CHAPTER 3

ANALYZING BIOLOGICAL PROCESS

3.1 INTRODUCTION

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.

3.2 BI-CLUSTERING ALGORITHM

DNA microarray data analysis is clustered both rows and columns concurrently of a data matrix to identify group of rows consistent with group of columns. This is known as Bi-clustering [60]. The algorithms based on these were broadly used in DNA microarray data analysis, text mining, exchange analysis, collaborative filtering, market research, information retrieval, electoral trends, and so on. Bi-clustering algorithms are the Co-clustering or two-mode clustering data mining technique where the Genes regulated by multiple processes concurrently. This type of algorithms was used to find a subset of genes that exhibit similar expression entities under a subset of circumstances in gene expression data analysis. Bi-clustering analysis is suitable for identifying and characterizing the biological factors that disturbing the patients with the equivalent gene subsets [61].

3.2.1 Bi-Cluster Structure

Different types of Bi-cluster structure are tested as follows. The way of structuring Bi-cluster is to find the single largest Bi-cluster and to delete rows and columns from the data, allowing exclusive row and column Bi-cluster was obtained. Certainly, the true challenge was to detect overlapping or at least non-exclusive Bi-cluster [62]. The non-exclusive Bi-cluster is a bi-cluster in which some rows or columns do not belong to any Bi-cluster at all and that the bi-clusters overlap in some places. The types of bi-clusters are as follows:

- Single Bi-cluster
- Exclusive row and column Bi-cluster
- Exclusive-rows or exclusive-columns Bi-cluster
- Non-overlapping non-exclusive Bi-cluster
- Arbitrarily positioned overlapping Bi-cluster
Figure 3.1 shows the various Bi-clusters structure.

A Single Bi-cluster shown in Figure 3.1(a) can either be only one Bi-cluster in the data matrix or K bi-clusters, where K is the number of bi-clusters which we expect to identify and is usually defined as apriori [62]. Even though the Bi-clustering algorithms can find more than one Bi-cluster, the target Bi-cluster is usually the one considered the best according to some criterion [61]. When the Bi-clustering algorithm assumes the existence of several bi-clusters in the data matrix, the following Bi-cluster structures shown in Figure 3.1(b) to Figure 3.1(f) can be obtained:

- Exclusive row and column bi-clusters (rectangular diagonal blocks after row and column reorder)
- Exclusive-rows bi-clusters
- Exclusive-columns bi-clusters
- Non-Overlapping non-exclusive bi-clusters
- Arbitrarily positioned overlapping bi-clusters

Forming a color image helps to identify several bi-clusters in a data matrix. These blocks of colored images are subsets of rows and subsets of columns with similar expression values; hence, it is a Bi-cluster. As shown in Figure 3.1(b), an image with some number K of rectangular blocks on the diagonal was produced by reordering the data matrix. Although this can be the first approach to extract relevant knowledge from gene expression data, it has long been recognized that such an ideal reordering, which would lead to such a Bi-cluster structure, will seldom exist in real data [61]. As shown in Figure 3.1(c), other Bi-clustering approaches assume that row belongs to one Bi-cluster and in the case of gene expression data; column belongs to several bi-clusters. When using opposite orientation of the data matrix exclusive-columns bi-clusters will be produced which means the columns of the data matrix can only belong to one Bi-cluster while the rows can belong to one or more bi-clusters which were shown in Figure 3.1(d). A non-exhaustive variation of the structure which was shown in Figure 3.1(e) was assumed by [62]. However, it is more likely that, in real data, some rows or columns do not belong to any Bi-cluster at all and that the bi-clusters overlap in some places. The structure, without any constrain, represents the data matrix as a sum of possibly overlapping bi-clusters as shown in Figure 3.1(f).

