Mean Squared Residue Based Biclustering Algorithms for the Analysis of Gene Expression Data

Computational Biology is the research area that contributes to the analysis of biological data through the development of algorithms which address significant research problems. The data from molecular -7iology includes DNA, RNA, Protein and Gene expression data. Gene Expression Data provides the expression level of genes under different conditions. Gene expression is the process of transcribing the DNA sequences of a gene into mRNA sequences which in turn are later translated into proteins. The number of copies of mRNA produced is called the expression level of a gene. Gene expression data is organized in the form of a matrix. Rows in the matrix represent genes and columns in the matrix represent experimental conditions. Experimental conditions can be different tissue types or time points. Entries in the gene expression matrix are real values. Through the analysis of gene expression data it is possible to determine the behavioral patterns of genes such as similarity of their behavior, nature of their interaction, their respective contribution to the same pathways and so on. Similar expression patterns are exhibited the genes participating in the same biological process. These patterns e immense relevance and application in bioinformatics and clinical research. These patterns are used in the medical domain for aid in more accurate diagnosis, prognosis, treatment planning, drug discovery and protein network analysis.

To identify various patterns from gene expression data, data mining techniques are essential. Clustering is an important data mining technique for the analysis of gene expression data. To overcome the problems associated with clustering, biclustering is introduced. Biclustering refers to simultaneous clustering of both rows and columns of a data matrix. Clustering is a global model whereas biclustering is a local model. Discovering local expression patterns is essential for identifying many genetic pathways that are not apparent otherwise. It is therefore necessary to move beyond the clustering paradigm towards developing approaches which are capable of discovering local patterns in gene expression data.

A bicluster is a submatrix of the gene expression data matrix. The rows and columns in the submatrix need not be contiguous as in the gene expression data matrix. Biclusters are not disjoint. Computation of biclusters is costly because one will have to consider all the combinations of columns and rows in order to find out all the biclusters. The search space for the biclustering problem is 2m±n where m and n are the number of genes and conditions respectively. Usually m+n is more than 3000. The biclustering problem is NP-hard. Biclustering is a powerful analytical tool for the biologist.

The research reported in this thesis addresses the problem of biclustering. Ten algorithms are developed for the identification of coherent biclusters from gene expression data. All these algorithms are making use of a measure called mean squared residue to search for biclusters. The objective here is to identify the biclusters of maximum size with the mean squared residue lower than a given threshold. All these algorithms begin the search from tightly coreegulated submatrices called the seeds . These seeds are generated by K-means clustering algorithm.

The algorithms developed can be classified as constraint based, _:reedy and metaheuristic. Constraint based algorithms uses one or more f the various constraints namely the MSR threshold and the MSR Difference threshold.The greedy approach makes a locally optimal choice each stage with the objective of finding the global optimum. In metaheuristic approaches Particle Swarm Optimization (PSO) and ariants of Greedy Randomized Adaptive Search Procedure (GRASP) are used for the identification of biclusters.

These algorithms are implemented on the Yeast and Lymphoma datasets. Biologically relevant and statistically significant biclusters are identified by all these algorithms which are validated by Gene Ontology ,iatabase. All these algorithms are compared with some other biclustering algorithms. Algorithms developed in this work overcome some of the -_,roblems associated with the already existing algorithms. With the help of some of the algorithms which are developed in this work biclusters with ery high row variance, which is higher than the row variance of any other algorithm using mean squared residue, are identified from both Yeast and Lymphoma data sets. Such biclusters which make significant change in the expression level are highly relevant biologically.