Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLASSIFICATION OF GENES FOR DISEASE IDENTIFICATION USING DATA MINING TECHNIQUES

[1] P M BOOMA, [2] DR.S.PRABHAKARAN

[1] Research Scholar, Department of Computer and Engineering, SRM University

[2] Professor, Department of Computer Science and Engineering, SRM University

E-mail: [1] boomak@gmail.com [2] prabakaran.mani@gmail.com

ABSTRACT

Scientists are nowadays providing great awareness about microarray gene expression dataset. Researches tell that data mining fails to recognize the most important biological associations between genes. Recently biological information mining using clustering techniques were used for the analytical evaluation of gene expression. There are many challenges exist in the existing methods. Optimization Algorithm for multi dimensional search space does not provide relational optimization result on varying gene expressional problems. The existing method solves the clustering problem, but bi-cluster based gene expression information was not extracted. A key point on existing work was to handle multi modal structure optimization problems with effective searching process, but it does not offer relational sequence optimized result on the associated gene data.

To overcome the above issues, the proposed research is developed. The objective of the proposed research is to extract the biological information and identify the relational sequences on gene expression to identify abnormal genes. The proposed techniques like biological process on physiological data, PCPHC and Bi-clustered Ant Optimized Feature Relational Sequencing extract the biological process information from gene expression datasets. These techniques are tested on various bench marked datasets called Cancer Gene Expression datasets Broad Institute repository for experimental evaluation of the proposed method with an existing method, which identifies and extracts the hidden information from datasets. Finally, the gene patterns were verified as normal or abnormal on the basis of simple pattern matching process.

Keywords: *Gene Expression Datasets, Heuristic Search, Hierarchical Clustering Model, Feature Relational Sequencing*

1. INTRODUCTION

In modern era, many new diseases evolve. The fast and efficient identification of such diseases are essential to save human life in time. Many diseases are caused by gene mutation. Hence, attacking the diseases the root level is the necessity of the time. Computer science and engineering serves a lot in the research related to disease and treatment by providing efficient tools and techniques in this research.

Data mining is one such field of Computer science and engineering in research. The research in clinical diagnosis and prognosis requires efficient and fast classification techniques, which in turn requires large amount of gene data generation and analyzing these large amounts of data. The large amount of gene data generation is obtained using microarray technique in which expression of thousands of genes are concurrently measured and we are in need of an efficient data mining technique for these large amounts of data.

Mining micro-array gene expression data is an imperative subject in bioinformatics in diagnosis of disease, drug development, genetic functional interpretation and gene metamorphisms etc. Recently biological information mining using clustering techniques were used for the analytical evaluation of gene expression.

A bi-clustering algorithm discussed in [1] analyses clustering process for gene expression dataset. In this paper, the authors developed an improved bicluster score that eliminated the bias and uphold the discovery the most significant bicluster in the dataset. They utilized this score within a novel biclustering approach based on the bottom up search strategy. They believed that the bottom-up search approach could better model the essential functional modules of the gene expression dataset. A set of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

heuristic algorithms based mainly on node removal to discover one bicluster or a set of biclusters is adopted in proposed works. Jaegyoon et al. [7] proposed a Noise-robust algorithm for identifying functionally associated biclusters from gene expression data.

Biclustering has been emerged as a powerful tool for identification of a group of co-expressed genes present in a gene expression dataset. The work in [8] described Bi-correlation clustering algorithm for determining a set of co-regulated genes. The Bi-correlation clustering algorithm produced a diverse set of biclusters of co-regulated genes over a subset of samples where all the genes in a bicluster had a similar change of expression pattern. So, biclusters with certain patterns are more interesting from a biological point of view. A bottom up algorithm detected one biclusters at a time. Similarly, the work on Correlation–based scatter search presented in [7] formed biclusters from gene expression data.

Considering the impact of this algorithm, it was quite promising enrichment method with regard to mean square residue. An initial bicluster was created or accepted as input and then it had extended by adding rows and columns. A set of biclusters were created with different initializations. Only a few passes over the data matrix were required to find a bicluster. The method was able to detect some very small biclusters as it added rows incrementally. As if there was no bicluster with 2 genes, it was detect bicluster with 3 genes. The limitation of the algorithm was that, even if it was generated bicluster, all biclusters were not found to be interesting.

DNA micro array knowledge procedures give the gene expression level of thousands of genes beneath numerous experimental. Simon [10] proposed an analysis of DNA microarray expression data as a power tools for studying biological mechanisms and for developing prognostic and predictive classifiers for identifying the patients who require treatment and the best candidates for specific treatments. The examination of data generated by micro-array technology in [6] was very practical to appreciate how the genetic information turn out to be practical gene products. Such examination through biclustering algorithms can establish a collection of genes which were processed beneath a set of tentative conditions.

Leo et al. [2] proposed a method for mining discriminative patterns for classifying trajectories on road networks. By analyzing the behavior of trajectories on road networks, they found that, in addition to the locations where vehicles had visited, the order of these visited locations was essential for improving classification accuracy. Based on analysis, this method challenged that (frequent) sequential patterns were good feature candidates since they preserve the order information.

Discovering, and classifying cancer types correctly was essential for performing the successful diagnosis and treatment of cancer. Certain Gene expression analysis with an integrated CMOS microarray as in Ta-chien D, Huang.,et al., 2009 [9] presented time-resolved fluorescence detection in successful diagnosis. Similarly, Miles A et al. [11] presented OpenFlyData, an exemplar data web, integrating gene expression data of the fruit fly Drosophila melanogaster. Combining heterogeneous data across distributed sources is an important requirement for in silico bioinformatics supporting translational research. One of the main disadvantages in class discovery from cancer data sets was that cancer gene expression profiles not only included a large number of genes, but also had contained a lot of noisy genes. To diminish the effect of noisy genes in cancer gene expression profiles. Zhiwen Y et al. [3] proposed a two new consensus clustering frameworks, namely triple spectral clustering-based consensus clustering (SC3) and double spectral clustering-based consensus clustering (SC2 N cut), for cancer discovery from gene expression profiles.