3.2.2 Types of Bi-Cluster

Different bi-clustering algorithms had different definitions of bi-cluster [61]. Below figure illustrates the examples of different types of bi-cluster following:

- Bi-cluster with constant values
- Bi-cluster with constant values on rows or columns
- Bi-cluster with coherent values
- Bi-cluster with coherent evolutions
Figure 3.2: Bi-cluster with constant values

Figure 3.2 shows a Bi-cluster with constant value structure. The Bi-clustering algorithm is used to find the constant values by rearranging the rows and columns of the matrix hence it can group collectively similar rows/columns and find Bi-clusters with similar values.

3.2.2.1 Bi-cluster with Constant Values

To find a constant Bi-cluster, reordering the rows and columns of the matrix is the common method for it, so similar rows/columns can be grouped together for finding the bi-clusters by means of similar values. When the data is in order this method is satisfactory. But if the data can be noisy, it can’t be a satisfied data. More sophisticated methods should be used. Variance is used to compute constant bi-clusters. So, a perfect bi-cluster is a matrix with variance 0. Also, it is used to avert the partition of the data matrix into bi-clusters with only one row and one column.

The variance [58] is used to evaluate the quality of each Bi-cluster \((i,j)\) is defined as follows:

\[
\text{var}(I, J) = \sum_{i \in I, j \in J} (a_{ij} - a_{IJ})^2
\]

(3.1)

Where \(a_{ij}\) and \(a_{IJ}\) represent the element in row and column.
From (3.1) a perfect Bi-cluster is a sub-matrix with variance equal to zero. Therefore, every single-row, single-column matrix in the data matrix, which equal to each element, is an ideal Bi-cluster since variance (I,J)=0.

<table>
<thead>
<tr>
<th></th>
<th>2.0</th>
<th>2.0</th>
<th>2.0</th>
<th>2.0</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

**Figure 3.3: Bi-cluster with constant values on rows**

### 3.2.2.2 Bi-clusters with Constant Values on Rows or Columns

Variance among the row and column are used to construct the bi-clusters. In order to complete the identification, first normalization of the columns and the rows has been done. There are some other algorithms, which can locate bi-clusters that have rows and columns with different approaches without using normalization step. Figure 3.3 shows the Bi-clusters with Constant Values on Rows and Figure 3.4 shows the Bi-clusters with Constant Values on Columns.
3.2.2.3 Bi-clusters with Coherent Values

All together, enhancement over the algorithms for the bi-clusters by means of constant values on rows or on columns is supposed to be considered for bi-clusters with coherent values on rows and columns. That means an advanced algorithm is required. This algorithm should have the analysis of variance among groups, with co-variance among both rows and columns. In the theorem presented by Cheng and Church’s, definition of a Bi-cluster is explained as a subset of rows and columns by means of the same score. The similarity score is used to compute the consistency of rows and columns.

The coherence of the additive and multiplicative model is mathematically expressed as follows’,

Perfect additive coherent values,

\[ a_{ij} = \mu + \alpha_i + \beta_i \]  \hspace{1cm} (3.2)

Perfect multiplicative model,

\[ a_{ij}^* = \mu^* \alpha_i^* \beta_i^* \]  \hspace{1cm} (3.3)

From (3.2) and (3.3) Where, \( \mu \) is the typical value within the Bi-cluster and \( \alpha_i \) is the adjustment for row \( i \) and \( \beta_i \) is the adjustment for column. Additive coherent
values, where each row or column is obtained by adding a constant to another row or column and multiplicative coherent values is obtained for each row or column is attained by multiplying other row or column by a constant value.

There were several Bi-clustering algorithms proposed for bioinformatics. Some of them are block clustering [62], Coupled Two-Way Clustering (CTWC) [62], Crossing Minimization [62] and Robust Bi-clustering Algorithm (RoBA) [62]. These algorithms were also known as Bi-Dimensional Clustering, Co-Clustering, Subspace Clustering or Two-Way Clustering in remaining application fields.

