ABSTRACT

In post genomic era, with the availability of the genome databases the focus of research has shifted from sequencing to eliciting knowledge from these databases. Data mining is a collection of techniques for discovering previously unknown, valid and useful patterns from large databases. Most of the data mining techniques are machine learning and Statistical techniques.

The objective of this research work was to explore the application of data mining techniques in DNA data analysis particularly in analysing gene expression data and promoter sequences. Gene expression is the process of transcribing a gene's DNA sequence into mRNA sequences which are later translated into a sequence of amino acids of proteins. Proteins are the cell's worker molecules. They do all sorts of jobs and play a crucial role in how the cell behaves. The number of copies of produced RNA is called the expression level of the gene. Gene expression databases contain information regarding the expression levels of genes in different experimental conditions.

Expression levels of many genes can be measured simultaneously either by using hybridization techniques like DNA microarray or by using sequencing techniques like SAGE. This is called Gene expression profiling. In order to draw meaningful inferences from gene expression data, it is important that each gene is surveyed under several different conditions. Such data sets may be analyzed using a range of methods with increasing depth of inference, such as cluster analysis, correlation analysis, and association analysis. Data mining includes a collection of such techniques. This research is mainly on the application of association rule mining in analysing gene expression data and developing an efficient data mining algorithm for mining regulatory motifs from the promoter sequences of co-expressed genes.
The state-of-art of existing association rule mining algorithms has been studied and found that the traditional column enumeration algorithms are not suitable for mining gene expression databases because of its wide structure. In gene expression databases genes are items and tissue samples are transactions. Also there will be large number of genes, usually in order of thousands expressed in a tissue sample while the number of samples available is very less. Thus this research lead to the development of a novel mining algorithm for gene expression databases named **GeneExpMiner**. After evaluating its performance with a column enumeration algorithm, **Apriori** and with a row enumeration algorithm, **Carpenter**, it is found that GeneExpMiner outperforms both them in mining gene expression databases.

Association rule mining is a technique that is commonly used for finding association between genes. Study of expressed genes is one of the major areas in cancer research. Studies reveal that there is not a single gene behind the initiation and prognosis of cancer, but a group of genes acting together. Thus this research proposed a novel application of association rule mining. The application proposed here is to find co-regulated signature genes in cancer, i.e., to find association between signature genes. The identified associations are represented in the form of association rules. Signature genes are set of genes differentially expressed in cancer tissue compared to normal tissue. Association rules of the form \textit{gene A [+] \rightarrow \{geneB[+], geneC[-], geneD[+]\}} has been generated. The meaning of this rule is that whenever gene A is over expressed in a pancreas cancer tissue, gene B and gene D are also over expressed and gene C is under expressed. These rules can be used for developing better therapeutics. For example, a medicine which can reduce the expression level of gene A may have a hidden effect of decreasing the expression level of gene B and gene D and increasing the expression level of gene C. The proposed data mining approach is demonstrated with SAGE data related to pancreas cancer.
Discovery of transcription factor binding sites or motifs is a much explored area of research in functional genomics. Regulatory motifs are short patterns of nucleotides, usually 5-20 bp long, found common in the promoter region of set of co-expressed genes, which control the expression or regulation of genes. Identification of these motifs gives insight into the regulatory mechanism of genes. The state-of-art of existing motif discovery algorithms has been studied and found that even though many motif discovery algorithms are reported there is not a single algorithm that can be considered as the best. Some of them are very expensive and NP-hard and some of them may fall into local optimum. Thus another objective of this research was to find an efficient data mining algorithm for mining regulatory motifs from the promoter sequences of co-expressed genes. The first proposed algorithm, MotifMiner, is a table driven greedy algorithm. It is an enumerative algorithm and the greedy approach is used to speed up the search. Even though the algorithm could identify many meaningful motifs from the test datasets, the greedy approach does not guarantees to give the global optimum solution. Thus the next objective was to develop an enumerative algorithm which is not NP-hard and guarantees to give global optimum solution. For this an emerging data mining technique called artificial immune system is applied. The second algorithm, AISMOTIF, is an artificial immune system based motif discovery algorithm, which overcomes the problems of MotifMiner. The algorithm could give all possible motif instances from the input promoter sequences in reasonable time.