Even though the Mining Discriminative patterns for classifying trajectories on road networks [2] improved the accuracy of the classification but this method was a not efficient and effective method for pattern-based classification. SC3 [3], presented spectral clustering for performing clustering on the gene and cancer sample dimension and finally partitioned the consensus matrix from multiple clustering solutions. But, the flaw was that this method was suitable for only cancer gene expression profiles.

Black Hole Phenomenon (BHP) [4] structured a novel type of heuristic algorithm that, during iteration, the candidate of best in nature is evaluated to be the black hole. Followed by this, it started extracting other candidates nearer to it

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and was referred to as stars and was able to solve the clustering problem. But, bi-cluster based gene expression information was not extracted. In recent decades, applying computer simulation for addressing complex systems has grown as a promising method that includes building thermal simulation programs of dynamic in nature, analyzing behaviors of energy for target oriented applications, analyzing gene expression data and so on.

An overview on simulation-based optimization methods in the building sector, aimed at clarifying recent advances and outlining potential challenges and problem in building design optimzation, is provided in [5]. Simulation-based optimization (SO) [5] method though provided multi-objective optimization into real world design, does not offer relational sequence optimized result on the associated gene data.

2. DATASETS USED

In order to evaluate the proposed method, following datasets taken from the repository [40] are used with 314 tumor and 90 normal cancer tissues for 16,063 genes.

Pancreas

In this dataset, expression levels of 63 tumor and 20 normal Pancreas tissues are measured using microarray technology. A selection of 4500 genes with highest intensity has been achieved by using proposed BPPD method. The data is pre-processed with log ratio for each sample with zero mean and one standard deviation.

Ovary

The Ovary dataset contains expression levels of 2021 genes for 56 tumor and 15 normal ovary cancer tissues. The dataset is available in [40]. In the original dataset, after normalization each sample has zero mean and one standard deviation.

Uterus

In Uterus cancer dataset, expression levels of 55 tumor and 15 normal tissues for 1500 genes are measured. After normalizing the genes with logarithmic ratio each sample has zero mean and one standard deviation.

Colorectal

The colorectal dataset we used is available in [40], with expression levels of 54 tumor and 18 normal samples. A selection of highest intensity genes of number 3454 has been obtained from proposed method.

Prostate

This dataset contains expressions of 88 tumor and 22 normal prostate cancer tissues for 4588 genes. It achieves zero mean and one standard deviation value after normalization.

3. THE METHODS

3.1 Analyzing Biological Process of Physiological Data using Heuristic Approach

The gene expression consists of collection of genes present in the datasets. Each gene consists of two types of pattern ie, physical pattern and logical pattern. The physical pattern provides information about physical structure of the gene on the gene expression datasets i.e, color, 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.

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. For identifying the biological changes on physiological data of gene expression datasets, a heuristic search is used to identify the quality solutions to emphasize the systematic process of analyzing the Biological Processes on Physiological Data (BPPD).

After identifying the physiological data on gene expression datasets, the heuristic search algorithm is used for identifying genes undergoing 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

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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

i) Partition of samples *S*

ii) Set of genes *G*

iii) Quality of the state Ω computed based on partition

An adjustment of the state would be

i) If gene $g \in G$, , insert *g* into *G*

ii) If gene $g \in G$, remove *g* from *G*

iii) For a sample s in *S*, move s to *S*' where *S* is not equal to *S'*

To identify the process of an adjustment to a state, compute the quality gain of the adjustment as per the alteration of the quality, i.e., $\Delta\Omega = \Omega' - \Omega$, where Ω and Ω' are the quality of the states before and after the adjustment, concurrently.

Heuristic Search algorithm proposed has two phases namely initialization phase and iterative adjusting phase. 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 and gene selection is performed.

The heuristic algorithm is inclined to the class of genes and form instruction measured in all iteration. 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.

The proposed Heuristic Search Algorithm is as follows

Algorithm 1 Heuristic Search Algorithm

Initialization phase

Read the Gene expression datasets and adopt a random initialization and calculate the Expression value

Iterative adjusting phase

for (i=0 ; i<*g* ; i++) // *g* for gene

Identify the physical and logical entity

Register a sequence of genes and samples arbitrarily

for (i=0 ; i<*g*/*S*; i++) // each gene or sample along the sequence,

if(*e*==*g*) // *e* is entity

Calculate $\Delta\Omega$ for the possible insert/remove;

else if(*e* == *S*)

Calculate $\Delta\Omega$ for the common reputation increase progression;

end if

if $\Delta\Omega$ >=0,

achieve the adjustment;

else if $\Delta\Omega$ < 0,

$$p = \exp\left(\frac{\Delta\Omega}{\Omega \times T_i}\right)$$

end if

//go to initialization process until biological process evaluation can be observed

Output the best state of identifying the biological process

After examining the physiological data, the biological processes of those data is identified through the heuristic algorithm. The biological processes occur only if the physiological data of a gene have met with some changes in their nature. In that case, the biological changes occur and those changes are identified by noting down the set of genes physiological data before changes has been made with those physiological data and performance of gene selection process in samples as shown in Figure 1 which could be done efficiently using Heuristic search algorithm.

