy of computational analysis. Hereafter many researches, clustering approaches from data mining field which is a part of machine learning have been applied to gene expression analysis. For example, *k*-means [7], hierarchical clustering [8], local self-organizing maps [9], local adaptive clustering [10] and many more were implemented in this analysis. The relationship between molecular mechanisms and dissimilar physiological states as well as gene expression signatures have been explored and identified by these approaches. Classical clustering approaches basically cluster the genes into mutually separate subsets, hence, the genes or conditions cannot fit into more than one cluster. On top of that, all the rows or all the conditions are taken into deliberation. Thus, certain genes may only be co-regulated and co-expressed under certain conditions and not in all conditions in the cellular processes. But, a gene may fit into multiple clusters because a gene may be able to participate in more than one molecular process [11, 12].

Received October 23, 2018; Revised February 27, 2019; Accepted March 20, 2019

Therefore, to overcome the limitation of traditional clustering, co-clustering becomes a substitute for the analysis of gene expression. Co-clustering [or bi-clustering] primarily clusters genes and samples simultaneously in order to identify subgroups of genes that show similar patterns under a certain subset of experimental conditions. The similarity in co-clustering, measured through the coherence of genes and samples in a co-cluster, rather than gene pairs or samples pairs function [11-13]. Furthermore, attaining of overlapping co-clusters is tolerable because in a different regulation pattern a gene can be involved according to the different measured group of conditions [14]. In finding of biologically significant patterns, co-clustering plays a vital role where those patterns are: with constant values in the whole co-cluster, with constant values in rows, with constant values in columns, with additive constant values and with multiplicative coherent values [15, 16].

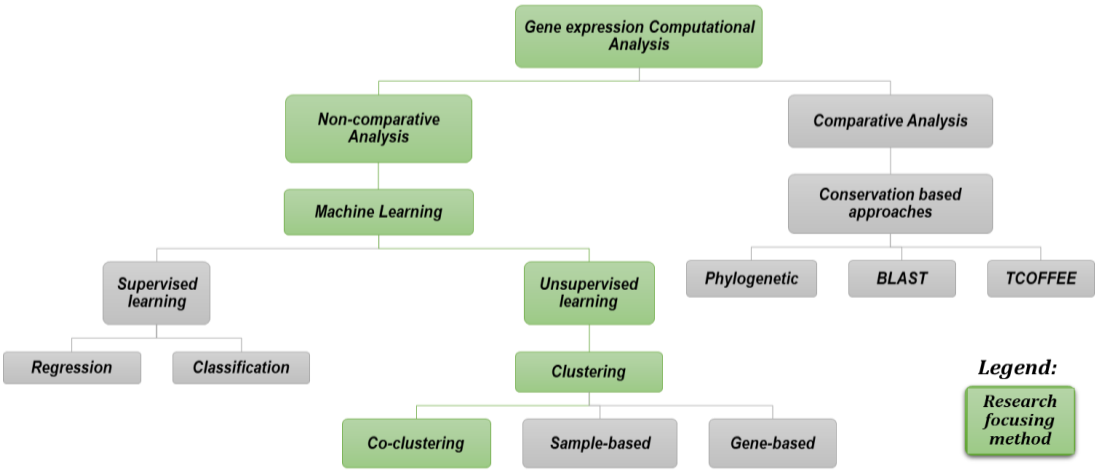

Figure 1. General view on computational analysis category

Previous research studies mentioned that co-clustering is often referred as bi-clustering, bi-dimensional clustering, two-way clustering or subspace clustering. At first, bi-clustering was introduced by Hartigan [17] whereas Cheng and Church [18] are the pioneers for applying bi-clustering in gene expression data analysis.At the early stage, co-clustering was implemented in other fields, for instance, search in database [17, 19], market search [20], target marketing [21, 22], text mining [23, 24] and analysing foreign exchange data [25].

