European Journal of Applied Sciences 9 (2): 88-102, 2017
ISSN 2079-2077

DOI: 10.5829/idosi.ejas.2017.88.102

Efficient Data Mining Technique for Disease Identification

P.M. Boom and S. Prabakaran

Department of Computer Science and Engineering, SRM University, Chennai, Tamil Nadu, India

Abstract: For extracting the relevant biological information to identify diseased samples, the present work is proposed. Primarily, a novel technique called heuristic search has been designed for analysis of the standard biological process on physiological data [BPPD] of the gene expression. The physiological data consists of two patterns of the gene expression datasets: physical patterns and logical pattern. The biological process of these patterns of gene expression datasets were analyzed through heuristic search identifies the biological changes of physiological data and is helpful in extracting more expressive genes. Finally, to classify the samples into normal or diseased, the relational sequence optimization among the biologically associated genes was performed using an efficient Bi-clustered Ant Optimized Feature Relational Sequencing (BAOFRS) method. BAOFRS method clustered similar relational features using Ant Optimized Medoids algorithms to improve the relational sequence based clustering on gene expressional data to obtain sequence-pairs and finally clusters normal and abnormal samples. The experimental evaluation was conducted using Cancer datasets which is derived from Broad Institute repository. The proposed method results were compared with existing bi-clustering method, HBH and SO method. The results show that, the proposed algorithms perform better and the diseased samples were extracted from dataset with eminent similarity value and accuracy.

Key words: Gene Expression Data · Bi-clustering · Heuristic Search · Biological Association · Feature Relational Sequencing

INTRODUCTION

Bio informative knowledge discovery is an essential one in many disease diagnosis, drug improvement, genetic functional interpretation, gene metamorphisms etc. For this purpose of knowledge discovery, tremendous amount of gene expression data is to be generated. Also, fast and novel analytical methods are needed to analyze this large amount of data. Performing research in traditional methods is not feasible for huge gene data. The data generation is achieved by various expertises. DNA Micro Array Technology [1] is the most promising technologies to create large amount of data.

DNA Microarray is the technology that allows the researchers to examine and to solve the difficulties which are non-traceable. Using this technology, the expression of several genes is analyzed successfully in a single process. High-density DNA microarrays are the powerful tools for functional genomics studies. They are used for computing the thousand expressions of genes. The microarray technology uses the chips with DNA sequences as probes. Many techniques are used to examine the large quantity of data created. Recently, biological information removal using clustering methods [2] were utilized for analytical estimation of gene expression. To design the methods to examine the large quantity of information in gene expression data, Bi-clustering algorithm is designed. It also presents local structures from gene expression data set. Traditional single cluster model is not capable to mine exact information from large and heterogeneous collection gene expression data. Data mining techniques extract the hidden information from datasets that does not recognize the biological links between genes [3]. In addition, techniques like optimization Algorithm [4] for multi dimensional search space does not provide relational optimization result on varying gene expressional problems. Therefore, there is a due need to develop new techniques to effectively address these issues. The present work concentrates to develop such new techniques to suit to address these issues.

Corresponding Author: P.M. Boom, Department of Computer Science and Engineering, SRM University, Chennai, Tamil Nadu, India.

MATERIALS AND METHODS

Analyzing Biological Changes: Mining micro-array gene expression data is a crucial subject matter in bioinformatics with widespread applications such as disease diagnosis, drug development, genetic functional interpretation, gene metamorphisms etc. In recent times, biological information mining using clustering techniques are used for the analytical evaluation of gene expression. To tap out the massive quantity of information enclosed in gene expression data, a Bi-clustering algorithm is used to explore local structures from gene expression data set. Since, traditional single cluster model is unable to mine precise information from large and heterogeneous collection gene expression data. So, the development of a new computational method is in need to improve the analysis of gene expression data sets, particularly to identify the genes expressing more in a biological process. Despite, the presence of existing Bi-clustering algorithm somehow manages this task, an efficient new Heuristic approach for analyzing standard biological process of gene expression data, particularly physiological data is discussed in this chapter. Before discussing this proposed Heuristic approach in detail, some necessary basic concepts are explained.

The physiological data consists of both physical and logical patterns of the gene expression datasets. The biological process of physical and logical pattern of gene expression datasets are analyzed through Heuristic search in the present work. After identifying the physiological data on gene expression datasets, the Heuristic search algorithm is used for identifying the biological process. Experimental evaluations are conducted for this proposed Heuristic Search based on analysis of Biological Process on Physiological Data with standard benchmark gene expression data sets from research repositories in terms of size of gene expression datasets, Heuristic search threshold and response time [5].

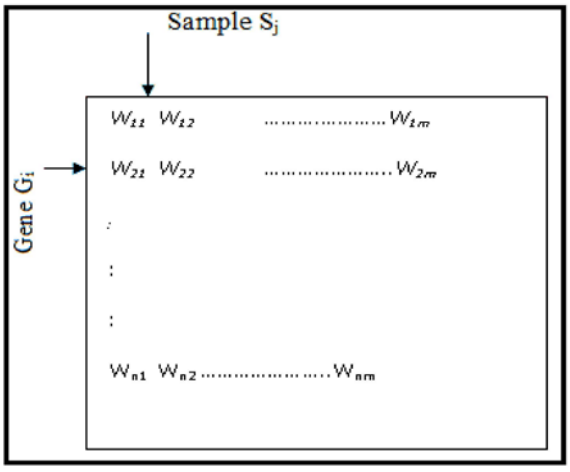
Gene Expression Datasets: Gene expression is the procedure by which information from a gene is utilized in the production of an efficient gene product. These goods are habitually proteins, but in non-regulatory genes such as ribosomal RNA (rRNA) genes, transfer RNA (tRNA) genes or small nuclear RNA (snRNA) genes, the product is a practical RNA. The progression of gene expression is utilized by all recognized multi cellular organisms, prokaryotes and viruses to produce the macromolecular machinery for life [6].

Gene expression data was generated by DNA chips and other microarray techniques. It was frequently presented as matrices of expression levels of genes under different conditions such as environments, individuals and tissues. One of the objectives in expression data analysis was to group genes according to their expression under multiple conditions, or to group conditions based on the expression of a number of genes. This was lead to discovery of regulatory patterns or condition similarities [7].

A micro-array research classically evaluates a huge amount of DNA sequences (genes, cDNA clones, or spoken sequence tags [ESTs]) under numerous conditions. These circumstances may be an instance series through a genetic process (e.g., the yeast cell cycle) or a compilation of diverse tissue samples (e.g., normal versus cancerous tissues). In this work, we focus on the analysis of biological process on physiological data on gene expression datasets. Likewise, we consistently submit to all varieties of tentative conditions as "*samples*" if no perplexity will be caused. A gene expression data set from a micro-array experiment can be symbolized by a real-valued expression matrix [8],

$$M = \{w_{ij} \mid 1 \le i \le n, i \le j \le\} \quad (1)$$

where the rows $(G = \{\vec{g}_1, \ldots\ldots, \vec{g}_n\})$ form the expression patterns of genes, the columns $(S = \{\vec{s}_1, \ldots, \vec{s}_m\})$ represent the expression profiles of samples and each cell w_{ij} is the measured expression level of gene i in sample j.



