University of Ibadan
Journal of Science and Logics in ICT Research

A Review on Biclustering Algorithms for Data Mining Analysis of Gene Expression Data

[1]Osuntokun, O. D., [2]Adeyemo, A. B., [3]Makolo, A. U.

University of Ibadan, Ibadan, Nigeria
[1]kemolyte2006@yahoo.com, [2]sesanadeyemo@gmail.com, [3]aumakolo@gmail.com,

Abstract

Data mining techniques have established their usefulness in extracting and bringing to light, novel and insightful discoveries from gene expression data. Over the past few decades, these approaches have been valuable for disease diagnosis, drug discovery, and understanding gene functions. Well known examples of these techniques include classification, dimension reduction analysis methods, association rules, clustering and biclustering. In recent years, as a state of the art data mining method, biclustering techniques have ascertained their indisputable efficacy for studying gene expression data. In existing literature, various studies have made attempt of classifying biclustering methods into different categories. In this paper, an extensive survey and classification of existing biclustering methods proposed in the last ten years was done. These methods were grouped into six categories namely probabilistic models, iterative greedy search, nature inspired models, linear algebra models, and hybrid approaches. It was found that hybrid, nature inspired models were particularly suited for solving complex, non-linear, and high dimensional problems such as biclustering when compared to other methods. Nature inspired methods have the ability to solve difficult problems using seemingly simple initial rules and conditions despite having minute or essentially no knowledge of the search space. However, it is known that they might have deficiencies that prevent them from finding optimal solutions. These deficiencies can be curtailed if they are hybridized with another search method. The reviewed studies were also grouped according to the intra and inter bicluster evaluation functions that were utilized to measure the coherence within biclusters and to measure the accuracy of the algorithms to extract real implanted biclusters in a matrix. It was revealed that most of the studies that used evaluation functions utilized the MSR and Jaccard index as their intra and inters bicluster evaluation functions. It was also deciphered from the review that most of the studies were focused on yeast expression data and a few other gene expression data sets. This study therefore proposes that more attention should be given to the study of other expression data set in order to enhance improved disease diagnosis, prognosis and disease prevention.

Keywords: *data mining, gene expression data, biclustering, gene expression analysis*

1. Introduction

Recent technological evolution has proffered the opportunity to fully sequentialize the genetic data of various living organisms. One of such technology is the microarray technology, which allows the serial assessment of the relative amount of mRNA expressed in several genes under different conditions [1]. This has generated an immense volume of experimental data. These data are usually arranged in matrix form where the i^{th} row depicts the i^{th} gene and the j^{th} column depicts the j^{th} condition [2]. In the last few decades, there has been tremendous upsurge in interest pertaining to the extraction of useful knowledge from data generated from gene expression data. This is due to the fact that the derived knowledge can be useful for disease diagnosis, drug diagnosis, drug discovery, gene function prediction, and disease classification. Several data mining techniques have proven their efficacy in analyzing gene expression data. Well established examples of such technologies include classification, dimension reduction analysis, association rules, clustering and biclustering. Clustering as a data mining technique assumes that genes in a group or cluster behave in the same manner in all experimental conditions. However, in the real sense, genes only

Osuntokun O. D., Adeyemo A. B. and Makolo. A. U. (2020). A Review of Biclustering Algorithms for Data Mining Analysis of Gene Expression Data. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 5 No. 1, pp. 64 - 76

behave similarly in certain conditions. Also, clustering techniques usually group genes into exactly one group. Whereas genes should be permitted to belong to more than one group because genes have the capability of performing more than one function. Biclustering techniques were developed to overcome the drawbacks of clustering algorithms. These techniques allow the extraction of clusters of genes that exhibit related or close expression values across specific conditions [3]. The biclustering technique gained its initial appearance in the research community through the work of Hartigan but Cheng and Church were the first to implement it for the analysis of gene expression data in the year 2000.

In the formal sense, a bicluster can be defined as follows:

Let $X = \{1,2, \ldots\ldots\ldots n\}$ describe a set of n genes and $Y = (1,2, \ldots. m)$ describe a set of m conditions and C (I, J) is a pair (I' J') such that I' $\in$ I and J' $\in$ J.

The biclustering challenge can be depicted as follows. Given a data matrix, unearth a set of biclusters B_{opt} associated with M such that:

$$f(B_{opt}) = \max_{B \in BC(m)} f(B)$$

Where f is an objective function depicting the quality of a set of biclusters and BC(M) is the set all possible biclusters related to M. Obviously, the biclustering task is a non-trivial data mining problem with a search space of size O $(2^{/A/+ /B/)}$. This means the biclustering challenge is an NP hard one.

2 Data Mining Techniques For Gene Expression Analysis

Classification Techniques are methods that are useful for categorizing data into different data groups or classes. Some Classifications known to be helpful for gene expression analysis include, Naïve Bayes, Support Vector Machine, Artificial Neural Network (ANN), etc. The techniques are typically used for purposes such as disease diagnosis and prognosis, disease subtype division, gene function determination and drug discovery [4].

a) Support Vector machine

Support vector machine adopts the concept of structural Risk Minimization with the central objective of locating the optimal hyper plane that divides the classes in the input space [5]. However, it usually requires long training time and it is computationally intensive. Devi Arockia Vanitha, Devaraj, and Venkatesulu [6] proposed an efficient gene classification strategy which utilizes Support Vector Machine. Informative genes are extracted by using mutual Information between gene sets and Class Label for training the SVM.