Furthermore, for the gene expression analysis various co-clustering techniques have been implemented [26]. Since Cheng and Church [18] are the discoverers of bi-clustering solution for the NP-hard problem for clustering of gene expression data. Greedy search was applied to overcome this problem to identify bi-clusters with low mean-squared residue score. Bi-cluster was found by removing and adding genes and conditions iteratively from gene expression data matrix for which mean-squared residue score is below the threshold value. Nevertheless, only one bi-cluster at a time has been produced from this iterative solution, so that makes difficulty to set a constant threshold. Thus, the iterative signature algorithm was developed in order to discover bi-clusters based on two-predetermined thresholds for rows and columns by [27].

Scientists also take a step ahead to identify multiple bi-clusters at a time when they develop bi-clustering based on graph theory [28, 29], information theory [23], statistical method [30] and matrix factorization [31] approaches. Numerous different co-clustering methods were developed with different types of algorithms such as factor analysis bi-clustering (FABIA), the concept of gene expression motifs (xMOTIFS), Bayesian co-clustering (BCC), and minimum sum-squared residue co-clustering (MSSRCC) and so on. However, all of these methods have their own advantages and disadvantages where it can be applied for not all but certain suitable situation. Moreover, most of the previous research did not integrate biological evidence, for instance, molecular interaction networks into the clustering process. Molecular interaction networks play a vital role in each and every life processes and the mechanisms of the diseases are able to discover through the understanding of these networks [5, 32]. Though, in recent

times, the significance of network-based information has been found to be very valuable but the usage of this information in the current methods are still poor [11, 12, 33].

Therefore, the motivation of this research is to improve cancer subtype identification by improving a method which is able to integrate molecular interaction networks with clustering process. This paper improves network-assisted co-clustering for the identification of cancer subtypes (iNCIS) algorithm. At first, this method calculates a weight for each gene as it is prominence to be utilised in the clustering process [34]. Basically, genes regulating a lot of other genes and showing extremely variable expression patterns will be deliberated as more significant in the clustering process. Additionally, implanting the gene weights into the co-clustering objective function is a dynamic part of the algorithm.

2. Research Method

In this research, a co-clustering method has been proposed by integrating prior knowledge of gene network interactions between genes with gene expression data and simultaneously cluster genes and samples into subtypes. Incorporating network structure in the clustering process will lead to a better selection of informative genes for clustering. Hence, it can be assumed that, more biologically significant co-clusters are produced.

2.1. Datasets and Tools

There were two datasets used in this research, they are large scale breast cancer (BRCA) [35] and glioblastoma multiforme (GBM) [36] from TCGA. The gene network was built from different sources such as Reactome [37], NCI-Nature Curated PID [38] and KEGG [39]. Co-clustering algorithm was implemented using MATLAB platform. Table 1 shows the details of gene expression data.

2.2. Assigning Weights to Genes

Gene expression datasets are known as high dimensional, due to large number of genes and low number of samples. Significant genes selection from this high-dimensional dataset is important for clustering. Thus, gene network interactions with gene expression data utilized to obtain genes that play vital roles among samples. In this stage, weights are assigns to genes which are attained from the GeneRank method. The directed graph is used because a gene that regulates many other genes should obtain larger weights. In specific, NMAD (normalized median absolute deviation) is used for measurement of gene expression variation. This formulation helps in stabilising the weight assigning process.

2.3. Weighted Co-clustering

After assigning weights to genes, the output such as gene weights and the sample of gene expression profile used for co-clustering interpretation. This method is based on Semi-Nonnegative Matrix Tri-Factorization (SNMTF) while Orthogonal Non-Negative Matrix Tri-factorization (ONMTF) imposed for the non-negative constraint. There are three key parts need to be considered in this stage. They are:

a. Objective

Let matrix X comprises d genes and *n* samples, then would like to group the genes into m clusters and group the samples into c clusters (subtypes). The objective function has been improved from the original which is *minimizing the sum-squared distance* between all entries of co-cluster and the centroid to *minimize the sum-squared residue* between entries and centroid of co-clusters.

b. Optimization

Once genes are assigned with weights, then the objective function will be contributing to optimize the matrix X. There will be iterations to decrease the value of objective function until convergence.

c. m and c selection

To obtain better results which are converging, iNCIS run for 50 times with randomly-set initiations and get a sample consensus matrix $\overline{X_a}$ and a gene consensus matrix $\overline{X_g}$. From every single run, an $n\,x\,n$ sample connectivity matrix X_a and a dxd gene connectivity matrix X_a are obtained. The range of entries is between 0 and 1, where 0 represent samples [gene] belong to the different clusters and 1 represent they belong to same clusters.