- Gene Expression Form

The unique gene expression datasets attained from an examining process contains missing values, noise and organized distinctions happening from the uncertain

procedure. Thus the gene expression datasets are generally formed with the given logical scheme of the expression levels [9].

Analyzing Biological Process on Gene Expression Datasets using Heuristic Search: The proposed work is efficiently designed for analyzing the biological process on physiological data which is present in the gene expression datasets using Heuristic search. Heuristic search is a method that finds the good solution in a reasonable time. The proposed work operates under two different operations. The first operation is to analyze the Gene Expression datasets. The second operation is to mine related biological data which is required for disease identification [10].

Analyzing Biological Process of Physiological Data using Heuristic Search (BPPD): The gene expression consists of collection of genes present in the datasets. Each gene consists of two types of pattern i.e., physical pattern and logical pattern. The physical pattern provides information about physical structure of the gene on the gene expression datasets i.e., colour, shape and structure of the gene based on its environment. The logical pattern provides information about the intelligence of the gene among all genes present in it and it also represent the gene reactions on all types of situations. The physical and logical patterns form a physiological data which provides all information about the genes [11].

Figure 2 shows the process of identifying the biological processes using this Heuristic search algorithm. Identifying the biological changes on genes based on physical and logical pattern is presented in this work. The biological process indicates the changes occurring in the genes when some foreign particles disturb the genes in the sample sequences. Heuristic search is used for identifying the biological changes on physiological data of gene expression datasets [12].

After identifying the physiological data on gene expression datasets, the Heuristic search algorithm is used for identifying the biological process. A Heuristic search algorithm sustains a collection of genes as the candidates of subjective genes and a division of samples as the candidates of gene expression datasets. The good quality will be possessed by repeatedly adjusting the candidate sets. A Heuristic search algorithm also measures two basic elements, a state and the distinct adjustments. Necessitate of the algorithm describes the following items [13]:

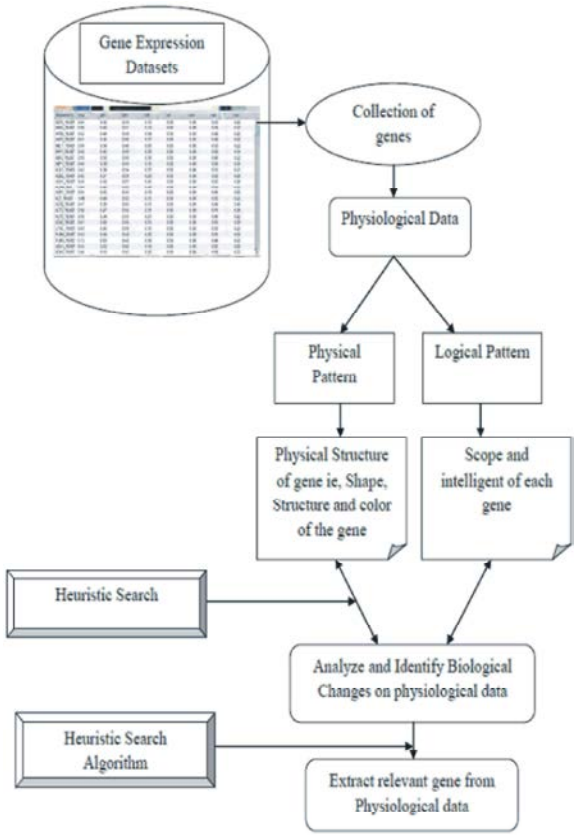


Fig. 2: Process of Heuristic based Analysis of Biological Process (BPPD)

- Partition of samples *S*
- Set of genes *G*
- Number of the states Ω computed based on partition (Depends on the entity of dataset and gene)

An adjustment of the state would be:

- Insertion of *g* into *G,* if gene $g \notin G$
- Removal of *g* from *G, i*f gene $g \in G$
- Movement of *s* to *S'* where *S* is not equal to *S',* for a sample s in *S*

The proposed Heuristic Search Algorithm is as follows:

Algorithm 1: Heuristic Search Algorithm
Input: Gene Expression datasets
Output: Relevant physiological data (Extract the Relevant gene names)
Process
Initialization phase
Input the Gene expression datasets
Adopt a random initialization
Calculate the Expression Level

Iterative adjusting phase
Initialize an element *I* as 0
For each gene *g*
Do
Identify the physical and logical entity
Register a sequence of genes and samples arbitrarily
Repeat until *i<g*
Increment *i* by 1
End For
For each gene *g* or sample *S* along the sequence
Do
Run *Heuristic Wrest Algorithm*
Repeat Until *i<g/S*
Increment *i* by 1
End For
Sort *P*
Select first 14 genes in a row *R*
End

The algorithm can also be explained as follows:

Step 1: Heuristic Search Algorithm initially takes Gene Expression datasets as input

Step 1.1: The algorithm has two phases namely initialization phase and iterative adjusting phase [14].

Step 2: In the initialization phase, an initial state is processed arbitrarily and the particular quality value is computed. Given a gene expression matrix *M* with *m* samples and *n* genes, the task is to identify the biological process on physiological data on Gene Expression datasets.

Step 3: During the Iterative adjusting phase, the physical and logical patterns are identified for all the genes in Gene Expression datasets (i to n genes) and Register a sequence of genes and samples arbitrarily.

Step 4: Call *Heuristic Wrest Algorithm* for each gene *g* or sample *S* along with the register sequence till reaching all gene or samples

Step 5: Sort the returned P in ascending order to extract the relevant gene

Step 6: The relevant Physiological data, which means the name of the genes relevant to our search in disease identification.

To identify the process and changes of the biological data, compute the actual and updated entity, i.e., $\Delta\Omega_P = \Omega_P' - \Omega_P$, where Ω_P' and Ω_P are the different structure before and after the transform, concurrently.

The following algorithm explains the process of extracting relevant 14 genes for each dataset as follows:

Algorithm 2: Heuristic Wrest Algorithm
Input: Gene *g* or sample *S* along the sequence
Output: The relevant P data, which means the name of the genes relevant to our search in disease identification.