Chen *et al.* [7] presented a Multi Kernel SVM based data mining system. Multi task systems such as associated rule extraction, feature selection, data fusion, decision rule extraction, and subclass discovery, class prediction were incorporated. ALL-AML leukaemia data set was utilized to examine the effectiveness of the system. The result of the study showed that data fusion and feature selection can enhance the efficiency of the SVM in tackling high noise and diverse set of data.

b) Random Forest Algorithm

Random Forest Algorithm makes use of an ensemble method that give rise to more than one model or trees. The trees are designed with the aid of bootstrap sample of the data chosen randomly. Random Forest possesses some characteristics that enhance their suitability for gene expression classification. These are its suitability for multiclass problems, non-over fitting, usefulness in cases where there are more variables than observations [8]. Random Forest is useful for effectively analyzing the problem of large features and small samples. This means Random Forest is useful for gene selection and ranking, as well as detecting and proffering an effective treatment for cancer [8].

c) Naïve Bayes Method

Krishnaiah, Narsimha, and Chandra [9] propose a model for quick discovery and accurate detection of diseases thereby helping doctors in their life saving endeavours using Naïve Bayes Classifier and Naïve Creedal Classifier 2, One dependency Augmented.

Several data mining techniques and methods which have been developed and used in data mining research include association, classification, clustering, prediction, and sequential patterns [3]. The focus of this work is the classification technique.

Association Rule

Association rule technique introduced in the year 1993 are useful for finding significant relations in biological data [10]. Gene expression analysis using association rules helps in describing the biological relationship between different genes under different experimental sample or condition [11]. They are very useful for extracting interesting patterns in data. Association rules helps to bring to light, the relationship among gene expression data and gives important decision rules for disease diagnosis [12]. Frequent pattern mining is a major undertaking of association rule mining. There exist different strategies which are utilized in association rule. One of them is the Apriori algorithm that needs large memory and uses exponential time for generating candidate solutions.

Vengateshkumar, Alagukumar and Lawrence [12] presented a Boolean Association rule Mining method (BARM) to generate frequent gene expression. The method helps to unearth the relationship among the gene expression data useful for giving out the optimum decision for disease diagnosis. The BARM was used to select significant genes and discover association rule with less memory and low computational time.

Dimension Reduction Method

Dimension reduction entails the process of minimizing the dimension of a data set to combat the problem posed by high dimensional data and to preserve crucial characteristics of the data to a very large extent [13]. Curse of dimensionality is the issues that come up due to the analysis of data with high dimensions. Attributes/features selection methods are used for discerning the most useful attributes needed for a particular task that has the capability of giving similar or better results compared to situations where all the attributes are used.

There are three types of Attribute/feature selection method. They are the filter, wrapper, and embedded approaches. The attribute subsets are chosen using an autonomous method in the filter approach and they extract features from the data void of any learning strategy, filters usually involve less computational cost and does not consider classifier, whereas the classification method to be used for analysis is used for attribute selection in the wrapper method and they utilize learning methods to evaluate the useful attributes. Wrappers usually demonstrate performance in selecting features because model hypothesis is considered by training and testing in the feature space. Wrappers are categorized into two main groups; deterministic wrappers and randomized wrappers. Embedded techniques ensemble the feature selection step and classifier design. Feature Extraction techniques can be grouped into linear and non-linear methods.

Oliver and Njeunje [14] utilized the principal component analysis technique as the linear dimension reduction method in gene expression analysis while Laplacian Eigen maps (LE) as non-linear reduction method. The methods are implemented using MATLAB due to the exceptional ability of MATLAB in handling matrix operations. The aim of the study was to demonstrate the superiority of Laplacian Eigen maps (LE) over PCA in preserving biologically relevant structures in cancer expression data set.

Wang and Vander Laan [15] proposed a targeted Maximum Likelihood Estimation procedure based on variable importance measurement (TMLE-VIM). The TMLE is an extension of Maximum likelihood Estimation method. It was demonstrated that TMLE-VIM can help obtain the shortest possible list with the most truly associated variables.

Hira and Gillies [16] presented a review of various ways of performing dimensionality reduction method on high dimensional data. It was asserted that removing excess features helps to enhance the quality of result. It was also explained that two main dimensionality reduction methods exist. They are the feature subset selection (Filter, Wrapper and Embedded), and the feature extraction method (Linear and Non-linear Method).

Hybrid Methods

Ha and Jo [17] proposed the hybrid method with association rules (Apriori) and classification trees (C5.0). It was designed for classification of Chest pain diseases. The algorithm was compared with SVM and NN. It was found that the hybrid algorithm had better performance.

Lavanya and Rani [18] presented a hybrid approach that combined the classifier CART with cascading feature selection and clustering to enhance the accuracy of classifier CART. It was confirmed that combining data mining techniques enhances data mining result. The algorithm was tested on breast cancer.

Nookala *et al*. [19] presented the result of the comparative analysis of 14 different classification algorithms such as J48, and tested their performance using three cancer data sets. They are the breast cancer, lymphoma, Leukaemia data sets. The result indicated that none of the methods outperformed the others.

3. Biclustering Techniques

3.1 Biclusters design based on their Structures

A group of biclusters can be in one of the following cases

1. Single bicluster (Fig 2.1 (a))
2. Exclusive rows and columns group of biclusters (Figure 2.1 (b))
3. Non-overlapping group of biclusters with checkerboard structure (Figure 2.1(c)
4. Exclusive rows group of biclusters (Figure 2.1 (d))
5. Exclusive columns group of biclusters (Figure 2.1 (e))
6. Non-overlapping group of biclusters with tree structure (Figure 2.1 (f))
7. Non-overlapping non-exclusive group of biclusters (Figure 2.1 (g))
8. Overlapping group of biclusters with hierarchical structure (Figure 2.1(h))
9. Or, arbitrarily positioned overlapping group of biclusters (Figure 2.1(i))