3.2.3 Clustering Vs Bi-Clustering

It is clear that Bi-clustering techniques generated local models, while clustering methods lead to global models [62]. A given gene cluster is defined using all the conditions, if a clustering algorithm on the rows of the gene expression matrix is used. Based on a subset of conditions, a Bi-clustering technique is assigned. Additionally, when a clustering algorithm is applied to the rows of the gene expression matrix, it assigns each gene to a single cluster. Alternatively, Bi-clustering techniques identify clusters that are not mutually exclusive or exhaustive. A gene is belonged to no cluster, one or more clusters.

Figure 3.5 illustrates the clustering and Bi-clustering of a gene expression matrix.
As shown in Figure 3.5, Clusters represent the disjoint strips in the matrix. A gene cluster contains all columns and a condition cluster contains all rows. Bi-clusters represent subjective subsets of rows and columns that shown here as rectangles. Since gene (condition) clusters were disjoints, the rows (columns) of the matrix has been rearranged so that each cluster is an adjacent strip. Similar rearranging of rows and columns shows all the bi-clusters as rectangles are generally not possible. This bias was addressed in [1].

In paper [1], proposed a bi-clustering algorithm for gene expression dataset. In this work, the author developed an improved Bi-cluster score that eliminated this bias and uphold the discovery the most significant bi-clusters in the dataset. The author utilized this score within a novel Bi-clustering approach based on the bottom up search strategy. The author believed that the bottom-up search approach better modeled the essential functional
modules of the gene expression dataset. A set of Heuristic algorithms based mainly on node removal to discover one Bi-cluster or a set of bi-clusters was presented in this work.

Choosing normal clustering approaches or Bi-clustering approaches depend on the objectives of analysis of gene expression. Common objectives followed when analyzing gene expression data include:

- Grouping of genes according to their expression under multiple conditions;
- Classification of a new gene, given its expression and the expression of other genes, with known classification;
- Grouping of conditions based on the expression of a number of genes; and
- Classification of a new sample, given the expression of the genes under that experimental condition.

For the purpose of getting genes combination with multiple conditions, the proposed method is developed. To enhance the gene biological process analysis, a Heuristic search algorithm for Bi-clustering was discussed in this work and presents a Heuristic search to identify the biological process on physiological data in gene expression datasets.

3.3 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.

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.

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,

\[ M = \{ w_{ij} \mid 1 \leq i \leq n, 1 \leq j \leq m \} \]  (3.4)

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

Generally, gene expression datasets attained from an examining process contains missing values, noise, and organized distinctions happening from the uncertain procedure.

3.4 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.

3.4.1 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.
3.4.2 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.

3.4.3 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.

3.4.4 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.

3.4.5 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.

3.5 ANALYZING BIOLOGICAL PROCESS ON GENE EXPRESSION DATASETS USING HEURISTIC SEARCH

The proposed work is designed for analyzing the biological process on physiological data in gene expression datasets using Heuristic search. Heuristic search is a method that finds the better solution in terms of a response time for searching the relevant gene from the sample. The proposed work operates under two different operations. The first operation is to analyze the Gene Expression datasets as shown in Figure 3.7. The second operation is to mine related biological data which is required for disease identification.
Figure 3.6: Architecture Diagram for Analyzing the Biological Process

Figure 3.6 illustrates the architecture diagram of the proposed analysis of biological process on gene expression datasets using Heuristic search. The first phase explains the process of Gene Expression datasets. The gene expression datasets consists of process by which information from a gene is employed in the separation of an efficient gene product. The second phase illustrates the Heuristic search based Analysis of Biological Process carried over with the physiological data present in the gene expression datasets.

3.5.1 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 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. The intelligence of the gene refers reactions of genes to various situations. The physical and logical patterns form a physiological data which provides all information about the genes.

Figure 3.7: Process of Heuristic based Analysis of Biological Process (BPPD)
Figure 3.7 shows the analysis of biological processes using Heuristic search algorithm 3.1. 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 related gene on physiological data of gene expression datasets.