Process
For all Samples S
For all genes g in Sample S
If genes $\notin$ Physical Entity *P*, then
Calculate the difference between 2 states,
$\Delta\Omega_P = \Omega_P' - \Omega_P$
//difference between actual gene structure and updated structure of the gene
If genes $\notin$ Logical Entity *L*, then
Calculate the difference between 2 states,
$\Delta\Omega_L = \Omega_L' - \Omega_L$
// difference between actual gene behaviour and updated behaviour of the gene
End If
End If
If $\Delta\Omega_P >= 0$ *AND* $\Delta\Omega_L >= 0$, **THEN**
Extract the gene
$$p = \exp\left(\frac{\Delta\Omega_L * \Delta\Omega_P}{\Omega_L \Omega_P \times T_i}\right)$$
//extracts the relevant gene
Else
Ignore the gene
End If
Return *p*
End For
End For
End

The Heuristic Wrest Algorithm can be explained as follows:

Step 1: Consider gene/Sample along the sequence which was obtained from Heuristic Algorithm
Step 2: For all samples and all genes in samples,

Step 2.1: Compute structure difference $\Delta\Omega_P$ based on Physical Entity *P, (g* $\notin$ *P),*

Table 1: List of selected genes for Ovarian Cancer using proposed framework

Name of the Gene
hum_alu_at (miscellaneous control)
ADP-ribosylation factor-like 2
Anne xin II (lipocortin II) pseudogene 2
Protein kinase mitogen- activated 13
EST: zv26h12.r1 Soares NhHMPu S1 Homo sapiens cDNA clone 754823 5' similar to contains Alu repetitive element;, mRNA sequence. (from Genbank)
TNNT2 gene exon 11
Alpha-1 collagen type I gene, 3' end
Tumor-associated 120 kDa nuclear protein p120, partial cds(carboxyl terminus)
Mitochondrial 16S rRNA gene (partial)
RPS14 gene (ribosomal protein S14) extracted from Human ribosomal protein S14 gene
RPL37 Ribosomal protein L37
Ribosomal protein L27a mRNA
RPS3A Ribosomal protein S3A
UBA52 Ubiquitin A-52 residue ribosomal protein fusion product 1

$$\Delta\Omega_P = \Omega_P' - \Omega_P$$

where Ω_P is a actual gene structure and Ω_P' is a updated structure of the gene

Step 2.2: Compute behaviour difference $\Delta\Omega_L$ based on Logical Entity *L, (g∉ L),*

$$\Delta\Omega_L = \Omega_L' - \Omega_L$$

where Ω_L is a actual gene behaviour and Ω_L' is a updated behaviour of the gene

Step 3: If ($\Delta\Omega_P >= 0$ && $\Delta\Omega_L >= 0$), then extract the n number of genes and compute *p,* which diminish the number of gene extracted,

$$p = \exp\left(\frac{\Delta\Omega_L * \Delta\Omega_P}{\Omega_L \Omega_P \times T_i}\right) \quad (2)$$

Step 3.1: Ignore the gene which is irrelevant

Step 4: Return *P*

To present each gene or check a sensible chance, all possible adjustments are formed subjectively at the enterprise of all iteration. Before Heuristics Search Algorithm proceeds for identifying the biological changes, the physical and logical patterns are analyzed and noted. After examining the physiological data, the biological changes of those data are identified thorough Heuristic Algorithm. The biological changes occur only if the physiological data of gene have met with some changes in their nature. The biological changes and extraction of the relevant gene was done through Heuristic Wrest Algorithm. In that case, the biological changes occur and those changes are identified by noting down the set of genes, which was done efficiently using Heuristic Wrest Algorithm and Heuristics Search Algorithm [15, 16].

Table 2: Comparison of Gene Expression Level for Proposed BPPD Method and Existing Bi-Clustering Approaches

		Gene Expression Level (%)	
Datasets	No. of Samples	Existing Bi-clustering algorithm	Proposed BPPD
Pancreas	63	16	22
Uterus	55	11	25
Ovary	56	21	25
Prostate	63	03	22
Colorectal	54	18	26

Performance Analysis of Proposed Method: In this work, we analyze the biological process of physiological data occurred on gene expression datasets using Heuristic search algorithm. The physical and logical pattern of each gene is first identified and then the biological processes of physiological data are identified using Heuristic search algorithm. An experimental evaluation is also being conducted to estimate the performance of the BPPD method with some metrics. The performance of the proposed BPPD is measured in terms of gene expression level, Heuristic search threshold and response time.

Gene Expression Level: One of the reasons to carry out a microarray experiment is to monitor the expression level of genes. Gene expression level is measured based on the gene count taken for experimental purpose.

$$Gene\,Expression\,Level(\%) = \left(\frac{No.of\ genes\,extracted}{No.of\ Samples}\right) * 100 \quad (3)$$

Table 2 describes the process by which physiological data from a gene is used in the synthesis from the gene expression datasets. The outcome of the proposed analysis of the biological process on physiological data present in the gene expression datasets using Heuristic search is compared with an existing bi-clustering algorithm.

Figure 3 depicts the process of identifying the retrieval of physiological data from each gene present in the gene expression datasets. In the proposed BPPD,

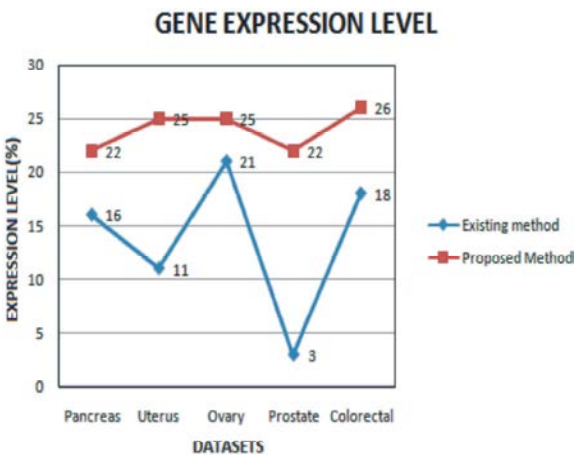


Fig. 3: Type of Datasets vs. Gene Expression Level

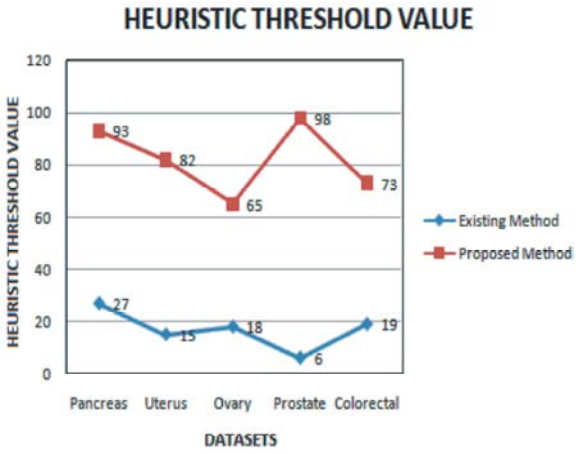


Fig. 4: Type of Datasets vs. Heuristic Search Threshold

the gene expression datasets are analyzed and the physiological entity for each gene is identified and processed. The gene expression level is high in the proposed BPPD since it used the Heuristic search algorithm which identifies the best solution for the biological change issues. Compared to an existing bi-clustering algorithm, the proposed BPPD outperforms well and the variance is 40-50% high.

Heuristic Search Threshold: In existing method, the time taken for searching genes from samples differs for each dataset because of different number of genes. In Pancreas for 10 genes it takes 36.4ms, in Uterus for 6 genes it takes 40.01ms, in Ovary for 12 genes it takes 67.88ms, in Prostate for 2 genes it takes 33.64ms and in Colorectal for 10 genes it takes 52.11ms. But in proposed method, only 14 genes from each dataset are used and so, searching time alone differs. For Pancreas 15.01ms, Uterus 17ms, Ovary 21.45, Prostate 14.33ms and Colorectal 19.09ms.

$$Heuristic\ Search Thershold(\%) = \left(\frac{No.of\ genes\ searched}{Time}\right) * 100 \quad (4)$$

Table 3 describes the process of Heuristic search method based on the size of data present in the gene expression datasets. Based on the table, the below graph is depicted.