Name of the Gene
Metallothionein isoform 2
KIAA0618 gene product
EST: zv16g03.s1 Soares NhHMPu S1 Homo sapiens cDNA clone 753844 3', mRNA sequence. (from Genbank)
RPS3 Ribosomal protein S3
Putative ubiquitin C-terminal hydrolase (UHX1) mRNA
EIF-2-associated p67 homolog mRNA
EST: zx10a05.s1 Soares total fetus Nb2HF8 9w

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Homo sapiens cDNA clone 786032 3' similar to contains Alu repetitive element;, mRNA sequence. (from Genbank)
Human kpni repeat mrna (cdna clone pcd-kpni-4), 3' end
Lysozyme gene (EC 3.2.1.17)
Lamin-Like Protein (Gb:M24732)
VIL2 Villin 2 (ezrin)
EST: zv26h12.r1 Soares NhHMPu S1 Homo sapiens cDNA clone 754823 5' similar to contains Alu repetitive element;, mRNA sequence. (from Genbank)
SM22-ALPHA HOMOLOG
UBA52 Ubiquitin A-52 residue ribosomal protein fusion product 1

Figure 1. List of selected genes for Pancreas cancer using proposed framework

3.1.1Performance Analysis

3.1.1.1Response Time

Heuristic search takes minimum response time to analyze gene expression differences. Response time is defined as time period required to analyze biological process of physiological data on gene expression dataset. Cancer Gene Expression datasets from Broad institute repository is taken in order to perform experimental evaluation of the proposed BPPD method in comparison with the existing biclustering algorithm in [1].

Table 1 shows the time taken to response the biological process identification procedures based on the size of data present in the gene expression datasets. The response time is measured in terms of seconds (secs).

Figure 2 depicts the response time in performing the search process at given interval of with respect to size of data.

Table 1. Tabulation For Response Time

Datasets	Size of the data (KB)	Response time (sec)	
		Existing biclustering algorithm (M1)	Proposed BPPD (M2)
Pancreas	71	50	30
Uterus	43	58	45
Ovary	46	64	57
Prostate	112	75	62
Colorectal	18	80	69

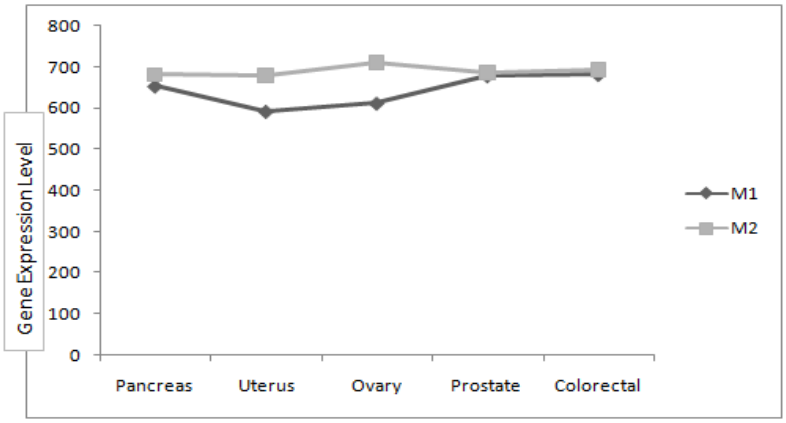

Figure 2. Measure Of Response Time

From Figure 2, we may observe that, response time in searching physiological data on gene expression dataset is reduced for the proposed BPPD(M2) method, as the physical and logical patterns of the genes are identified at first step using heuristic search. The proposed BPPD in analyzing the biological process on physiological data in the gene expression datasets using Heuristic search consumes 20-30% less response time when compared to an existing bi-clustering algorithm(M1) [1]. Bi-clustering algorithm [1] takes more time even for clustering process resulting in higher response time. Thus, BPPD

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

justified better performance in analyzing the physiological data among gene expression datasets using heuristic search algorithm in minimum time interval.

3.2 An Improved Pearsons Correlation Proximity Based Hierarchical Clustering Model for Mining Biological Association

During the initial stage, gene expression dataset was taken as input. With the given input gene expression dataset, heuristic search was applied on both physical and logical pattern in order to analyze the biological process. With the application of heuristic search algorithm, the proposed method identified differentially expressed genes between different classes in gene-expression datasets. The process is continued with the identification of biological association between genes. For this purpose, an improved Pearson's Correlation Proximity-based Hierarchical Clustering (PCPHC) model was developed and applied on these genes.

The improved PCPHC modal examines higher rate of expression levels between genes, for measuring biological association between genes simultaneously. PCPHC model captures the biological fact that follow a hierarchical clustering model for identifying the biological analysis between genes. Moreover, a global PCPHC (GL-PCPHC) model performs pattern growing method to mine GL-PCPHC patterns and discover significant biological associations between genes. Additionally, the Seed Augment algorithm design average linkage methods on rows and columns to expand a seed PCPHC model into a maximal GL-PCPHC pattern and to identify association between the clusters.

The first part in PCPHC is to apply hierarchical clustering model to the input gene dataset with the similarity between two genes, expression pattern being measured using the improved Pearson's Correlation Proximity. The second part in PCPHC is to address the pattern growing method by applying Seed Augment algorithm which uses average linkage method on rows and columns to obtain maximal GL-PCPHC model with genes consisting of similar expression patterns is constructed and the biological association between genes are measured.

Gene expression data consists of information from a gene is utilized in the separation of an efficient gene product. The PCPHC model uses microarray initially to estimate the expression level of genes. The identification of biological process for the physiological data present in the gene expression datasets is performed using Heuristic search. A heuristic search algorithm provides a collection of genes as the candidate's subjective genes and a division of samples as candidates of gene expression datasets.

3.2.1 Identification of Similar Gene Expression Pattern

Due to the experimental complexity involved in the expression matrix, gene expression data comprises of a huge amount of noise. Improved PCPHC modal significantly extracts relevant information by removing certain level of noise from gene expression data.

Hierarchical clustering is used to combine similar objects into clusters. Each row and/or column is considered as cluster in the beginning. In hierarchical clustering, the two most similar clusters are grouped and it continues clustering process until all objects are in the same cluster. Gene based hierarchical clustering for gene expression dataset produces a hierarchical series of clusters which is illustrated by a tree, called dendrogram. The leaves of dendrogram for gene based hierarchical clustering, not only generate the formation of the clusters but also record the similarity between the clusters.

By removing the dendrogram at certain level, a specific number of clusters for gene dataset are obtained. The results of the hierarchical clustering are plotted as a dendrogram that signifies the clusters and relation between the clusters. Genes are combined together to form clusters and clusters are combined together to form a higher level cluster by an inter-cluster distance. Hierarchical clustering is an unsupervised clustering method. Based on how the hierarchical dendrogram is formed, the Hierarchical clustering algorithms is further divided into two approaches namely: agglomerative approaches and divisive approaches.