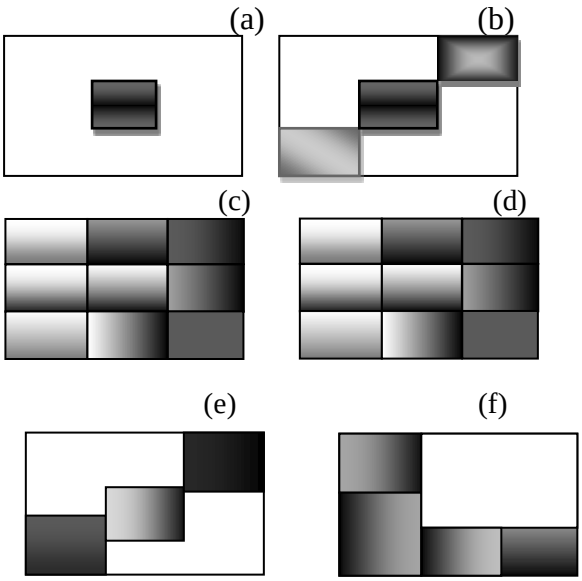


Figure 1.: Probable design of a set of biclusters [2]

3.2 Bicluster Taxonomy based on Gene Expression Patterns

Constant values: A bicluster which possess constant values depicts subgroup of genes with comparable expression values in a subgroup of conditions. This situation can be described with: $bij = \pi$.

Constant values on rows or columns: A bicluster which has constant values along the rows/columns depicts a subgroup of genes/conditions with comparable expression levels across a subgroup of conditions/genes. This means that expression levels might differ from gene to gene or from condition to condition. It can also be expressed either in an additive or multiplicative way:

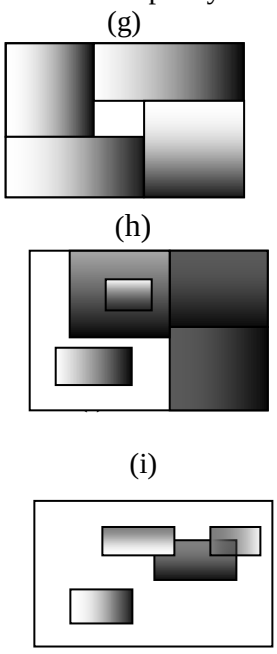
-Additive: $bij=\pi+\beta i$, $bij=\pi+\beta j$

- Multiplicative: $bij=\pi\times\alpha i$, $bij=\pi\times\alpha j$

Coherent values on both rows and columns: This class of biclusters exhibit more intricate relationships between genes and conditions, which could be in an additive or multiplicative way.

- Additive: $bij=\pi+\beta i+\beta j$
- Multiplicative: $bij=\pi\times\alpha i\times\alpha j$

Coherent evolutions: This describes a subset of genes that behaves similarly in an increasing and decreasing manner over a subgroup of conditions without taking cognizance of their actual expression values. In this situation, data in the bicluster do not adopt any mathematical model.



3.3 Overview of Biclustering Algorithms

Biclustering techniques are renowned as excellent data mining algorithms effective for finding important relationships in a data matrix. These relationships are biclusters defined as subsets of genes which exhibit similar characteristics over some subset of conditions. Hartigan introduced the biclustering concept in the year 1972 but it was put to use for gene expression analysis for the first time by Cheng and Church in the year 2000.

In recent years, a number of biclustering techniques have been developed, which are proven to be helpful in uncovering information needed to discover unknown functions of genes, develop vaccines and drugs for treatment of various diseases.

Biclustering Algorithms have been classified into different approaches by some studies in the literature.

A study by Sri and Porkodi [20] presented a survey of biclustering algorithms used for the analysis of gene expression data, evaluation functions and various types of biclusters. The methods were classified into five main groups. They are Iterative Row and Column Clustering Combination, Greedy Iterative Search, Divide and Conquer, and Exhaustive Bicluster Enumeration and Distribution parameter.

In another study by Saber and Elloumi [21], biclustering algorithms were categorized based on the strategies that are employed by the biclustering techniques for extracting useful information from data matrix. The approaches are Iterative Row and Column Clustering Combination Approach, Greedy Iterative Search Approach, Exhaustive Bicluster Enumerative Approach, Distribution Parameter Identification Approach and Divide and Conquer approach. In the study, it was made known that Iterative Row and Column Clustering Combination Approach use a clustering technique on the rows and columns of a data matrix distinctly. The result of the technique is then merged to obtain good biclusters. Examples of the biclustering approaches are Croki2, CroBin and DCC. In addition, Greedy Iterative Search Approach was depicted as involving the construction of sub matrices of a data matrix by inserting/deleting a row and column into or away from a given data submatrix to optimize or proffer the best solution to a given optimization problem.

The procedure is terminated when there is cessation of the availability of any row and column suitable for being inserted to or deleted from the data matrix. However, it was iterated that it is possible to take erroneous decisions and lose meaningful biclusters even though the greedy iterative search approach is usually fast. Examples are BicFinder, Maximum Similarity Bicluster (MSB), and Order Preserving Submatrix (OPSM). In the case of Exhaustive Bicluster Enumeration Approach, all likely groups of biclusters are detected with the focal aim of keeping the best one. However, the method is computationally costly when the time and the amount of resources required are taken into consideration. Examples are Order Preserving Clusters (OPC), and Correlated Patterns Biclustering (CPB).

The Distribution Parameter Identification Approach obtain the distribution parameters with the help of a statistical model which reduces a certain parameter in an iterative manner. They exhume optimal biclusters even though they derive good result by restricting the volume of the biclusters. Examples are Qualitative Biclustering Algorithm (QUBIC), Factor Analysis for Bicluster Acquisition (FABIA). The Divide and Conquer Approach is started by representing the whole data matrix by a bicluster, next it is split into two biclusters then this strategy is done recursively till a certain number of biclusters is attained which agrees with the pre-specified properties. The technique is fast but it might disregard good quality biclusters by dividing them before discovering them. An example is BiBit. In the study, it was concluded that no biclustering method is faultless and that extraction of large biclusters from large micro array data remains a challenge which need constant attention.

