Journal of Theoretical and Applied Information Technology
30th November 2013. Vol. 57 No.3

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

ANALYZING BIOLOGICAL PROCESS ON GENE EXPRESSION DATASETS USING HEURISTIC SEARCH

[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

Mining micro-array gene expression data is an imperative subject matter in bioinformatics with extensive applications. Bio informative knowledge discovery from DNA microarrays become more essential in various disease diagnosis, drug development, genetic functional interpretation, gene metamorphisms etc., Recently biological information mining using clustering techniques were used for the analytical evaluation of gene expression. The expression of Nnumerous genes can be scrutinized concurrently using DNA micro-array technology. To develop the massive quantity of information enclosed in gene expression data, revision of existing work presented a biclustering algorithm, which presents local structures from gene expression data set. However, traditional single cluster model unable to mine precise information from large, and heterogeneous collection gene expression data. So the development of a new computational method is presented in this work to improve the analysis of gene expression data sets. In this work we first introduce the heuristic search for the standard biological process on physiological data of the gene expression. The physiological data consists of both physical and logical patterns of the gene expression datasets and the biological process of physical and logical pattern of gene expression datasets are analyzed through Heuristic search. Experimental evaluations are conducted for our heuristic search based analysis of biological process on physiological data with standard benchmark gene expression data sets from research repositories such as UCI in terms of size of gene expression datasets.

Keywords: *Gene Expression Datasets, Physiological Data, Biological Process, Heuristic Search*

1. INTRODUCTION

High-density DNA microarrays are one of the most powerful tools for functional genomic studies. They are used for measuring expression of thousands of genes concurrently. The researchers normally modified the clustering approaches to recognize collection of genes that share same expression profiles or those symbolizing numerous physiological conditions. In the same gene-expression test, one may recognize a group of hundred genes going to be better candidates for class discriminants. To recover the task of all inter-gene interactions one should investigate all probable subsets of the few hundred individual genes. This outcome would be in a computationally interactable set of the candidates that must be investigated by an algorithm.

Gene expression profiles surrender quantitative and semi-quantitative data on the absorption of protein articulated by the equivalent genes in a definite condition and time, particularly nRNA. One of the elevated throughput data applications is to understand or reverse engineer gene authoritarian networks (GRNs) between genes utilizing a choice of arithmetical approaches. But different expression profiles could not be sufficient for swearing the rising number of concluding algorithms. Simulation processes could be a one of several procedures utilized for the specified inference algorithm.

Compared with the conventional approach to genomic research, which has possessed on the local investigation and set of data on distinct genes, micro-array technologies have made it possible to supervise the appearance levels for tens of thousands of genes in analogous.

Progress in gene expression micro-array approaches above the last decade or so contain made it is probable to determine the appearance levels of thousands of genes over many investigational conditions. As datasets enlarge in size, nevertheless, it becomes progressively more improbable that genes will preserve correlation transversely under the full set of conditions building clustering problematic. A set of heuristic

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

algorithms based mainly on node removal to discover one bicluster or a set of biclusters is presented in [4].

DNA micro array knowledge procedures give the gene expression level of thousands of genes beneath numerous experimental conditions. The examination of data generated by micro-array technology [10] is 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 are processed beneath a set of tentative conditions. Similarly Scatter Search [1] is an evolutionary technique that is supported on the development of a small set of solutions which are selected consistent with quality and assortment measures. Scatter Search also uses an assessment based on linear correlations [2] between genes to appraise the eminence of biclusters.

Correlation Coefficient among two arbitrary variables [6] may be utilized for learning the linear dependency among two genes. This reality has provoked the exercise of measures supported on proposed correlations among genes [5]. In [7] the association coefficient is utilized for creating biclusters with a greedy algorithm. In [8] an account of novel algorithm supported on a tree structure for biclustering is accessible and it uses an assessment Function.

