LOVELY PROFESSIONAL UNIVERSITY

FUZZY BASED DATA MINING APPROACH TO IDENTIFY STRUCTURAL PATTERNS IN PROTEINS

A Dissertation proposal submitted by

Shivika Sharma (11008527)

to

School of Computer Science Engineering

In partial fulfilment of requirement for the

Award of the Degree of

Master of Technology in Computer Science

Under the Guidance of

Asst. Prof. Ms. Avinash Kaur

Under the Co-Guidance of

Asst. Prof. Harshpreet Singh

(May, 2015)

ABSTRACT

The progress of bioinformatics generates a large volume to data that needs to be analysed in order to identify various structures. Proteins, involved in several important tasks in living organism, are a mixture of amino acids which have different structures and patterns. The number of primary structures solved and stored in databases are growing faster than our capability to solve these tertiary structures using different experimental methods. The analysis of bioinformatics data has seen major shifts from traditional data mining approach to hybrid approaches. Fuzzy method is proposed as an algorithm for mining proteomic data. Fuzzy association rule helps to recognize the uncertainties and vagueness in patterns of the protein structure. The methodology considers the perception and cognitive uncertainty of subjective decisions allowing the usage of imprecise description of protein data.

ACKNOWLEDGEMENT

The satisfaction that accompanies the half completion of the task would be incomplete without the mention of the people whose ceaseless co-optation made it possible, whose constant guidance and encouragement crown all efforts with success.

I am grateful to Mrs. Avinash Kaur (Asst Prof) and Mr. Harshpreet Singh (Asst Prof) for the inspiration and the constructive suggestions that help me in completing the assignment of preparing the said project within the time stipulated. I would like to thank my Parents and God. With their support and well wishes I am able to complete this project in time.

Shivika Sharma

DECLARATION

I hereby declare that the dissertation entitled, “Fuzzy based data mining approach to identify the structural patterns of proteins” submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: __________ Investigator: Shivika Sharma

Registration No: 11008527

CERTIFICATE

This is to certify that Shivika Sharma has completed M.Tech dissertation titled “FUZZY BASED DATA MINING APPROACH TO IDENTIFY STRUCTURAL PATTERNS IN PROTEINS” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma. The dissertation is fit for the submission and the partial fulfilment of the conditions for the award of M.Tech Computer Science and Engineering.

Date:

Signature of Advisor

Name: Avinash Kaur

Designation: Assitant professor

UID:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

This chapter gives a brief overview of big data and bioinformatics analysis process and also introduces about the challenges of big data in bioinformatics.

1.1 BIG DATA

Big data is a new technology that deals with huge quantities of information generated from digital data. The internet exemplifies a large place where heavy amounts of data are appended regularly. There are 100 Terabytes of data that are reorganized in a day through multiple sources like social websites such as Facebook, twitter etc. This data is very large and complex therefore cannot be processed or analysed easily. Big data can be sized into peta, exa and in zettabyte. Collection of large data sets of unstructured data is known as big data. An important aspect of Big Data is that it cannot be handled with standard data management techniques due to the inconsistency and unpredictability of the possible combinations.

Big Data can be measured using 4V's (figure 1) which are elaborated as follows **(Ularu, 2012):**

Volume: various data sources do not contain huge volumes of data rather number of data sources grown to be in the millions. The massive scale and growth of unstructured data outstrips traditional storage and analytical solutions.

Velocity: there are many sources where data is being collected, many of the data sources are very dynamic and the number of data sources is also being exploded. It refers to the time in which big data are being processed.

Variety: data sources are diverse. Different types of data are required and stored.

Veracity: refers to the degree in which whatever information is used that the information is relevant to take decisions.

Big data concerns with the large volume, growing data sets with multiple sources. Big data can be characterized by its heterogeneous and autonomous sources. These sources, generating huge amount of data that are distributed geographically with a decentralized control. The aim is to explore the evolving relationships and reducing the complexity among the data. Managing the huge amount of data is a challenging task. More challenging is structuring the data, which involves analysis and storage of the data.

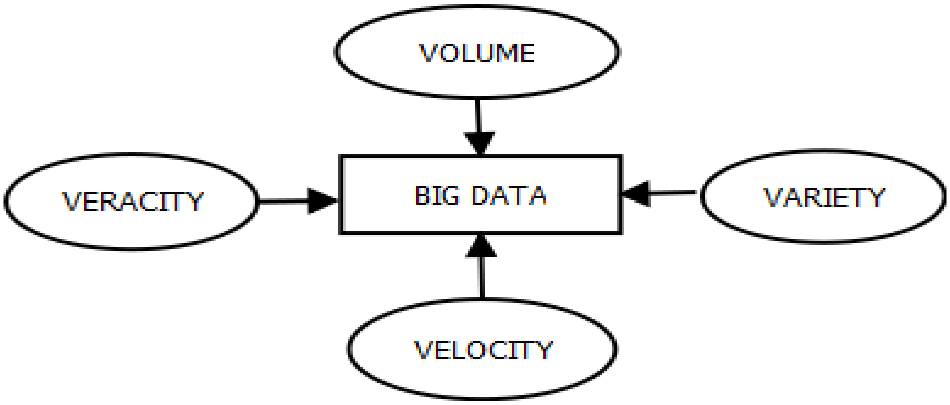


Figure 1.1: Four V's of big data

Of first understanding big, data is a complex and trivial task. It is very important that the data used should be properly analysed and should be in a properly structured way.

1.1.1 Process Of Big Data

Big data means large volumes of data that cannot be managed and analyse in a direct way. Steps to process the big data is as follows [figure-2]:

Extraction of data:

It is the process of getting qualitative or quantitative information from valuable sources like databases. It is act of analysing and gathering a piece of knowledge which means retrieving of data from the relevant data sources. In this process the data extracted from the data sources are usually unstructured or poorly structured and then converted into understandable structure.

Extracting and manipulation of data can be done by

Data mining: the analysis step of the Knowledge Discovery of Databases (KDD) process is a computational process to analyse patterns of data according to different approaches and categorize it into useful information.

Text mining: it is the process of deriving valuable structured information though linguistic and lexical analysis and statistical pattern learning techniques for intelligence, exploratory research and investigation.

Forecasting: it is a process of predicting results about events that have not yet occurred. Statements are produced regarding the outcomes of the occurrences that are yet to come in future including the factors of risk and uncertainty.

Data optimization: it is the process of making the best and most effective use of resources that contains huge sizes of data which is becoming ubiquitous. A data optimization technique rearranges or rewrites to improve the efficiency of retrieval or processing and thus solve optimization problems of data of unprecedented sizes.

The more fundamental challenge of Big data is extracting the important or relevant data. In many situations knowledge extraction process has to be very efficient. The challenge is not to extract meaningful information of data, but to gain knowledge; look for patterns and to make sense of the data that whatever data is extracted must be useful in any way.

For extracting many different approaches like statistical and graph theoretical methods, data mining and machine learning methods are used. Statistical methods are used to identify relevant patterns (Hirsh, 2008).Nowadays Knowledge discovery in data mining is used to extract the data which is a whole process or chain of extracting the relevant data from different data sources. Integration of data from various data sources are performed which result in the common entities which frequently encountered in KDD and called Merge/Purge problem **(J. Zhou, 2013).** This problem is difficult to solve both in scale and accuracy.

Many approaches for handling the large amount of data can be used. There are some data mining environments that produces a large amount of data that should be analysed and should be in a properly structured way. Some of the different algorithms and methods implemented to handle big data are GLC++ which deals with the large sample type of data and PROXIMUS, is used for compression of transaction sets **(C. Yadav, 2013).**

Some extraction techniques are

- Statistical methods **(Hirsh, 2008):** Used for identifying relevant patterns.
- HCI-KDD **(Andreas Holzinger, 2014)**: Enable end users to find usable information.
- BIO-CAT **(Jie Zhou, 2013)**: Helps user for pattern recognition for biological image.
- Apriori and AprioriTid **(Rakesh Agrawal, 1993)**: Applicable for large databases for mining association.

Table 1.1: Different extraction techniques for the processing of data

Sno	References	Extraction techniques	Findings
1	Zhou J, et.al	Statistical methods	Used for identify relevant patterns
2.	Andreas Holzinger, et.al	HCI-KDD	Enable end users to find useable information.
3.	Jie Zhou	BIO-CAT	User friendly platform for pattern recognition for biological image
4.	Rakesh Agrawal, et.al	Apriori and AprioriTid	Applicable for large databases for mining association
5.	S. Arya, et.al	Nearest Neighbor Search	when searching object is in High Dimensional space

Processing of data:

Data processing means collecting and manipulating the data to produce important information.

In this era of technology, data processing is done through computers. Any type of raw information which is in human readable form is fed to computer so as to convert it into machine readable form and thus process it and convert it back to human readable form. Earlier in 19th century and 20th century the data was either Manually Processed or Automatically Processed. In Manual Processing, complete manual methods were carried out as individuals were appointed at that time to carry out processing of data to produce fully detailed reports. In the case of Automatic Processing, the use of many independent equipment's was started to carry out data processing. Individuals that were appointed for the same started using such equipment's that could generate results faster and easier. But with the advancement in technology, the overhead has reduced to large extent and the accuracy and efficiency of data processing has increased tremendously. Due to the innovation of computers and super computers, several pieces of equipment's that were used earlier in data processing were left no longer in use. An analysis part comes in the processing of big data. Processing of data has a number of applications which includes ensuring clean, correct and useful data, data sorting, decreasing detailed information into structured main points, separation of data into multiple classes and the interpretation and demonstration of data

Processing means collecting and manipulating the data to produce an important information. After extracting the data from the multiple data sources refining and analysing of data is done. There are many techniques for analysing the different types of data like clustering, classification etc. Big data means data that cannot be handled and processed in a straightforward way. Earlier data are saved, then loaded to some disk and interactively perform one or more analysis on data. But now in the today's world of big data the interactivity, processing of data is very fast.

Some extraction techniques are:

- Cluster analysis **(Anoop Kumar Jain, 2012)**: It is not an algorithm but a general task of partitioning set of objects and then combines equivalent objects into groups. As we all know that the data available is either unstructured or very poorly structured. So, there is a need to organize the fetched data and arrange it in a proper fashion to make it knowledgeable and understandable. It is a common technique for statistical data analysis that is used in a number of fields like recognition of patterns, image analyzing and processing, studying biological data for the innovation of useful information
- Hierarchical clustering **(Lan Yu, 2010)**: hierarchical clustering means to make the clusters in the form of hierarchy. It produces a nested sequence of clusters. This clustering used objects based on their distances to form clusters. Clusters are formed in top-down and bottom-up fashion. It is based on the idea of the objects being more related to nearby objects as compared to distant objects. In this algorithm the distance between two objects is computed through various methods and then these objects are connected to form clusters. There are two strategies of hierarchical clustering which are:

  Hierarchical Agglomerative Clustering: This is a bottom-up approach which starts with a single item and in each successive repetition it amalgamates the closest pair of clusters under some comparable criteria until all the items are in a single cluster.The working of this algorithm is stated as follows:

  1. For any N items, each item is assigned a cluster, thus forming N clusters such that each cluster contains exactly one item.

2. Now we determine the distances between each cluster to every other cluster and combine the clusters having the shortest distance amongst them thus leaving N-1 number of clusters behind.
3. Now we calculate the distances between the new cluster and every already existing cluster. This can be done in three different clustering methods named as single linkage, complete linkage and average linkage clustering.
4. Repeat step 2 and step 3 until all the items are combined into one cluster having N items.

Some of the examples of hierarchical clustering are CURE, CHAMELEON, BIRCH

Hierarchical Divisive Clustering: This is a top-down approach which starts with a single cluster containing all items and in each successive repetition it separates the clusters with the farthest distance under some comparable criteria until all the clusters are left with one single item each. This strategy is generally not offered and is very rarely used because it proves to be very sluggish for large amount of data. The working of this algorithm is stated as follows:

1. For a cluster of size N, it constitutes N number of items.
2. Now we start dividing the cluster into daughter cluster that have the maximum distance to the parent cluster.
3. We continue to partition the parent cluster step by step such at each step the cluster having the maximum distance gets separated from the parent cluster until the parent cluster gets partitioned into N number of clusters each of size 1.

- Central based clustering **(Yun Ling, 2010):** Here clustering is focused on the central vector. K-means is the algorithm used for this type of clustering. K-means helps to partition the data space into a structure known as voronoi diagram.
- Distribution based clustering **(Vignesh T. Ravi, 2009)**: It is based on the distribution models. This model use the concept of sampling arbitrary objects from a distribution. It is very efficient method, it also helps to generate complex models therefore user can have option to choose best model for his use. One of the example of distribution based clustering are DBCLASD which means distribution based clustering of large spatial databases used to discover clusters of this type of algorithm.

- Density based clustering **(J. Prabhu, 2010)**: Used for connecting purposes within a certain distance. It helps to remove the noise in the dataset. Features of this algorithm are:

  It handles the clusters of arbitrary shapes present in the dataset.

  It requires only one scan of the data.

  It needs density parameters initialized.

  DESCRY is the new algorithm used for mining the large datasets. It is used to identify the clusters of data which have different shapes and sizes. Some of the examples of this clustering are DBSCAN, DENCLUE and OPTICS.
- Bayesian networks **(Nir Friedman, 2000)**: Used for analyzing biological patterns. it is basically used for the analysis of gene expression data. It differentiates the expression levels of proteins. It helps to create the networks from the database itself.
- Cloud resource monitoring **(Vincent C. Emeakaroha, 2011):** Used for searching bio-medical relevant patterns. Therefor used for workflow applications. With the rapid development in recent years of high-throughput technologies in the life sciences, huge amounts of data are being generated and stored in databases. An analysis of these large-scale data in a search for bio medically relevant patterns remains a challenging task. So a cloud resource monitoring technique and knowledge management strategy is used.
- Sequence analysis **(K. Raza, 2012)**: Collect sequence and stored in a structured database. Temporally it was used for market basket analysis. Different sequence analysis techniques are SPADE, aprioriAll and GSP are used for mining the different frequent sequences of the itemsets.

Table 1.2: Data mining using different clustering technique

S.no	Reference	Analysis techniques	Findings
1	Anoop Kumar Jain	Cluster analysis	Used for statistical data analysis
2	Lan Yu	Hierarchical clustering	Used objects based on their distances to form clusters
3	Yun Ling & Hangzhou	Central based clustering	It is based on its central vector

4	Vignesh T. Ravi	Distribution based clustering	It is based on the distribution models
5	J. Prabhu et.al	Density based clustering	Used for connecting purposes within a certain distance
6	Nir Friedman, et.al	Bayesian networks	Used for analyzing biological patterns
7	Gowtham Atluri, et.al	Association analysis	Used to mine the large data
8	Vincent C. Emeakaroha, et.al	Cloud resource monitoring	Used for searching bio-medical relevant patterns
9	Khalid Raza	Sequence analysis	Collect sequence and stored in structured database

Managing of data:

Data Management is a process of controlling the information that has been extracted and processed. Managing data is an integral part of research and analysis procedure. Managing data means to govern the large amount of structured and unstructured data. There is a need of data management process because of the following reasons:

- Sometimes, same data gets stored in multiple locations, sometimes in different file formats, leading to data redundancy that cause issues regarding space of the disk that holds the data. So data should be managed so as to avoid data redundancy.
- Due to data redundancy, the problem of data inconsistency arises during managing information and data. By data inconsistency we mean data items having different values at different locations at the same time.
- We need our data to be isolated so that it becomes easily accessible as compared to data that is scattered in various files.
- By managing our data we need to make sure that the data remains consistent even if there is a system failure. This can be done by ensuring proper recovery and backup mechanisms.