3.2.1.1 Agglomerative algorithms

Agglomerative algorithm is a bottom-up approach. Initially, agglomerative algorithm

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

observes each data object as an individual cluster. At each step, combine the closest pair of clusters until all the groups are combined into one cluster. Single link, complete link, and minimum-variance are the various measures of cluster proximity that is used to obtain different combine strategies.

3.2.1.2 Divisive algorithms

Divisive algorithm is a top-down approach and this approach starts with one cluster containing all the data objects. They iteratively divide clusters until each cluster contains only one data object or definite stop condition is met. For divisive approaches, the essential problem is to decide how to split clusters at each step.

In improved PCPHC model, each log ratio factor of the gene expression matrix is colored on the basis of the ratio of fluorescence measure whereas the rows of the gene expression matrix are reordered on the basis of the hierarchical dendrogram structure with the help of a constant node-ordering. Once the cluster is being formed, the original gene expression matrix is transformed into a colored table where higher patches of color denote the genes sharing similar expression patterns.

Given an expression matrix $EM = (R, C)$ with R representing rows and C denoting the columns, the methodology followed to mine PCPHC patterns is carried out as follows. The gene expression datasets is represented as vectors as given below

$$V_i = \{v_{i,j} \mid where\ j\ lies\ between\ 1\ and\ f\} \quad (4.1)$$

Where $v_{i,j}$ represents the value of j[th] feature f for the i[th] gene value and f denotes the features.

The proximity level between the two gene values v_i and v_j is obtained with the corresponding vectors v'_i and v'_j using the improved Pearson's coefficient for a measurable dimension (dim) as given:

$$\sum = 1\ (v_{i\,dim} - \mu_{vi})\ (v_{j\,dim} - \mu_{vj}) / \sqrt{v_{i\,dim} - \mu_{vi}}\ \sqrt{v_{j\,dim} - \mu_{vj}} \quad (4.2)$$

Where μ_{vi} and μ_{vj} are the average values of two vectors v_i and v_j respectively and $\dim$ *where* $\dim i = 1,2,..,n$.

The hierarchical clustering algorithmic steps denote the way to measure similar gene expression pattern using the proximity Pearson's Correlation. The proximity level between two gene value measures using improved Pearson's correlation is viewed as a dynamic value with *n* observations. Moreover, measures the similarity between two genes by evaluating the exponential relationship between the distributions of the two dynamic gene variables.

Hierarchical clustering algorithmic steps to measure the similar gene expression pattern is as follows.

Algorithm 2 Hierarchical clustering algorithm

Begin
Input gene expression dataset, threshold δ
Let the gene expression data be represented as vectors
$V_i = \{v_{i,j} \mid$ where *j* lies between *1* and *f*$\}$
Let the Expression Matrix be *EM* = *(R, C)* and $w_{i,j}$ be the log ratio factor
Let v_i and v_j be two gene values and vector,v_i and v_j
Let μ_{vi} and μ_{vj} be average values for two vectors v_i and v_j,
Measure the proximity level using
$\sum = 1\ (v_{i\,dim} - \mu_{vi})\ (v_{j\,dim} - \mu_{vj}) / \sqrt{v_{i\,dim} - \mu_{vi}}\ \sqrt{v_{j\,dim} - \mu_{vj}}$
for $R = \{r_1, ..., r_n\}$
Form the expression patterns for rows
for $C = \{c_1, ...c_n\}$
Form the expression profiles of samples
Obtain log ratio factor $W_{i,j}$
end for
end for
If $(W_{i,j} < \delta)$
Similarity gene expression patterns are obtained
else if $(W_{i,j} > \delta)$
Similarity gene expression patterns are not obtained
end if

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2.2 Biological Association Between Genes

The second part of PCPHC is to measure the biological association between genes using the Seed Augment algorithm. Seed Augment algorithm employs average linkage method on rows and columns to expand seed PCPHC model into GL-PCPHC model and to identify association between the clusters. This, in result, helps in the efficient measure of biological association between the genes. The distance between the two gene expression data is defined as the mean distance between all genes of one group with all the genes of another group.

$$Dist(Gen1, Gen2) = 1 / (N_{Gen1} * N_{Gen2}) (\sum \sum dist(v_i, v_j)) \quad (4.3)$$

Where, $v_i \in Gen1$ and $v_j \in Gen2, i = 1, 2, \ldots, N_{Gen1}$ and $j = 1, 2, \ldots, N_{Gen2}$. The seed augment algorithm is defined below

Algorithm 3: Seed Augment Algorithm

Begin
Input: Similarity gene expression patterns obtained from algorithm (1), GL-PCPHC (P, Q, C_P, e_{max} and h_{min})
//**Expression Matrix Column Expansion**
for $C_i \in C - P$
do
  for O_i in $O(P, Q)$
do
  $O_i = COLEXPANSION(e_{max}$ and $h_{min})$
  end for
end for
// Expression Matrix Row Expansion
for $Ri \in R - Q$
do
  for O_i in $O(P, Q)$
    $O_i = ROWEXPANSION(e_{max}$ and $h_{min})$
    end for
  end for
    Obtain distance between two genes from (4)
  end

Seed Augment algorithm expands a seed PCPHC by rows and columns which takes the best possible column to associate the current pattern. The row expansion procedure scans the remaining rows that are not incorporated in the current pattern and associate the pattern by those rows that support the best order. Procedures in different order lead to different pattern association strategies. In order to generate effective association, the column-centric strategy and the row-centric strategy is adopted in GL-PCPHC. The Seed Augment algorithm works with respect to the size of the expression matrix size. As the average linkage method associates with the PCPHC patterns, the thresholds values are set to e_{max} and h_{min} to be both *0* for fair comparison, which means that the seed augment algorithm also mines PCPHC patterns.