Another study by Pontes, Giraldez and Aguilar-Ruiz [1] presents an analysis of biclustering methods by classifying them into two groups which are methods that utilize an evaluation strategy and those that do not employ an evaluation strategy. It was asserted in the study that great number of the biclustering techniques in literature carried out their solution search using optimization methods. A conclusion was reached that the use and development of an effective heuristic as well as appropriate fitness function is

important for providing direction for the search process in order to obtain eye-opening biclusters in the microarray data set. Nonetheless, there exist other biclustering methods which carry out their search process without utilizing any evaluation measures or fitness function for extracting biclusters.

In the study, evaluation-based measures were grouped as Iterative greedy search methods, stochastic search methods, nature inspired metaheuristics methods, and clustering based methods. Non-metric based Biclustering methods were classified as graph-based approaches, one-way clustering-based approaches, probabilistic models, linear-algebra based approaches, and optimal re-ordering of rows and columns. The importance of nature inspired techniques for Biclustering were highlighted because it constitutes the most explored field in stochastic strategies. Biclustering Techniques that are graph based are Statistical and Algorithmic Method for Bicluster Analysis (SAMBA), and QUBIC. Probabilistic models are Plaid models, Bayesian Model, Optimal reordering of rows and columns (OPSM).

A study by Oghabian *et al.* [22] categorized Biclustering methods based on the biclusters found and the mathematical models used to find them. The methods were divided into four categories which include Correlation Maximization Biclustering Methods (CMB), Variance Minimization Biclustering Methods (VMB), Two- way Clustering Methods (TWC) and Probabilistic and generative (PGM) methods. CMB methods look out for subsets of genes and samples where the expression values of genes or samples correlate significantly among the samples, examples are Cheng and Church, Flexible Overlapped Biclustering technique. VMB methods have low variance all through the chosen genes, condition or entire matrix, for example, Xmotif searches for biclusters with constant gene expression by enforcing the condition that the expression values of each gene are within a very small range, each gene exhibit a near constant expression level for a subset of sample.

Two-way clustering methods (TWC) discover the homogenous subset of genes and samples (biclusters) by iteratively performing one-way clustering on the genes and samples. PGM methods employ probabilistic techniques to discover genes or samples that are co-expressed across a subset of sample or gene in the data matrix.

A study by Freitas *et al.* [2] classified Biclustering algorithms into two groups namely Systematic Search Algorithms and Stochastic Search Algorithms (otherwise known as Metaheuristics algorithms). Systematic Algorithms are divide and conquer approaches which repeatedly divide the original problem into similar structures to the original problem. The results of the sub problems are then joined together as solutions to the original problem. Greedy Iterative Search Approach builds a solution in a stepwise manner using a given quality criterion. Decisions made at each step are based on information at hand without much concern about the repercussion the decision may have in future, example is the Maximum Similarity Biclustering Algorithm.

Bicluster Enumerative Approach tries to itemize all the solutions for the initial problem. As described above, this technique is proficient in getting the best solution but it is costly in terms of computational time and memory space. Stochastic Search Biclustering methods are either based on Neighbourhood Search Approach, Evolutionary Computation Approach, or Hybrid Approach. Neighbourhood Search Approach begins with an initial solution and then moves iteratively to a neighbouring solution due to help of neighbourhood exploitation strategy. Example is Cheng and Church Algorithm.

Evolutionary Computation Approach is based on the natural evolutionary process such as population, reproduction, mutation, recombination and selection. Candidate solutions of the problems are sampled by a set of individuals in a population. A fitness evaluation method is utilized to measure the quality of each individual solution. The hybrid approach also called the memetic approach combines both the neighbourhood search and evolutionary methods taking advantage of the complementary nature of the evolutionary and neighbourhood search methods.

3.4 Biclustering Techniques

According to the various classification criteria of gene expression techniques in literature, Biclustering techniques on gene expression data can generally be classified into probabilistic models, nature inspired models, iterative greedy search models, graph-based models, linear

algebra-based models and hybrid approaches. These methods will be discussed in details to show their distinctiveness in extracting biologically relevant biclusters in gene expression data.

3.4.1 Probabilistic Models

Chekouo and Murua [23] presented a probabilistic model known as Penalized plaid biclustering model. In the study, Hard-Expectation Maximization was preferred against Expectation maximization for parameter estimation since the number of parameters increase exponentially with the number of bicluster K in the bicluster Model.

Joung *et al.* [24] used a probabilistic co-evolutionary model which combines co-evolutionary search with population based probabilistic search for discovering coherent patterns in gene expression data set. However, the algorithm could only extract biclusters with 2000 as its maximum size. The PCOBA decomposes the entire search space into subcomponents to extract hidden patterns in the data. The algorithm was tested using real yeast microarray data set and synthetic data.

Zhao *et al.* [25] asserted that a query-based strategy is beneficial for biclustering of gene expression data. For this reason, they developed a ProBic, a query based Probabilistic Relational biclustering strategy that employs the use of prior distributions to extract information contained in the gene set. The algorithm was applied on Eschenchia Coli compendium. Its performance was compared to other similar algorithm in terms of bicluster expression quality, robustness against noisy seed set, and biological relevance.

3.4.2 Correlation Based Methods

Cha *et al.* [26] utilized a correlation-based biclustering approach to find co-expressed gene set in five neurodegenerative diseases and three psychiatric disorders. They found 4,307 gene sets correlatively expressed in multiple brain diseases and 3409 gene sets exclusively specified in individual based diseases. They performed function enrichment analysis of the gene sets using Composite function annotation enriched by protein complex which demonstrated many new possible functional bases and neurological processes that are common or specific for the eight diseases.