- Data management also includes that once the data or the information is managed properly it should be accessible to all those persons who are authorized to do so.
- Through data management techniques, we can ensure the integrity of our information and resources and thus, contribute in a rightful manner.
- Managing our data saves lots and lots of time and resources and will do the same in future.

Techniques that are used for analysis purpose is Cloud resource monitoring technique and a knowledge management strategy that is used to manage computational resources. The challenge here is not to extract data but to make knowledge and the data should be meaningful. Many different approaches like statistical **(Raymer, 2003)** and graph theoretical method **(Shelokar, 2013)** can be used to manage and to process data. There is a multimode data manager tool that helps to manage data **(Viceconti, 2007)**. There is a necessity to manage large volumes of data referred to as big data and keep on doing managing new data so that in the coming time we don't have to face any kind of problems. Effective data management practices include the following:-

- Every individual should be responsible enough of his/her own tasks and duties in managing data.
- Define and decide thoroughly to smartly conclude that how the data should be stored and backed up.
- Determine how to deal with the data when it goes through the modification stage.

Storing data:

Storing means to save structured data for the future use. With the fastest growing volume of biological data, a diversity of data sources has been created to facilitate data management, accessibility and analysis. Researchers need to be skilful in interrogating these data sources and in the use of extracting information for further data analysis. Integration of data from different sources like distributed, heterogeneous and voluminous data sources turns out to be an obstacle for big biological data **(Z. Zang, 2011).**

There are two types of storage options:-

- Local based storage: - this option states that we can save and store our data and information on our computers, external storage devices such as hard disks, USB etc., local network access server and thus data can be retrieved or accessed everyday conveniently.
- Cloud based storage: - This option states that we can upload our data and information on remote servers which are located at very distant locations. This option takes away the

problem of access and control issues as the data stored on these remote servers is not usually retrieved or accessed back.

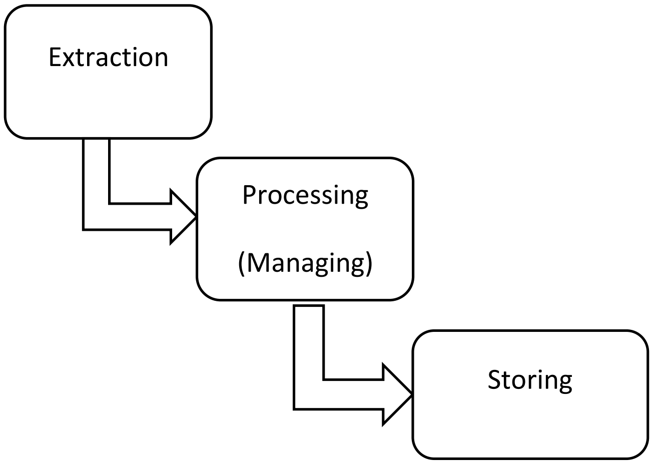


Figure 1.2 Steps to process the data

1.2 Data Mining

Data mining is the process of examine the large amount of database to generate new information. It is the process to analyse the data and generate some important information in terms of rules or patterns. Data mining is a powerful technology with great potential to help persons focus on the most important information in their data warehouses. Different tools and techniques are used to extract patterns and information that are hidden in the large databases. These tools can predict future trends and behaviours and has the capability to answer desired questions that were traditionally too time consuming to resolve. Data mining is the integral part of KDD. Data mining plays an important tasks in the field of research and practical applications. The main challenges to the data mining are as follows

- Huge datasets and high dimensionality
- Understandability of the patterns
- Redundant data
- Incomplete data and data integration

Above issues of data mining can be solved by various other techniques of data mining. Some of the techniques of data mining are:

1.2.1 Statistical methods:

From the word itself statistics means collecting, analysing and presenting the data. Statistics helps to abstract the knowledge from the database. Typically it differs from traditional statistics on the bases of size of data set and the data that is initially collected for the data mining analysis. Some of the statistical issues in data mining are:

- On the basis of size of the data.
- The curse of dimensionality and approaches to address it
- Assessing uncertainty
- Automated analysis
- Algorithms in data analysis in statistics
- Visualization
- Scalability
- Sampling

1.2.2 Machine learning:

As statistical methods have some disadvantages. Statistical method faces difficulty incorporating subjective information in their models and it also faces problem of interpreting the results. Therefore machine learning produces the better predictive accuracy. It is free from the parametric and structural assumptions and results in the good performance. Some of the machine learning techniques are neural networks, genetic algorithm, support vector machines, decision tree induction.

1.2.3 Association rules:

Association rule mining helps in finding the frequent patterns, associations, relationships, or structures among sets of items or objects in transactional databases, relational databases and other information warehouses. Some of the applications of association rule miming are market basket data analysis, cross-marketing, catalogue-design, loss-leader data analysis etc.

Some properties of association rules are

1 It helps to express how items or objects are co related and how they tend to group together.

2 It is very simple to understand that is it is comprehensible.

3 It helps to provide useful information (utilizability).

4 Efficient discovery algorithms exist.

Different types of association rules based on

(a) Types of values handled

(b) Levels of abstraction involved

(c) Dimensions of data involved

Association rule mining is the two-step process

1 Find the frequent item sets

2 Use the frequent item sets to generate (strong) association rules that satisfy the minimum support and the minimum confidence.

Therefore it includes two most important quality measures that are support and confidence.

Support: The support of the rule X→Y is the percentage of transactions in T that contain X∩Y. The support of the rule is represented by the formula

$$\mathrm{Supp}(\mathrm{X} \rightarrow \mathrm{Y}) = \frac{|X \rightarrow Y|}{n}$$

Where n is the total number of transactions

Confidence: It describes the percentage of the transactions containing X which contain Y.

$$\mathrm{Conf}(\mathrm{X} \rightarrow \mathrm{Y}) = \frac{|X \cap Y|}{|X|}$$

1.2.4 Apriori algorithm:

Apriori algorithm helps to enumerate all of the frequent item sets. It is used for mining the frequent item sets for Boolean association rules. Apriori means to use prior knowledge of frequent item sets properties. The algorithm includes two key processes that are connecting step and pruning step.

1. In connecting step : In this step the algorithm firstly scans the database to form all sets of possible combination for field that have support equal to or minimum to it.
2. In pruning step: In this the algorithm considers only minimum value for confidence and discard the other value sets.

1.2.5 Fuzzy association rule:

Fuzzy association rule is the new approach that is used to mine quantitative data frequently present in the databases. Fuzzy association rule uses fuzzy logic to convert numerical attributes to fuzzy attributes. A fuzzy data mining method that is Fuzzy Frequent Pattern growth (FFP-growth) algorithm can be used to mine the large datasets. It treats each data item as a linguistic variable and its portioned is based on the linguistics value.

In this method there are two phases. One is to scan the large item sets and the other is to establish a fuzzy FP-tree by scanning the databases again and again. Then one conditional pattern and FP-tree will be extracted from each node in fuzzy FP-tree to make the fuzzy association rules

1.2.6 Rough sets techniques:

Rough set theory proposed in 1982 by Zdzislaw Pawlak. This theory concerned with the classification and analysis of the incomplete information or knowledge of the data. It deals with the approximation of the sets or deals with the approximation of lower and upper spaces of sets.

Some of the applications of the rough sets are pattern recognition, emergency room diagnostic medical, power system security analysis, spatial and meteorological pattern classification, intelligent control systems and measure the quality of a single set.

Data mining also use different tools to extract different types of data. Some of the top six tools of data mining are Weka, KEEL, R(revolution), KNIME, RAPIDMINER, and ORANGE.

Table 1.3: Top six tools of data mining

Sno.	Tool Name	Features
1.	RAPID MINER	It used a client/server model. Basically used for business, industries, researchers etc. It includes multiple new aggregation functions.
2.	KNIME	Open source data analytics and integration platform. Scalable and high-extensible. Easy to try.
3.	R	It is a statistical computing. Used for data error handling. Numerical problems can easily integrated.
4.	KEEL	User friendly graphical interface. Cluster discovery. Includes regression and pattern mining.
5.	WEKA	Suitable for machine learning schemas Best for mining association rules.
6.	ORANGE	Best for data visualization. Scripting interface, large toolbox. Includes set of components for data pre-processing.

1.3 Bioinformatics Data

By the name itself Bioinformatics means information about the biological data. Bioinformatics is using various techniques and concepts from informatics, statistics, mathematics, biochemistry, physics, and linguistics. In the field of biology and medicines it has many practical applications. Bioinformatics deals with the creation and maintenance of a database to store biological information. Development of this type of database includes design issues as well as an interface where researchers can access existing data and revised data.

The consequently fast growing volume of biological data, a variety of data sources that is databases and web servers have been created to facilitate data management, accessibility and analysis. One must have knowledge how to extract, refine and integrate data. It helps in storing, extracting, analysing and utilizing information from biological sequences and molecules.

In bioinformatics file sizes often exceeds 100 GB, persistent and transportation is the current big challenge. Therefore, transportation of data from one location to another is a big problem. There are some other issues of Bioinformatics data that are security issues and how to store large data.

Some of the different fields of bioinformatics are **(Khalid Raza, 2012):**

1. **Sequence analysis**: during medical analysis, it is important to check the sequences, therefore this operation helps to find which biological sequences are alike and which are different. It is the process of subjecting the DNA, RNA or peptide sequence to sequence alignment. For sequence analysis, BLAST, FASTA tools can be used.
2. **Genome annotation**: it is the process of finding the genes and other biological features in DNA. For genomic analysis SLAM, MEME/MAST tools can be used.
3. **Analysis of protein expression**: Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample. Bioinformatics is very much involved in making sense of protein microarray and HT MS data. For protein Pfam, ProDom tools can be used.
4. **Analysis of mutations in cancer**: in this analysis of the affected cells is done. In this new algorithms and software tools are created for analyzing the affected genes.
5. **Comparative genomics**: in comparative genomics, gene finding is an important task. This explores the differences and similarities in the proteins and RNA.

Bioinformatics cites to the use of managing genetic data using different computational and statistical techniques. Bioinformatics helps to detect the different diseases produces an enormous amount of data that is related to molecular science. For storing the large amount of biological data, parallel clustering concept can be used **(V. Olman, 2009). (Lin Dai, 2012)** provided with the brief introduction about cloud based resources in bioinformatics. They use the concept of cloud computing. Storing and analysis of biological data is the difficult task, so cloud based services can be used. In future to analyse the bioinformatics data, regression techniques **(M. Kim, 2012)** can be used. Test prioritization can be used to analyse the bioinformatics data.

In bioinformatics the data are generally in the form of different structural patterns. Some of the structured patterns are genes, RNA, DNA, etc.. For this DDP **(J.Wang, 2013)** can be used to increase the speed and accuracy of different data. Different sequence mapping tools are used to process big sequence data like cloudbursts, cloudAligner and cloudblast which help to increase the speed and performance accuracy.

Therefore Bioinformatics is a vast topic which have the different fields.

1.3.1 Challenges faced in processing bioinformatics

There are many challenges that are faced during the processing of bioinformatics data.

- The first challenge is how to extract relevant data from multiple data sources. Attainment data is the challenge which is faced during the extraction process.
- The second main task is to mechanically generate the right metadata.
- Data analysis is a more puzzling task than identifying, understanding and citing data.
- If users cannot understand the analysis process therefore having the ability to analyze big, data is of limited value. Ultimately, decision maker provided with the result of analysis.
- Big data problems are primarily at application and system level.

In bioinformatics the data are in the unstructured form, therefore understanding the unstructured clinical data is very important. There are some data which are in the form of the genomic structure and for analyzing it, computational sources can be used. Earlier the biggest challenge faced by the researchers are to make solutions for sequencing DNA. But in the coming years, it is solved very quickly. Today human body genomes are sequenced and results can be found very quickly. Nowadays the biggest challenge is to understand and how to achieve the huge about of data which is created from DNA arrangements.

As the biological data are in the different structured and unstructured form so different types of data have different challenges. Some of the challenges that are faced by the protein science area:

- There are many hypothetical proteins.
- Deviation in structures and functions.
- Analysis of proteins with low sequence analysis.

Similarly, when analysis of gene expression data is done, there are many challenges that are faced by researchers. Management of single data is very difficult. Many sources and commercial packages help researchers to store and manage gene expression data.

CHAPTER 2
REVIEW OF LITERATURE

This chapter introduces the studies that have been already been used to extract, process and to integrate data. It introduces the techniques and tools that have been used by researchers.

Nir Friedman, et.al [14] 2000 proposed a Bayesian network based model for analysis of biological pattern **(Nir Friedman, 2000)**. Gene expression patterns were analysed by using the Bayesian network concept. The two approaches that were used in this research work are a novel search algorithm and a hybrid approach. The main concern was to extract features of the gene expression data and to interact between the genes. Some problems were Statistical aspects of interpreting the results, algorithmic complexity issues in learning from the data, and the choice of local probability models. Methods for expression analysis can be improved by developing the theory for learning local probability models which are suitable for the type of interactions that appear in expression data. Incorporating biological knowledge as prior knowledge to the analysis, helps to improve the search heuristics.

N. Jacq, et.al. 2003 [19] developed a bioinformatics tool which help in storing the biological data **(N. Jacq, 2003)**. Grid applications were used for appending and analysing the biological data. The grid tool was used to store and normalize the data. An interface was used to predict the molecular functions and to identify the sequences formed in biology. BLAST algorithm was also used which helped to store sequences and also compare each datum. Tool still leads to disadvantages in managing all data, as it is unable to cover the duplication of the data.

Gowtham Atluri, et.al, 2009 [15] proposed the different types of association patterns and some of their applications in Bioinformatics **(Gowtham Atluri, 2009)**. Challenges which are needed to be addressed to make association analysis-based techniques are more applicable to a number of interesting problems in Bioinformatics. Basically two types of patterns were used that were observed-Finite item set pattern and Association rule pattern. Some of the types of patterns are Traditional Frequent patterns, Hyper clique Patterns, Error-Tolerant Patterns and Discriminative Pattern Missing. Association analysis has proven to be a powerful approach for analyzing traditional market basket data, and has even been found useful for some problems in Bioinformatics in a few instances. However, there are a number of other important problems in Bioinformatics, such as finding biomarkers using dense data like SNP data and real-valued data like gene-expression data, where Finite item set and Association rule pattern techniques could prove to be very useful, but cannot currently be easily and effectively applied. An important example of patterns which are not effectively captured by the traditional association analysis framework and its current extensions, is a group of genes that are co-expressed together across a subset of conditions in a gene expression data set.

Chanchal Kumar,et.al. 2009 [18] reviewed the limitations of biochemical methods and other related technologies that are only applicable on single type proteins **(Chanchal Kumar, 2009)**. Mass spectrometry when combined with other innovative strategies and advanced experimental and computational methods enables to carry out a large scale study on proteins of cellular level. Another technique name SILAC was used when two or more biological states were needed to be compared. Software named MaxQuant was used for parallel processing of complex data sets. These datasets were the result of a combination of contemporary mass spectrometry and advanced spectrometry. Thus, MaxQuant generates a multidimensional matrix that would contain data regarding proteomes. There are challenges one has to face such as mapping while carrying out such kind of analysis. There are certain structures like BioMart that help in resolving mapping issues, but still it has errors and inconsistencies while producing results.

Ashish Mangalampalli, et, al 2009 proposed an algorithm for fuzzy association rule mining which is used to mine the data from the very large dataset **(Ashish Mangalampalli, 2009)**. Fuzzy association rules are the technique which uses fuzzy logic to convert numeric attributes to fuzzy attributes. Today the popular fuzzy association rule mining algorithms is fuzzy apriori, but this algorithm is very slow so a new fuzzy ARM algorithm is developed for large datasets. This algorithm is very fast and efficient for large datasets. It also includes an operative method to convert crisp data into the fuzzy dataset. The algorithm used is the two phased multiple tidlist-style processing in which tidlists are represented in the form of byte-vectors. Byte-vector representation of tidlists and fast compression of tidlists add a lot to the proficiency in performance.

Vincent C. Emeakaroha, et.al 2011 [16] proposed a Cloud resource monitoring technique and a knowledge management strategy to manage computational resources **(Vincent C. Emeakaroha, 2011)**. The main contributions were the introduction of cloud management techniques applicable to workflow applications and the optimization of scientific workflow application to support their successful completions based on Cloud techniques. As in biomedical science, there are many challenges like the matching of patterns, sequences, etc. Workflow applications were used to the cloud resource monitoring concept which provides us with the on demand promising. Knowledge management was used to monitor the information to allocate the resources during runtime. Structure of the cloud management techniques were also proposed. The top-hat tool was used in the analysis of RNA sequence data.

Rashmi Rameshwari, et.al 2011 [17] reviewed various bioinformatics tools and software's used in visualization to interact with proteomic data and the interaction between the similar data **(Rashmi Rameshwari, 2011)**. Emphasis was made on producing highly efficiency and productive data in terms of biological networks by analysing different commercial software's such as Ingenuity Pathway Analysis, MetaCore and Pathway studio. Advancement in technology has led to many other software's like Cytoscape with the capability of producing high throughput but are prone to errors and thus need to improve. The proposal arises the issues that such tools or software's should act as a complete package thus possessing the capability to produce desired results by analysing the data.

Diogo Stelle, et.al, 2011, [21] proposed various data mining techniques which can be used to form patterns of protein data **(Diogo Stelle, 2011)**. A methodology was used to stores the structure of proteomic data. Data mining is the process to find an relevant information in large datasets. The different areas of data mining techniques are association rules detection, clustering, outlier detection, classification and others. The association rule helps in the mining of hydrophobic profile or patterns to form specific structures of proteins. These patterns can be used in order to help the techniques of protein structure prediction. This work contributes for two areas: prediction of protein structure and protein folding.

Vivian F. López, et.al 2012 [20] proposed various data mining methods for recognizing patterns in biology **(Vivian F Lopez, 2012)**. Combining various techniques of association analysis with classical sequential algorithms, helps to generate grammatical structures. These structures are then converted to Context-Free Grammars. An application of compiler generator, named GAS 1.0 allows to measure the complexity of the obtained grammar automatically from textual data which helps to reduce the learning time for compiler generation. Talking of Context-Free Grammar, there are certain techniques that are used to find relations between attributes of data sets.

Anoop Kumar Jain, et, al 2012 [22] reviewed the processes of cluster analysis, which is one of the methods that is used for statistical data analysis, including different fields like machine learning, pattern recognition, image analysis, information retrieval and bioinformatics **(Anoop Kumar Jain, 2012)**. It is an iterative process of knowledge discovery or iterative multi-objective optimization. The different clustering techniques that were taken under study were Hierarchical clustering, centrally based clustering, Distribution based clustering and Density based clustering. Agglomerative algorithms were also reviewed, which first divides the cluster separately and then helps to combines it to form large clusters. Divisive algorithms used to splits-up the clusters differently and then used for the making of smaller clusters. Issues regarding various clustering techniques that can be used in the mining process of bioinformatics data were highlighted. A survey also says that by using these clustering techniques time taken for retrieving relevant data is less.

Chanchal Yadav, et.al, 2013 [12] proposed a survey on data mining techniques **(Chanchal Yadav, 2013)**. The traditional data mining techniques have been applied and various algorithms were used to manage the unstructured data. A concept of Clustering technique can be applied to analyze large data sets. A new algorithm GLC++ was used for the large - sample type of data and PROXIMUS was used for compression of transaction sets. The issues regarding big data has been discussed. Security is the main concern in big data. An insight into how different algorithms can help us to handle unstructured data has been proposed.

C.L. Philip Chen, et.al, 2014 [10] provided with a brief survey on big data applications, challenges, techniques and technologies **(C.L. Philip, 2014)**. Big data change the method that are used in the business, managements and researches. Data intensive science is coming into the world that purposes is to provide a tools that need to handle the big data problems. With the increase in the data intensive applications the world of science has rehabilitated. Therefore data-intensive is viewed as the new science paradigm in which one should know how to handle the large sets of data. It deals with the challenges, technologies and problems that are faced by big data. A brief discussion was done on the big data problems in commerce and business, society administration and scientific research fields. Techniques used were statistics, data mining, machine learning, and pattern recognition and optimization methods. The concept of cloud computing was discussed, which is the emerging technology that came into existence.

Akhil Kumar Das, et.al, 2014 [11] proposed a fuzzy mining technique for the estimation of a gene expression data **(Akhil Kumar Das, 2014)**. The main aim was to cluster those genes which are having same profiles. As cellular data are difficult to analyze, it is in the gene form and are very complex, therefore for analyzing the gene structured data, a concept of gene clustering was used. KDD helps to convert the low level data into high level data. Fuzzy association patterns were used to combine the microarray genes. Different steps that involved to estimate the gene expression data are Fuzzyfication, Association Based Clustering, Weight Assignment and Gene Function Prediction. By applying these technique the data could be differentiated into low, medium and high qualitative terms.

Andreas Holzinger, et.al, 2014 [13] proposed the process for mining the biological data and also discussed about the challenges faced during the extraction process **(Andreas Holzinger, 2014).** A brief discussion was done on the statistical and graph theoretical methods, data mining and machine learning methods. KDD was also used for mining the data. Firstly, integration of different data sets were done. Then the cleaning of the data and preprocessing of data was done by applying data mining methods. Although machine intelligence is the best way to deal with large amounts of data. But there are also many challenges that are faced in computational science. One of the challenges is in the guaranteeing the results, if a result has to be ensured from different sources then that result can be changed.

Kalpana Rangra, et, al 2014 reviewed the study of data mining tools **(Kalpana, 2014).** As data mining has number of applications ranging from marketing, artificial intelligence research, biological science etc. Due to its widespread use a large number of data mining tools have been developed over decades. Data mining offer us with effective means to access various types of data and information. Different applications of data mining also helps in decision making. Data mining provide many techniques to extract relevant data. Therefore in this paper theoretical analysis is done on the different data mining tools. Some of the tools are Weka, KEEL, R(revolution), KNIME, RAPIDMINER, ORANGE etc.

CHAPTER 3
PRESENT WORK

3.1 Problem Formulation

The work presented in the dissertation is based on the Fuzzy Association rule to mine proteins data. Here we extract rules to associates patterns to specific secondary structures of proteins. The time taken to scan data is very less and it form sets of combination on the bases of support and confidence. The main emphasis is to propose a fuzzy data mining technique to find the fuzzy association rules using fuzzy partition method and FP-growth on the proteins datasets.

3.2 Objectives of the study

The objectives of this dissertation proposal are as follows:

1. Study of big data and bioinformatics data

2. Study and analysis of various evolved approaches for mining bioinformatics data.

3. To model and propose an algorithm for mining proteomics using fuzzy association rule.

4. To implement the proposed algorithm in java and then validate its executive efficiency for proteomics data.

3.3 Scope of the study

As Bioinformatics is demarcated mostly as the study of the essential structure of biological data. It is the combination of biology and new computational resources. Here we attain large information of biological data, but still there is a more scope of information in the field of bioinformatics. A lot new types of software's are required and the existing systems are still prone to errors. Some software's that already exist are not able to provide sufficient and consistent information.

Some of the limitations accounted as:

- Analysis of the protein data in selecting the data mining method for bioinformatics data is very complex as bioinformatics data have different patterns and unstructured and to analyze them in a structured form is very difficult.

- The area of concern is the process of data input and the user interface that is used during the analysis process as analyzing of proteomics is a difficult task rather to amplify proteins than DNA.
- Biological data are enormously difficult because each data is made up of millions of cells and genes. Therefore the data we have is very complex.
- Data analysis and data mining of all data is a very staggering task due to the inconsistency and vagueness in the data generated.

To uncover the uncertainties in the data fuzzy computations can be considered for effective evaluation of the proteomic data.

3.4 Research Methodology

Research is a consistent method and orderly search for new and useful information on a particular subject. Research Methodology is one of the ways to interpret and resolve a research problem. It adds contribution to the existing knowledge.

In order to achieve the mentioned objectives, the effective methodologies are considered to complete this task as:

1. In order to achieve the objective “Study of big data and bioinformatics data” a comprehensive literature survey was carried out and different challenges were observed that exist during extraction and managing of bioinformatics data.

2. In order to achieve the objective “Study and analysis of various evolved approaches for mining bioinformatics data.” Fuzzy association rule is used to analyse and mine the proteomics data.

3. In order to achieve the objective “To model and proposed an algorithm for mining proteomics using fuzzy association rule.” the work is lead as

Step 1: Firstly input the protein dataset.

Step 2: Next convert the protein data into the binary format using file converter.

Step3: Make a schema file according to the protein data and then load the binary dataset.

Step4: The load data is mined for making the rules and sets by applying the proposed algorithm in which the concepts of Association rule Mining has been used.

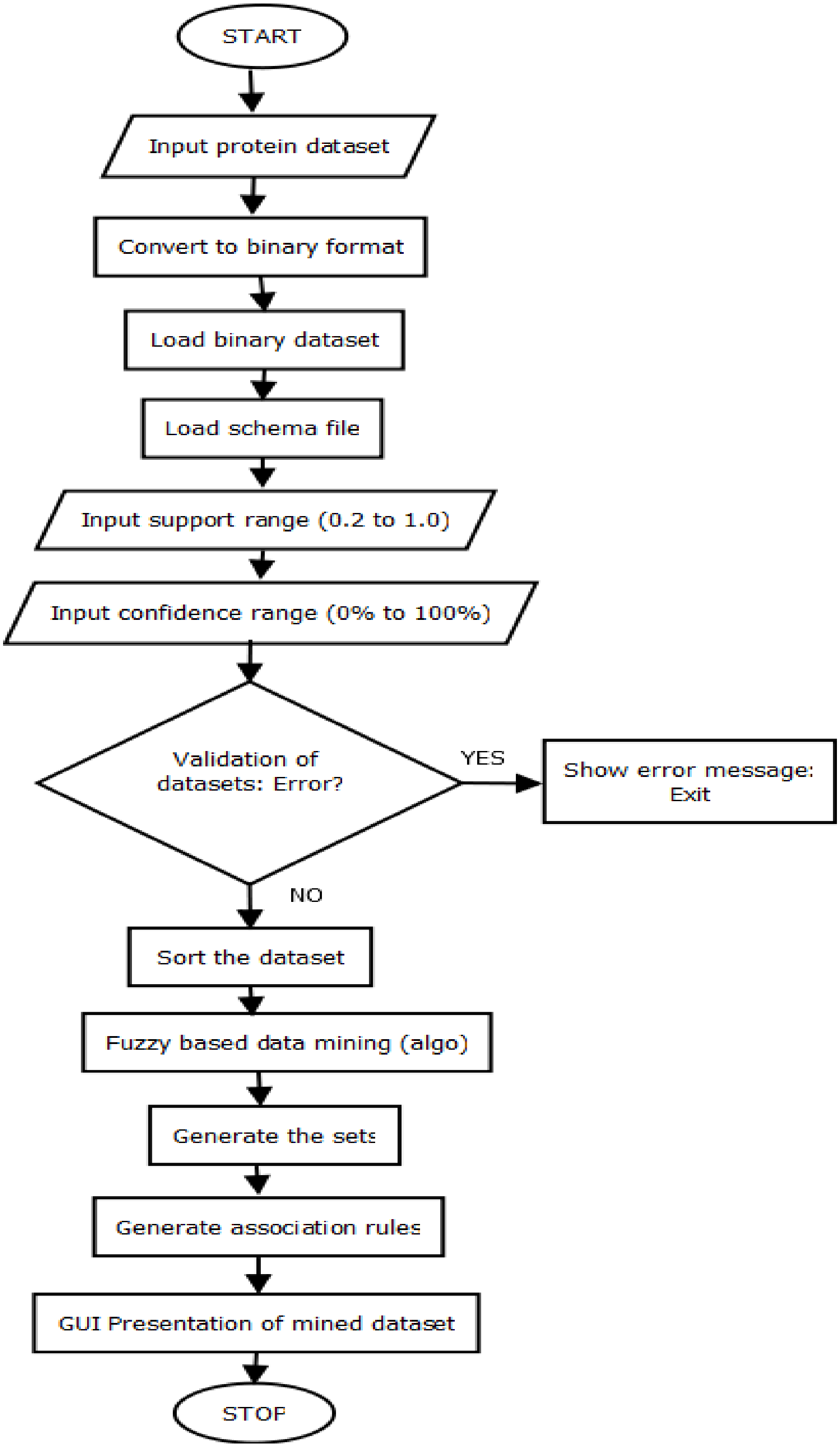

START
Input protein dataset
Convert to binary format
Load binary dataset
Load schema file
Input support range (0.2 to 1.0)
Input confidence range (0% to 100%)
Validation of datasets: Error?
YES
Show error message: Exit
NO
Sort the dataset
Fuzzy based data mining (algo)
Generate the sets
Generate association rules
GUI Presentation of mined dataset
STOP

Figure3.1: Flow of the methodology

Association rule mining: It is the method of achieving the relationship between the variables. It helps in finding the patterns and associations. As it is a two-step process that are find the frequent item sets and used the item sets based on the support and confidence value. Let *T* transaction contains an item *x* if $x \in T$. Let an item set *X* occurs in a transaction *T* if $X \subseteq T$. Let a dataset *D* of transactions and item set *X* be given. The dataset cardinality is denoted by |*D*|. The count of *X* in *D* is denoted by $count^D(X)$. It is the number of transactions in *D* that contains *X*. The support of X in D is denoted by $\sup port^D(X)$. It is the percentage of transactions in D that contain X.

$$\sup port^D\,(\mathrm{X}) = \frac{|\{T \in D \;|X \subset T\}\;|}{|D\;|} \quad \ldots(i)$$

An association rule is a pair as, $X \to Y$ where X and Y are two item sets and $X \bigcap Y = \Phi$. The item set X is called the antecedent of the rule. The item set Y is called the consequent of the rule.

- Support: Support can be defined as, $\sup port^D(X \to Y) = \sup port^D\left(X \bigcup Y\right)$
- Confidence: Confidence of the rule is defined as percentage of transactions in D containing X that also contain Y.

$$confidence^D(X \to Y) = \frac{\sup port^D\left(X \bigcup Y\right)}{\sup port^D(X)} = \frac{count^D\left(X \bigcup Y\right)}{count^D(X)} \quad \ldots(ii)$$

Step5: Input the support range between (0.2 to 1.0) and confidence range between (0% to 100%). Next validation of datasets should be analysed that is format of data item and schema file must be according to the protein datasets.

Step6: Sorting is done and fuzzy data mining algorithm is applied which form the sets and generate the different association rules which are based on the minimum support and confidence value.

Fuzzy association rule mining: it decompose into two following phases:

Phase I: this finds the all frequents fuzzy item sets <P, A> from the input protein data.

Phase II: this phase generates the all fuzzy association rules that are based on the fuzzy confident and support value. It is straightforward and takes less time comparing to the above step. If the <P, A> is a frequent fuzzy item set, therefore the rules formed are P' is A' $\rightarrow$ A\A' , in which P' and A' are non-empty subsets of P and A. This (\) sign denotes the subtraction operator between two sets.

Suppose the inputs of the algorithm used are a database P with attribute set M and record set R, and fconf and fsup.

The output are the fuzzy confident association rules.

The algorithm used is as follows:

```
BEGIN
    $(P_f, M_f, R_f)$= fuzzyMaterialization of (P,M,R)
    $F_1$ =Count($P_f$, $M_f$, $R_f$, fsup)
    k=2
    while ($F_{k-1} \neq \phi$ ) {
        $C_k$ = join($F_{k-1}$);
        $F_k$ = checking ( $C_k$, $P_f$, fsup);
        $F = F \cup F_k$;
        k= k+1;
    }
    GenerateRules (f, fconf);
END
```

Some of the notations used in the fuzzy association rule mining algorithm are:

- P : a transactional database
- M : attribute set in P
- R : record set in P
- P_f : output database after applying fuzzification over the original database
- M_f : set of fuzzy attributes in P_f.
- R_f : set of fuzzy record in P_f

- C_k : set of fuzzy k-item set candidate
- F_k : set of frequent fuzzy k-itemsets
- F : set of all frequent fuzzy item sets from database P_f
- fsup : fuzzy support
- fconf : fuzzy confidence

The above algorithm uses the following subprograms:

- **(P_f, M_f , R_f)= fuzzyMaterialization of (P,M,R)**: this function is used to convert the original database into the fuzzified database $P_{f.}$ After M and R are also converted to M_f and $R_{f.}$
- **F_1 =Count(P_f, M_f , R_f, fsup):** this function generate the all frequent fuzzy 1- item sets. All elements in F_1 must have support equal or greater to fsup.
- **C_k** = **Join**(**F_{k-1}**) : this function used to generate set of all fuzzy candidate k- item sets that are discovered in the previous step.
- **F_k = checking (C_k, P_f, fsup):** this part firstly scan the whole database and check candidate itemset whose support is smaller than fsup.
- **GenerateRules (f, fconf):** It generates rules possible confident fuzzy association rules.

Step7: At last GUI presentation of mined datasets are showed.

CHAPTER 4

RESULTS AND DISCUSSIONS

The results obtained after using the proposed algorithm are as follows:

Step1: Firstly the protein datasets are used.

CAATTGA
AGCTAGC
GTTAATG
TTGTTAT
CTTCCTC
ACATCTA
ATCTAGA
ATTATTG
TAGCTAA
CAATTGA
AGCTAGC
GTTAATG
TTGTTAT
CTTCCTC
ACATCTA
ATCTAGA
ATTATTG
TAGCTAA
CAATTGA
AGCTAGC
GTTAATG

TTGTTAT
CTTCCTC
ACATCTA
ATCTAGA
ATTATTG
TAGCTAA
CAATTGA
AGCTAGC
GTTAATG
TTGTTAT
CTTCCTC
ACATCTA
ATCTAGA
ATTATTG
TAGCTAA
CAATTGA
AGCTAGC
GTTAATG
TTGTTAT
CTTCCTC
ACATCTA

ATCTAGA
ATTATTG
TAGCTAA
CAATTGA
AGCTAGC
GTTAATG
TTGTTAT
CTTCCTC
ACATCTA
ATCTAGA
ATTATTG
TAGCTAA
CAATTGA
AGCTAGC
GTTAATG
TTGTTAT
CTTCCTC
ACATCTA
ATCTAGA
ATTATTG
TAGCTAA

Step 2: Next convert the protein data into the binary format using file converter. Therefore it shows the numeric values for each alphabetic letter. For C it shows <1, 0.3> , for A it shows <2,0.1>, for T it shows <4,2> and for G it shows <6,0.7>. Here for example in this numeric value <1,0.3>, the 1 is the id of C and 0.3 is the value of C respectively

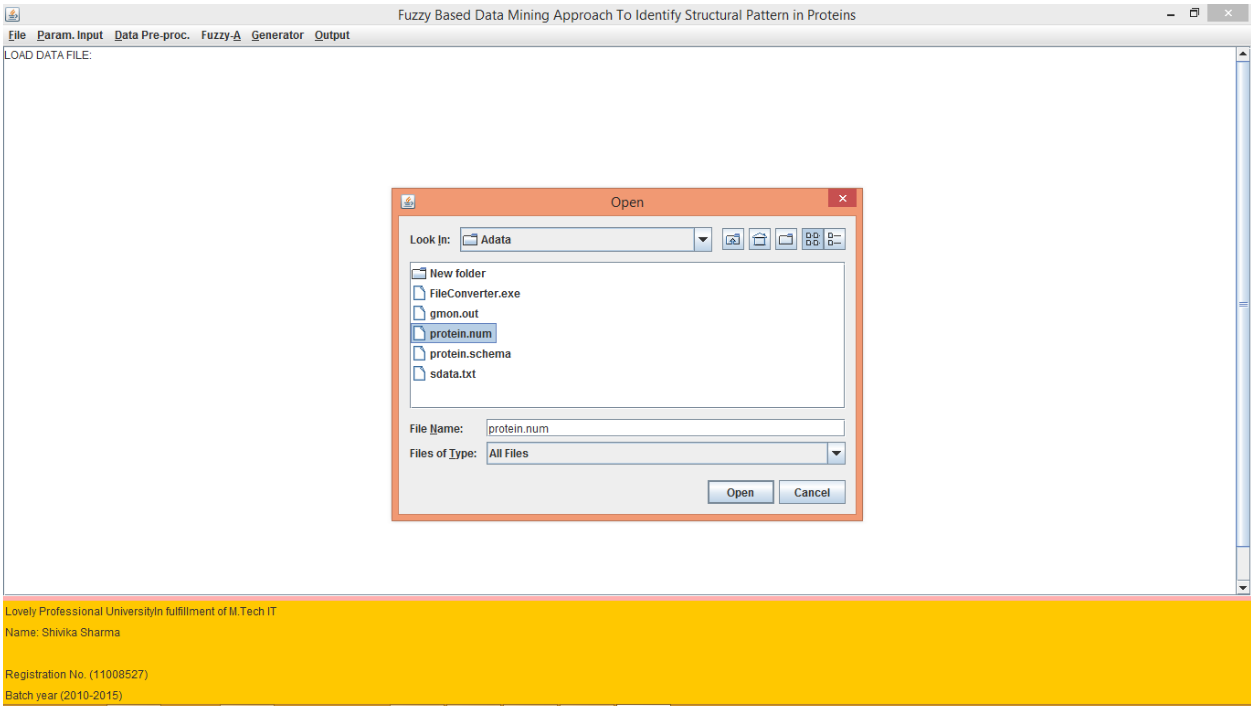
<1,0.3> <2,0.1> <3,0.1> <4,2> <5,2> <6,0.7> <7,0.1>
<1,0.1> <2,0.7> <3,0.3> <4,2> <5,0.1> <6,0.7> <7,0.3>
<1,0.7> <2,2> <3,2> <4,0.1> <5,0.1> <6,2> <7,0.7>
<1,2> <2,2> <3,0.7> <4,2> <5,2> <6,0.1> <7,2>

<1,0.3> <2,2> <3,2> <4,0.3> <5,0.3> <6,2> <7,0.3>
<1,0.1> <2,0.3> <3,0.1> <4,2> <5,0.3> <6,2> <7,0.1>
<1,0.1> <2,2> <3,0.3> <4,2> <5,0.1> <6,0.7> <7,0.1>
<1,0.1> <2,2> <3,2> <4,0.1> <5,2> <6,2> <7,0.7>

```
<1,2> <2,0.1> <3,0.7> <4,0.3> <5,2> <6,0.1>
<7,0.1>
<1,0.3> <2,0.1> <3,0.1> <4,2> <5,2> <6,0.7>
<7,0.1>
<1,0.1> <2,0.7> <3,0.3> <4,2> <5,0.1> <6,0.7>
<7,0.3>
<1,0.7> <2,2> <3,2> <4,0.1> <5,0.1> <6,2>
<7,0.7>
<1,2> <2,2> <3,0.7> <4,2> <5,2> <6,0.1>
<7,2>
<1,0.3> <2,2> <3,2> <4,0.3> <5,0.3> <6,2>
<7,0.3>
<1,0.1> <2,0.3> <3,0.1> <4,2> <5,0.3> <6,2>
<7,0.1>
<1,0.1> <2,2> <3,0.3> <4,2> <5,0.1> <6,0.7>
<7,0.1>
<1,0.1> <2,2> <3,2> <4,0.1> <5,2> <6,2>
<7,0.7>
<1,2> <2,0.1> <3,0.7> <4,0.3> <5,2> <6,0.1>
<7,0.1>
<1,0.3> <2,0.1> <3,0.1> <4,2> <5,2> <6,0.7>
<7,0.1>
<1,0.1> <2,0.7> <3,0.3> <4,2> <5,0.1> <6,0.7>
<7,0.3>
<1,0.7> <2,2> <3,2> <4,0.1> <5,0.1> <6,2>
<7,0.7>
<1,2> <2,2> <3,0.7> <4,2> <5,2> <6,0.1>
<7,2>
<1,0.3> <2,2> <3,2> <4,0.3> <5,0.3> <6,2>
<7,0.3>
<1,0.1> <2,0.3> <3,0.1> <4,2> <5,0.3> <6,2>
<7,0.1>
<1,0.1> <2,2> <3,0.3> <4,2> <5,0.1> <6,0.7>
<7,0.1>
<1,0.1> <2,2> <3,2> <4,0.1> <5,2> <6,2>
<7,0.7>
<1,2> <2,0.1> <3,0.7> <4,0.3> <5,2> <6,0.1>
<7,0.1>
<1,0.3> <2,0.1> <3,0.1> <4,2> <5,2> <6,0.7>
<7,0.1>
<1,0.1> <2,0.7> <3,0.3> <4,2> <5,0.1> <6,0.7>
<7,0.3>
<1,0.7> <2,2> <3,2> <4,0.1> <5,0.1> <6,2>
<7,0.7>
<1,2> <2,2> <3,0.7> <4,2> <5,2> <6,0.1>
<7,2>
<1,0.3> <2,2> <3,2> <4,0.3> <5,0.3> <6,2>
<7,0.3>
<1,0.1> <2,0.3> <3,0.1> <4,2> <5,0.3> <6,2>
<7,0.1>
<1,0.1> <2,2> <3,0.3> <4,2> <5,0.1> <6,0.7>
<7,0.1>
<1,0.1> <2,2> <3,2> <4,0.1> <5,2> <6,2>
<7,0.7>
<1,2> <2,0.1> <3,0.7> <4,0.3> <5,2> <6,0.1>
<7,0.1>
<1,0.3> <2,0.1> <3,0.1> <4,2> <5,2> <6,0.7>
<7,0.1>
<1,0.1> <2,0.7> <3,0.3> <4,2> <5,0.1> <6,0.7>
<7,0.3>
<1,0.7> <2,2> <3,2> <4,0.1> <5,0.1> <6,2>
<7,0.7>
<1,2> <2,2> <3,0.7> <4,2> <5,2> <6,0.1>
<7,2>
<1,0.3> <2,2> <3,2> <4,0.3> <5,0.3> <6,2>
<7,0.3>
<1,0.1> <2,0.3> <3,0.1> <4,2> <5,0.3> <6,2>
<7,0.1>
<1,0.1> <2,2> <3,0.3> <4,2> <5,0.1> <6,0.7>
<7,0.1>
<1,0.1> <2,2> <3,2> <4,0.1> <5,2> <6,2>
<7,0.7>
<1,2> <2,0.1> <3,0.7> <4,0.3> <5,2> <6,0.1>
<7,0.1>
<1,0.3> <2,0.1> <3,0.1> <4,2> <5,2> <6,0.7>
<7,0.1>
<1,0.1> <2,0.7> <3,0.3> <4,2> <5,0.1> <6,0.7>
<7,0.3>
<1,0.7> <2,2> <3,2> <4,0.1> <5,0.1> <6,2>
<7,0.7>
<1,2> <2,2> <3,0.7> <4,2> <5,2> <6,0.1>
<7,2>
```

Step3: Make a schema file according to the protein data and then load the binary dataset and schema file on the framework that is used in this work.

protein_A
protein_B
protein_C
protein_D
protein_E
protein_F
protein_G



Look In: Adata

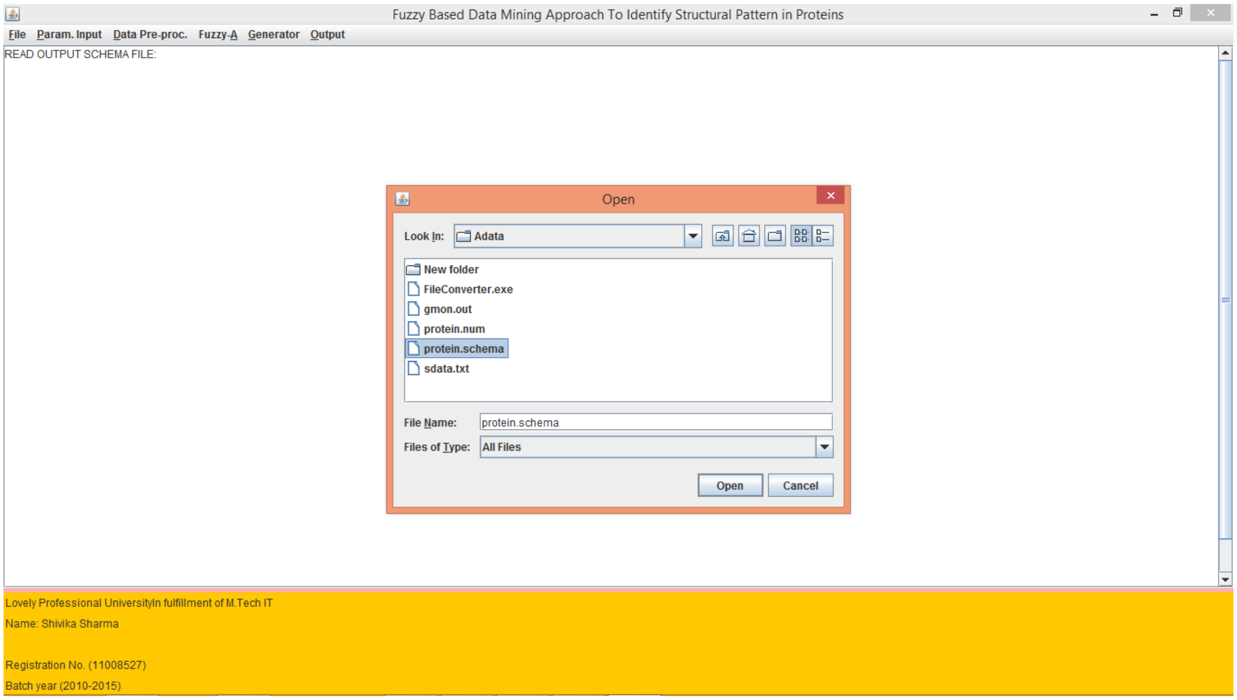
File Name: protein.num

Files of Type: All Files

Lovely Professional UniversityIn fulfillment of M.Tech IT
Name: Shivika Sharma

Registration No. (11008527)
Batch year (2010-2015)

Figure 4.1: Load binary data of protein



```
READ OUTPUT SCHEMA FILE:
```

Lovely Professional UniversityIn fulfillment of M.Tech IT

Name: Shivika Sharma

Registration No. (11008527)

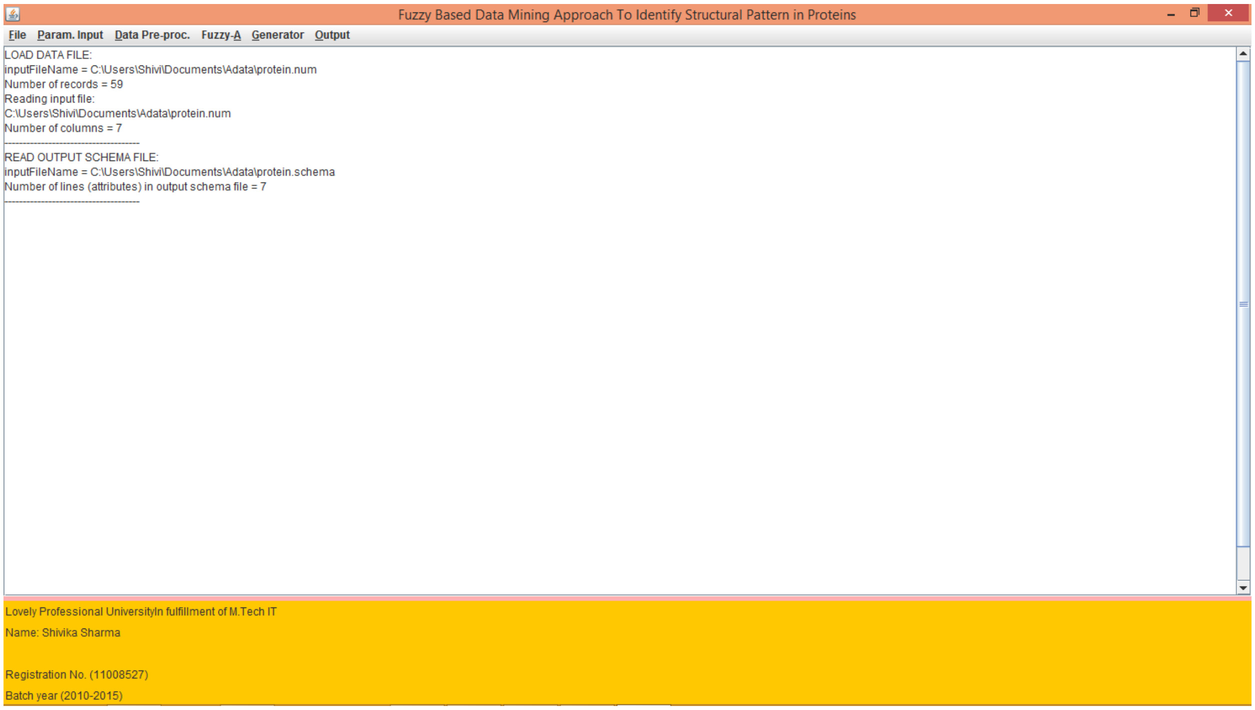
Batch year (2010-2015)

Figure 4.2: Load schema file of protein

In below figure the number of records are 59 and number of column are 7.

Number of attributes in output schema file are 7.

Step4: The load data is mined for making the rules and sets by applying the proposed algorithm in which the concepts of Association rule Mining has been used.



Fuzzy Based Data Mining Approach To Identify Structural Pattern in Proteins
File
Param. Input
Data Pre-proc.
Fuzzy-A
Generator
Output

```
LOAD DATA FILE:
inputFileName = C:\Users\Shivi\Documents\Adata\protein.num
Number of records = 59
Reading input file:
C:\Users\Shivi\Documents\Adata\protein.num
Number of columns = 7
----------------------------------------
READ OUTPUT SCHEMA FILE:
inputFileName = C:\Users\Shivi\Documents\Adata\protein.schema
Number of lines (attributes) in output schema file = 7
----------------------------------------
```

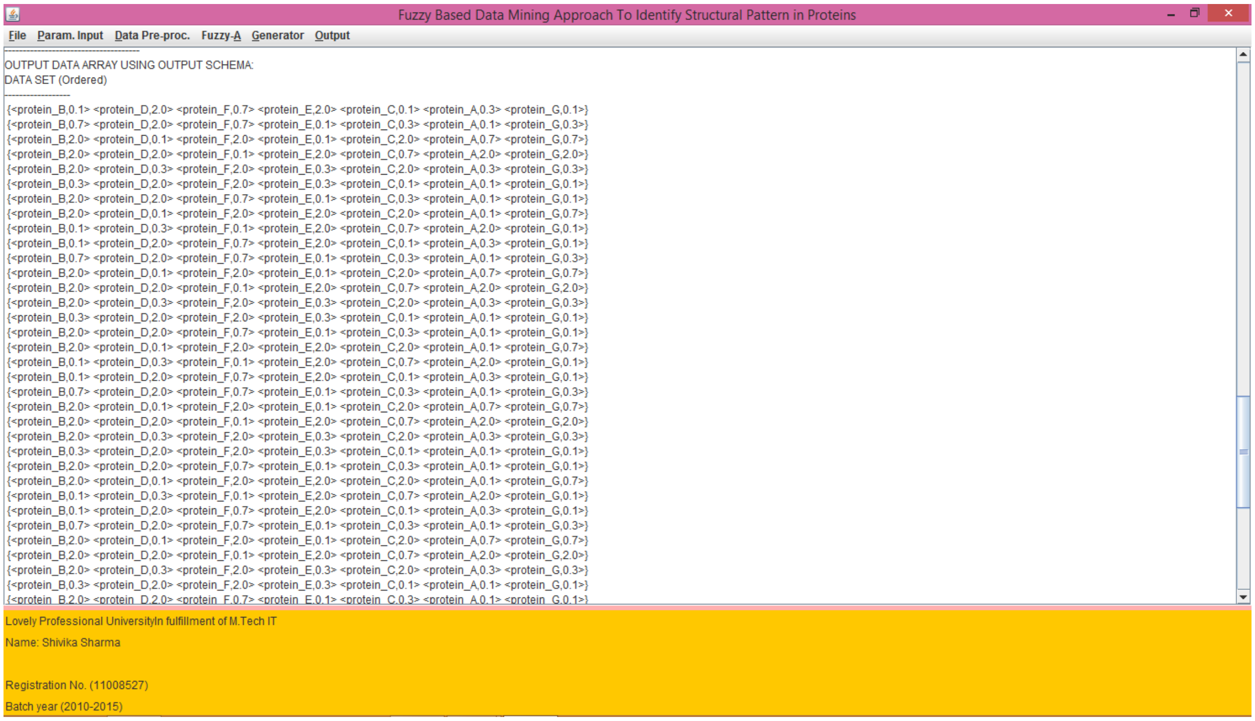
Lovely Professional UniversityIn fulfillment of M.Tech IT

Name: Shivika Sharma

Registration No. (11008527)

Batch year (2010-2015)

Figure 4.3: Reading output files


Fuzzy Based Data Mining Approach To Identify Structural Pattern in Proteins
File
Param. Input
Data Pre-proc.
Fuzzy-A
Generator
Output

```
--------------------------------------
OUTPUT DATA ARRAY USING OUTPUT SCHEMA:
DATA SET (Ordered)
------------------
{<protein_B,0.1> <protein_D,2.0> <protein_F,0.7> <protein_E,2.0> <protein_C,0.1> <protein_A,0.3> <protein_G,0.1>}
{<protein_B,0.7> <protein_D,2.0> <protein_F,0.7> <protein_E,0.1> <protein_C,0.3> <protein_A,0.1> <protein_G,0.3>}
{<protein_B,2.0> <protein_D,0.1> <protein_F,2.0> <protein_E,0.1> <protein_C,2.0> <protein_A,0.7> <protein_G,0.7>}
{<protein_B,2.0> <protein_D,2.0> <protein_F,0.1> <protein_E,2.0> <protein_C,0.7> <protein_A,2.0> <protein_G,2.0>}
{<protein_B,2.0> <protein_D,0.3> <protein_F,2.0> <protein_E,0.3> <protein_C,2.0> <protein_A,0.3> <protein_G,0.3>}
{<protein_B,0.3> <protein_D,2.0> <protein_F,2.0> <protein_E,0.3> <protein_C,0.1> <protein_A,0.1> <protein_G,0.1>}
{<protein_B,2.0> <protein_D,2.0> <protein_F,0.7> <protein_E,0.1> <protein_C,0.3> <protein_A,0.1> <protein_G,0.1>}
{<protein_B,2.0> <protein_D,0.1> <protein_F,2.0> <protein_E,2.0> <protein_C,2.0> <protein_A,0.1> <protein_G,0.7>}
{<protein_B,0.1> <protein_D,0.3> <protein_F,0.1> <protein_E,2.0> <protein_C,0.7> <protein_A,2.0> <protein_G,0.1>}
{<protein_B,0.1> <protein_D,2.0> <protein_F,0.7> <protein_E,2.0> <protein_C,0.1> <protein_A,0.3> <protein_G,0.1>}
{<protein_B,0.7> <protein_D,2.0> <protein_F,0.7> <protein_E,0.1> <protein_C,0.3> <protein_A,0.1> <protein_G,0.3>}
{<protein_B,2.0> <protein_D,0.1> <protein_F,2.0> <protein_E,0.1> <protein_C,2.0> <protein_A,0.7> <protein_G,0.7>}
{<protein_B,2.0> <protein_D,2.0> <protein_F,0.1> <protein_E,2.0> <protein_C,0.7> <protein_A,2.0> <protein_G,2.0>}
{<protein_B,2.0> <protein_D,0.3> <protein_F,2.0> <protein_E,0.3> <protein_C,2.0> <protein_A,0.3> <protein_G,0.3>}
{<protein_B,0.3> <protein_D,2.0> <protein_F,2.0> <protein_E,0.3> <protein_C,0.1> <protein_A,0.1> <protein_G,0.1>}
{<protein_B,2.0> <protein_D,2.0> <protein_F,0.7> <protein_E,0.1> <protein_C,0.3> <protein_A,0.1> <protein_G,0.1>}
{<protein_B,2.0> <protein_D,0.1> <protein_F,2.0> <protein_E,2.0> <protein_C,2.0> <protein_A,0.1> <protein_G,0.7>}
{<protein_B,0.1> <protein_D,0.3> <protein_F,0.1> <protein_E,2.0> <protein_C,0.7> <protein_A,2.0> <protein_G,0.1>}
{<protein_B,0.1> <protein_D,2.0> <protein_F,0.7> <protein_E,2.0> <protein_C,0.1> <protein_A,0.3> <protein_G,0.1>}
{<protein_B,0.7> <protein_D,2.0> <protein_F,0.7> <protein_E,0.1> <protein_C,0.3> <protein_A,0.1> <protein_G,0.3>}
{<protein_B,2.0> <protein_D,0.1> <protein_F,2.0> <protein_E,0.1> <protein_C,2.0> <protein_A,0.7> <protein_G,0.7>}
{<protein_B,2.0> <protein_D,2.0> <protein_F,0.1> <protein_E,2.0> <protein_C,0.7> <protein_A,2.0> <protein_G,2.0>}
{<protein_B,2.0> <protein_D,0.3> <protein_F,2.0> <protein_E,0.3> <protein_C,2.0> <protein_A,0.3> <protein_G,0.3>}
{<protein_B,0.3> <protein_D,2.0> <protein_F,2.0> <protein_E,0.3> <protein_C,0.1> <protein_A,0.1> <protein_G,0.1>}
{<protein_B,2.0> <protein_D,2.0> <protein_F,0.7> <protein_E,0.1> <protein_C,0.3> <protein_A,0.1> <protein_G,0.1>}
{<protein_B,2.0> <protein_D,0.1> <protein_F,2.0> <protein_E,2.0> <protein_C,2.0> <protein_A,0.1> <protein_G,0.7>}
{<protein_B,0.1> <protein_D,0.3> <protein_F,0.1> <protein_E,2.0> <protein_C,0.7> <protein_A,2.0> <protein_G,0.1>}
{<protein_B,0.1> <protein_D,2.0> <protein_F,0.7> <protein_E,2.0> <protein_C,0.1> <protein_A,0.3> <protein_G,0.1>}
{<protein_B,0.7> <protein_D,2.0> <protein_F,0.7> <protein_E,0.1> <protein_C,0.3> <protein_A,0.1> <protein_G,0.3>}
{<protein_B,2.0> <protein_D,0.1> <protein_F,2.0> <protein_E,0.1> <protein_C,2.0> <protein_A,0.7> <protein_G,0.7>}
{<protein_B,2.0> <protein_D,2.0> <protein_F,0.1> <protein_E,2.0> <protein_C,0.7> <protein_A,2.0> <protein_G,2.0>}
{<protein_B,2.0> <protein_D,0.3> <protein_F,2.0> <protein_E,0.3> <protein_C,2.0> <protein_A,0.3> <protein_G,0.3>}
{<protein_B,0.3> <protein_D,2.0> <protein_F,2.0> <protein_E,0.3> <protein_C,0.1> <protein_A,0.1> <protein_G,0.1>}
{<protein_B,2.0> <protein_D,2.0> <protein_F,0.7> <protein_E,0.1> <protein_C,0.3> <protein_A,0.1> <protein_G,0.1>}
```

Lovely Professional UniversityIn fulfillment of M.Tech IT

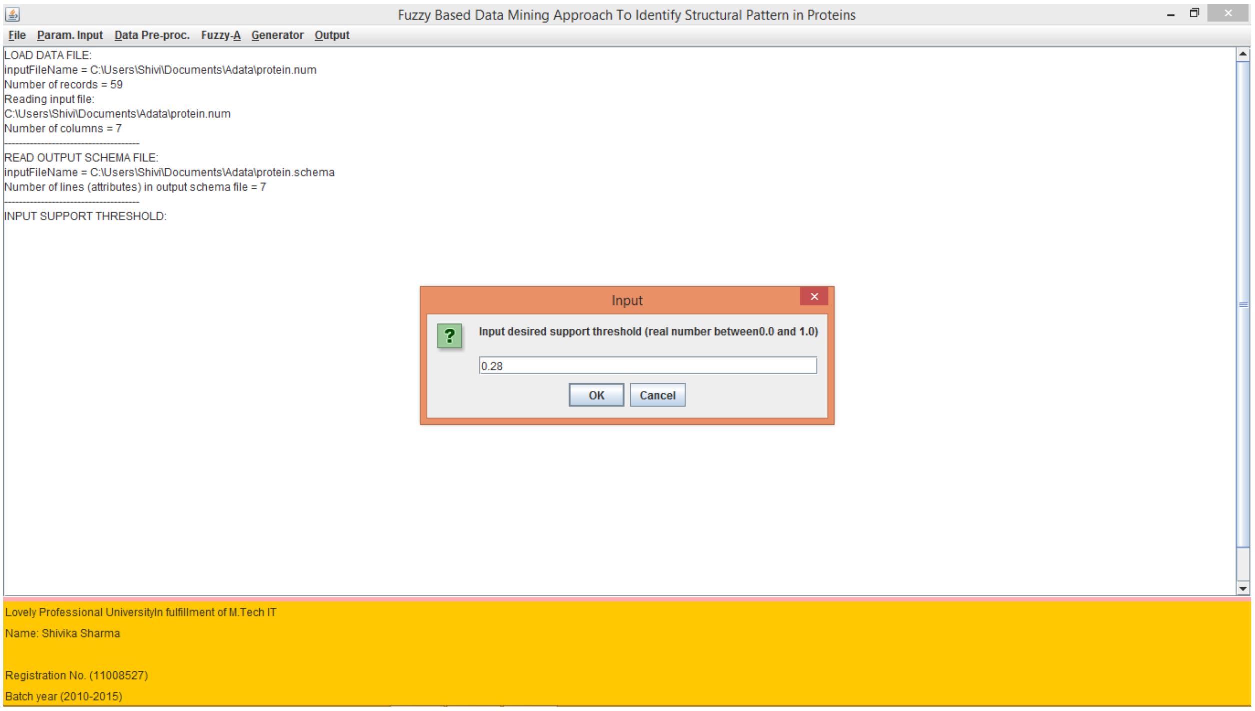
Name: Shivika Sharma

Registration No. (11008527)

Batch year (2010-2015)

Figure 4.4: Output data array using schema file

Step5: Input the support range between (0.2to1.0) and confidence range between (0%to100%). Next validation of datasets should be analysed that is format of data item and schema file must be according to the protein datasets.


Fuzzy Based Data Mining Approach To Identify Structural Pattern in Proteins
File
Param. Input
Data Pre-proc.
Fuzzy-A
Generator
Output
Input
?
OK
Cancel

```
LOAD DATA FILE:
inputFileName = C:\Users\Shivi\Documents\Adata\protein.num
Number of records = 59
Reading input file:
C:\Users\Shivi\Documents\Adata\protein.num
Number of columns = 7
--------------------------------------
READ OUTPUT SCHEMA FILE:
inputFileName = C:\Users\Shivi\Documents\Adata\protein.schema
Number of lines (attributes) in output schema file = 7
--------------------------------------
INPUT SUPPORT THRESHOLD:
```

Input desired support threshold (real number between0.0 and 1.0)

0.28

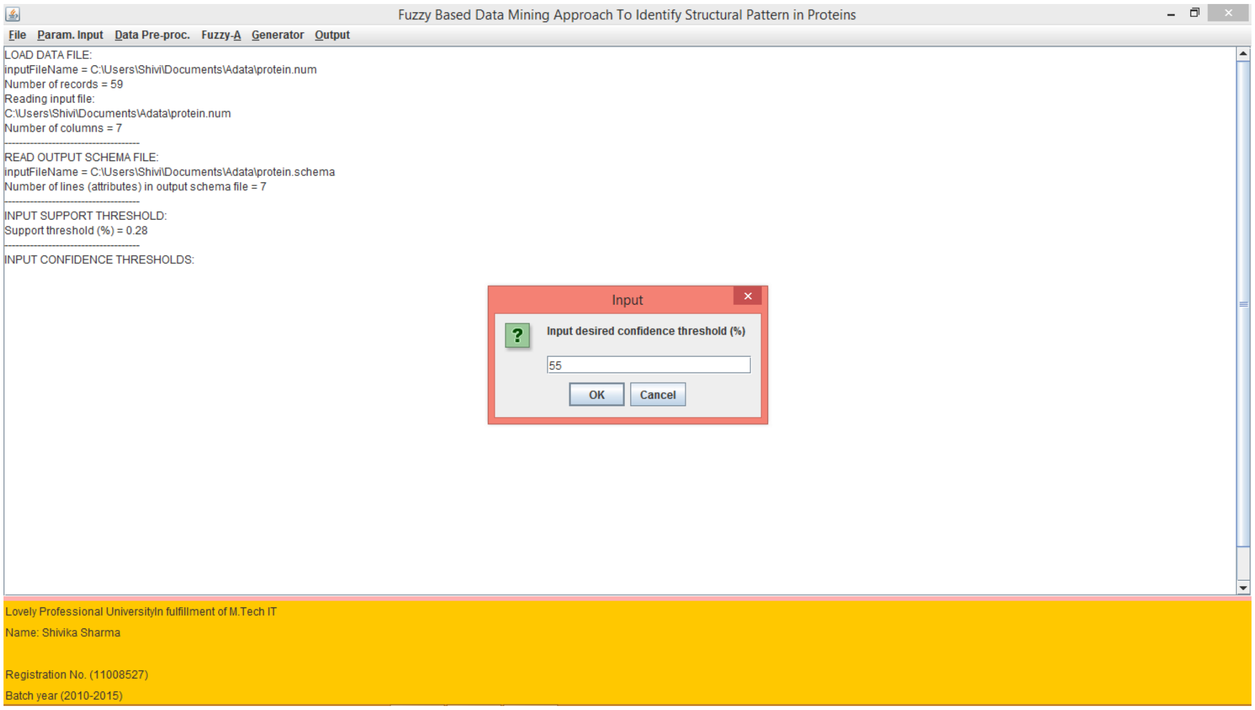
Lovely Professional UniversityIn fulfillment of M.Tech IT

Name: Shivika Sharma

Registration No. (11008527)

Batch year (2010-2015)

Figure 4.5: Input desired support


Fuzzy Based Data Mining Approach To Identify Structural Pattern in Proteins
File
Param. Input
Data Pre-proc.
Fuzzy-A
Generator
Output
Input
OK
Cancel

```
LOAD DATA FILE:
inputFileName = C:\Users\Shivi\Documents\Adata\protein.num
Number of records = 59
Reading input file:
C:\Users\Shivi\Documents\Adata\protein.num
Number of columns = 7
-----------------------------------
READ OUTPUT SCHEMA FILE:
inputFileName = C:\Users\Shivi\Documents\Adata\protein.schema
Number of lines (attributes) in output schema file = 7
-----------------------------------
INPUT SUPPORT THRESHOLD:
Support threshold (%) = 0.28
-----------------------------------
INPUT CONFIDENCE THRESHOLDS:
```

Input desired confidence threshold (%)

55

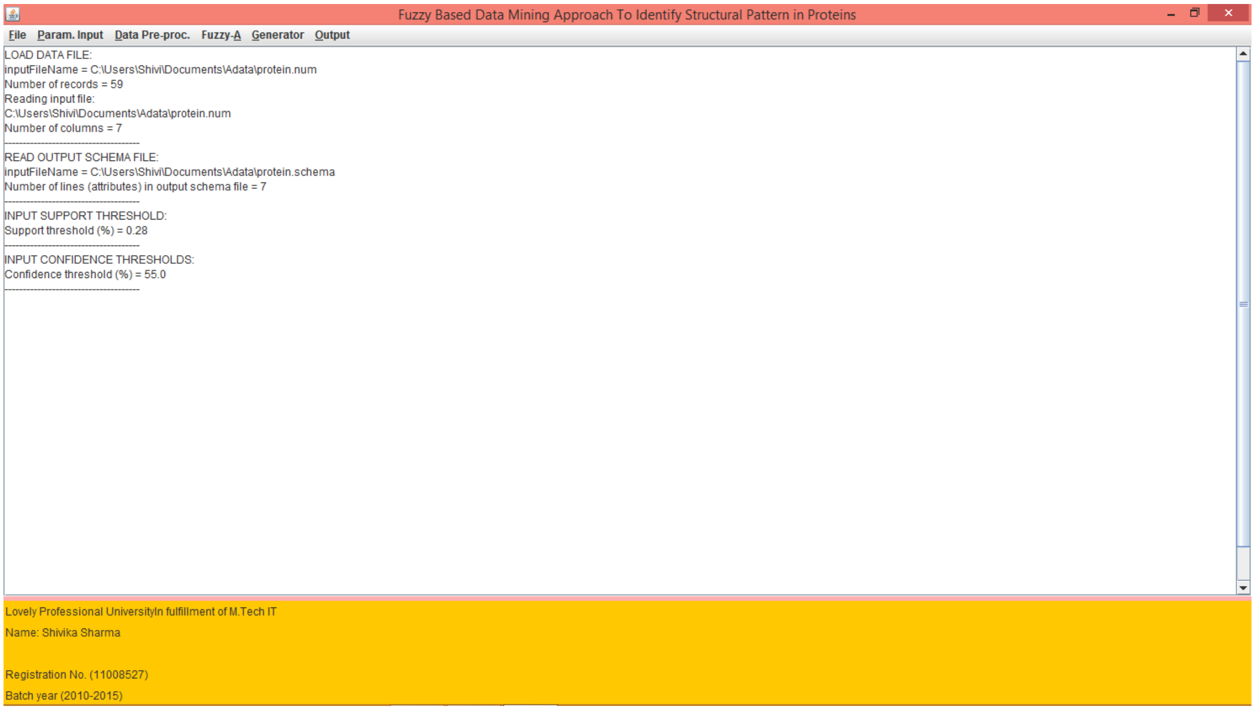
Lovely Professional UniversityIn fulfillment of M.Tech IT

Name: Shivika Sharma

Registration No. (11008527)

Batch year (2010-2015)

Figure 4.6: Input desired confidence


Fuzzy Based Data Mining Approach To Identify Structural Pattern in Proteins
File
Param. Input
Data Pre-proc.
Fuzzy-A
Generator
Output

```
LOAD DATA FILE:
inputFileName = C:\Users\Shivi\Documents\Adata\protein.num
Number of records = 59
Reading input file:
C:\Users\Shivi\Documents\Adata\protein.num
Number of columns = 7
-----------------------------------
READ OUTPUT SCHEMA FILE:
inputFileName = C:\Users\Shivi\Documents\Adata\protein.schema
Number of lines (attributes) in output schema file = 7
-----------------------------------
INPUT SUPPORT THRESHOLD:
Support threshold (%) = 0.28
-----------------------------------
INPUT CONFIDENCE THRESHOLDS:
Confidence threshold (%) = 55.0
-----------------------------------
```

Lovely Professional UniversityIn fulfillment of M.Tech IT

Name: Shivika Sharma

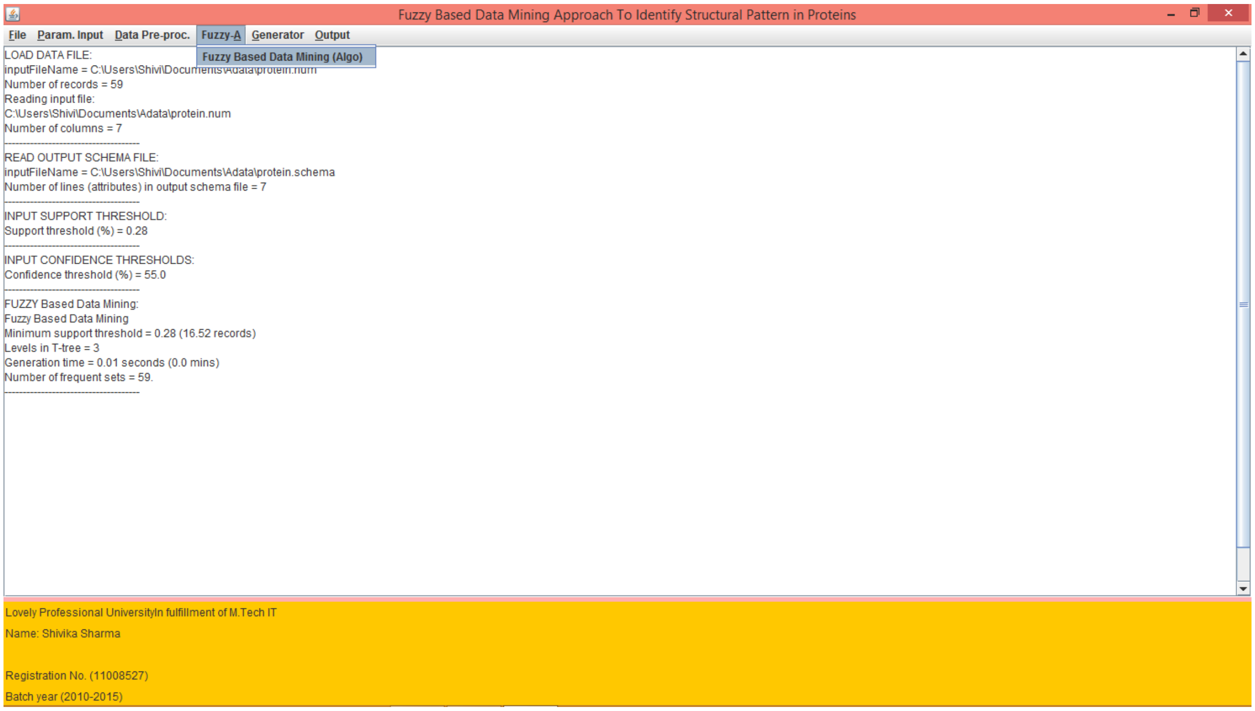
Registration No. (11008527)

Batch year (2010-2015)

Figure 4.7: Reading files with support and confidence

Step6: Sorting is done and fuzzy data mining algorithm is applied which form the sets and generate the different association rules which are based on the minimum support and confidence value.

By applying the fuzzy based data mining algorithm the number of frequent sets formed are 59 and generation time is 0.01 seconds. The minimum support threshold is 0.28 which form 16.52 records.


Fuzzy Based Data Mining Approach To Identify Structural Pattern in Proteins
File
Param. Input
Data Pre-proc.
Fuzzy-A
Generator
Output
Fuzzy Based Data Mining (Algo)

```
LOAD DATA FILE:
inputFileName = C:\Users\Shivi\Documents\Adata\protein.num
Number of records = 59
Reading input file:
C:\Users\Shivi\Documents\Adata\protein.num
Number of columns = 7
-----------------------------------
READ OUTPUT SCHEMA FILE:
inputFileName = C:\Users\Shivi\Documents\Adata\protein.schema
Number of lines (attributes) in output schema file = 7
-----------------------------------
INPUT SUPPORT THRESHOLD:
Support threshold (%) = 0.28
-----------------------------------
INPUT CONFIDENCE THRESHOLDS:
Confidence threshold (%) = 55.0
-----------------------------------
FUZZY Based Data Mining:
Fuzzy Based Data Mining
Minimum support threshold = 0.28 (16.52 records)
Levels in T-tree = 3
Generation time = 0.01 seconds (0.0 mins)
Number of frequent sets = 59.
-----------------------------------
```

Lovely Professional UniversityIn fulfillment of M.Tech IT
Name: Shivika Sharma

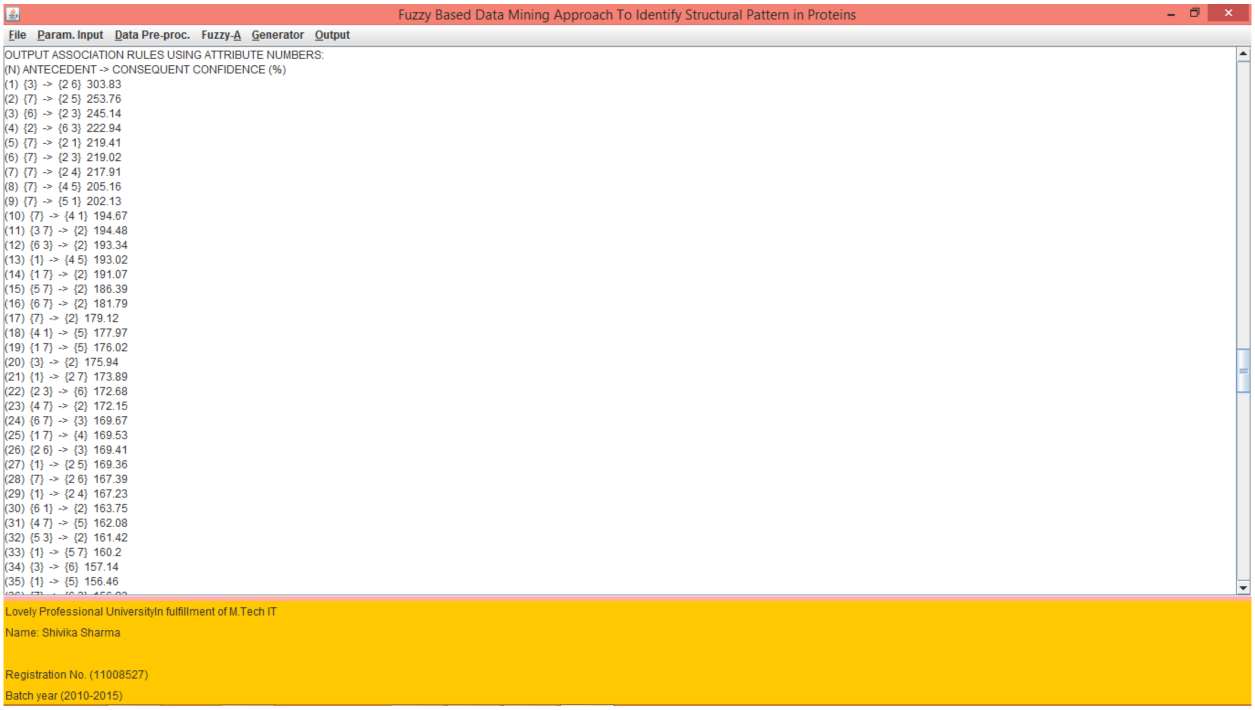
Registration No. (11008527)
Batch year (2010-2015)

Figure 4.8: Apply Fuzzy Association Rule

According to minimum confidence that is 55.0, it forms 193 rules.

(N) ANTECEDENT -> CONSEQUENT CONFIDENCE (%)
(1) {3} -> {2 6} 303.83
(2) {7} -> {2 5} 253.76
(3) {6} -> {2 3} 245.14
(4) {2} -> {6 3} 222.94
(5) {7} -> {2 1} 219.41
(6) {7} -> {2 3} 219.02
(7) {7} -> {2 4} 217.91
(8) {7} -> {4 5} 205.16
(9) {7} -> {5 1} 202.13
(10) {7} -> {4 1} 194.67
(11) {3 7} -> {2} 194.48
(12) {6 3} -> {2} 193.34
(13) {1} -> {4 5} 193.02
(14) {1 7} -> {2} 191.07
(15) {5 7} -> {2} 186.39
(16) {6 7} -> {2} 181.79
(17) {7} -> {2} 179.12
(18) {4 1} -> {5} 177.97
(19) {1 7} -> {5} 176.02
(20) {3} -> {2} 175.94
(21) {1} -> {2 7} 173.89
(22) {2 3} -> {6} 172.68
(23) {4 7} -> {2} 172.15
(24) {6 7} -> {3} 169.67
(25) {1 7} -> {4} 169.53
(26) {2 6} -> {3} 169.41
(27) {1} -> {2 5} 169.36
(28) {7} -> {2 6} 167.39
(29) {1} -> {2 4} 167.23
(30) {6 1} -> {2} 163.75
(31) {4 7} -> {5} 162.08
(32) {5 3} -> {2} 161.42
(33) {1} -> {5 7} 160.2
(34) {3} -> {6} 157.14
(35) {1} -> {5} 156.46
(36) {7} -> {6 3} 156.23
(37) {4 3} -> {2} 155.57
(38) {1} -> {4 7} 154.29
(39) {4 1} -> {2} 154.19
(40) {4 7} -> {1} 153.8
(41) {6 1} -> {3} 153.22
(42) {3 1} -> {2} 150.86
(43) {5 7} -> {4} 150.69
(44) {5 7} -> {1} 148.47
(45) {3} -> {2 5} 147.95
(46) {6} -> {2} 144.69
(47) {4 1} -> {7} 142.26
(48) {2 7} -> {5} 141.66
(49) {2 1} -> {7} 140.67
(50) {6 5} -> {2} 140.53
(51) {5} -> {2 3} 138.75
(52) {3 7} -> {6} 138.73
(53) {2 1} -> {5} 137.01
(54) {1} -> {2 3} 136.66
(55) {7} -> {5} 136.14
(56) {2 1} -> {4} 135.28
(57) {6 5} -> {3} 134.66

(58) {7} -> {5 3} 133.28
(59) {2} -> {6} 131.59
(60) {5} -> {2 7} 130.6
(61) {2} -> {3} 129.1
(62) {2 5} -> {3} 128.68
(63) {6} -> {3} 126.79
(64) {7} -> {4} 126.57
(65) {5 3} -> {6} 125.84
(66) {5} -> {4 1} 125.35
(67) {1} -> {2} 123.61
(68) {5 1} -> {4} 123.36
(69) {2 7} -> {1} 122.49
(70) {2 7} -> {3} 122.27
(71) {3 1} -> {5} 121.89
(72) {2 7} -> {4} 121.65
(73) {2 5} -> {7} 121.12
(74) {3} -> {2 7} 120.19
(75) {3 7} -> {5} 118.35
(76) {5} -> {4} 116.92
(77) {5} -> {2 4} 116.44
(78) {3} -> {6 5} 115.34
(79) {7} -> {1} 114.83
(80) {5} -> {2 6} 112.88
(81) {7} -> {3} 112.61
(82) {2 1} -> {3} 110.55
(83) {1} -> {5 3} 110.41
(84) {5} -> {2 1} 109.98
(85) {2} -> {5 3} 108.56
(86) {1} -> {4} 108.45
(87) {5 1} -> {2} 108.24
(88) {5} -> {6 3} 108.17
(89) {2 5} -> {4} 107.99
(90) {5} -> {2} 107.82
(91) {4 5} -> {1} 107.2
(92) {4 3} -> {5} 105.66
(93) {5} -> {4 7} 105.59
(94) {2 5} -> {6} 104.69
(95) {5} -> {1 7} 104.03
(96) {4} -> {2} 103.62
(97) {5 1} -> {7} 102.39
(98) {2} -> {5 7} 102.19
(99) {2 5} -> {1} 102.0
(100) {4} -> {5 1} 101.93
(101) {5} -> {1} 101.6
(102) {2} -> {4} 99.7
(103) {4 5} -> {2} 99.58
(104) {7} -> {3 1} 99.02
(105) {5 7} -> {3} 97.9
(106) {6} -> {2 5} 97.12
(107) {3 1} -> {6} 96.18
(108) {4} -> {5} 95.08
(109) {4} -> {2 5} 94.69
(110) {3} -> {2 1} 94.63
(111) {2 7} -> {6} 93.45
(112) {1} -> {2 6} 93.11
(113) {6} -> {5 3} 93.06
(114) {7} -> {6} 92.08
(115) {3} -> {5} 91.65
(116) {2 4} -> {5} 91.37
(117) {4 3} -> {1} 91.29
(118) {4} -> {2 7} 91.2
(119) {2} -> {4 5} 91.1
(120) {1} -> {7} 91.01
(121) {1} -> {3} 90.58
(122) {4 5} -> {7} 90.3
(123) {6} -> {4} 89.71
(124) {2} -> {1 7} 88.35
(125) {2} -> {6 5} 88.32
(126) {4} -> {2 1} 88.31
(127) {2} -> {3 7} 88.19
(128) {2 4} -> {7} 88.01
(129) {3 7} -> {1} 87.92
(130) {2} -> {4 7} 87.75
(131) {4 3} -> {7} 87.35
(132) {1} -> {6 3} 87.12
(133) {3 1} -> {7} 86.63
(134) {1 7} -> {3} 86.23
(135) {2} -> {5 1} 86.05
(136) {5} -> {3} 85.95
(137) {4} -> {5 7} 85.86
(138) {3} -> {6 7} 85.74
(139) {6 7} -> {5} 85.6
(140) {2 4} -> {1} 85.22
(141) {2} -> {4 1} 84.97
(142) {4} -> {6} 84.8
(143) {2} -> {5} 84.36
(144) {2 3} -> {5} 84.09
(145) {5 3} -> {1} 83.41
(146) {4 3} -> {6} 83.31
(147) {7} -> {4 3} 82.83
(148) {4} -> {1 7} 81.48
(149) {3} -> {2 4} 80.96
(150) {5} -> {6} 80.32
(151) {4 6} -> {2} 79.92
(152) {5 3} -> {7} 79.8
(153) {7} -> {6 5} 78.82
(154) {1} -> {3 7} 78.47
(155) {6 5} -> {4} 76.49
(156) {3} -> {5 1} 76.45
(157) {3 1} -> {4} 75.74
(158) {2 1} -> {6} 75.32
(159) {6} -> {2 7} 74.12
(160) {3 7} -> {4} 73.55
(161) {6 3} -> {5} 73.4
(162) {3} -> {5 7} 73.14
(163) {2} -> {7} 72.13
(164) {5} -> {3 1} 71.7
(165) {6} -> {2 4} 71.7
(166) {5 1} -> {3} 70.57
(167) {5} -> {7} 70.06
(168) {2} -> {3 1} 69.44
(169) {6} -> {3 7} 69.17
(170) {6} -> {5} 69.1
(171) {1} -> {4 3} 68.61
(172) {5} -> {3 7} 68.59
(173) {2 3} -> {7} 68.31
(174) {4} -> {2 6} 67.77
(175) {2} -> {6 7} 67.41
(176) {2 6} -> {5} 67.12
(177) {4 7} -> {3} 65.44
(178) {2 4} -> {6} 65.4
(179) {2} -> {4 6} 65.21
(180) {4 1} -> {3} 63.26
(181) {2} -> {1} 62.81
(182) {3} -> {1} 62.72
(183) {3} -> {7} 61.8
(184) {4} -> {2 3} 61.75
(185) {5} -> {4 6} 61.45
(186) {3} -> {6 1} 60.33
(187) {5 3} -> {4} 59.99
(188) {2 4} -> {3} 59.59
(189) {2} -> {4 3} 59.41
(190) {4 6} -> {5} 58.92
(191) {5 7} -> {6} 57.89
(192) {4} -> {1} 57.27
(193) {1} -> {6} 56.86


Fuzzy Based Data Mining Approach To Identify Structural Pattern in Proteins
File
Param. Input
Data Pre-proc.
Fuzzy-A
Generator
Output

```
OUTPUT ASSOCIATION RULES USING ATTRIBUTE NUMBERS:
(N) ANTECEDENT -> CONSEQUENT CONFIDENCE (%)
(1) {3} -> {2 6} 303.83
(2) {7} -> {2 5} 253.76
(3) {6} -> {2 3} 245.14
(4) {2} -> {6 3} 222.94
(5) {7} -> {2 1} 219.41
(6) {7} -> {2 3} 219.02
(7) {7} -> {2 4} 217.91
(8) {7} -> {4 5} 205.16
(9) {7} -> {5 1} 202.13
(10) {7} -> {4 1} 194.67
(11) {3 7} -> {2} 194.48
(12) {6 3} -> {2} 193.34
(13) {1} -> {4 5} 193.02
(14) {1 7} -> {2} 191.07
(15) {5 7} -> {2} 186.39
(16) {6 7} -> {2} 181.79
(17) {7} -> {2} 179.12
(18) {4 1} -> {5} 177.97
(19) {1 7} -> {5} 176.02
(20) {3} -> {2} 175.94
(21) {1} -> {2 7} 173.89
(22) {2 3} -> {6} 172.68
(23) {4 7} -> {2} 172.15
(24) {6 7} -> {3} 169.67
(25) {1 7} -> {4} 169.53
(26) {2 6} -> {3} 169.41
(27) {1} -> {2 5} 169.36
(28) {7} -> {2 6} 167.39
(29) {1} -> {2 4} 167.23
(30) {6 1} -> {2} 163.75
(31) {4 7} -> {5} 162.08
(32) {5 3} -> {2} 161.42
(33) {1} -> {5 7} 160.2
(34) {3} -> {6} 157.14
(35) {1} -> {5} 156.46
```

Lovely Professional UniversityIn fulfillment of M.Tech IT

Name: Shivika Sharma

Registration No. (11008527)

Batch year (2010-2015)

Figure 4.9: Output of association rules using attribute numbers

Next the output T-tree as a textual list is given as follows in which different support values are shown on the bases of node and item number.

Format: [N] {I} = S, where N is the node number, I is the item set and S the support.

[1]{1}= 0.64
[1] {2} = 1.25
[2] {4} = 1.21
[2.1] {2 4} = 1.25
[3] {6} = 1.14
[3.1] {2 6} = 1.65
[3.2] {4 6} = 1.02
[3.2.1] {2 4 6} = 0.82
[4] {5} = 0.98
[4.1] {2 5} = 1.06
[4.2] {4 5} = 1.15
[4.2.1] {2 4 5} = 1.14
[4.3] {6 5} = 0.79
[4.3.1] {2 6 5} = 1.11
[7] {7} = 0.51
[7.1] {2 7} = 0.9
[7.2] {4 7} = 0.64
[7.2.1] {2 4 7} = 1.1
[7.3] {6 7} = 0.47
[7.3.1] {2 6 7} = 0.85
[7.4] {5 7} = 0.69
[7.4.1] {2 5 7} = 1.28
[7.4.2] {4 5 7} = 1.04
[4.3.2] {4 6 5} = 0.6
[5] {3} = 0.92
[5.1] {2 3} = 1.62
[5.2] {4 3} = 0.48
[5.2.1] {2 4 3} = 0.75
[5.3] {6 3} = 1.45
[5.3.1] {2 6 3} = 2.8
[5.3.2] {4 6 3} = 0.4
[5.4] {5 3} = 0.84
[5.4.1] {2 5 3} = 1.36
[5.4.2] {4 5 3} = 0.51
[5.4.3] {6 5 3} = 1.06
[6] {1} = 0.64
[6.1] {2 1} = 0.79
[7.4.3] {6 5 7} = 0.4
[7.5] {3 7} = 0.57
[7.5.1] {2 3 7} = 1.11
[7.5.2] {4 3 7} = 0.42
[7.5.3] {6 3 7} = 0.79
[7.5.4] {5 3 7} = 0.67
[7.6] {1 7} = 0.58
[7.6.1] {2 1 7} = 1.11
[7.6.2] {4 1 7} = 0.98
[6.2] {4 1} = 0.69
[6.2.1] {2 4 1} = 1.07
[6.3] {6 1} = 0.36
[6.3.1] {2 6 1} = 0.59
[6.4] {5 1} = 1.0
[6.4.1] {2 5 1} = 1.08
[6.4.2] {4 5 1} = 1.23
[6.5] {3 1} = 0.58
[6.5.1] {2 3 1} = 0.87
[6.5.2] {4 3 1} = 0.44
[6.5.3] {6 3 1} = 0.56
[6.5.4] {5 3 1} = 0.7
[7.6.3] {5 1 7} = 1.02
[7.6.4] {3 1 7} = 0.5

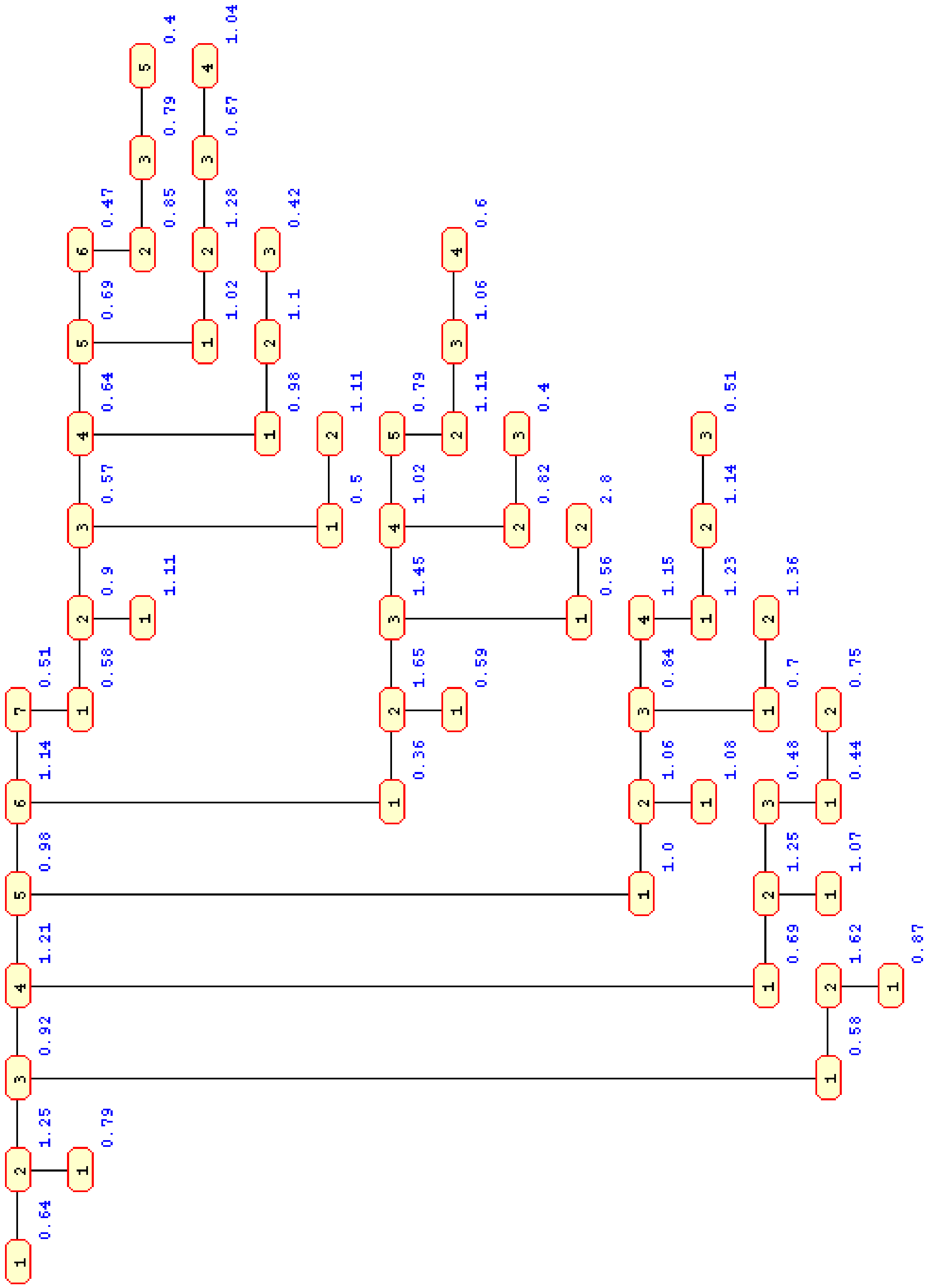
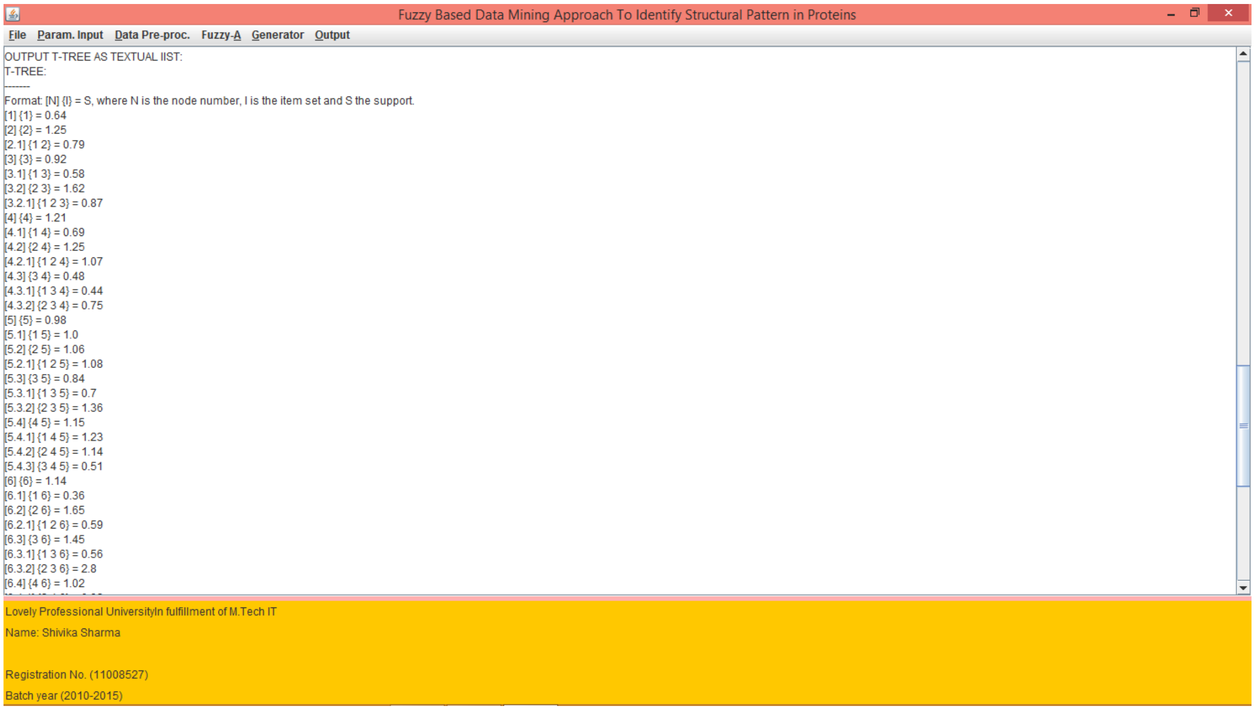


Figure 4.10 GUI presentation of mined datasets

Step7: At last GUI presentation of mined datasets are shown in the figure 4.10.


Fuzzy Based Data Mining Approach To Identify Structural Pattern in Proteins
File
Param. Input
Data Pre-proc.
Fuzzy-A
Generator
Output

```
OUTPUT T-TREE AS TEXTUAL IIST:
T-TREE:
-------
Format: [N] {I} = S, where N is the node number, I is the item set and S the support.
[1] {1} = 0.64
[2] {2} = 1.25
[2.1] {1 2} = 0.79
[3] {3} = 0.92
[3.1] {1 3} = 0.58
[3.2] {2 3} = 1.62
[3.2.1] {1 2 3} = 0.87
[4] {4} = 1.21
[4.1] {1 4} = 0.69
[4.2] {2 4} = 1.25
[4.2.1] {1 2 4} = 1.07
[4.3] {3 4} = 0.48
[4.3.1] {1 3 4} = 0.44
[4.3.2] {2 3 4} = 0.75
[5] {5} = 0.98
[5.1] {1 5} = 1.0
[5.2] {2 5} = 1.06
[5.2.1] {1 2 5} = 1.08
[5.3] {3 5} = 0.84
[5.3.1] {1 3 5} = 0.7
[5.3.2] {2 3 5} = 1.36
[5.4] {4 5} = 1.15
[5.4.1] {1 4 5} = 1.23
[5.4.2] {2 4 5} = 1.14
[5.4.3] {3 4 5} = 0.51
[6] {6} = 1.14
[6.1] {1 6} = 0.36
[6.2] {2 6} = 1.65
[6.2.1] {1 2 6} = 0.59
[6.3] {3 6} = 1.45
[6.3.1] {1 3 6} = 0.56
[6.3.2] {2 3 6} = 2.8
[6.4] {4 6} = 1.02
```

Lovely Professional UniversityIn fulfillment of M.Tech IT

Name: Shivika Sharma

Registration No. (11008527)

Batch year (2010-2015)

Figure 4.11: Output T-Tree as a textual list

.

Chapter 5
CONCLUSION

Fuzzy method can be used to efficiently analyse the bioinformatics data. Some of the fuzzy applications are intelligent control and pattern classification which can be incorporated for structuring the proteomic data. This method help a user to formulate his knowledge and to clarify the process to retrieve and exploit the proteomic information in a most simple way. Association mining is used to express the relationships among the objects and can be used to mine the proteomics data from bioinformatics.

This dissertation work presents a methodology to extract rules that associates to specific secondary structures of proteins. This work demonstrated that the rules fuzzy based association algorithm is a good technique to study this type of bioinformatics problem.

Various extraction and analysis techniques are also discussed for mining and analysing bioinformatics data. Bioinformatics is a vast topic in which various research work are done by researchers. For mining proteomics data using fuzzy association rule can be used.

Chapter 8

REFERENCES

Andreas, H., Dehmer, M., & Jurisica, I. (2014). Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics*, *15*(Suppl 6), I1

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (Vol. 22, No. 2, pp. 207-216). ACM.

Atluri, G., Gupta, R., Fang, G., Pandey, G., Steinbach, M., & Kumar, V. (2009). Association analysis techniques for bioinformatics problems. In*Bioinformatics and Computational Biology* (pp. 1-13). Springer Berlin Heidelberg

Atluri, G., Gupta, R., Fang, G., Pandey, G., Steinbach, M., & Kumar, V. (2009). Association analysis techniques for bioinformatics problems. In*Bioinformatics and Computational Biology* (pp. 1-13). Springer Berlin Heidelberg.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314-347

Dai, L., Gao, X., Guo, Y., Xiao, J., & Zhang, Z. (2012). Bioinformatics clouds for big data manipulation. *Biology direct*, *7*(1), 43.

Emeakaroha, V. C., Łabaj, P. P., Maurer, M., Brandic, I., & Kreil, D. P. (2011, November). Optimizing bioinformatics workflows for data analysis using cloud management techniques. In *Proceedings of the 6th workshop on Workflows in support of large-scale science* (pp. 37-46). ACM

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of computational biology*, *7*(3-4), 601-620.

Hirsh, H. (2008). Data mining research: Current status and future opportunities.*Statistical Analysis and Data Mining: The ASA Data Science Journal*, *1*(2), 104-107.

Jacq, N., Blanchet, C., Combet, C., Cornillot, E., Duret, L., Kurata, K. I., ... & Breton, V. (2004). Grid as a bioinformatic tool. *Parallel Computing*, *30*(9), 1093-1107.

Jain A. K. and Maheshwari S. , (2012) , "Survey of recent clustering techniques in data mining," *International Archive of Applied Sciences and Technology,* vol. 3, no. 2, pp. 68-75.

Rangra, K., & Bansal, K.L., (2014). Comparative study of data mining tools, "International Journal Of Advanced Research In Computer Science and Software Engineering", vol. 4, no. 6, pp. 216-223.

Kim, M., Cobb, J., Harrold, M. J., Kurc, T., Orso, A., Saltz, J., ... & Navathe, S. B. (2012, July). Efficient regression testing of ontology-driven systems. In*Proceedings of the 2012 international symposium on software testing and analysis* (pp. 320-330). ACM.

Kumar, C., & Mann, M. (2009). Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS letters*, *583*(11), 1703-1712.

Ling, Y., & Ye, C. (2009, July). Fast Co-clustering Using Matrix Decomposition. In *Information Processing, 2009. APCIP 2009. Asia-Pacific Conference on*(Vol. 2, pp. 201-204). IEEE.

López, V. F., Aguilar, R., Alonso, L., & Moreno, M. N. (2012). Data mining for grammatical inference with bioinformatics criteria. *Expert Systems with Applications*, *39*(3), 2330-2334.

Mangalampalli, A., & Pudi, V. (2009, August). Fuzzy association rule mining algorithm for fast and efficient performance on very large datasets. In *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on* (pp. 1163-1168). IEEE.

Olman, V., Mao, F., Wu, H., & Xu, Y. (2009). Parallel clustering algorithm for large data sets with applications in bioinformatics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *6*(2), 344-352.

Prabhu, J., Sudharshan, M., Saravanan, M., & Prasad, G. (2010, August). Augmenting rapid clustering method for social network analysis. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on* (pp. 407-408). IEEE.

Rameshwari, R., & Prasad, T. V. (2011). Systematic and Integrative Analysis of Proteomic Data using Bioinformatics Tools. *arXiv preprint arXiv:1211.2743*.

Ravi, V. T., & Agrawal, G. (2009, May). Performance issues in parallelizing data-intensive applications on a multi-core cluster. In *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*(pp. 308-315). IEEE Computer Society.

Raymer, M. L., Doom, T. E., Kuhn, L. A., & Punch, W. F(2003). Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 33(5), pp 802-813

Shelokar, P., Quirin, A., & Cordón, Ó. (2013). A multiobjective evolutionary programming framework for graph-based data mining. Information Sciences,237, pp 118-136.

Stelle, D., Barioni, M. C., & Scott, L. P. (2011). Using data mining to identify structural rules in proteins. *Applied Mathematics and Computation*, *218*(5), 1997-2004.

Ularu, E. G., Puican, F. C., Apostu, A., & Velicanu, M. (2012). Perspectives on Big Data and Big Data Analytics. *Database Systems Journal*, *3*(4), 3-14.

Viceconti, M., Taddei, F., Montanari, L., Testi, D., Leardini, A., Clapworthy, G., & Jan, S. V. S. (2007) Multimod Data Manager: A tool for data fusion.Computer methods and programs in biomedicine, 87(2), pp 148-159,

Wang, J., Crawl, D., Altintas, I., Tzoumas, K., & Markl, V. (2013). Comparison of Distributed Data-Parallelization Patterns for Big Data Analysis: A Bioinformatics Case Study. In *Proceedings of the Fourth International Workshop on Data Intensive Computing in the Clouds (DataCloud)*.

Yadav, C., Wang, S., & Kumar, M. (2013). Algorithm and approaches to handle large Data-A Survey. *arXiv preprint arXiv:1307.5437*.

Yu, L. (2010, August). Applying clustering to data analysis of Physical Healthy Standard. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on* (Vol. 6, pp. 2766-2768). IEEE.

Zhang, Z., Townsend, J. P., Yu, J., Cheung, K. H., & Bajic, V. B.(2011) Data integration in bioinformatics: current efforts and challenges. INTECH Open Access Publisher,

Zhou, J., Lamichhane, S., Sterne, G., Ye, B., & Peng, H. (2013). BIOCAT: a pattern recognition platform for customizable biological image classification and annotation. *BMC bioinformatics*, *14*(1), 291.