Empirical studies show that the application of the PCPHC model improves the quality of the associated patterns. The Seed Augment algorithm to produce GL-PCPHC model is also significantly more efficient than the state-of-the art PCPHC mining method that proves the robustness of the Seed Augment algorithm.

3.2.3 Performance Analysis

Once the underlying biological associations between genes for identification of neurodegenerative diseases have been established, biological associations between neurodegenerative diseases are examined. The availability of systematically generated biological fact with the application of improved Pearson's Correlation Proximity-based Hierarchical Clustering (PCPHC) model provides insight into the biological associations for the neurodegenerative disease associations and helps to identify presence of genes involved in disease pathogenesis. **3.2.3.1 Biological Association between Genes**

The measure of Biological association accuracy rate proves better performance of PCPHC model. Accuracy rate in analyzing biological association between genes is judged by comparing numerous capacities from the same or different sources in gene expression data, to attain improved percentage.

$$Accuracy\ Rate = \frac{No.of\ correctly\ associated\ biological\ information}{Total\ no.of\ gene\ levels} \quad (4.4)$$

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2.Tabulation Of Accuracy Rate In Measure Of Biological Association Between Genes

Datasets	No. of Samples	Biological Association Accuracy (%)		
		Existing MDP method (M1)	SC3 (M2)	PCPHC model (M3)
Pancreas	63	75	77	87
Uterus	55	78	80	87
Ovary	56	76	78	88
Prostate	63	78	80	87
Colorectal	54	79	80	88

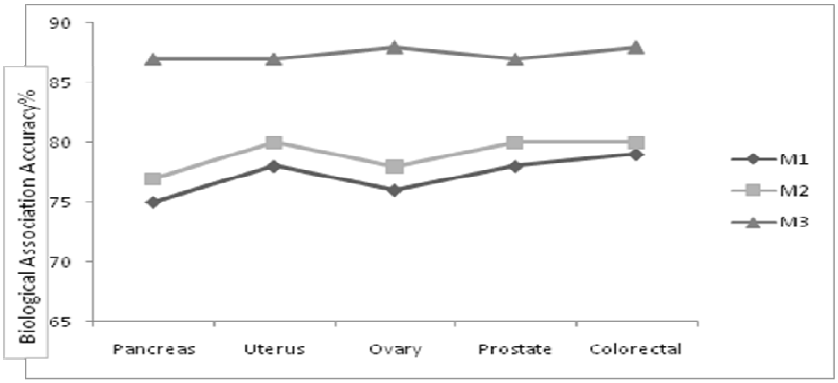

Figure 3. Accuracy Rate In Measuring Biological Association Between Genes

The improved Pearson's correlation proximity-based hierarchical clustering (PCPHC) model is compared with the existing Mining Discriminative Patterns for Classifying Trajectories (MDP) method in [2] and triple spectral clustering-based consensus clustering (SC3) in [3]. Using Genbank database, the biological association accuracy rate is calculated. The measure of accuracy rate in analyzing biological association between genes is tabulated in Table 2.

Figure 3 visualizes the accuracy rate of existing MDP method(M1) [2], model [3] measuring biological association between genes and is compared against the PCPHC(M3) model in terms of percentage (%).

Linear Prefix Tree (LPTree) structure organizes the candidate linear orders and improves the accuracy rate to 15 % when compared with the MDP method [2] and 8% better than the SC3 model(M2) [3]. The Accuracy rate using SC3 model is better than the MDP method because the SC3 model uses separately a consensus function to split the consensus matrix constructed from multiple clustering solutions resulting in increased accuracy rate. The Seed Augment algorithm rapidly updates candidate using the col expansion and row expansion function. Moreover, the Seed Augment algorithm applies average linkage method on rows and columns to expand seed PCPHC model into GL-PCPHC model and to identify association between the clusters. This in result helps in the efficient measure of biological association between the genes. Performing the expansion helps generate the candidate linear orders and count the supports effectively in PCPHC model

3.3 Bi-Clustered Ant Optimized Feature Relational Sequence Model for Improving Similarity Value

An effective pattern matching model called as Bi-clustered Ant Optimized Feature Relational Sequencing (BAOFRS) method was designed for gene expression data. The BAOFRS method was implemented to identify the relational sequences for minimizing the bi-clustering time by improving the similarity score level. 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.

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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 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 related features are obtained through K-mers relational knowledge algorithm, and then the Jaccard similarity coefficient is used in BAOFRS method to identify the similarity value between the features. Jaccard similarity coefficient helps in identifying the similarity value between larger gene dataset of different attribute sizes. Jaccard coefficient holds the relative sequence to identify the similarity value and provides an efficient optimization process.

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 groups the set of sequences of similar gene data features.

Multi-dimensional gene expression data is used for the BAOFRS experimental work. Gene expressions with the biological information consists of different products (i.e.,) features. Gene regulation serves as a substrate for evolutionary change in BAOFRS and relates all the extracted features to produce optimized result. 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 BAOFRS method to find the similarity value among the features. 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 groups similar (i.e.,) relative features using Ant optimized bi-clustering scheme. The ant optimized bi-clustering uses the K-medoids algorithm to improve the clustering efficiency for gene expression data.

3.3.1 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,

$$\rho = (f_1, f_2, f_3, f_4, \dots f_n) \quad (5.1)$$

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

$$\rho' = (f_i, f_{i+1}, f_{i+2}, \dots \dots f_{i+k-1}) \quad (5.2)$$

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

$$Kmer\ relation\ (R) = \sum_{f=1}^{d} \alpha_k \quad (5.3)$$

The relation 'R' contains the relational features. α_k is the identified related features and is 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 as follows.

Algorithm 4: K-mer relational algorithm

Intialize f=1
for (K>=1)
$$\rho = (f_1, f_2, f_3, f_4, \dots f_n)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

// for identifying relation

$$\rho' = (f_i, f_{i+1}, f_{i+2}, \dots \dots f_{i+k-1})$$

//used on all feature relation K-mers

Relation 'R' identified with discrete probability division scheme

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

// Sequence 'l' features relation identified for effective mining of gene data

end for

K-mer relation identifies the features relation for effective mining of gene data. The relation *R* is identified with discrete probability division scheme. With the obtained related features, the BAOFRS method finds the similarity value among them. The similarity value is identified through Jaccard similarity coefficient which is briefly explained in section 5.3.