3.4.3 Graph based Models

Dennito *et al.* [27] presents a biclustering approach based on factor graph whose objective function can be solved effectively using Max-sum algorithm. The biclustering approach was formulated as an incremental search for the largest bicluster, with preference for biclustering solution containing coherent entries, consideration for only valid assignment according to a bicluster, and the most important biclusters are those containing high valued entries. The approach was tested and compared with two synthetic and two (yeast and breast tumor) real data sets.

Roy *et al.* [28] developed a pattern based co-regulated biclustering method for gene expression data using tree to group, expand and merge genes according to their patterns capable of finding positively and negatively co regulated patterns. It uses a Biclust tree that requires a pass over the entire data set for finding a set of biologically relevant biclusters. Its performance was compared to 3 other popular algorithms using 4 yeast data set and BicAT tool.

Kanungo, Sahoo and Gore [29] utilizes one layer fixed weighted bipartite graph crossing minimization for biclustering technique for gene expression data. The gene expression data set was modelled as a weighted bipartite graph between gene and condition. The algorithm was implemented in C++ and it was compared with other well-known algorithms.

3.4.4. Metaheuristics Biclustering Techniques

Rengeswaran, Mathaiyan and Kandasamy [30] combined mutation, which is an important component of Genetic Search with Cuckoo Search metaheuristics optimization technique to generate biologically relevant biclusters. The algorithm attempted to extract the largest possible biclusters with lower Mean Square Residue with high gene variance. The performance of the algorithm was compared with Binary Particle Swarm Optimization and Shuffled Frog Leaping (SFL). The mutation operator was used to help the Cuckoo search algorithm escape local minima. The overlapping rate of the biclusters was found using Jaccard index. Yeast data, Arabidopsis thaliana expression data, and rat CNS were the benchmark gene expression data set that was utilized for performing comparative assessment of the algorithms.

Gowri *et al.* [31] proposed a Venus flytrap based biclustering method. The algorithm is inspired by the swift closure nature of the Venus flytrap leaves. The quality of the biclusters were analyzed using Average Correlation Value (ACV). The efficiency of the algorithm was tested using the yeast data set. Each bicluster is depicted as a binary string having length n = number of genes + number of conditions. Initially, K-means clustering method is employed to group the expression data into clusters that possess 30 rows and clusters that possess 3 columns which are combined to form the 30x 3 co-clusters. Thereafter, the Venus flytrap algorithm is used to extract biclusters to form the co-clusters. The algorithm was implemented using MATLAB and compared with and PSO-SA algorithms, Particle Swarm Optimization (PSO), and Simulated Annealing (SA).

Premalatha and Balamurugan [32] proposes an algorithm for biclustering gene expression data using Modified Harmony Search. The harmony search is a metaheuristics algorithm inspired by the improvisation process of musicians to find a good harmony in terms of aesthetics. Levy flight was included in the algorithm to increase diversity of the solutions. Parameters like Pitch adjusting rate is changed dynamically to escape from local minima. A comparative analysis of the algorithm was done with the biclustering with deterministic frequent pattern mining, Flexible Overlapped Biclustering Algorithm (FLOC), single objective genetic algorithm, and multi objective evolutionary algorithm. The algorithm was coded in MATLAB and tested using yeast stress data, rat CNS and Arabidopsis thaliana data. MSR was utilized in the algorithm in extracting the biclusters.

Yin and Liu [33] developed an algorithm for biclustering gene expression data using the co evolution cuckoo search. The algorithm was implemented in MATLAB 2012b. The Virtual Error, ACV and Average Spearman's Rho (ASR) were used as their cost function that assess the quality of the extracted biclusters.

Kanungo and Jaiswal [34] developed a PSO based algorithm which comprises of two phases namely the seed finding phase and seed growing phase. The algorithm was implemented in MATLAB and Yeast Saccharomyces cerevisae cell cycle expression data set was used for evaluating the efficiency of the technique. The MSR was used as the fitness function to measure the degree of coherence of the biclusters that were found.

Maatoux *et al.* [35] presented an innovative evolutionary algorithm which is built on a cross over method called (EBAcross). The algorithm uses a local search method to extract biclusters with good quality biclusters. Selection operation was done using some fitness functions such as MSR and ACV. Also, with the aid of cross over operator, biclusters were found based on discretization and standard deviation function, and mutation operator was used to improve diversification of biclusters. The algorithm's performance was examined using yeast cell cycle and saccharomyces cerevisae expression data**.** Benchmark algorithms such as ISA, Bimax, CC, and OPSM was employed for a comparative assessment with the newly proposed evolutionary algorithm.

To and Liew [36] developed a genetic algorithm based biclustering algorithm for detection of linear biclusters. It uses hyper plane to explain the direct relationships amid rows (genes) and columns (conditions) in a submatrix. The biclustering goal was achieved by determining GA control parameters such as probabilities of cross over, population size, number of generations, reproduction, and mutation. The algorithm used genetic algorithm to extract linear coherent submatrices in a data. The three GO categories and KEGG pathway was used to determine the biological enrichment of the obtained. biclusters were known to be significantly enriched biologically using. The Jaccard index was used for interbicluster evaluation.

3.4.5 Hybrid Approaches

Saber and Elloumi [37] employed the divide and conquer technique as well as the Iterative row approaches to develop some biclustering algorithms. The algorithms are the BiBin Alter, BiBin Cons and BiBin Sim for biclustering gene expression data. The algorithms were compared with OPSM, Bimax, ISA and CC. The efficiency of the algorithm was measured based on their average MSR (Mean Square Residue), average size, average gene, average condition and maximum size.

Thangavel *et al.* [38] presented a hybrid PSO-SA model for biclustering gene expression data. They emphasized that algorithms like Kmeans, greedy,

Genetic Algorithm, PSO and SA may not be sufficient to find correlations between genes dues to their inherent pros and cons. The degree of overlapping, average number of gene, average number of conditions, average volume was used to evaluate the quality of the generated biclusters. Yeast, Colon Cancer data set and breast data set was used to test the algorithm.