Hereafter, Figure 2 shows the summarized methodology of this research in order to generate cancer subtype. As a first step, gene network information collected from specific database which shows 0 as absence of gene interaction and 1 shows presence of interaction among genes. This network information is then integrated with gene expression to obtain gene weights by GeneRank algorithm. And then, this output together with gene expression data used as input in incise algorithm to produce cancer subtypes. From the clusters predicted which is later used as class information to classify the cancer genes. Finally, cancer subtypes and genes are validated for functional association and drug target genes.

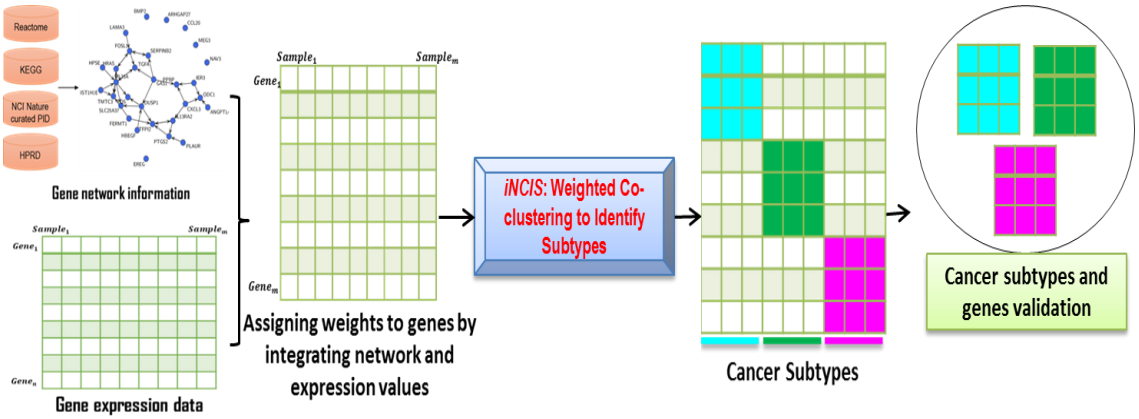

Figure 2. General flow of methodology

3. Results and Analysis

In this section, it is explained the results of research and at the same time is given the comprehensive discussion. There were mainly two cancer gene expression datasets were implemented with the co-clustering algorithm.

3.1. BRCA

BRCA dataset contains 17,814 genes across 547 samples. This gene expression profile integrated with gene network information and for 8726 genes weights are trained in both resources. Therefore, 8726 weighted genes and 547 samples were taken as the input for the co-clustering algorithm besides α = 0.85 was set. On top of that, $c = 5$ and $m = 8$ has been chosed for 50 runs. Once the subtypes are obtained, SVM used to train a classifier so that it can help for the patient diagnosis in future. Thus, 35 genes with largest weights were selected.

3.2. GBM

The second dataset used is Glioblastoma multiforme which contains 11,861 genes and 202 samples. Then gene expression is incorporated with gene network information to train weight for each of the 7,183 genes in both sets. In addition, α = 0.85 was set. After assigning weights to the 7,183 genes and 202 samples were implemented in co-clustering in where $c = 4$ and $m = 7$ are fixed. From here, four subtypes have been identified. Results were obtained shown in Table 2 for both datasets.

Table 1. Gene Expression Datasets

Dataset	Genes	Sample
BRCA	17814	547
GBM	11861	202

Table 2. Genes Selected from GeneRank Method

Dataset	Genes	Sample
BRCA	8726	547
GBM	7183	202

3.3. Discussion

As mentioned above, the parameter values were fixed. This is because; the experiment was conducted with multiple values for m and c. Among these values, one set of best value was chosen according to average cophenetic correlation coefficient. This calculation was done to evaluate to cluster stability over 50 runs. The results were shown in Tables 3 and 4 for BRCA

and GBM respectively. Thus, from the result, the highest value considers the best parameter to be selected. In addition, the input value of α is also tested on various range from 0.1 ≤ α ≥ 0.9. It has been observed that, generally α=0.85 has better performance compared to other ranges.