Gene expression data are symbolized by thousands of measured genes on only a little tissue samples and represented by association rule based classifiers [12]. This can direct more to potential over-fitting and dimensional curse or even to an absolute failure in examination of micro-array data. In [3], enlargement of a hybrid particle swarm optimization (PSO) and tabu search (HPSOTS) strategy for gene selection for tumor classification was discussed and the customized particle swarm optimization is used in[11]. The assimilation of tabu search (TS) as a local enhancement method facilitates the algorithm HPSOTS to overleap local optima and explain reasonable performance. To resolve the gene expression datasets problems, [9] planned a Crossing Minimization Biclustering Algorithm (CMBA) was discussed to compact with the specific issues.

To enhance the gene biological process analysis, in this work, a heuristic search algorithm was discussed and we present a Heuristic search to identify the biological process on physiological data in gene expression datasets which is describes briefly under section 3.

2. BACKGROUND

2.1 Microarray

The two major types of micro-array experiments are the *cDNA micro-array* and *oligo nucleotide arrays* (abbreviated *oligo chip*). Even though differences in the particulars of their experimentation protocols, both types of experiments engage three general vital procedures:

Chip manufacture: A micro-array is an undersized chip, against which tens of thousands of DNA molecules (*probes*) are detached in permanent grids. Each grid cell relates to a DNA sequence.

Target preparation, labeling and hybridization: Naturally, two mRNA illustrations (a test sample and a control sample) are repeal transcribed into cDNA (*targets*), tagged using either glowing dyes or radioactive iso-topics, and then hybridized with the explores on the surface of the chip.

The scanning process: Chips are examined to examine the indication intensity that is release from the tagged and hybridized objectives. Usually, both cDNA micro-array and oligo chip researches determine the expression level for each DNA progression by the ratio of gesture intensity among the test model and the manage sample, consequently, data sets resulting from both methods distribute the same genetic semantics.

2.2Gene 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 rRNA genes or tRNA genes, the product is a practical RNA. The progression of gene expression is utilized by all recognized multicellular organisms, prokaryotes and viruses to produce the macromolecular machinery for life.

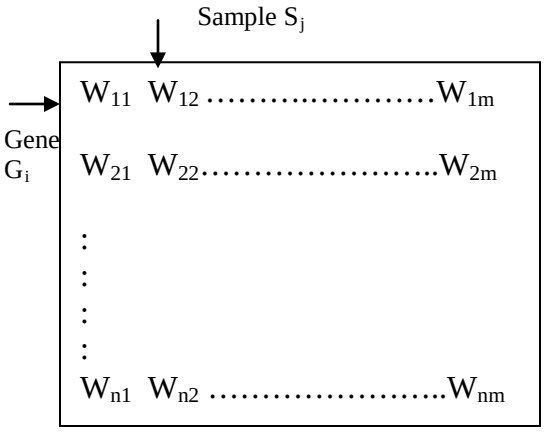
A micro-array research classically evaluates a huge amount of DNA sequences (genes, cDNA clones, or spoken sequence tags [ESTs]) under numerous conditions. These circumstances may be an instance series through a genetic process (e.g., the yeast cell cycle) or a compilation of diverse tissue samples (e.g., 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

Journal of Theoretical and Applied Information Technology
30th November 2013. Vol. 57 No.3

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

$$M = \{w_{ij} | 1 \le i \le n, i \le j \le\} \quad \text{..... Eqn 1}$$

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



The unique gene expression datasets attained from a examining process contains missing values, noise, and organized distinctions happening from the tentative procedure. Thus the gene expression datasets are normally formed with the given logical schema of the expression levels.

3. PROPOSED ANALYZING BIOLOGICAL PROCESS ON GENE EXPRESSION DATASETS USING HEURISTIC SEARCH

3.1 The Method

The present work is efficiently designed for analyzing the biological process on physiological data present in the gene expression datasets using Heuristic search. The proposed work [BPPD] operates under two different operations. The first operation is to analyze the Gene Expression datasets. The second operation is to extract the biological process on physiological data on Gene expression datasets. The architecture diagram of the proposed analysis of biological process on gene expression datasets using heuristic search is shown in fig 3.1.