Abohamad, Korayem and Moustafa [39] incorporated the greedy search approach as the primary search heuristic in an immune inspired algorithm and developed a Clonal selection algorithm for biclustering gene expression data. (BicCSA). The performance of BicCSA is compared to other local search-based methods and immune inspired methods. The algorithm which was used for discovering genetic pathway seem to be particularly effective for extracting samples and genes having more coherent values and reject those that are random noise.

Das and Idicula [40] developed a greedy search binary PSO algorithm. They highlighted the fact that greedy search has the potential of being entrapped in the local minima but PSO has the feature for escaping local minima and finding global minima. Three steps were taken to attain the biclustering of gene expression data. The result of the Greedy Search Algorithm is used to initialize the bicluster. They explained that gene participating in the same biological process will have similar expression pattern. MSR was used to measure the coherence of the bicluster.

Kavitha and Arulanand [41] presented a biclustering algorithm that utilizes the Hybrid Nelder-Mead method as its search heuristic. Nelder-Mead is a local search method sensitive to the choice of initial points and does not assure the attainment of global optimum. It is a simplex search utilized for problems whose derivates may not be known. For this reason, a hybrid method was proposed with levy flight and Tabu search.

3.4.6 Linear Search Approaches

Nepomuceno *et al.* [42] highlight the fact that most studies used the Mean Square Residue as their Cost or Fitness function. However, they described the inability of the MSR to extract scaling patterns that are meaningful and interesting from a biological point of view. They detected scaling and shifting patterns using linear correlation among genes but detected only positive correlation. The Gene Ontology Database was adopted to perform comparative assessment of the algorithm with Cheng and Church ISA, OPSM, Bimax, Motifs and Samba using Gene Ontology Database.

Yun and Yi [43] proposes a biclustering approach useful for mining biclusters using clustered seeds known as Correlated and Large number of individual seeds (BICLIC). The algorithm comprises of four phases which are extracting seed biclusters, expanding seed biclusters, filtering less correlated genes and condition and checking and removal of duplicated biclusters. The algorithm was compared to three existing biclustering algorithm: BCCA, CPB, and QUBIC. Performance comparison was done with using simulated and real data set.

3.4.7 Clustering based Approaches

Tsai and Chiu [44] present an extension of Simultaneous Clustering and Attribute Distribution (SCAD). The objective function of the method was to minimize the residues within all biclusters based on MSR model. It adopts Kmeans to partition the data into K clusters which converge towards an optimal solution. However, this method is burdened with the challenge of assigning the right number of k biclusters. The performance of the algorithm was tested using yeast data set.

Tyagi-Tiwari *et al.* [45] present a parallel biclustering algorithm based on MATLABMPI (MFCM). The method is based on data parallelism and it was able to find biclusters by using gene entropy to filter the clustered data. The algorithm was tested on well-known cell cycle of yeast Saccharomyces cerevisiae, breast cancer subtype and leukaemia data. The algorithm is based on four steps gene clustering, sample clustering, gene centre finding and gene centre reduction by gene entropy filter. It utilizes fuzzy C means method for biclustering using message passing model in MATLAB environment.

3.4.8 Other Approaches

Roy, Bhattacharyya and Kalita [46] describes the nature of meaningful patterns which can may be detected in expression data and how the power of biclustering techniques can be harnessed in extracting fascinating gene patterns having close expression pattern. It was asserted that the patterns play prominent roles in deciphering gene interactions, functions of genes, and disease

targets. In the study, it was explained that shifting patterns detected in a data matrix exhibit similar trend but in terms of distance between them, they can be far from each other. Also, scaling patterns have multiplicative distance among them, and coherent patterns are group of genes that show similar pattern tendency across different conditions and co regulated patterns which could be positively co regulated or negatively co regulated.

Huoari, Ayadi and Yahia [47] proposed a biclustering algorithm called NBic-ARM (Negative Biclusters using Association Rule Mining) which uses Generic Association rules to extract negative correlated genes. The algorithm has four steps which are pre-processing of gene expression data matrix, extracting biclusters of positive correlations using generic association rules, extracting negative correlated biclusters and extracting maximal correlated genes. Human B- cell Lymphoma, Yeast cell cycle, and Saccharomyces expression data were used in verifying the statistical and biological relevance of the algorithm. The statistical relevance of the algorithm was tested using p value criterion. Gene ontology (GO) annotation was also used to ascertain the biological relevance of the generated biclusters.

4. Conclusion

Data mining algorithms are valuable tools for analysis of gene expression data. Well known examples of these methods are classification, dimension reduction, association rules, clustering and biclustering algorithms. The various biclustering techniques have opened new frontiers of knowledge in genomics by being helpful for unearthing functionally related genes among other benefits. However, the bioinspired algorithms have been shown to be better suited for complex and hard problems such as a biclustering which are non-linear and highly dimensional. This is because they possess the quality of proffering solutions to intricate relationship from non-complex rules without having full knowledge of the search space.

Although, these algorithms usually have some deficiencies which can deter them from finding optimal solution to the problem that they have been deployed to solve. It found that some of these deficiencies can be ameliorated when they are hybridized with another algorithm. These were demonstrated in studies that derived better solutions using such hybridized bioinspired algorithms. It was also deduced from the review that many of the studies focused on gene expression data set such as yeast data in analyzing their algorithm. It is believed that the results of the biclustering analysis of this study will provide useful knowledge for utilizing biclustering algorithms for disease diagnosis and prognosis. In addition, this study proposes that more attention needs to be given to the analysis of gene expression data of various disease-causing organisms especially the neglected ones.

References

[1] Pontes, B. Giraldez, R. Aguilar-Ruiz, J. S. (2015)Biclustering on expression data: A review *Journal of Biomedical Informatics*, Vol. 57, pp.163–180, http://dx.doi.org/10.1016/j.jbi.2015.06.028.