Figure 4 illustrates the process of identifying the Heuristic search threshold value based on number of data present in the gene expression datasets. In the proposed BPPD, the physiological data is first identified and the process of those physiological data is noted. Then the biological process of those physiological data is identified based on Heuristic search algorithm. The Heuristic search

Table 3: Comparison of Heuristic Search Threshold Level for Proposed BPPD Method and Existing Bi-Clustering Approaches

		Heuristic Threshold Value (%)	
Datasets	No. of Samples	Existing Bi-clustering algorithm	Proposed BPPD
Pancreas	63	27	93
Uterus	55	15	82
Ovary	56	18	65
Prostate	63	06	98
Colorectal	54	19	73

threshold is measured in terms of how far the best solution has been identified based on physiological data. Compared to an existing bi-clustering algorithm that clusters the genes alone without knowing its biological processes, the proposed scheme used Heuristic search algorithm for identifying the biological process for each gene present in the datasets and it outperforms well and variance is 70% high in the proposed BPPD.

Response Time: In existing method, the time taken for searching relevant genes from samples differs for each dataset because of different number of genes.

$$\text{Re}spnse\ Time(ms) = No.of\ relevant\ genes\ searched * (\text{Re}sponse\ End\ Time - \text{Re}sponse\ Start\ Time) \quad (5)$$

Table 4 shows the time taken to response the biological process identification procedures based on the size of data present in the gene expression datasets. Based on the table, the below graph is depicted.

Figure 5 depicts the time taken to response the search process at given interval of time based on number of data. In the proposed BPPD, the time taken to response

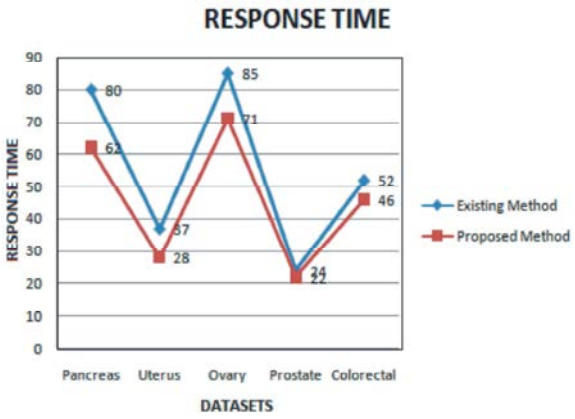


Fig. 5: Type of Datasets vs. Response Time

Table 4: Comparison of Response Time Level for Proposed BPPD Method and Existing Bi-Clustering Approaches

		Response time (sec)	
Datasets	No. of Samples	Existing Bi-clustering algorithm	Proposed BPPD
Pancreas	63	80	62
Uterus	55	37	28
Ovary	56	85	71
Prostate	63	24	22
Colorectal	54	52	46

the Heuristic search process is limited since the physical and logical patterns of the genes are identified at first step. The response time is measured in terms of seconds (secs). Compared to an existing which consumes more time even for clustering process, the proposed analyzing the biological process on physiological data present in the gene expression datasets using Heuristic search consumes less response time and provide an accurate value related to it and the variance is 20-30% high in the proposed BPPD.

Finally, it is being observed that the proposed scheme used Heuristic search algorithm for identifying the standard biological processes on physiological data in gene expression datasets. The physiological data are first analyzed among the gene expression datasets and the biological process of those physiological data is identified using Heuristic search algorithm in a less interval of time.

Identification of Diseased Samples: In a multi dimensional search space, an optimization algorithm does not provide relational optimization result on varying gene expressional problems. Existing Heuristic Algorithm with Black Hole (HBH) phenomenon solves the clustering problem, but the bi-cluster based gene expression information was not extracted. A key issue on existing work was to handle multi modal structure optimization problems with effective searching process. Simulation-Based Optimization (SO) method provided multi-objective optimization into real world design, but it could not offer relational sequence optimized result on the associated gene data. To address this issue, Bi-clustered Ant Optimized Feature Relational Sequence (BAOFRS) method is proposed in this paper. This method is designed using an ant optimized relational sequences on gene expressional data. Relational sequences are identified by the features and also compute the similarity values between the sequences are also computed. BAOFRS method uses the K-mers relational knowledge algorithm to improve the clustering efficiency and also to identify the relational sequences on gene expression dataset. Then, Jaccard similarity coefficient in BAOFRS method computes the similarity value between the feature vectors. Using Ant Optimized Medoids Algorithms, BAOFRS method group similar relational features to improve the relational sequence based clustering on gene expressional data to obtain sequence-pairs.

After obtaining the number of sequence-pairs and obtaining the estimates of the sequence-pairs, a simple pattern matching process is performed to evaluate the sequence-pairs into ‘normal’ or ‘abnormal’. BAOFRS method minimizes the bi-clustering time on performing the relational sequence bi-clustering. The performance of the proposed method is compared with the some existing methods and it is measured in terms of relational sequence result rate, bi-clustering efficiency rate and similarity score level using Cancer gene expression dataset. These existing methods are explained in the following sections before proceeding further.

Ant Optimized Feature Relational Sequence on Gene Expression Data: In this work, an effective pattern matching model called as Bi-clustered Ant Optimized Feature Relational Sequencing (BAOFRS) method has been designed for gene expression data. It is implemented to identify the relational sequences by minimizing the bi-clustering time by improving the similarity score level. That is, the main objective of Bi-clustered Ant Optimized Feature Relational Sequencing (BAOFRS) method is to identify the relationship sequence between features of the gene expression data.

The relationship between different features from vast dataset is taken into consideration to identify the relations on different types of features. The relation sequence is identified in BAOFRS method using the K-mers relational

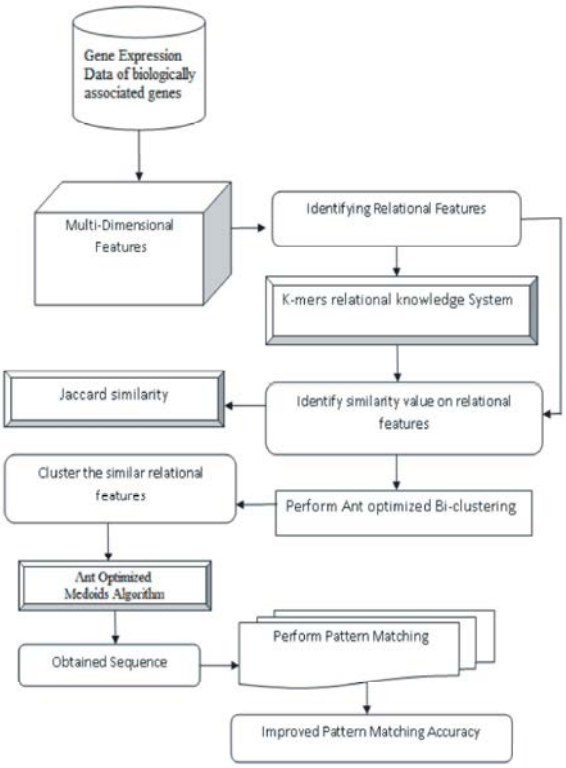


Fig. 6: Architectural Diagram of BAOFRS Method

knowledge algorithm. K-mers relational knowledge uses the multi-dimensional 'K' features (i.e.,) attributes as input to relate and order (i.e.,) sequences in BAOFRS method. K-mer identifies the relationship on biological amino acid information using discrete probability division with K-mer combination of features.