3.3.2 Jaccard Similarity Value Identification

Once the K-mer related features are identified using relational knowledge system, the next 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 BAOFRS method is taken as the similarity value points on the feature set. The similarity value on features is computed as,

$$Jaccard\ Similarity\ (S) = \frac{\sum_{f=1}^{d} x_f * y_f}{\sum_{f=1}^{d} x_f^2 * \sum_{f=1}^{d} y_f^2 - \sum_{f=1}^{d} x_f * y_f} \quad (5.4)$$

The gene data sample denoted by '*x*' and '*y*' in the BAOFRS method is used for similarity value identification. Jaccard Similarity improves the score level on the relative sequence features '*f*' with '*d*' representing the relative knowledge system to identify the similarity value on relative sequence feature on gene expression data.

3.3.3 Bi-Clustering Ant Optimized

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 themselves to be efficient. Ant-based bi-clustering is a biologically stimulated data clustering technique. Clustering task aims at the unsupervised classification of patterns in different groups. Ant colonies formulate some powerful nature-inspired heuristics for solving the 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.

The ant optimized rule is used in BAOFRS method to group the relational similarity value in some type of ants. The bi-clustering of grouping behavior of ants (i.e.,) similar features designed the algorithm based on K-medoids method. The ant optimized clustering of features is given as,

$$AOC = \frac{K_1}{K_1 + f} + \frac{K_2}{K_2 + f} \quad (5.5)$$

K_1 and K_2 are the *'K'* 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. K-medoids is developed in BAOFRS method minimize the distance between the bi-clustering points, 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.

K-medoids demonstrate better performance on clustering of relative feature sequences of the gene data. Once the K-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 K-medoids because the features are clustered on particular distance.

The K-medoids algorithmic steps is described as follows,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Algorithm 5 : K-medoids algorithm

Begin
Input: Input of features '*f*', Knowledge system with '*d*' system
Output: '*K*' clusters with minimal processing time

for(*i*=0; $i <= (\rho = (f_1, f_2, f_3, f_4, \ldots f_n))$ **;** ***i*++)**

Chooses gene relative data features for clustering

Compute $AOC = \frac{K_1}{K_1+f} + \frac{K_2}{K_2+f}$

end for
//Repeat step 1 to 3 on each clustering work
end

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 has greater influence on the similarity score level, BAOFRS method simplifies the evaluation of gene patterns by considering all 1484 instances into two groups: genes with and without the top threshold '*T*', being observed as 'normal' or as 'abnormal' gene patterns. The genes with the top '*T*' gene patterns were considered as 'normal' genes and that below '*T*' gene patterns were recorded as 'abnormality'. The threshold value of '*T*' is evaluated by measuring the relationship between the feature vectors.

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 is performed on the biological processes of genes. The protein sequence pattern matching analysis test whether the set of interesting genes is enriched with normal or abnormal type of gene when compared against all other genes on the microarray.

3.3.4 Performance Analysis

One situation where the identification of shared patterns is of significant importance occurs when there are a set of functionally related sequences in neurodegenerative diseases. This can be tested by identifying the similarity value to obtain sequence pairs. The proposed work first searches the relational features in the diseased patient with neurodegenerative diseases as provided by the user and obtains the relational features using K-mers relational knowledge System.

3.3.4.1 Similarity Score Level

A measure of significance is performed using Jaccard similarity coefficient that identifies the similarity value on relational features in neurodegenerative diseases. The significance of pattern for neurodegenerative diseases is calculated to identify the similarity value between the two gene data samples using Jaccard similarity. The similarity score level measure is evaluated in table 3. The proposed BAOFRS method in similarity score level measure is compared against existing Heuristic algorithm with Black Hole (HBH) phenomenon in [4] and Simulation-based optimization (SO) method in [5]. A comparative study on similarity score level is given in Table 5.1.

Table 3. Tabulation Of Similarity Score Level

Datasets	No. of Samples	Similarity Score Level (%)		
		HBH phenol-menon (M1)	SO method (M2)	BAOFRS method (M3)
Pancreas	63	67	65	72
Uterus	55	69	68	78
Ovary	56	71	70	81
Prostate	63	73	72	83
Colorectal	54	76	74	84

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

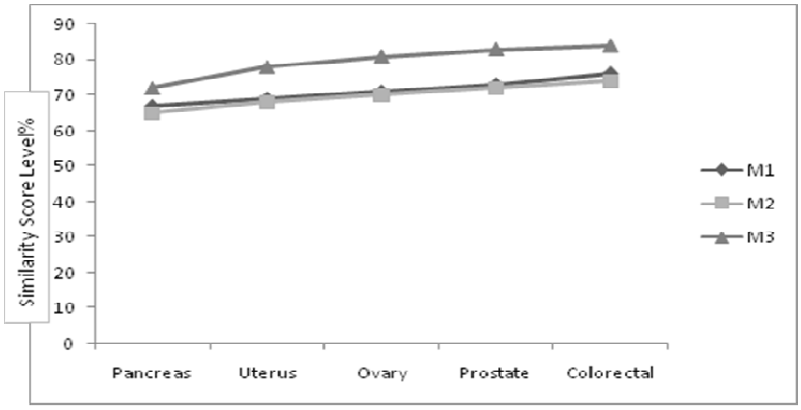

Figure 4. Measure Of Similarity Score Level

Table 3 illustrates the analysis of similarity score level of BAOFRS(M3) method with respect to number of features ranging between 10 and 80 that measures the similarity value between feature vectors is obtained using equation (5.5). Similarity score level is measured in terms of percentage (%) and it is compared with the existing methods in [4] and [5]. Using cancer gene expression datasets, the similarity score level is calculated.