The first phase describes the process of Gene Expression datasets. The gene expression datasets consists of process by which information from a gene is utilized in the separation of an efficient gene product.

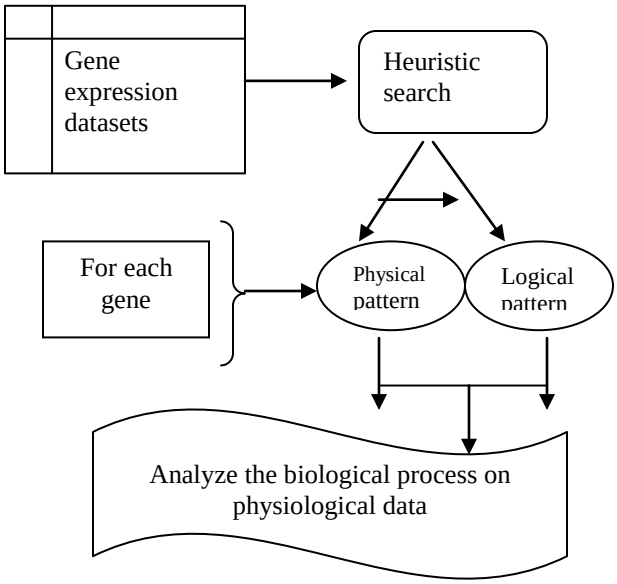


Fig 3.1 Architecture Diagram For Analyzing The Biological Process On Physiological Data (BPPD)

The second process describes the process of identifying the biological process carried over with the physiological data present in the gene expression datasets using Heuristic search.

From the above fig (Fig 3.1), the entire process of the proposed BPPD is briefly depicted. From the Gene expression datasets, the physiological data are analyzed and extracted and the biological process carrying over with those data are analyzed using Heuristic search process which identifies the best solution for Gene expression datasets. The table below describes the notation description used in the system.

3.2 Heuristic based analyzing biological process of physiological data

The gene expression consists of collection of genes present in the datasets. Each gene consists of two types of entitites.

One is physical entity and another one is logical entity. The physical entity provides an information about color, shape and structure of the gene based on its environment i.e., physical structure of the gene on the gene expression

Journal of Theoretical and Applied Information Technology
30th November 2013. Vol. 57 No.3