After identifying the physiological data on gene expression datasets, the Heuristic search algorithm is used to extract the relevant gene. Heuristic Search algorithm is basically a guided search, which is widely used to find the optimal solution. Among the multiple physiological data present in the gene expression data set, the Heuristic Search algorithm is applied on it to search and extracts the related gene and it produces high quality solutions. So, this algorithm is more appropriate for finding the best and optimal genes for identifying the disease.

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 better quality of proposed biological process is obtained 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:

- Partition of samples $S$
- Set of genes $G$
- Number of the states $\Omega$ 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 $G$, if 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 3.1: 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 gene Expression Level

**Iterative adjusting phase**

Initialize an element \(i\) and \(n\) as 0

For each gene \(g_n\)

Do

Identify the physical and logical entity

Register a sequence of genes and samples arbitrarily

Repeat until \(i < n\)  // \(n\) is the number of gene

Increment \(i\) by 1

End For

For each gene \(g_n\) or sample \(S\) along the sequence

Do

Run *Heuristic Wrest Algorithm*

Repeat Until \(i < n/S\)  // \(S\) is the total number of samples

Increment \(i\) by 1

End For

Sort \(P\)

Select first 14 genes in sorted list \(P\)

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.

Step 2: In the initialization phase, an initial state is processed and calculates the gene expression level. 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 in random manner.

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 structure of the gene by using the size transformation in gene, i.e., $\Delta \Omega_p' = \Omega_p' - \Omega_p$, where $\Omega_p'$ and $\Omega_p$ are the different structure before and after the transform, concurrently.

The 14 relevant attributes are selected from gene to analyze the relevant Physiological data, which means the name of the genes that are relevant to our search in disease identification. The Heuristic Wrest Algorithm is applied to select the relevant genes and to ignore the irrelevant gene using structural difference and behavioral difference measurement.
The following algorithm explains the process of extracting relevant 14 genes for each dataset as follows:

**Algorithm 3.2: 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**

1. **Initialize** Physical P entity and Logical L entity as 0
2. **For** all Samples S
3.   **For** all genes g in Sample S
4.     **If** genes $\notin$ Physical Entity $P$, then
5.        Calculate the difference between 2 states,
6.        $\Delta \Omega_P = \Omega'_P - \Omega_P$ // difference between actual gene structure and updated structure of the gene
7.     **If** genes $\notin$ Logical Entity $L$, then
8.        Calculate the difference between 2 states,
9.        $\Delta \Omega_L = \Omega'_L - \Omega_L$ // difference between actual gene behavior and updated behavior of the gene
10. **End If**
11. **End If**
12. **If** $\Delta \Omega_P \geq 0 \text{ AND } \Delta \Omega_L \geq 0$, THEN
13.    Extract the gene
14.    $P = \exp\left(\frac{\Delta \Omega_L \times \Delta \Omega_P}{\Omega_L \Omega_P \times T_i}\right)$ // extracts the relevant gene
15. **Else**
16.    Ignore the gene
17. **End If**
18. **Return** $P$
19. **End For**
20. **End For**
The Heuristic Wrest Algorithm can be explained as follows:

Step 1: Consider gene and 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 $P$,
   
   $$(g \notin P),$$
   
   $\Delta \Omega_p = \Omega_p' - \Omega_p$
   
   Where $\Omega_p$ is an actual gene structure and $\Omega_p'$ is a updated structure of the gene.

   Step 2.2: Compute behavior difference $\Delta \Omega_L$ based on Logical Entity $L$,
   
   $$(g \notin L),$$
   
   $\Delta \Omega_L = \Omega_L' - \Omega_L$
   
   Where $\Omega_L$ is a actual gene behavior and $\Omega_L'$ is a updated behavior of the gene.