Table 3. Cophenetic Correlation Coefficient for BRCA

	$m=7$	$m=8$	$m=9$
$c=4$	0.931	0.930	0.944
$c=5$	0.930	**0.948**	0.942
$c=6$	0.924	0.942	0.947

Table 4. Cophenetic Correlation Coefficient for GBM

	$m=6$	$m=7$	$m=8$
$c=4$	0.911	**0.920**	0.904
$c=5$	0.903	0.904	0.906
$c=6$	0.902	0.894	0.902

As the first stage of analysis, rand index (RI) and F1-measure was adopted to evaluate the quality of clustering. From the Table 5 and Figure 3, it has been observed that, iNCIS performs better than other methods under both RI and F1-measure. It was a fairly comparison between NCIS and NetBC which is known as Network aided Bi-Clustering. Comparison was done on fairly node where the parameter setting for all the three methods iNCIS, NCIS and NetBC are the same. As an objective function for NetBC, the researcher adopts sum-squared residue from MSSRCC (minimum sum-squared residue co-clustering) [40]. Though the objective function for iNCIS and NetBC is similar, but other processes inside the core algorithm is differ. It can be concluded, iNCIS in an effective co-clustering method to identify cancer subtypes. The formula to measure RI and F1-measure are stated as follows.

$$RI = \frac{np_1 + np_4}{np_1 + np_2 + np_3 + np_4}$$

Gene expression matrix, $\boldsymbol{C} = \{\boldsymbol{C_1}, \dots, \boldsymbol{C_d}\}$

Cluster, $\boldsymbol{C'} = \{\boldsymbol{C_{1'}}, \dots., \boldsymbol{C_{d'}}\}$

np_1 : Number of pairs samples that are both in the same clusters of C and C'
np_2 : Number of pairs of samples that are in same clusters of C but in different clusters of C'
np_3 : Number of pairs of samples that are in different clusters of C but in same clusters of C'
np_4 : Number of pairs of samples that are in different clusters of C but in different clusters of C'

$$F1 - measure = \frac{2 * \mathrm{Pr} * Re}{\mathrm{Pr} + Re}$$

Sensitivity: The probability that classification result is positive when the gene pairs are interacting:

$$\frac{TP}{TP + FN}$$

Specificity/Recall [Re]: The probability that classification result is negative when the gene pairs are non-interacting:

$$\frac{TN}{TN + FP}$$

Precision [Pr]: The probability that the gene pairs are interacting when the classification result is positive:

$$\frac{TP}{TP + FP}$$

False-positive rate: The probability that classification result is positive when the gene pairs are non-interacting:

$$1 - specificity$$

Table 5. Cancer Subtypes and Rand Index Comparison

Datasets	Methods	Subtypes	Number of genes	Rand Index	F1-measure
BRCA	iNCIS	5	35	**0.9890**	**0.9839**
	NCIS	5	35	0.6400	0.7350
	NetBC	5	20	0.7540	0.8350
GBM	iNCIS	4	30	**0.9059**	**0.8616**
	NCIS	4	30	0.6750	0.6550
	NetBC	4	28	0.7325	0.7150

Further analysis need to be carried out regarding biological validation on gene subtypes obtain from iNCIS. Genes can be verified through DAVID database and PAM50. Beside this, the pathways enriched by each type cancer genes are also possible to analyse. After that, both of this cancer datasets are used for classification of subtypes. On top of that, since the true class of samples are unknown, clinical features analysis is required to evaluate the efficiency of clustering algorithm.