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

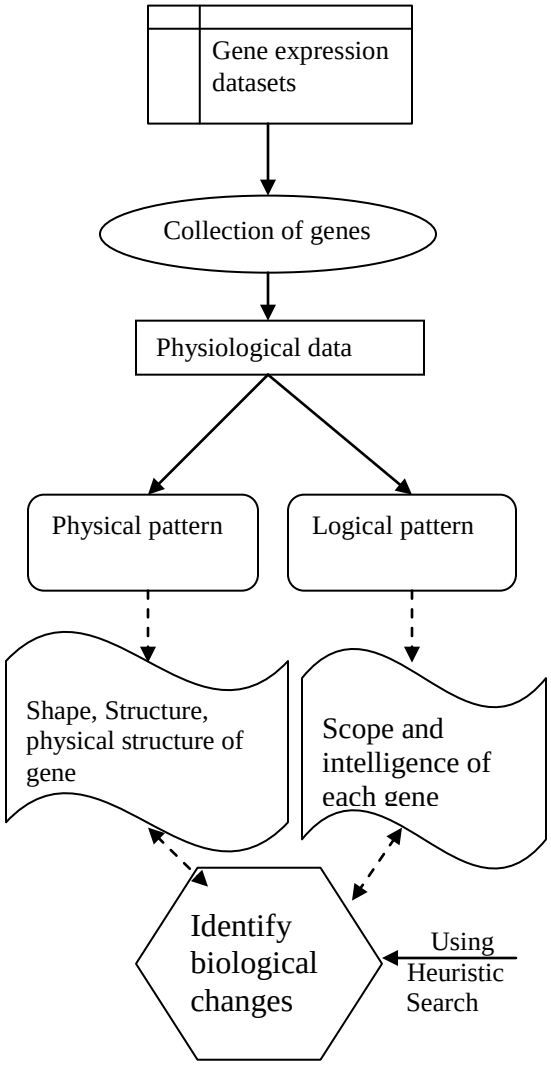


Fig 3.2Process Of Heuristic Based Analysis Of Biological Data

datasets. The logical entity 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 entities form a physiological data which provides all information about the genes.

Table 1 Parametric Description

Parameter	Description
n	Number of genes
m	Number of samples
W_{ij}	Each cell in gene expression
g_i	i^{th} Gene
s_j	j^{th} Sample

In this work we are going to present a technique to identify the biological changes on genes based on physical and logical entity. 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. The process of identifying the biological processes using this heuristic search algorithm is shown in fig 3.2.

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

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 $g \notin G$, 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.

The algorithm has two phases: 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.

Algorithm

Initialization phase

Step 1: **Read:** Gene expression datasets

Step 2: Adopt a random initialization and calculate the value

Iterative adjusting phase

Step 3: For each gene g

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

Step 4: Identify the physical and logical entity
Step 5: End Forw
Step 6: Register a sequence of genes and samples arbitrarily
Step 7: For each gene or sample along the sequence, do
Step 8: if the entity is a gene,
Step 9: Calculate $\Delta\Omega$ for the possible insert/remove;
Step 10: Else if the entity is a sample,
Step 11: Calculate $\Delta\Omega$ for the common reputation increase progression;
Step 12: if $\Delta\Omega$ >=0, then achieve the adjustment;
Step 13: Else if $\Delta\Omega$ < 0, then achieve the modification with probability

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

Step 14: go to 1), until biological process evaluation can be observed;
Step 15: Output the best state of identifying the biological process

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 proceed for identifying the biological changes, the physical and logical patterns are analyzed and noted. After examining the physiological data, then the biological processes of those data is identified thorough heuristic algorithm based a gene G and samples S. The biological processes occur only if the physiological data of 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 which could be done efficiently using Heuristic search algorithm.

4. EXPREIMENTAL EVALUATION

An experimental evaluation is conducted for the proposed analyzing the biological process on physiological data present in the gene expression datasets using Heuristic search to estimate its performance for Gene expression datasets. The Yeast Gene Expression datasets is derived from UCI repository for experimental evaluation of the proposed BPPD with an existing biclustering algorithm, which identify only the local structures from gene expression data set.

The yeast gene expression datasets consists of 8 attributes and 1484 instances with a classification associated tasks. The attributes used here for the evaluation of gene expression datasets are Sequence Name(Accession number for the SWISS-PROT database), mcg (McGeoch's method for signal sequence recognition), gvh (von Heijne's method for signal sequence recognition), alm (Score of the ALOM membrane spanning region prediction program), mit (Score of discriminate analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins), erl (Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen)), Binary attribute, pox (Peroxisomal targeting signal in the C-terminus), vac (Score of discriminate analysis of the amino acid content of vacuolar and extra-cellular proteins), nuc (Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins).

At first, the physiological data is first analyzed and the biological processes are also being observed and proceed. The heuristic search algorithm is used with the analysis of biological process on physiological data and the performance of the proposed BPPD is measured in terms of
i) Size of gene expression datasets,
ii) Heuristic search threshold,
iii) response time

5. RESULTS AND DISCUSSION

In this work, we have seen how the biological process of physiological data occurred on gene ecpression datasets using Heusristic search algorithm. The physical and logical pattern of each gene is first identified and then the biological processes of physiological data is identified using Heuristic search algorithm. An experimental evalaution is also being consucted to estimate the performance of the proposed BPPD with some metrics. The below table and graph describes the performance of the proposed BPPD using Heuristic search algorithm.