Step 3: If $(\Delta \Omega \geq 0 \&\& \Delta \Omega_L \geq 0)$, then extract the relevant genes and compute $P$, which diminish the number of gene extracted,

   $$P = \exp\left(\frac{\Delta \Omega_L \cdot \Delta \Omega_P}{\Omega_L \cdot \Omega_P \cdot T_i}\right)$$  \hspace{1cm} (3.5)

   From (3.5) $T_i$ is response time taken to extract the relevant gene, exp is the exponential function

   Step 3.1: Ignore the gene which is irrelevant

Step 4: Return $P$

To present each gene, all possible biological changes 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 behavioral changes in the person’s body. 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. Table 3.1 illustrates the selected genes list for pancreas cancer using Heuristic search Algorithm.

**Table 3.1: List of selected genes for Pancreas Cancer using proposed framework**

<table>
<thead>
<tr>
<th>Name of the Gene</th>
<th>EST Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metallothionein isoform 2</td>
<td>KIAA0618 gene product</td>
</tr>
<tr>
<td>KIAA0618 gene product</td>
<td>EST: zv16g03.s1 Soares NhHMPu S1 Homo sapiens cDNA clone 753844 3’, mRNA sequence. (from Genbank)</td>
</tr>
<tr>
<td>RPS3 Ribosomal protein S3</td>
<td>Putative ubiquitin C-terminal hydrolase (UHX1) mRNA</td>
</tr>
<tr>
<td>Putative ubiquitin C-terminal hydrolase (UHX1) mRNA</td>
<td>EIF-2-associated p67 homolog mRNA</td>
</tr>
<tr>
<td>EIF-2-associated p67 homolog mRNA</td>
<td>EST: zx10a05.s1 Soares total fetus Nb2HF8 9w Homo sapiens cDNA clone 786032 3’ similar to contains Alu repetitive element;, mRNA sequence. (from Genbank)</td>
</tr>
<tr>
<td>Human kpni repeat mrna (cdna clone pcd-kpni-4), 3’ end</td>
<td>Lysozyme gene (EC 3.2.1.17)</td>
</tr>
<tr>
<td>Lysozyme gene (EC 3.2.1.17)</td>
<td>Lamin-Like Protein (Gb:M24732)</td>
</tr>
<tr>
<td>Lamin-Like Protein (Gb:M24732)</td>
<td>VIL2 Villin 2 (ezrin)</td>
</tr>
<tr>
<td>VIL2 Villin 2 (ezrin)</td>
<td>EST: zv26h12.r1 Soares NhHMPu S1 Homo sapiens cDNA clone 754823 5’ similar to contains Alu repetitive element;, mRNA sequence. (from Genbank)</td>
</tr>
<tr>
<td>VIL2 Villin 2 (ezrin)</td>
<td>SM22-ALPHA HOMOLOG</td>
</tr>
<tr>
<td>SM22-ALPHA HOMOLOG</td>
<td>UBA52 Ubiquitin A-52 residue ribosomal protein fusion product 1</td>
</tr>
</tbody>
</table>
3.6 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 with the parameters such as gene expression level, heuristic search threshold and response time for searching the relevant gene.

3.6.1 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. For the experimental consideration, the gene expression dataset contains the more number of attributes. Despite, gene expression level is a metric widely used in the literature [64] to compare the quality of the genes. Gene expression level is defined as follows:

\[
Gene\ Expression\ Level(\%) = \left( \frac{\text{No. of genes extracted}}{\text{No. of Samples}} \right) \times 100
\]  

(3.6)

Table 3.2: Comparison of Gene Expression Level for Proposed BPPD Method and Existing Bi-Clustering Approaches