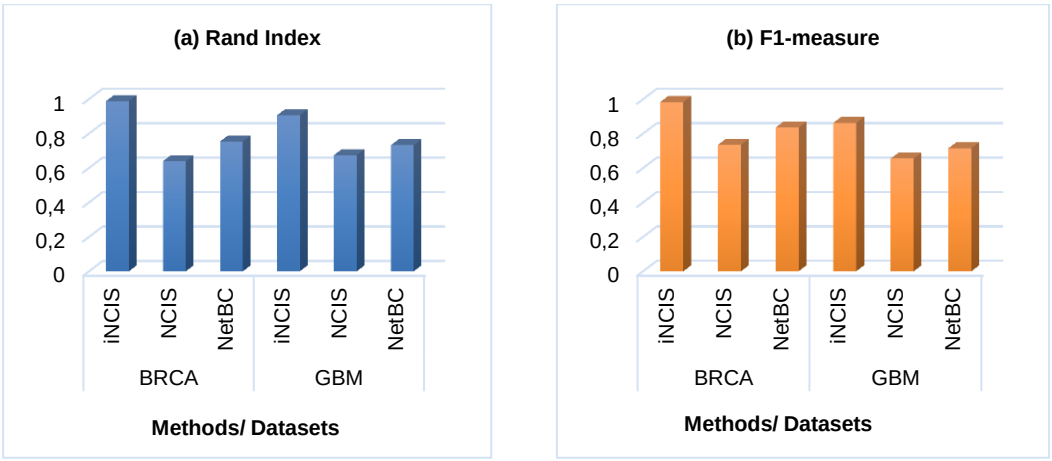

Figure 3. (a) Rand index (RI) and (b) F1-measure of different co-clustering methods on two datasets

4. Conclusion

Cancer subtypes play a significant role in diagnosis and prognosis of cancer disease. This paper presented an improved co-clustering method called iNCIS, to obtain cancer subtypes from high-dimensional gene expression. On top of that, gene network information integrated in this co-clustering process in order to improve identification of more sample subtypes. The key advantage of this network information integration is bi-product, the gene weights which define the genes' characters in the network and able to differentiate the patients. Further analysis need to be carried out on clinical features to evaluate effectiveness of the method. Beside this, optimal parameter tuning for c, m and α in iNCIS is needed to be designed to improve the results. Results show that, iNCIS is beneficial to comprehensively detect cancer subtypes and the key genes involved in each subtypes.

Acknowledgements

The author would like to show gratitude to Dr. Afnizanfaizal from Synthetic Biology Research Group and Dr. Zuraini from School of Computing, Faculty of Engineering, University of Technology Malaysia for their pearls of wisdom during the course of this research and thanks as well as to the anonymous reviewers for their so-called insights.

References

[1] Brazma A, Vilo J. Gene expression data analysis. *FEBS Letters*. 2000; 480(1): 17-24.
[2] Kallioniemi O-P, Wagner U, Kononen J, Sauter G. Tissue microarray technology for high-throughput molecular profiling of cancer. *Human molecular genetics*. 2001; 10(7): 657-62.