The below table (table 5.1) describes the process by which physiological data from a gene is used in the synthesis from the gene expression datasets. The outcome of the proposed analysis of the biological process on physiological data present in the gene expression datasets using Heuristic search is compared with an existing bi-clustering algorithm.

Journal of Theoretical and Applied Information Technology
30th November 2013. Vol. 57 No.3

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

Table 5.1 Size Of Data Vs. Gene Expression Level

No. of genes	Gene Expression level	
	Proposed BPPD	Existing biclustering algorithm
25	20	10
50	42	15
75	56	22
100	68	28
125	79	36

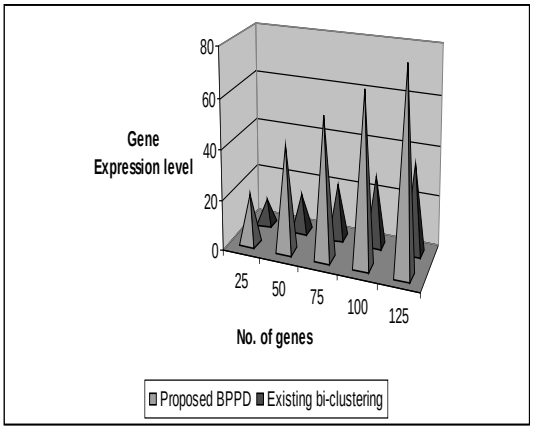


Fig 5.1 Size Of Data Vs. Gene Expression Level

Fig 5.1 describes the process of identifying the retrieval of physiological data from each gene present in 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 best solution for the biological change issues. Compared to an existing bi-clustering algorithm, the proposed BPPD outperforms well and the variance is 40-50% high.

Table 5.2 Size Of Data Vs. Heuristic Search Threshold

Size of data	Heuristic search threshold	
	Proposed BPPD	Existing biclustering algorithm
25	15	9
50	29	15
75	35	24
100	48	30
125	60	35

The above table (table 5.2) describes the process of heuristic search method based on the size of data present in 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.

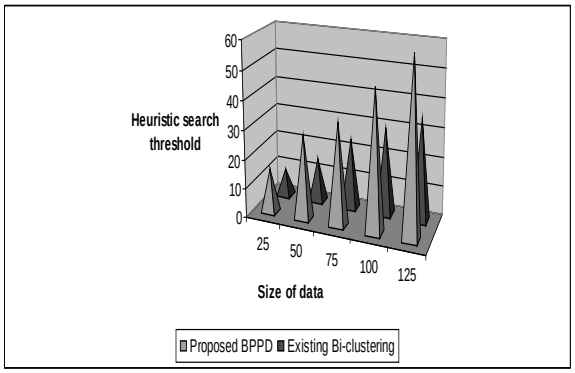


Fig 5.2 Size of data vs. Heuristic search threshold

Fig 5.2 describes 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 in terms of how far the best solution has been identified based on physiological data. Compared to an existing bi-clustering algorithm clusters the genes alone without knowing its biological processes, the proposed scheme used Heuristic search algorithm for identifying the biological process for each gene present in the datasets and it outperforms well and variance is 70% high in the proposed BPPD.

The table (table 5.3) describes the time taken to response the biological process identification procedures based on the size of data present in 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.