<table>
<thead>
<tr>
<th>Datasets</th>
<th>No. of Samples</th>
<th>Gene Expression Level (%)</th>
<th>Existing Bi-clustering algorithm</th>
<th>Proposed BPPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pancreas</td>
<td>63</td>
<td>16</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Uterus</td>
<td>55</td>
<td>11</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Ovary</td>
<td>56</td>
<td>21</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>63</td>
<td>03</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Colorectal</td>
<td>54</td>
<td>18</td>
<td>26</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2 provides the comparison of gene expression level for proposed BPPD method and existing bi-clustering approaches 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. While considering the Prostate dataset, the gene expression level is obtained very low i.e 03 % using existing Bi-clustering algorithm. This is because number of irrelevant genes is present in the dataset. However, the proposed BPPD method reduces the irrelevant genes present in the dataset and improves the gene expression level using Heuristic search algorithm.

![Gene Expression Level](image)

**Figure 3.8: Type of Datasets vs. Gene Expression Level**

Figure 3.8 depicts the measurement of gene expression level percentage based on the gene expression datasets. In the proposed BPPD, 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 better solution for the biological change issues. The biological changes occurred and it is identified by noting down the set of genes, obtained using Heuristic
Wrest Algorithm and Heuristics Search Algorithm. Compared to an existing bi-clustering algorithm, the proposed BPPD outperforms well and the variance is 40-50% high.

### 3.6.2 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.

\[
Heuristic\ Threshold\ Value(\%) = \left( \frac{\text{No. of genes searched}}{\text{Time}} \right) \times 100
\]  

(3.7)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>No. of Samples</th>
<th>Existing Bi-clustering algorithm</th>
<th>Proposed BPPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pancreas</td>
<td>63</td>
<td>27</td>
<td>93</td>
</tr>
<tr>
<td>Uterus</td>
<td>55</td>
<td>15</td>
<td>82</td>
</tr>
<tr>
<td>Ovary</td>
<td>56</td>
<td>18</td>
<td>65</td>
</tr>
<tr>
<td>Prostate</td>
<td>63</td>
<td>06</td>
<td>98</td>
</tr>
<tr>
<td>Colorectal</td>
<td>54</td>
<td>19</td>
<td>73</td>
</tr>
</tbody>
</table>

Table 3.3: Comparison of Heuristic Search Threshold Level for Proposed BPPD Method and Existing Bi-Clustering Approaches

Table 3.3 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. From the tabulated value, while considering the different number of samples with different data set, the Heuristic Threshold Value is increased but not linear. This is because
for different sample value, irrelevant genes are presented in the searching process. The number of genes differs for each dataset. In existing method, the whole genes present in a datasets were used. When all the genes in a datasets were selected, the time complexity arises. In order to overcome this issue, the propped BPPD method uses only 14 genes.

Figure 3.9: Type of Datasets vs. Heuristic Search Threshold

Figure 3.9 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 threshold is measured as the number of relevant genes is searched from the different samples per unit time. An existing bi-clustering algorithm clusters the genes alone without knowing its biological processes but the proposed scheme used Heuristic search algorithm to identify the biological process for each gene present in the datasets and it provides 70% high Heuristic Search Threshold value. If we use all genes in datasets the computational overhead for searching and time for search increases.
3.6.3 Response Time

The response time is defined as the difference between response end time and start time which is multiplied with the number of relevant gene searched from samples. Response time is measured in terms of seconds (sec).

\[
\text{Response Time (ms)} = \text{No. of relevant genes searched} \times (\text{Response End Time} - \text{Response Start Time})
\]  

Table 3.4: Comparison of Response Time Level for Proposed BPPD Method and Existing Bi-Clustering Approaches

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

Table 3.4 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 3.10: Type of Datasets vs. Response Time

Figure 3.10 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 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). The existing Bi-clustering algorithm consumes more time for clustering process. But in the proposed BPPD method analyzing the biological process on physiological data present in the gene expression datasets using Heuristic search which consumes less response time of 20-30% compared to existing Bi-clustering algorithm.

Finally, the proposed scheme uses 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 with minimum response time.

3.7 SUMMARY

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.