The optimized relational sequence is clustered using Ant optimized bi-clustering to group similar relational features. The similar relational features are clustered using the ant rule and, as a result, the bi-clustering time is also minimized in BAOFRS method. Relational Sequence Bi-clustering group the set of sequences of similar gene data features. The architectural diagram of Bi-clustered Ant Optimized Feature Relational Sequencing method is illustrated in Figure 6.

As illustrated in Figure 6, multi-dimensional gene expression data is used for the BAOFRS experimental work. Gene expressions with the biological association consists of different products (i.e.,) features. The relational feature sequence is identified using the K-mers relational knowledge system and applied widely on multi-dimensional gene data objects to relate the features. The features used to relate in BAOFRS method contains the special set of properties with specific domain.

The identified relationship is used in the proposed method to find the similarity value. The similarity value is identified using the Jaccard similarity coefficient. The similarity coefficient of the dimensional vector is expressed as '0' or '1'. The similarity on features is represented as'1' and if dissimilarity occurs on features, then it is represented as '0'. The similarity value clearly group similar (i.e.,) relative features using Ant optimized bi-clustering scheme. The ant optimized bi-clustering uses the Ant Optimized Medoids algorithm to improve the clustering efficiency for gene expression data.

K-Mers Relational Knowledge System: The first step in Bi-clustered Ant Optimized Feature Relational Sequencing is to identify the relational features. The relative features 'f' uses the K-mers relational knowledge system, where K>=1 on all cases with K being the sequence of features (i.e.,) attributes of length 'l'. The relational sequence of features is denoted as,

$$\rho=\left(f_1,f_2,f_3,\ldots f_n\right) \tag{6}$$

where $f_1 \ldots f_n$ are the set of features on the multi-dimensional gene structure whereas the sub features are described as,

$$\rho^{'}=(f_i,f_{i+1},f_{i+2},\ldots\ldots f_{i+k-1}) \tag{7}$$

where f_i denotes the sub features in BAOFRS method. The discrete probability division on the 'K' knowledge system $|\rho| = d$, then the K-mer relation is identified as,

$$K\,mer\,relation\,(R)=\sum_{f=1}^{d}\alpha_k \tag{8}$$

where, $\alpha_k=(x_{k_1},x_{k_2},\ldots\ldots\ldots x_{k_{nk}})$ and $nk=d-k+1$.

The relation 'R' contains the relational features; α_k is the identified related features and represented in the set as $\{f_1,f_3,f_6\}$ and $\{f_2,f_4,f_5\}$. The probability of either '0' or '1' is used for relating the features. Discrete division of features in BAOFRS method is used for relating the gene data. The relational sequence on the gene data helps to easily mine the information from the multi-dimensional space. K-mer relational knowledge system is explained through algorithmic step.

Algorithm 3: K-mer Relational Algorithm
Input: Biologically associated genes
Output: Relationally sequenced gene
Method

Intialize f as 1 // gene feature

For K>=1

Compute relational sequence of features ρ for all features $(f_1, f_2, f_3, \ldots f_n)$

Compute relational sequence of features ρ for all sub features $(f_i, f_{i+1}, f_{i+2}, \ldots\ldots f_{i+k-1})$

K-mers Relation 'R' identified with discrete probability division scheme using ρ, ρ', d

$R = \sum_{f=1}^{d} \alpha_k$ //Sequence 'l' features relation identified for effective mining of gene data

End For

Return R

End

The above step evidently explains K-mer relation knowledge as follows:

Step 1: Assign the relative features '$f = 1$'

Step 2: With the attributes of length 'l', $K>=1$ on all cases with K being the sequence of features, find the relation of features and sub-features,

$$\rho = (f_1, f_2, f_3, \ldots f_n) \text{ and } \rho' = (f_i, f_{i+1}, f_{i+2}, \ldots\ldots f_{i+k-1})$$

Step 3: Using K – mers relation R discrete probability,

$$R = \sum_{f=1}^{d} \alpha_k$$

where $\alpha_k = \rho * \rho'$ the features are related.

Finally, K – mers system is used in BAOFRS method to relate the features for a sequence length 'l'. To mine gene data effectively, K-mer relation identifies the features relation. The relation R is identified with discrete probability division scheme. With the related features obtained, the BAOFRS method finds the similarity value. The similarity value is identified through Jaccard similarity coefficient which is briefly explained in section 5.4.2.

Jaccard Similarity Value Identification: Once the K-mer related features are identified using relational knowledge system, the second step in BAOFRS is to find the similarity value between the two gene data samples using Jaccard Similarity. Jaccard index is a name used for comparing similarity, dissimilarity and distance of the data set. The Jaccard Similarity measure the degree to which the common value occur on gene features. The unique gene feature composition in the proposed method is taken as the similarity value points on the feature set. The similarity value on features is computed as,

$$Jaccard\ Similarity\ S_{ij} = \frac{\sum_{f} x_{f_1} x_{f_2} x_{f_3} \ldots\ldots x_{f_n}}{\sum_{f} x^2_{f_1} + \sum_{f} x^2_{f_2} + \sum_{f} x^2_{f_3} + \ldots\ldots \sum_{f} x^2_{f_n} - \sum_{f} x_{f_1} x_{f_2} x_{f_3} \ldots\ldots x_{f_n}} \quad (5.4)$$

where, x denotes the sample gene data, f denotes a number of relative features, n is the total number of genes in a sample(n=14), x_{f_1} denotes a number of features that positive for 1st gene, x_{f_2} denotes a number of features that positive for 2nd gene, vice versa and $x_{f_1} x_{f_2} x_{f_3} \ldots\ldots x_{f_n}$ denotes a number of features positive for all 14 relevant genes.

Jaccard Similarity improves the score level on the sequence features 'f' to identify the similarity value on relative sequence feature on gene expression data.

Ant Optimized Bi-clustering: The final step in BAOFRS is to cluster the similar relational features of the gene data using the ant optimized bi-clustering operation. Ant optimized Bi-clustering is a type of clustering algorithm that imitates the behavior of ants. Inspired by the food-searching behavior of real ants, the ant optimized system algorithms are developed to be efficient. Ant-based bi-clustering is a biologically stimulated data clustering technique. This clustering task aims at the unsupervised classification of patterns in different groups. Ant colonies formulate some powerful nature-inspired Heuristics for solving such clustering problems. Ant-based bi-clustering algorithms are used in a wide variety of applications such as Gene expression data analysis, Knowledge discovery in DNA chip analysis data and web usage mining etc.,

Figure 7 illustrates the results of the clustering by ant optimized bi-clustering and it is represented by color. The ant optimized rule is used in BAOFRS method to group the relational similarity value in some type of ants. The ant optimized clusters based on features is given as,

$$AOC = \frac{n_1}{n_1 + f} + \frac{n_2}{n_2 + f} \quad (9)$$

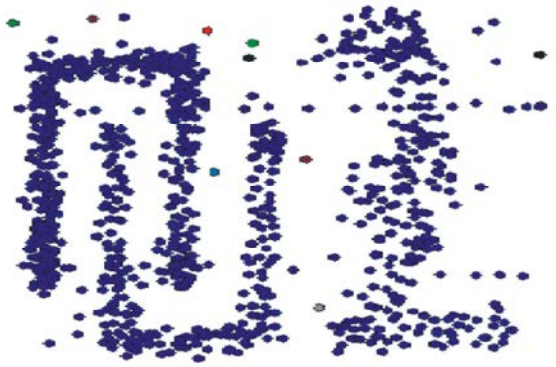


Fig. 7: Clustering Results by Ant Optimized Bi-clustering

where n_1, n_2 are the 'n' medoids of grouping on the relative features and f denotes the relative sequence features on gene.