[2] Freitas, A. V. Ayadi, W. Elloumi, M. Oliveira, J. Oliveira, J. and Hao, J. (2013). Survey on Biclustering of Gene Expression Data. In Biological Knowledge Discovery Handbook (eds M. Elloumi and A. Y. Zomaya). doi:10.1002/9781118617151.ch25.

[3] Ayadi, W. Elloumi, M. Hao, J. K. (2014). Systematic and Stochastic biclustering, Algorithms for microarray data analysis, pages 1-30, http://www.researchgate.net/publication/2813244.

[4] Tarek, S. Abd Elwahab, R. Shoman, M. (2017). Gene expression-based cancer classification, *EgyptianInformatics Journal*, Vol. 18, pp 151-159, http://dx.doi.org/10.1016/j.eij.2016.12.001.

[5] Aydadenta, H. and Adiwijaya, (2018). On the classification techniques in data mining for microarray data classification, International Conference on Data and Information Science, IOP Conf. Series: Journal of Physics: Conf. Series 971 (2018) 012004, doi :10.1088/1742-6596/971/1/012004.

[6] Devi Arockia Vanitha, C. Devaraj, D., Venkatesulu M. (2015). Gene Expression Data Classification using Support Vector Machine and Mutual Information-based Gene Selection, *Procedia Computer Science*, vol. 47, pp. 13 – 21.

[7] Chen, Z. Li, J. Wei, L. Xu, W. Shu, Y. (2011). Multiple-kernel SVM based multiple-task oriented data mining system for gene

expression data analysis, *Expert Systems with Applications*, doi:10.1016/j.eswa.

[8] Ram, M. Najafi, A. Shakeri, M. T. (2017). Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest, *Iran Journal of Pathology*, 12 (4): 339-347.

[9] Krishnaiah, V., Narsimha, G., Chandra, S. (2013). Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques International Journal of Computer Science and Information Technologies, Vol. 4, No 1, page 39-45.

[10] Arundhathi, T. Adilakshmi T. (2017). Role of Association Rule Mining in DNA Microarray Data - A Research, International Journal on Future Revolution in Computer Science & Communication Engineering, Vol. 3, no.12, pp. 443 – 447.

[11] Alagukumar, S. Lawrance, R. (2015). Selective Analysis of Microarray Data using Association Rule Mining, Procedia Computer Science, vol. 47, pp. 3 – 12.

[12] Vengateshkumar, R. Alagukumar, S. Lawrance, R. (2017). Boolean Association Rule Mining on Microarray Cancer Gene Expression Data using Gene Expression Intervals, *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 5, no. 1.

[13] Das, S. (2011). Mean Squared Residue Based Biclustering Algorithms for the analysis of gene expression data. PHD thesis, Department of Computer Science, Cochin University of Science and Technology, Kochi, India.

[14] Olivier, F. Njeunje, N. (2014). Linear and Non-linear Dimension Reduction Applied to Gene Expression Data of Cancer Tissue Samples, *Applied Mathematics, Statistics, and Scientific Computation*, University of Maryland - College ParK.1-25.

[15] Wang, H. and Van der Laan, M. J. (2011). Dimension reduction with gene expression data using targeted variable importance measurement, *BMC Bioinformatics*, 12:312 http://www.biomedcentral.com/1471-2105/12/312.

[16] Hira, Z. M. and Gillies, D. F. (2015). A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data, Advances in Bioinformatics, Volume 2015, Article ID 198363, 13 pages, http://dx.doi.org/10.1155/2015/198363.

[17] Ha, S. H. and Joo, S. H. (2010). A Hybrid DataMining Method for the Medical Classification of Chest Pain. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, Vol.4, No.1.

[18] Lavanya, D. Rani, K. U. (2013). A Hybrid Approach to Improve Classification with Cascading of Data Mining Tasks, *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, Vol. 2, Issue 1.

[19] Nookala, G. K. M. Pottumuthu, B. K. Orsu, N. Mudunuri, S.B. (2013). Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification, *International Journal of Advanced Research in Artificial Intelligence (IJARAI)*, Vol. 2, No.5.

[20] Sri, N. K. Porkodi, R. (2019). An Extensive Survey on Biclustering Approaches and Algorithms for Gene Expression Data, *International Journal Of Scientific & Technology Research* Vol. 8, no 9, ISSN 2277-8616.

[21] Saber H.B. and Elloumi M. (2014) A new survey on Biclustering of Microarray data, Computer Science and Information Technology, PP.165-183 CS&IT-CSCP DOI:10.5121/csit.2014.41314.

[22] Oghabian, A. Kilpinen, S. Hautaniemi, S. Czeizler, E. (2014). Biclustering Methods: Biological Relevance and Application in Gene Expression Analysis. PLoSONE 9(3): e90801. doi:10.1371/journal.pone.0090801.

[23] Chekouo, T., Murua A. (2015). The penalized biclustering model and related algorithms, *Journal of Applied Statistics*, vol 2 no. 6 pages 1255-1277,https://doi.org/10.1080/02664763.2014.999647.

[24] Joung, J. G., Kim, S. J. Shin, S.Y. Zhang, B.T. (2012), A probabilistic coevolutionary biclustering algorithm for discovering coherent patterns in gene expression dataset, *BMC Bioinformatics*,13 (Suppl 17): S12 shttp://www.biomedcentral.com/1471-2105/13/S17/S12.

[25] Zhao, H. Cloots, L. Van den Bulcke, T. Wu, Y. Smet, R. D. Storms, V. Meysman, P. Engelen, K. Marchal, K. (2012). Query-based

biclustering of gene expression data using Probabilistic Relational Models, *BMC Bioinformatics*, http://www.biomedcentral.com/1471-2105/12/S1/S37 2011,12 (Suppl 1): S37.