Figure 4 visualizes the similarity score level for BAOFRS method, HBH phenomenon (M1) [4] and SO method (M2) [5] versus increasing number of features from n = 10 to N = 80.

The similarity score level improvement returned by gene expression data over HBH and SO increases gradually as the number of features gets increased. For example, for n = 20, the percentage improvement of BAOFRS compared to HBH is 11.53 % and compared to SO is 12.82 %, whereas for n= 30 the improvements are around 12.34% and 13.58 % compared to HBH and SO method respectively. 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 6 -12 % when compared to HBH and 9 – 13 % when compared to SO method.

4. OVERALL FRAMEWORK OF EFFICIENT DATA MINING TECHNIQUES FOR CLASSIFICATION OF GENES AND DISEASE IDENTIFICATION

The consolidated framework of heuristic search analysis on biological process, an improved Pearson's Correlation Proximity-based Hierarchical Clustering (PCPHC) model and Bi-clustered Ant Optimized Feature Relational Sequence (BAOFRS) method is described in this section. The process starts by providing a physiological data set that includes both physical and logical pattern to analyze the biological process using Heuristic search. With the application of heuristic search, the proposed method identifies biological association between genes.

Next, to discover the most significant biological associations between genes an improved Pearson's Correlation Proximity-based Hierarchical Clustering (PCPHC) model is designed. A global LOCS (Gl-LOCS) model adopts pattern growing method to mine Gl-LOCS patterns. By applying Seed Augment algorithm, two different growing strategies are applied on rows and columns in order to expand a seed LOCS model into a maximal Gl-LOCS pattern that efficiently identifies biological association on clusters.

Finally, relationship sequence between features of the gene expression data are analyzed to identify the relations on different features using K-mers relational knowledge algorithm. Next, Jaccard similarity coefficient is used to obtain the similarity value between larger gene dataset of different attribute sizes. Similar relational features are grouped using K-medoids and Ant optimization to reduce the biclustering time. The overall diagram starting from heuristic search analysis on biological process, followed by PCPHC model and finally performing BAOFRS method is depicted in Figure 5.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

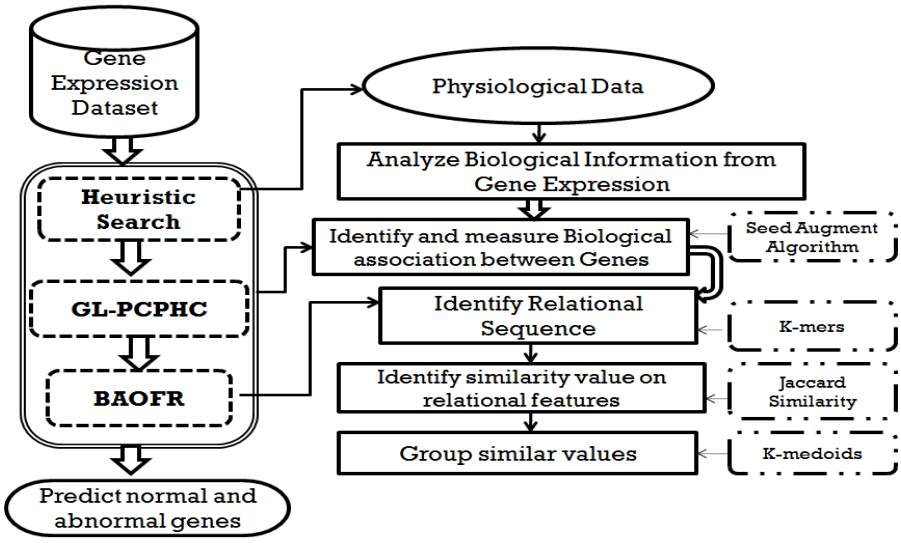

Figure 5 Overall Framework Of Efficient Data Mining Techniques For Classification Of Genes And Disease Identification

4.1 OVERALL PERFORMANCE ANALYSIS

Among 412 samples, 314 are tumor and 98 normal tissues were processed and 296 tumor and 94 normal tissues pass the quality control criteria and used for analysis. Table 4 and Figure 6 show the accuracy rate comparison with 6 methods namely Biclustering Algorithm, Existing MDP method, SC3, HBH phenomenon, SO and proposed BAOFRS. The proposed method reached accuracy of 94.66% in identifying the tumor and normal tissue.

Finally, similar relational features of the gene data are clustered using k-medoids and Ant optimization to effectively evaluate the gene into patient affected with Cancer disease normal or patient not affected with Cancer disease.

Table 4. Comparison Of Classification Accuracy Results Of Cancer Dataset (%)

Dataset	No. of samples			Classification Accuracy based on various methods(%)											
	Training Set	Test Set	PD	Training Set						Test Set					
				M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6
Prostate	46	20	2	38.5	45.5	65.1	55.4	77	94.2	38.5	44.8	54.9	55.4	71.2	90
Ovary	41	15	0	44.3	49.1	68.3	64.1	76.5	95	33.4	51.1	58.4	63.1	76.5	92.7
Uterus	41	14	0	58.7	58.7	67.5	68.3	75.5	88.4	55.9	58.7	63.3	68.3	72.9	89.1
Colorectal	39	14	1	61	62.9	66.2	68.2	72.1	91.1	60.5	62.9	66.2	64.2	71.1	91.4
Pancreas	44	17	2	59.9	59.9	69	68.6	71.8	85.9	57.1	61.2	65.6	69.9	76.8	94.9

Journal of Theoretical and Applied Information Technology
31st January 2016. Vol.83. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

