ABSTRACT

Bioinformatics is an application for computational method which is helpful to work out biological problems. Exploring genetic data for various clinical diagnosis and prognosis in bioinformatics take a new turn after the development of microarrays. Mining of these microarrays to tap out hidden information about genes is a hot research nowadays. However, researches feel that data mining fails to recognize the most important biological associations between genes. So, this issue made the scientists to study more about the gene expression datasets.

Biological information mining using clustering techniques were used for the analytical evaluation of gene expression. There are many challenges in the existing clustering methods. The existing methods solve the clustering problem. But, they do not explore the relevant data pertained to be the problem. Bi-Clustering algorithm manages this issue to some extent. But, they too could not mine the biological association among genes effectively. Another key point on existing work was to handle multi modal structure optimization problems with effective searching process. The existing methods provided multi-objective optimization into real world design, but do not offer relational sequence optimized result on the associated gene data.

To overcome the above issues and for extracting the relevant biological information, the present research is proposed. The main challenges lie in creation of minimum response time, reducing bi-clustering time, increasing biological association
efficiency and relational sequential result rate. The objective of the proposed research is to extract the biological information and to identify the relational sequences on gene expression with minimum response time and bi-clustering time for identifying diseased samples.

Primarily, a novel technique called heuristic search has been designed for analysis of the standard biological process on physiological data [BPPD] of the gene expression. The physiological data consists of two patterns of the gene expression datasets: physical patterns and logical pattern. The biological process of these patterns of gene expression datasets were analyzed through heuristic search. This method improved the performance rate of analyzing the biological process and identifying the biological changes of physiological data and is helpful in extracting more expressive genes. It was compared with the performance of existing bi-clustering algorithm. When comparing to existing method, the proposed method extracts only 14 relevant genes for each dataset which is the main advantage for escalating the performance rate. The experimental results show that the proposed BPPD method affords better performance rate in analyzing the biological process and mine the relevant gene of physiological data. In this stage, the genes articulating more were selected.

Next, the biological association among various genes selected in the previous stage was analyzed for finding biologically associated genes to proceed further. A hierarchical clustering based on the proximity measure of Pearson correlation co-efficient (PCHPC) was designed and tested on standard benchmark cancer data sets. The results were compared with Mining Discriminative Patterns (MDP) and Triple Spectral Clustering (SC3). The results divulge that the proposed GL-PCPHC ascertains
biological association between genes in lesser execution time and provides better pattern quality level based on the significance level.

Finally, to classify the samples into normal or diseased, the genes were subjected to be identified as diseased or normal. For this purpose, the relational sequence optimization among the biologically associated genes was performed using an efficient Bi-clustered Ant Optimized Feature Relational Sequencing (BAOFRS) method. BAOFRS method clustered similar relational features using Ant Optimized Medoids algorithms to improve the relational sequence based clustering on gene expressional data to obtain sequence-pairs. The proposed BAOFRS method was administered on the biologically associated genes obtained from the previous stage. The results were compared with the Heuristic algorithm with Black Hole (HBH) and Simulation-based optimization (SO). The evaluation of gene patterns into normal or abnormal genes were measured by following a simple pattern matching process using optimal T value acquired using the relationship between feature vectors. BAOFRS method minimizes the bi-clustering time on performing the relational sequence bi-clustering and increases the similarity score level when compared to the existing methods.

The proposed techniques like BPPD, PCPHC and BAOFRS extort the biological process information from gene expression datasets which is obliging in identifying the diseased sample. The experimental evaluation was conducted using Cancer datasets which is derived from Broad Institute repository. The proposed method results were compared with existing bi-clustering method, MPD method, SO3, HBH and SO method. The results show that, the proposed algorithms perform better in
comparison with the standard beach marking approaches and the diseased samples were extracted from gene expression dataset with eminent similarity value and accuracy.