[26] Cha, K. Hwang, T. K. Yi,,G. S. (2015). Discovering transnosological molecular basis of human brain diseases using biclustering analysis of integrated gene expression data, *BMC Medical Informatics and Decision making*, 15 (Suppl 1): S7 http://www.biomedcentral.com/1472-6947/15/S1/S7.

[27] Denitto, M. Farinelli, A. Figueiredo, M.A.T.,and Bicego, M. (2017). A biclustering approach based on factor graphs and the max-sum algorithm, *Pattern Recognition*, Vol. 62, Pages 114-124, https://doi.org/10.1016/j.patcog.2016.08.033 .

[28] Roy, S. Bhattacharyya, D. K, Kalita, J. K. (2013). CoBi: Pattern Based Co-Regulated Biclustering of Gene Expression Data, 1-37.

[29] Kanungo, S. Sahoo G. Gore, M. M. (2011). BiCross: A Biclustering Technique for Gene Expression Data using One Layer Fixed Weighted Bipartite Graph Crossing Minimization, *International Journal of Computer Applications*, Vol. 29, No.4, pp. 0975 – 8887.

[30] Rengeswaran, B. Mathaiyan, N. and Kandasamy, P. (2017). Cuckoo Search with Mutation for Biclustering of Microarray Gene Expression Data, *The International Arab Journal of Information Technology*, vol. 14, no.3, pp. 300-306.

[31] Gowri, R., Sivabalan, S., Rathipriya R. (2016). Biclustering Using Venus Flytrap Optimization Algorithm. In: Behera H., Mohapatra D. (eds) *Computational Intelligence in Data Mining*, Volume 1, Advances in Intelligent Systems and Computing, vol 410. Springer, New Delhi.

[32] Premalatha, K. Balamurugan, R. (2016). A modified Harmony Search Method for biclustering Microarray Data, *International Journal of data mining and Bioinformatics*, vol. 16, no. 4, pp. 269-289.

[33] Yin, L. Liu Y. (2015). Biclustering of the Gene Expression Data by Coevolution Cuckoo Search *Int.J.Bioautomation*, Vol.19, no.2, pp. 161-176.

[34] Kanungo, S. Jaiswal, S. (2015). A Framework for Mining Coherent Patterns Using Particle Swarm Optimization based Biclustering, *I.J. Intelligent Systems and Applications*, Vol.11, pp. 33-40, DOI: 10.5815/ijisa.2015.11.05.

[35] Maˆatouk, O. Ayadi, W. Bouziri, H and. Duva, B. (2014). Evolutionary Algorithm Based on New Crossover for the Biclustering of Gene Expression Data, Springer International Publishing Switzerland, LNBI 8626, pp. 48–59.

[36] To, C. Liew, A. W-C. (2014). Genetic Algorithm based detection on general linear biclusters, Proceedings of 2014 International Conference on Machine Learning and Cybernetics (ICMLC), Lanzhou, China, Vol. 2, https://doi.org/10.1109/ICMLC.2014.7009667.

[37] Saber, H. B. and Elloumi, M. (2015). A New study on biclustering Tools, On Biclusters Validation Evaluation and Functions, *International Journal of Computer Science & Engineering Survey (*IJCSES), Vol.6, No.1, pp 1-13, Doi:10.5121/Ijcses.2015.6101.

[38] Thangavel, K., Bagyamani,J., and Rathipriya, R. (2011). Novel Hybrid PSO-SA Model for Biclustering of Expression Data, International Conference on Communication Technology and System Design, *Procedia Engineering*, Vol. 30, pp. 1048 – 1055, doi:10.1016/j.proeng.2012.01.962.

[39] Abohamad, W., Korayem, M. and Moustafa, K. (2010) Biclustering of DNA microarray data using artificial immune system, 10th International Conference on Intelligent Systems Design and Applications, pp. 1223-1228. https://www.researchgate.net/publication/215642421.

[40] Das, S. Idicula, S. M. (2010) Greedy Search-Binary PSO Hybrid for Biclustering Gene Expression Data, *International Journal of Computer Applications*, Vol. 2, No. 3, pp. 0975 –8887.

[41] Kavitha, D. Arulanand N. (2016). A hybrid Nelder Mead method for biclustering gene expression data. *International Journal of Technology Enhancements and Emerging Engineering Research*, VOL 4, ISSUE 2, pp. 21-26, ISSN 2347-4289.

[42] Nepomuceno, J. A. Troncoso, A. Aguilar-Ruiz, J. S. (2011). Biclustering of Gene Expression Data by Correlation-Based Scatter Search, *Biodata Mining*, Vol. 4, No. 3.

[43] Yun, T. and Yi, G. S. (2013). Biclustering for the comprehensive search of correlated gene expression patterns using clustered seed expansion, *BMC Genomics*, 14:144, pp. 1-15, http://www.biomedcentral.com/1471-2164/14/144.

[44] Tsai, C.Y. and Chiu, C.C. (2011). A Novel Microarray Biclustering Algorithm, World Academy of Science, Engineering and Technology, *International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering*, Vol.4, No 5.

[45] Tyagi-Tiwari, D. Das, S. Jha, M. & Srivastava, N. (2015). Biclustering using Parallel Fuzzy Approach for Analysis of Microarray Gene Expression Data, *International Journal of Computer Science and Security (IJCSS)*, Vol. 9, no. 5, pp. 253-265.

[46] Roy, S., Bhattacharyya, D. K., and Kalita, J. K. (2015). Analysis of Gene Expression Patterns Using Biclustering, *Methods in Molecular Biology*, DOI 10.1007/7651_2015_280 Springer Science+Business Media, New York.

[47] Houari, A. Ayadi, W., Yahia S. B. (2017). Mining Negative Correlation Biclusters from Gene Expression Data using Generic Association Rules, 21th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017 Procedia Computer Science, 112 pp. 278-287.