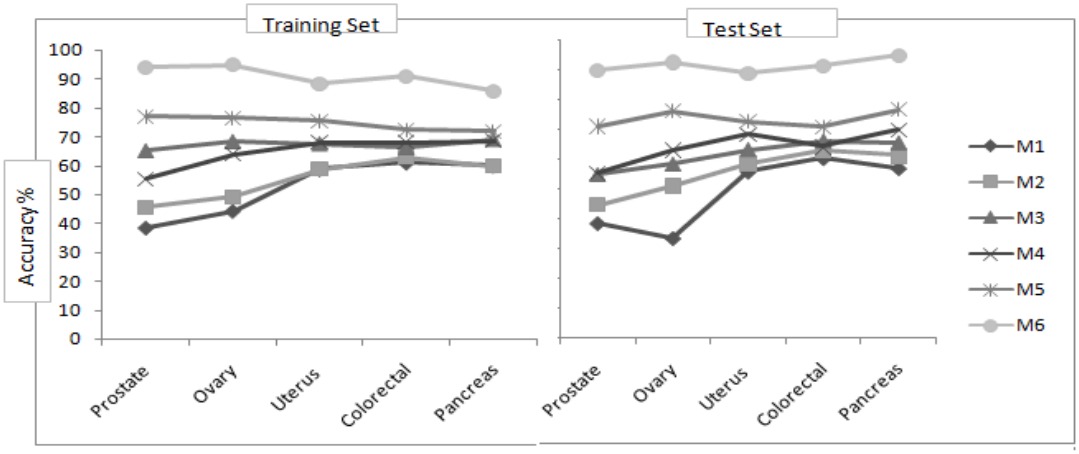

Figure 4.2 Accuracy Rate Of Cancer Dataset

M1: Biclustering Algorithm
M2: Existing MDP method
M3: SC3
M4: HBH phenomenon
M5: SO
M6: BAOFRS

.

5. CONCLUSION

Identifying normal or abnormal genes is important for clinical analysis and diagnosis. In this work, a novel framework for analyzing gene data was designed and developed. For this, initially, Bio-information from gene expression data was evaluated with the establishment of analyzing biological process using heuristic search [BPPD]. BPPD method identified the biological process on physiological data using heuristic search algorithm in rough set theory for gene-expression data analysis. This method extracted the biological process on gene expression data. The proposed method used heuristic search algorithm for identifying the biological process and processed based on two phases. The first phase was initialization phase and another was iterative adjustment phase. With respect to these two phases, the biological process of each gene and gene selection for a dataset is identified in terms of physiological data on gene expression datasets. Experimental evaluations are conducted for heuristic search based analysis of biological process on physiological data with standard benchmark gene expression data sets from research repositories such as broad institute in terms of size of gene expression datasets. Experimental results of the proposed BPPD method was conducted on Cancer gene expression dataset. Compared to an existing bi-clustering algorithm, heuristic search algorithm provides better performance rate in analyzing the biological process of physiological data in 20-30% reduced time.

Proximity measure of improved Pearson's Correlation (PCPHC) based on hierarchical clustering model was proposed to monitor the higher rate of expression level between genes selected in the previous stages. The clustering model measured the biological association between genes simultaneously. It is an efficient model which exhaustively mine incorporating the similarity diversion strategy. A global PCPHC (GL-PCPHC) model adopts pattern growing method to mine GL-PCPHC patterns and discover significant biological associations between genes. PCPHC model allowed linear orders and Seed Augment algorithm adopts two different growing strategies on rows and columns in order to expand a seed PCPHC model into a maximal GL-PCPHC pattern. Simulations results of this model attained improved gene expression data with 10.085% effective biological association based on feature set.

Finally, issues in presence of extract the bi-cluster based gene expression information was addressed with proposed Bi-clustered Ant Optimized Feature Relational Sequencing (BAOFRS) method. The features used to identify the relational sequences also computed the similarity value between the sequences BAOFRS method used the K-mers relational knowledge sequence to 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 K-medoids algorithm for effective clustering of relative feature sequences

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of the gene data. Finally, the gene patterns were verified as normal or abnormal on the basis of simple pattern matching process. BAOFRS method minimizes the bi-clustering time on performing the relational sequence bi-clustering. The usage of BAOFRS method increases the similarity score level by 13% when compared to the existing methods.

REFERENCES

[1] Sadiq Hussain, Hazarika.G.C, "Improved Biclustering Algorithm For Gene Expression Data", Journal of Theoretical and Applied Information Technology, 32(1),pp. 1-7, 2011.

[2] Lee J, Han J, Xiaolei Li and Cheng H, "Mining Discriminative Patterns for Classifying Trajectories on RoadNetworks," IEEE Transactions On Knowledge And Data Engineering, 23(5), pp. 713-726, 2011.

[3] Zhiwen Y, Le L, Jane Y, Hau-San W, Guoqiang H,"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),pp. 1751-65, 2012.

[4] Hatamlou A., "Black hole: A new heuristic optimization approach for data clustering," Information Sciences, Elsevier journal., 222, pp. 175-84, 2013.

[5] Nguyen A., Reiter S., Rigo.P, "A review on simulation-based optimization methods applied to building performance analysis," Applied Energy, Elsevier journal.,113, pp. 1043-1058, 2014.

[6] Chen B et. al., "Inferring gene regulatory networks from multiple time course gene expression datasets", IEEE International Conference on Systems Biology, 8(3), pp. 12-17, 2011.

[7] Jaegyoon, Yoon Y, Park S," Noise-robust algorithm for identifying functionally associated biclusters from gene expression data", Information Sciences, Elsevier, 181, pp. 435-449, 2011.

[8] Bhattacharya A, De RK ' Bi-correlation clustering algorithm for determining a set of co-regulated genes. Bioinformatics, 25(21), pp. 2795-2801, 2009.

[9] Huang.D, Paul.S, Gong.P, et.al., "Gene expression analysis with an integrated CMOS microarray by time-resolved fluorescence detection," Biosensors and Bioelectronics , 26(1), pp. 2660-2665, 2010.

[10] Simon.R, "Analysis of DNA microarray expression data," Best Practice & Research Clinical Haematology, 22, pp. 271–282, 2009.

[11] Miles A, Zhao J, Klyne G.,et.al.," OpenFlyData: An exemplar data web integrating gene expression data on the fruit fly Drosophila melanogaster", Journal of Biomedical Informatics ,43, pp. 752–761, 2010.

[12] http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi