CHAPTER – 2

LITERATURE SURVEY

2.1. INTRODUCTION

In recent years, rapid developments in genomics and proteomics have generated a large amount of biological data. Sophisticated computational analysis is required to draw conclusions from these data. Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological data using information technology and computer science. The importance of this field of inquiry will grow as large quantities of genomic, proteomic, and other data are generated and integrated.

A particular active area of research in bioinformatics is the application and development of data mining techniques to solve biological problems. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Data mining is one stage in an overall knowledge discovery process. Mining bioinformatics data is an emerging area at the intersection between bioinformatics and data mining.

This thesis presents the efficiency of data mining algorithms on large biological datasets. Since the work presented in this thesis builds on some prior work, the related research work is explained in detail below.

2.2. LITERATURE REVIEW

Data mining techniques are the result of a long process of research and product development. The evolution of these techniques began when business data was first stored on computers. It continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.
In 1996 Fayyad UM, Gregory Piatetsky-Shapiro & Smuth P (Fayyad UM, et.al, 1996) clarified the relation between knowledge discovery and data mining. They provided an overview of the KDD process and basic data mining methods. They have explained though the various algorithms and applications of data mining might appear quite different they share many common components.

Some of the data mining techniques are association, classification, prediction and clustering. Bing Liu, Wynne Hsu & Yiming Ma (Bing Liu, et.al, 1998) have tried to integrate classification and association rule mining. Classification rule mining aims to discover a small set of rules in the database that forms an accurate classifier. Association rule mining finds all the rules existing in the database that satisfy some minimum support and minimum confidence constraints. They have shown the integration by focusing on mining a special subset of association rules called Class Association Rules. Experimental results show that the classifier built this way would be more accurate than the normal classification rules.

Applications of data mining to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction. There has been research in the area of applications of data mining for biological databases.

Han J (Han J, 2002) has shown how data mining can help bio-data analysis. Recent progress in data mining research has led to the development of numerous efficient and scalable methods for mining interesting patterns in large databases. Recent progress in biology, medical science, and DNA technology has led to the accumulation of tremendous amounts of bio-medical data that demands for in-depth analysis. The question is how to bridge the two fields, data mining and bioinformatics, for successful mining of bio-medical data He has analyzed how data mining can help bio-medical data analysis and outline some research problems that may motivate the further developments of data mining tools for bio-data analysis.

Almuallim H and Dietterich TG (Almuallim H, et.al, 2002) have introduced FOCUS-2 an algorithm that exactly implements the MIN-FEATURES bias. It is an
improvement over their previous FOCUS algorithm. They have also introduced algorithms such as Mutual-Information-Greedy, Simple-Greedy and Weighted-Greedy. They have proved that Weighted-Greedy algorithm provides an excellent and efficient approximation of MIN-FEATURES bias.

Rattanakronkul N, Wattarujeekrit T & Vaiyamai K (Rattanakronkul N, et.al, 2003) have developed a new system called ProsMine for automatically predicting protein structural class from sequence, based on using a combination of data mining techniques. They had created a previous protein structural class prediction system, where only enzyme proteins can be predicted. ProsMine can predict the structural class for all of the proteins. Their idea, based on the lattice theory, was to discover the set of Closed Sequences from the protein sequence database and use those appropriate Closed Sequences as protein features. A sequence is said to be “closed” for a given protein sequence database if it is the maximal subsequence of all the protein sequences in that database. Efficient algorithms have been proposed for discovering closed sequences and selecting appropriate closed sequence for each protein structural class. Experimental results, using data extracted from SWISS-PROT and CATH databases, showed that ProsMine yielded better accuracy.

The literature survey carried out during research shows that there is a need to analyze the different data mining techniques to understand the efficiency and problems associated with the techniques when they work on large data.

Lucas P (Lucas P, 2004) has discussed the current role of data mining and Bayesian methods in biomedicine and health care. Bayesian networks and other probabilistic graphical models have begun to emerge as methods for discovering patterns in biomedical data and also as a basis for the representation of the uncertainties underlying clinical decision-making. At the same time, techniques from machine learning are being used to solve biomedical and health-care problems.

Abdelghani Bellaachia & Erhan Guven (Abdelghani Bellaachia & et.al., 2005) present an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. They have investigated the accuracy of three data mining techniques: Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree.
The result of the experiments conducted shows that the C4.5 decision tree algorithm is more accurate than the others.

Sabbagh A & Darlu P (Sabbagh A, et.al, 2006) have shown that data mining methods, such as multifactor dimensionality reduction and neural networks, appear as promising tools to improve the efficiency of genotyping tests in pharmacogenetics with the ultimate goal of pre-screening patients for individual therapy selection with minimum genotyping effort.

Subramanyam RBV & Goswami A (Subramanyam RBV, et.al, 2006) has explained use of fuzzy quantitative association rules. The concept of fuzzy sets is one of the most fundamental and influential tools in the development of computational intelligence. They have proposed the fuzzy pincer search algorithm that generates fuzzy association rules by adopting combined top-down and bottom-up approaches. A fuzzy grid representation is used to reduce the number of scans of the database and our algorithm trims down the number of candidate fuzzy grids at each level. It has been observed that fuzzy association rules provide more realistic visualization of the knowledge extracted from databases.

Jiawei Han, Hong Cheng, Dong Xin, Xifeng Yan (Jiawei Han, et.al, 2007) have researched the current status and future directions of frequent pattern mining. Frequent pattern mining has been a focused theme in data mining research for over a decade. Abundant literature has been dedicated to this research and tremendous progress has been made, ranging from efficient and scalable algorithms for frequent item set mining in transaction databases to numerous research frontiers, such as sequential pattern mining, structured pattern mining, correlation mining, associative classification, and frequent pattern-based clustering, as well as their broad applications. They have provided a brief overview of the current status of frequent pattern mining and discuss a few promising research directions. They believe that frequent pattern mining research has substantially broadened the scope of data analysis and will have deep impact on data mining methodologies and applications in the long run. However, there are still some challenging research issues that need to be solved before frequent pattern mining can claim a cornerstone approach in data mining applications.
Xiaohua Hu & Yi Pan (Xiaohua Hu, et.al, 2007) have brought together the ideas and findings of data mining researchers and bioinformaticians by discussing cutting-edge research topics such as, gene expressions, protein/RNA structure prediction, phylogenetics, sequence and structural motifs, genomics and proteomics, gene findings, drug design, RNAi and microRNA analysis, text mining in bioinformatics, modelling of biochemical pathways, biomedical ontologies, system biology and pathways, and biological database management in their book.

Gowtham Atluri, Rohit Gupta, Gang Fang, Gaurav Pandey, Micheal Steinbach & Vipin Kumar (Gowtham Atluri, et.al, 2009) have worked on association analysis techniques for bioinformatics. Association analysis is one of the most popular analysis paradigms in data mining. Despite the solid foundation of association analysis and its potential applications, this group of techniques is not as widely used as classification and clustering, especially in the domain of bioinformatics and computational biology. They present different types of association patterns and discuss some of their applications in bioinformatics. They present a case study showing the usefulness of association analysis-based techniques for pre-processing protein interaction networks for the task of protein function prediction. Some of the challenges that need to be addressed to make association analysis-based techniques more applicable for a number of interesting problems in bioinformatics are also discussed by them.

Sondes Fayech, Nadia Essoussi and Mohamed Limam (Sondes Fayech, et.al, 2009) have worked on partitioning clustering algorithms for protein sequence data sets. Genome-sequencing projects are currently producing an enormous amount of new sequences and cause the rapid increasing of protein sequence databases. The unsupervised classification of these data into functional groups or families, clustering, has become one of the principal research objectives in structural and functional genomics. Computer programs to automatically and accurately classify sequences into families become a necessity. A significant number of methods have addressed the clustering of protein sequences and most of them can be categorized in three major groups: hierarchical, graph-based and partitioning methods. Among the various sequence clustering methods in literature, hierarchical and graph-based approaches have been widely used. Although partitioning clustering techniques are extremely used in other fields, few applications have been found
in the field of protein sequence clustering. It is not fully demonstrated if partitioning methods can be applied to protein sequence data and if these methods can be efficient compared to the published clustering methods.

Ashish Mangalampalli & Vikram Pudi (Ashish Mangalampalli, et.al, 2009) have worked on fuzzy association rule mining algorithm for fast and efficient performance on very large datasets. They have presented a novel fuzzy ARM algorithm, for very huge datasets, as an alternative to fuzzy Apriori, which is the most widely used algorithm for fuzzy ARM. Through their experiments, they have shown that the algorithm is 8-19 times faster than fuzzy Apriori. This considerable speed up has been achieved because novel properties like two-phased tidlist-style processing using partitions, tidlists represented in the form of byte-vectors, effective compression of tidlists, and a tauter and quicker second phase of processing.

Golriz Amooee, Behrouz Minaei-Bidgoli, Malihe Bagheri-Dehnavi (Golriz Amooee, et.al, 2012) have compared various data mining prediction algorithms for fault detection. The industrial companies try to improve their efficiency by using different fault detection techniques. Their strategy is to process and analyze previous generated data to predict future failures. They detected wasted parts using different data mining algorithms and compared the accuracy of the algorithms of neural networks, Bayesian network and Logistic regression. A combination of thermal and physical characteristics has been used and the algorithms were implemented on Ahanpishegan's current data to estimate the availability of its produced parts.

Ning Zhong, Yuefeng Li & Sheng-Tang Wu (Ning Zhong, et.al, 2012) have researched on effective pattern discovery for text mining. Many data mining techniques have been proposed for mining useful patterns in text documents. Most existing text mining methods adopted term-based approaches and they all suffer from the problems of polysemy and synonymy. People have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. They present an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant
and interesting information. The proposed technique of theirs uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. The experimental results show that the proposed model outperforms not only other pure data mining-based methods and the concept based model, but also term-based state-of-the-art models, such as BM25 and SVM-based models.

Syed Shajahaan S, Shanthi S & ManoChitra V (Syed Shajahaan & et.al., 2013) have shown the application of data mining techniques to model breast cancer data. Breast cancer poses a serious threat and is the second leading cause of death in women today and most common cancer in developed countries. As breast cancer recurrence is high, good diagnosis is important. Many studies have been conducted to analyze Breast Cancer Data. They explore the applicability of decision trees to predict the presence of breast cancer. They have analyzed the performance of conventional supervised learning algorithms viz. Random tree, ID3, CART, C4.5 and Naive Bayes and their experimental results prove that Random Tree serves to be the best one with highest accuracy.

Ravi Kumar G, Ramachandra GA & Nagamani K (Ravi Kumar G & et.al., 2013) have tried to predict breast cancer using data mining techniques. They worked on different classification techniques such as Decision Tree, Neural Networks, Naïve Bayes, Logistic Regression, Support Vector Machine and K-Nearest Neighbor on breast cancer data. The accuracy of classification techniques is evaluated based on the selected classifier algorithm. They found the performance of SVM showed the highest level of comparison with other classifiers. SVM classifier is suggested for diagnosis of Breast Cancer disease based classification to get better results with accuracy, low error rate and performance.

Vikas Chaurasia & Saurabh Pal (Vikas Chaurasia & et.al., 2013) studied different classifiers such as Classification and Regression Tree, Iterative Dichotomized3, Decision Table and the experiments were conducted to find the best classifier for predicting the patient of heart disease. Observation shows that CART performance is having more accuracy, when compared with other two classification methods. The best algorithm based on the patient’s data is CART Classification with accuracy of 83.49% and the total time taken to build the model is at 0.23 seconds. CART classifier has the lowest average error at 0.3 compared to others. These results suggest that among the machine
learning algorithm tested, CART classifier has the potential to significantly improve the conventional classification methods used in the study.

Vikas Chaurasia & Saurabh Pal (Vikas Chaurasia & et.al., 2014) also studied different prediction algorithms to predict breast cancer survivability. They used three methods of REPTree, Radial Basis Function Network and Simple Logistic. The best algorithm they found based on the patient’s data was Simple Logistic Classification with accuracy of 74.47% and the total time taken to build the model was at 0.62 seconds. The results suggest that among the machine learning algorithms tested, Simple logistic classifier has the potential to significantly improve the conventional classification methods used in the study.

Padmapriya B & Velmurugan T (Padmapriya B & et.al., 2014) studied the SEER Public-Use data of breast cancer and analyzed it under C4.5 and ID3. They found the accuracy of C4.5 was higher in comparison to other classification techniques applied for the same. They concluded that the performance of C4.5 is better than the other algorithms.

Peter Adebayo Idowu, Kehinde Oladipo Williams, Adeniran Ishola Oluwaranti and Jeremiah Ademola Balogun (Peter Adebayo Idowu & et.al, 2015) focused at using two data mining techniques to predict breast cancer risks in Nigerian patients using the naïve bayes’ and the J48 decision trees algorithms. The performance of both classification techniques was evaluated in order to determine the most efficient and effective model. The J48 decision trees showed a higher accuracy with lower error rates compared to that of the naïve bayes’ method while the evaluation criteria proved the J48 decision trees to be a more effective and efficient classification techniques for the prediction of breast cancer risks among patients of the study location.

Arutchelvan K & Periyasamy R(Arutchelvan K & et.al., 2015) has proposed a cancer prediction system based on data mining techniques. This system estimates the risk of the breast cancer in the earlier stage. The system is validated by comparing its predicted results with patient’s prior medical information. The main aim of this model is to provide the earlier warning to the users and it is also cost efficient to the user. A prediction system is developed to analyze risk levels which help in prognosis. This research helps in
detection of a person’s predisposition for cancer before going for clinical and lab tests which is cost and time consuming.

Data mining in bioinformatics is hampered by many facets of biological databases, including their size, number, diversity and the lack of a standard ontology to aid the querying of them as well as the heterogeneous data of the quality and provenance information they contain. Another problem is the range of levels the domains of expertise present amongst potential users, so it can be difficult for the database curators to provide access mechanism appropriate to all. The integration of biological databases is also a problem. Data mining and bioinformatics are fast growing research area today. It is important to examine and develop new data mining methods for scalable and effective analysis of biological data.