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

Table 5.3 Size Of Data Vs. Response Time

Size of data	Response time (secs)	
	Proposed BPPD	Existing biclustering algorithm
25	30	50
50	45	58
75	57	64
100	62	75
125	69	80

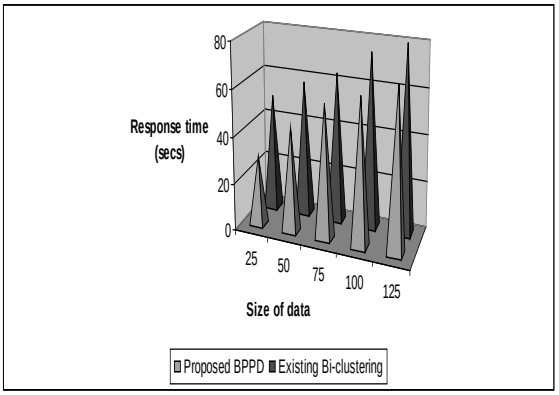


Fig 5.3 Size Of Data Vs. Response Time

Fig 5.3 describes 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). Compared to an existing which consumes more time even for clustering process, the proposed analyzing the biological process on physiological data present in the gene expression datasets using Heuristic search consumes less response time and provide an accurate value related to it and the variance is 20-30% high in the proposed BPPD.

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

6. CONCLUSION

In this paper, we introduced a novel method of identifying the biological process on physiological data using heuristic search algorithm in rough set theory for gene-expression data analysis. The proposed method is based on the heuristic search algorithm for identifying the biological process and processed based on two phases, one is initialization phase and another is iterative adjustment phase. Based on these two phases, the biological process of each gene is 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 using Heuristic search algorithm and estimated the performance of the proposed BPPD in terms of response time and heuristic search threshold. Compared to an existing bi-clustering algorithm, the proposed heuristic search performs better and the performance rate is 70-80% high in the proposed BPPD for analyzing the biological process of physiological data.

REFERENCES:

[1] Nepomuceno et al. “Biclustering of Gene Expression Data by Correlation-Based Scatter Search”, BioData Mining 2011 Jan 24;4(1):3.

[2]Juan A. Nepomuceno et, al., “Correlation–based scatter search for discovering biclusters from gene expression data”, proceeding of the 8th European conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBIO’10

[3] Qi Shen, Wei-Min Shi et. Al., „Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data“, Science direct on Computational Biology and Chemistry 32 (2008) 53–60

[4] Sadiq Hussain, Prof. G.C. Hazarika, “Improved Biclustering Algorithm For Gene Expression Data”, Journal of Theoretical and Applied Information Technology ,15th October 2011. Vol. 32 No.1

[5] Nepomuceno JA, Troncoso A, Aguilar-Ruiz “JS: Evolutionary metaheuristic for biclustering based on linear correlations among genes” SAC 2010 : Proceedings of the 2010 ACM Symposium on Applied Computing (SAC), Sierre, Switzerland, March 22-26, 2010, 1143-1147.

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

[6] Nepomuceno JA, Troncoso A, Aguilar-Ruiz , "JS: Correlation-Based Scatter Search for Discovering Biclusters from Gene Expression Data" EvoBIO 2010 : Proceedings of the 8th European Conference on Evolutionary Computation, Machine Learning and Data Mining, Istanbul, Turkey, April 7-9, 2010, 122-133.

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

[8] Ayadi W, Elloumi M, Hao JK " A biclustering algorithm based on a Bicluster Enumeration Tree : application to DNA microarray data." BioData Mining 2009, 2:9.

[9] Ismail, I.H. et. Al., "Biclustering gene expression datasets — an efficient technique", 2010 The 7th International Conference on Informatics and Systems (INFOS), 28-30 March 2010

[10] Bo-Lin Chen et. Al., "Inferring gene regulatory networks from multiple time course gene expression datasets", 2011 IEEE International Conference on Systems Biology (ISB)

[11] Mohamad, M.S. et. Al., "A Modified Binary Particle Swarm Optimization for Selecting the Small Subset of Informative Genes From Gene Expression Data", IEEE Transactions on Information Technology in Biomedicine, Nov. 2011

[12] Iwen, M.A. et. Al., "Scalable Rule-Based Gene Expression Data Classification", : IEEE 24th International Conference on Data Engineering, 2008. ICDE 2008.