The bi-clustering of features in BAOFRS method improves the efficiency rate on the mining of multi dimensional gene data objects from the larger dataset. Ant optimized bi-clustering of the gene data is represented through the Figure 7.

The proposed algorithm is developed in BAOFRS method to minimize the distance between the bi-clustering points; thereby the bi-clustering time is reduced on the gene dataset. Medoids is taken as the relative features used for clustering with the minimal processing time. The Proposed Ant Optimized Medoids algorithmic steps are described as, follows;

Algorithm 3: Ant Optimized Medoids algorithm

Input : Input of features '*f*', Samples *'S'*

Output : '*n*' clusters (normal and abnormal gene) with minimal processing time

Method

Arbitrary Choose *n* objects as initial Medoids

For
Initialize *i* as 0
Do
Chooses gene relative data features for clustering *'n_1'*
Repeat until *i*<= ρ // ρ set of features on the multi-dimensional gene structure
Increment by *1*
End For
For
Initialize j as 0
Do
Chooses gene relative data features for clustering *'n_2'*
Repeat until *j* <= ρ' // ρ' set of features on the multi-dimensional gene structure
Increment by *1*
End for

Compute Ant Optimized Clustering

$AOC = \frac{n_1}{n_1 + f} + \frac{n_2}{n_2 + f}$ // to group the relational similarity value in the type of ants

Choose random object

Swap and compute cost T

If T<0 then

Cl_a = The gene is Abnormal

Else If

Cl_n = Normal genes

End If

Return $n(CL_a, CL_n)$

End

The above algorithmic step explains the proposed Ant Optimized Medoids system as follows:

Step 1: With the input related features *f* from sample *S,* cluster all the related features from ρ and ρ' as n_1 and n_2

Step 2: Calculate the similarity value related to the relevant genes,

$$AOC = \frac{n_1}{n_1 + f} + \frac{n_2}{n_2 + f}$$

Step 3: To identify the normal and abnormal gene randomly choose n objects, then swap and compute the Cost T

Step 4: If *T<0* then

CL_a = gene is Abnormal

else

CL_n = gene is Normal

End

In BAOFRS method, Medoids demonstrate better performance on clustering of relative feature sequences of the gene data. Once the Medoids have been selected, each non-selected gene features is grouped with the medoids to which it is the most similar to attain the higher clustering efficiency rate in Bi-clustered Ant Optimized Feature Relational Sequencing. Bi-clustered Ant Optimized extracts the gene data (i.e., sequence-pairs) effectively while using the Medoids because the features are clustered on particular distance.

After obtaining the number of sequence-pairs and obtaining the estimates of the sequence-pairs, a simple pattern matching process is performed to evaluate the sequence-pairs into 'normal' or 'abnormal'. As large absolute sequence values have greater influence on the similarity score level, BAOFRS method simplifies the evaluation of gene patterns by considering it into two groups: genes above and below 'T', being observed as 'normal' or as 'abnormal' gene patterns. The genes above 'T' that is '1' gene patterns were considered as 'normal' genes and that below 'T' (ie) '0' gene patterns were recorded as 'abnormality'.

The separation of sequence-pairs into 'normal' and 'abnormal' categories facilitates the use of protein sequence pattern matching analysis for biological inferences on each sequence-pair. The protein sequence pattern matching analysis was performed on the biological processes of genes. The protein sequence pattern matching analysis tested whether the set of interesting genes was enriched with normal or abnormal type of gene when compared against all other genes on the microarray.

Experimental Result and Analysis: Bi-clustered Ant Optimized Feature Relational Sequencing (BAOFRS) method works on the JAVA platform. The performance analysis of the proposed method was analyzed with the help of cancer gene expression datasets. The proposed BAOFRS method was compared against the existing Heuristic algorithm with Black Hole (HBH) phenomenon in [4] and Simulation-based optimization (SO) method in [5]. The factors are experimented and the measures of the result percentage were analyzed with the help of table and graph values. Results are presented for different number of features considering the gene expression data. The experiment was conducted on the factors such as bi-clustering time, relational sequence result rate and similarity score level.

The proposed algorithms were administered on selected genes as described in chapter 3. The datasets used in this work briefly listed here again:

- Pancreas
- Uterus
- Ovary
- Prostate
- Colorectal

Bi-Clustering Time: Bi-clustering time (BCT) defines the time taken to cluster rows *R* and columns *C* of gene expression data *GeneExpData* it is measured in terms of milliseconds (ms) (ie.,) the bi-clustering time is defined as,

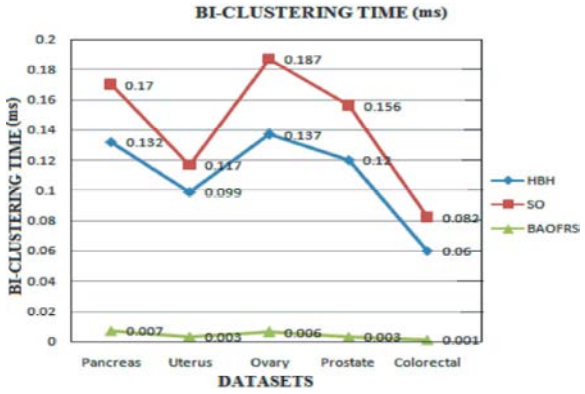


Fig. 8: Measure of Bi-clustering Time

Table 5: Comparison of Bi-Clustering Time for Different Approaches

		Bi-Clustering Time(ms)		
Datasets	No. of Samples	HBH phenomenon	SO method	BAOFRS method
Pancreas	63	0.132	0.17	0.007
Uterus	55	0.099	0.117	0.003
Ovary	56	0.137	0.187	0.006
Prostate	63	0.12	0.156	0.003
Colorectal	54	0.06	0.082	0.001

$$BCT = Time(GeneExpData_{R,C}) \quad (10)$$

Table 5 describes the bi-clustering time for our proposed method using the Ant Optimized Medoids against standard HBH phenomenon and SO method.

The bi-clustering time was calculated using cancer gene expression datasets. Based on the above table, the graph is depicted in Figure 5.7.