[3] Ben-Dor A, Friedman N, Yakhini Z. *Class discovery in gene expression data*. Proceedings of the fifth annual international conference on Computational biology. 2001: 31-38.
[4] D'haeseleer P. How does gene expression clustering work?. *Nature biotechnology*. 2005; 23(12): 1499.
[5] Abdullah A, Deris S, Hashim SZM, Mohamad MS, Arjunan SNV, editors. *An improved local best searching in Particle Swarm Optimization using Differential Evolution*. 2011 11th International Conference on Hybrid Intelligent Systems (HIS). 2011: 115-120.
[6] Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*. Proceedings of the National Academy of Sciences. 2001; 98(19): 10869-74.
[7] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nature genetics*. 1999; 22(3): 281.
[8] Eisen MB, Spellman PT, Brown PO, Botstein D. *Cluster analysis and display of genome-wide expression patterns*. Proceedings of the National Academy of Sciences. 1998; 95(25): 14863-8.
[9] Benabdeslem K, Allab K. Bi-clustering continuous data with self-organizing map. *Neural Computing and Applications*. 2013; 22(7): 1551-62.
[10] Domeniconi C, Gunopulos D, Ma S, Yan B, Al-Razgan M, Papadopoulos D. Locally adaptive metrics for clustering high dimensional data. *Data Mining and Knowledge Discovery*. 2007; 14(1): 63-97.
[11] Liu Y, Gu Q, Hou J, Han J, Ma J. A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics*. 2014; 15(1): 37.
[12] Yu G, Yu X, Wang J. Network-aided Bi-Clustering for discovering cancer subtypes. *Scientific Reports*. 2017; 7(1): 1046.
[13] Xu T, Le TD, Liu L, Wang R, Sun B, Li J. Identifying Cancer Subtypes from miRNA-TF-mRNA Regulatory Networks and Expression Data. *PLOS ONE*. 2016; 11(4): e0152792.
[14] Pontes B, Giráldez R, Aguilar-Ruiz JS. Biclustering on expression data: A review. *Journal of Biomedical Informatics*. 2015; 57(Supplement C): 163-80.
[15] Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*. 2004; 1(1): 24-45.
[16] Padilha VA, Campello RJGB. A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics*. 2017; 18(1): 55.
[17] Hartigan JA. Direct clustering of a data matrix. *Journal of the american statistical association*. 1972; 67(337): 123-9.
[18] Cheng Y, Church GM. *Biclustering of Expression Data*. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. 2000; 8: 93-103.
[19] Agrawal R, Gehrke JE, Gunopulos D, Raghavan P. *Automatic subspace clustering of high dimensional data for data mining applications*. 6,003,029 (Google Patents). 1999.
[20] Gaul W, Schader M. A new algorithm for two-mode clustering. *Data analysis and information systems*. 1996: 15-23.
[21] Wang H, Wang W, Yang J, Yu PS, editors. *Clustering by pattern similarity in large data sets*. Proceedings of the 2002 ACM SIGMOD international conference on Management of data. 2002: 394-405.
[22] Hofmann T, Puzicha J. *Latent class models for collaborative filtering*. IJCAI. 1999; 99.
[23] Dhillon IS, Mallela S, Modha DS. *Information-theoretic co-clustering*. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. Washington DC. 2003: 89-98.
[24] Lewis DD, Yang Y, Rose TG, Li F. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*. 2004; 5(Apr): 361-97.
[25] Lazzeroni L, Owen A. Plaid models for gene expression data. *Statistica sinica*. 2002: 61-86.
[26] Tanay A, Sharan R, Shamir R. Biclustering algorithms: A survey. Handbook of computational molecular biology. 2005; 9(1-20): 122-4.
[27] Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical review E*. 2003; 67(3): 031902.
[28] Denitto M, Farinelli A, Bicego M, editors. *Biclustering Gene Expressions Using Factor Graphs and the Max-Sum Algorithm*. Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015: 925-931.
[29] Kluger Y, Basri R, Chang JT, Gerstein M. Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome Research*. 2003; 13(4): 703-16.
[30] Shan H, Banerjee A. *Bayesian co-clustering*. Data Mining, 2008 ICDM'08 Eighth IEEE International Conference. 2008: 530-539.
[31] Carmona-Saez P, Pascual-Marqui RD, Tirado F, Carazo JM, Pascual-Montano A. Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC Bioinformatics*. 2006; 7(1): 78.
[32] Ismail MA, Deris S, Mohamad MS, Abdullah A. A Newton Cooperative Genetic Algorithm Method for In Silico Optimization of Metabolic Pathway Production. *PLOS ONE*. 2015; 10(5): e0126199.

[33] 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. *Briefings in Bioinformatics*. 2018: 1-16.
[34] Algfoor ZA, Sunar MS, Abdullah A. A new weighted pathfinding algorithms to reduce the search time on grid maps. *Expert Systems with Applications*. 2017; 71: 319-31.
[35] Network CGA. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012; 490(7418): 61.
[36] Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell*. 2010; 17(1): 98-110.
[37] Croft D, O'kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*. 2010; 39(suppl_1): D691-D7.
[38] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. *Nucleic acids research*. 2008; 37(suppl_1): D674-D9.
[39] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*. 2011; 40(D1): D109-D14.
[40] Cho H, Dhillon IS. Coclustering of Human Cancer Microarrays Using Minimum Sum-Squared Residue Coclustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2008; 5(3): 385-400.