Figure 8 illustrate the bi-clustering time based on the features in gene expression data. It is obvious from Figure 8 that, the proposed BAOFRS method performs relatively well when compared to two other methods HBH and SO. The bi-clustering time reduced by 17-59% when compared to HBH. Moreover, Ant Optimized Medoids used for K-mers relational knowledge mine significant features on gene expression dataset by minimizing the bi-clustering time by 30-81% when compared to SO. The results reported here confirm that with the increase in the number of features, as the time consumed to perform bi-clustering also increases. The process was repeated till 80 features for experimental purposes.

Relational Sequence Result Rate: The resultant relational sequence result rate *RSSR* (measured in terms of%) is obtained by the summation of relational sequence of features, sub features and the difference obtained from K-mers relation is given as follows:

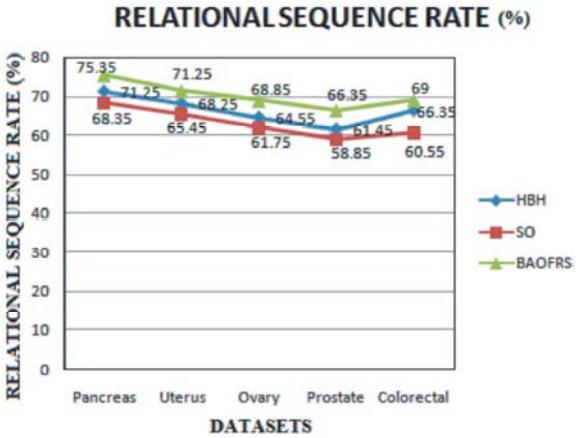


Fig. 9: Measure of Relational Sequence Result Rate

Table 6: Comparison of Relational Sequence Result Rate for Different Approaches

		Relational sequence rate (%)		
Datasets	No. of Samples	HBH phenomenon	SO method	BAOFRS method
Pancreas	63	71.25	68.35	75.35
Uterus	55	68.25	65.45	71.25
Ovary	56	64.55	61.75	68.85
Prostate	63	61.45	58.85	66.35
Colorectal	54	66.35	60.55	69

Table 7: Comparison of Pattern Quality Level for Different Approaches for Different Approaches

		Pattern Quality Level (Score Points)		
Datasets	No. of Sample	HBH phenomenon	SO method	BAOFRS method
Pancreas	63	75	78	82
Uterus	55	66	69	80
Ovary	56	75	78	86
Prostate	63	72	75	82
Colorectal	54	60	63	79

$$RSSR = \rho + \rho' - (R) \quad (5.7)$$

where,

ρ - Relational sequence of features

ρ' - Relational sequence of sub features

R- K-mer relation

Table 6 compares our proposed BAOFRS methods against HBH phenomenon and SO method in terms of the relational sequence rate using K-mers.

In order to maximize the relational sequence result rate, features and sub features are obtained and k-mer relational features identify the related features. In the experimental setup, the pattern size considered ranges from 100 to 385 KB. The results of 8 different patterns are listed in Table 5.2. The resultant relational sequence result rate using our method BAOFRS offer comparable values than the existing methods. The relational sequence result rate was computed using cancer gene expression datasets.

The targeting results of relational sequence result rate using BAOFRS method with two state-of-the-art methods in Figure 9 is presented for visual comparison based on the varied pattern sizes.

Our method differs from the HBH phenomenon and SO method in that we have incorporated the features to identify the relational sequences that also evaluate the similarity value between the sequences by improving the relational sequence result rate by 5-12% than when compared to HBH. In addition with the application of K-mers relational knowledge algorithm, the relational sequence result rate is improved by 9-12% when compared with the SO method.

Pattern Quality Level: Pattern Quality level is defined as the amount of accuracy in pattern mining based on the gene expressions. It is measured in terms of score points.

*Pattern Quality Level (%) = No. of related features * (Improved pattern level / Current Pattern Level)* (5.8)

Table 7 compares our proposed BAOFRS methods against HBH phenomenon and SO method in terms of the pattern quality level with respect to the number of features.

The pattern quality level of the proposed method was compared with the two existing methods and values are calculated using gene expression dataset. The Table 5.3 may be visualized as in Figure 10.

Figure 10 illustrates the pattern quality level using cancer gene expression datasets with respect to the number of features. As illustrated in the Figure 5.9, the pattern quality level increases. This is because with the highest similarity score level obtained using the proposed BAOFRS method; threshold value was evaluated measuring the relationship between the feature vectors increasing the pattern quality level by 8-24% when compared with HBH. The pattern quality level observed in BAOFRS method is less than the other two existing methods namely HBH and SO. This is because by applying Jaccard Similarity, the score level is improved on the relative sequence features during similar value identification on gene expression data. As a result, the pattern quality level is improved by 8-24% when compared to HBH and 4-20% than SO method.

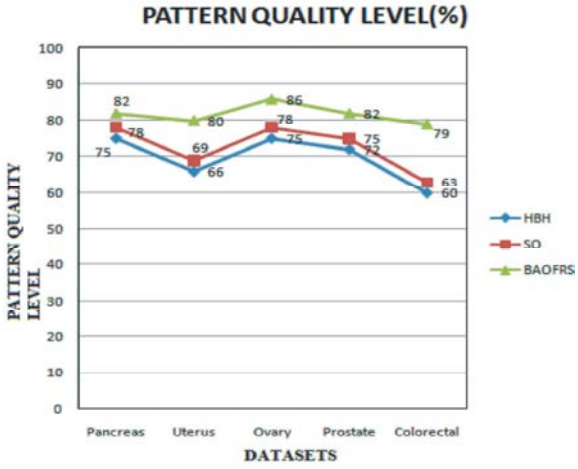


Fig. 10: Measure of Pattern Quality Level

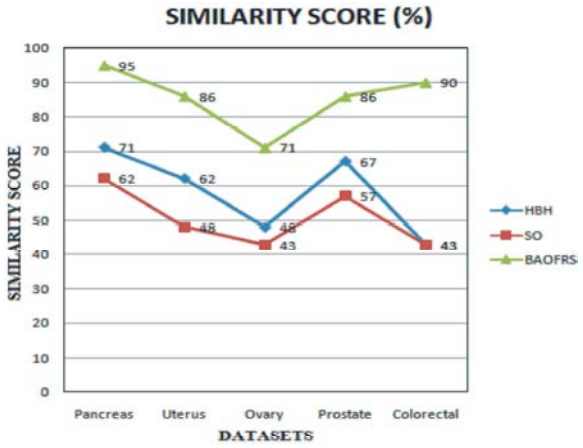


Fig. 11: Measure of Similarity Score Level

Table 8: Comparison of Similarity Score Level for Different Approaches

		Similarity Score Level (%)		
Datasets	No. of Samples	HBH phenol menon	SO method	BAOFRS method
Pancreas	63	71	62	95
Uterus	55	62	48	86
Ovary	56	48	43	71
Prostate	63	67	57	86
Colorectal	54	43	43	90

Similarity Score Level: Table 8 compares our proposed BAOFRS methods against HBH phenomenon and SO method in terms of similarity score level of BAOFRS method with respect to number of features using the following equation:

*Similarity Score (%) = (Similar Samples / No. of features) * 100* (5.9)

Similarity score level was measured in terms of percentage (%) and it was compared with the existing methods. Based on the table 5.4, the below graph is depicted. Using cancer gene expression datasets, the similarity score level is calculated.

Figure 11 shows the similarity score level using for BAOFRS method, HBH phenomenon and SO method versus increasing number of features for n = 14 for proposed method and n=16,063.

The similarity score level for HBH and SO shows low percentage because of the deficiency in extracting the relevant gene. BAOFRS similarity score level increases, the reason is that the similarity score level for BAOFRS method is evaluated using Jaccard similarity coefficient where the Jaccard similarity coefficient efficiently identifies the similarity on gene expressional data by applying ant optimized bi-clustering by improving the similarity score level by 25-53% when compared to HBH and 35-53% when compared to SO method.

CONCLUSION

In this work, a novel method of identifying the biological changes on physiological data using Heuristic search algorithm in rough set theory for gene expression data analysis is introduced. The proposed method is based on the Heuristic search algorithm. This algorithm identifies the biological changes and processes based on two phases, one is initialization phase and another is iterative adjustment phase. Based on these two phases, the biological changes of each gene are identified in terms of physiological data on gene expression datasets. The experimental results showed that the proposed BPPD method can identify differentially expressed genes among different classes in gene-expression datasets. To identify differentially expressed genes, Heuristic search algorithm is used. Then the performance of the proposed BPPD is estimated in terms of response time and Heuristic search threshold. The proposed Heuristic search performs better and the performance rate is 70-80% high in the proposed BPPD for analyzing the biological process and identifying relevant genes of physiological data compared to an existing bi-clustering algorithm. In the next chapter, we will discuss the method of identifying the biological association among selected genes in this process.

Bi-clustered Ant Optimized Feature Relational Sequencing (BAOFRS) method was proposed in this work to identify the relationship sequence between features of the gene expression data. The BAOFRS method employed the K-mers relational knowledge System to efficiently

identify the relational features. Jaccard similarity coefficient was applied to identify the similarity value on relational features. With the similar relational features obtained, Ant optimized Bi-clustering was performed using Ant optimized Medoids algorithm for effective clustering of relative feature sequences of the gene data. Finally, the gene patterns were verified as normal or abnormal on the basis of simple pattern matching process. The abnormal genes in every dataset imply that those genes were affected by disease. Experimental evaluation was done effectively with the cancer gene expression dataset to estimate the performance of the proposed BAOFRS method. Performance evaluation revealed that the proposed BAOFRS method reduce the bi-clustering time, improves the relational sequence result rate and increase the similarity score level compared with the existing HBH and SO methods.

REFERENCES

1. Hussain Sadiq and G.C. Hazarika, 2011. Improved Biclustering Algorithm For Gene Expression Data, Journal of Theoretical and Applied Information Technology, 32(1): 1-7.
2. Lee, J., J. Han, Li Xiaolei and H. Cheng, 2011. Mining Discriminative Patterns for Classifying Trajectories on Road Networks, IEEE Transactions On Knowledge And Data Engineering, 23(5): 713-726.
3. Zhiwen, Y., L. Le, Y. Jane, W. Hau-San and H. Guoqiang, 2012. SC (3): Triple spectral clustering-based consensus clustering framework for class discovery from cancer gene expression profile, IEEE Transactions on Computational Biology and Bioinformatics, 9(6): 1751-65.
4. Hatamlou, A., 2013. Black hole: A new Heuristic optimization approach for data clustering, Information Sciences, Elsevier Journal, 222: 175-84.
5. Nguyen, A., S. Reiter and P. Rigo, 2014. A review on simulation-based optimization methods applied to building performance analysis, Applied Energy, Elsevier Journal, 113: 1043-1058.
6. Chen, B., 2011. Inferring gene regulatory networks from multiple time course gene expression datasets, IEEE International Conference on Systems Biology, 8(3): 12-17.
7. Jaegyoon Yoon, Y. and S. Park, 2011. Noise-robust algorithm for identifying functionally associated biclusters from gene expression data, Information Sciences, Elsevier, 181: 435-449.
8. Bhattacharya, A. and R.K. De, 2009. Bi-correlation clustering algorithm for determining a set of co-regulated genes. Bioinformatics, 25(21): 2795-2801.
1. Huang, D., S. Paul and P. Gong, 2010. Gene expression analysis with an integrated CMOS microarray by time-resolved fluorescence detection, Biosensors and Bioelectronics, 26(1): 2660-2665.
2. Simon, R., 2009. Analysis of DNA microarray expression data,Best Practice & Research Clinical Haematology, 22: 271-282.
3. Mohamad, M.S., 2011. A Modified Binary Particle Swarm Optimization for Selecting the Small Subset of Informative Genes from Gene Expression Data, IEEE Transactions on Information Technology in Biomedicine, 15(6): 813-822.
4. Mohammadi, A., M.H. Saraee and M. Salehi, 2011. Identification of disease-causing genes using microarray data mining and gene Ontology, BMC Medical Genomics, 4(12): 1-9.
5. Ayadi, W., M. Elloumi and J.K. Hao, 2009. A biclustering algorithm based on a Bicluster Enumeration Tree: application to DNA microarray data, BioData Mining, pp: 2-9.
6. Bo-Lin Chen, 2011. Inferring gene regulatory networks from multiple time course gene expression datasets, IEEE International Conference on Systems Biology (ISB), pp: 12-17.
7. https://en.wikibooks.org/wiki/An_Introduction_to_Molecular_Biology/Genexpression
8. Iwen, M.A., W. Lang and J. Patel, 2008. Scalable Rule-Based Gene Expression Data Classification Extended Version - BST Classification, IEEE International Conference on Data Engineering (ICDE), pp: 1-12.
9. http://www.broadinstitute.org/ cgi-bin/ cancer/datasets.cgi
10. Bui Lam, T., Omar Soliman and Hussein A. Abbass, 2010. A Modified Strategy for the Constriction Factor in Particle Swarm Optimization, Volume 4828 of the series Lecture Notes in Computer Science, spinger, pp: 333-344.
11. Cheng, Y. and G.M. Church, 2000. Bi-clustering of expression data, The 8th International Conference on Intelligent Systems for Molecular Biology, La Jolla, CA, pp; 93-103.
12. Panteris E1, S. Swift, A. Payne and X. Liu, 2007. Mining pathway signatures from microarray data and relevant biological knowledge, Journal of Biomedical Information, 40(6): 698-706.

13. Bombini Grazia, Nicola Di Mauro, Stefano Ferilli and Floriana Esposito, 2010. Classifying Agent Behaviour through Relational Sequential Patterns, Volume 6070 of the series Lecture Notes in Computer Science, pp: 273-282.
14. Saber, H. ben and M. elloumi, 2015. An enumerative biclustering algorithm for DNA microarray data, pp: 132-138.
15. Madeira Sara, C. and Arlindo L. Oliveira, 2004. Biclustering Algorithms for Biological Data Analysis: A Survey, INESC-ID TEC. REP. 1/2004.
16. Marck, J.A., 2009. Short-cut method of solution of geodesic equations for Schwarzchild black hole, Classical and Quantum Gravity, 13: 393-402. A