# FUZZY INFERENCE SYSTEM FOR PREDICTING PROTEIN FUNCTIONAL ACTIVITY USING DATA MINING

*Dissertation submitted in partial fulfilment of the requirements for the Degree of*

## MASTER OF TECHNOLOGY
## in

**COMPUTER SCIENCE AND ENGINEERING**

By

**SAHIL SHARMA**
**11501832**

Supervisor
**SHEVETA**
**16856**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab (India)

Month- April Year- 2017

# ABSTRACT

The progress of bioinformatics generates a large volume of data that needs to be analyzed in order to identify various structures. Proteins, involved in several important tasks in a living organism. arc a mixture of amino acids which have different structures and patterns. The number of primary structures solved and stored in databases is growing faster than our capability to solve these tertiary structures using different experimental methods and so is the error occurring in them is more. To have efficient and effective protein prediction we need to have accurate results that can be obtained by minimizing the errors in the proteins. The analysis of bioinformatics data has seen major shifts from traditional data mining approach to hybrid approaches. Fuzzy inference system is proposed with and relief algorithm for decreasing error rate in the prediction of proteomic data. The fuzzy 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 the imprecise description of protein data.

# DECLARATION

I hereby declare that the research work reported in the dissertation entitled " FUZZY INFERENCE SYSTEM FOR PREDICTING PROTEIN FUNCTIONAL ACTIVITY USING DATA MINING” in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Ms. Sheveta. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University’s Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

**Sahil Sharma**

**11501832**

# SUPERVISOR’S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled “**FUZZY INFERENCE SYSTEM FOR PREDICTING PROTEIN FUNCTIONAL ACTIVITY USING DATA MINING”**, submitted by **Sahil Sharma** at **Lovely Professional University, Phagwara, India** is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

**Sheveta**

**16856**

**Date**

**Counter Signed by:**

1) **Concerned HOD:**
   HoD’s Signature: ________________

   HoD Name: ____________________

   Date: ___________________

2) **Neutral Examiners:**

   **External Examiner**

   Signature: _______________

   Name: __________________

   Affiliation: ______________

   Date: ___________________

   **Internal Examiner**

   Signature: _______________

   Name: __________________

   Date: ___________________

# ACKNOWLEDGEMENTS

The satisfaction that accompanies on the 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 Ms. Sheveta (Asst Prof) for the inspiration and the constructive suggestions that help me in completing the Dissertation 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

# TABLE OF CONTENTS

# LIST OF TABLES

**TABLE NO. TABLE DESCRIPTION PAGE NO**

# LIST OF FIGURES

**FIGURE NO** **FIGURE DESCRIPTION** **PAGE NO**

# CHAPTER 1

# INTRODUCTION

## 1.1 DATA MINING

It implies collecting or "mining" useful information from a ton of data. Data Mining is an examination of finding new interesting cases and association in the huge quantity of data. It is determined as "a way toward discovery important new connections, for examples, also patterns by delving into a lot of information put away in warehouses[1]. Data mining is additionally in some cases called Knowledge Discovery in Databases (KDD)[2]. Information mining is not limited to particular industry. It need carry out innovations and the excitement to inspect the possibility of covered information that dwells in the information. Data Mining approaches have all the earmarks of being preferably suited for Bio Informatics because it is data rich, yet does not have an extensive speculation of life's relationship at the sub-nuclear level. The broad databases of organic data make both difficulties and open doors for improvement of novel KDD strategies. Mining natural information removes helpful learning from enormous data sets accumulated in science, and in other related life sciences regions, for example, pharmaceutical and neuroscience.

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

1. Huge datasets and high dimensionality
2. Understandability of the patterns

3. Redundant data
4. 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

### 1.1.1 STATISTICAL METHODS

From the word, itself Statistics means collecting, analyzing and presenting the data. Statistics helps to abstract the knowledge from the database. Normally it differs from conventional statistics on the basis of the size of data set and data that is originally collected for the data mining analysis[3]. Some of the statistical issues in data mining are

1. On the basis of size of the data
2. The curse of dimensionality and approaches to address it
3. Assess of  uncertain data
4. Automated analysis of Data
5. Algorithms used in data analysis in statistics
6. Visualizing
7. Scalable
8. Sampling technique

**Table 1.1 Data Mining Tools**

| **S.no** | **Tool Name** | **Features** |
|---|---|---|
| 1 | RAPID MINER | It used a client/server model.<br>Basically used for business, industries, researchers etc.<br>It includes multiple new aggregation functions. |
| 2 | KNIME | Open source data analytics and integration platform.<br>Scalable and high –extensible.<br>Easy to try. |
| 3 | R | It is statistical computing.<br>Used for data error handling. |

| | | |
|---|---|---|
| | | The numerical problem can easily integrate. |
| 4 | KEEL | User-friendly graphical interface.<br>Cluster discovery.<br>Includes regression and pattern mining. |
| 5 | WEKA | Suitable for machine learning schemas<br>Best for mining association rules. |
| 6 | ORANGE | Best for data visualization.<br>Scripting interface, large toolbox.<br>Includes set of compoents for data pre-processing. |

## 1.2 DATA MINING TASK

Two "irregular state" basic goals in data mining, basically, are desire and explanation[4]. The essential mining has proper data, all of which incorporates mining imperative new cases from data, which are:

i. Classification: Classification means taking in a limit that the maps (arranges) a data thing into one of a couple predefined or already existing classes

ii. Estimation: Given some information, concocting an esteem for some obscure persistent variable.

iii. Prediction: Same as arrangement and estimation with the exception of that the records are ordered by future conduct or evaluated future esteem.

iv. Affiliation rules: Determining which things go together, additionally called reliance demonstrating.

v. Grouping: Segmenting a people into different subgroups or bunches.

vi. Portrayal and discernment: Representing the data using representation methodologies.

## 1.3 INTRODUCTION TO BIOINFORMATICS

With more natural data produced, the most squeezing assignment of BioInformatics has gotten to be to investigate and decipher different sorts of information, including nucleotide and amino corrosive successions, protein structures, quality expression profiling and relevance information mining methods of highlight era, include determination, and highlight combination with learning calculations to handle the issues of illness phenotype grouping and patient survival forecast from quality expression profiles, and the issues of utilitarian site expectation from DNA arrangements[5].

Throughout late decades fast changes in genomic and other nuclear research developments and progressions in information propels have joined to convey an epic measure of information related to sub-nuclear science. The fundamental target of BioInformatics is to fabricate the appreciation of natural techniques. A portion of the terrific territory of research in BioInformatics incorporates:

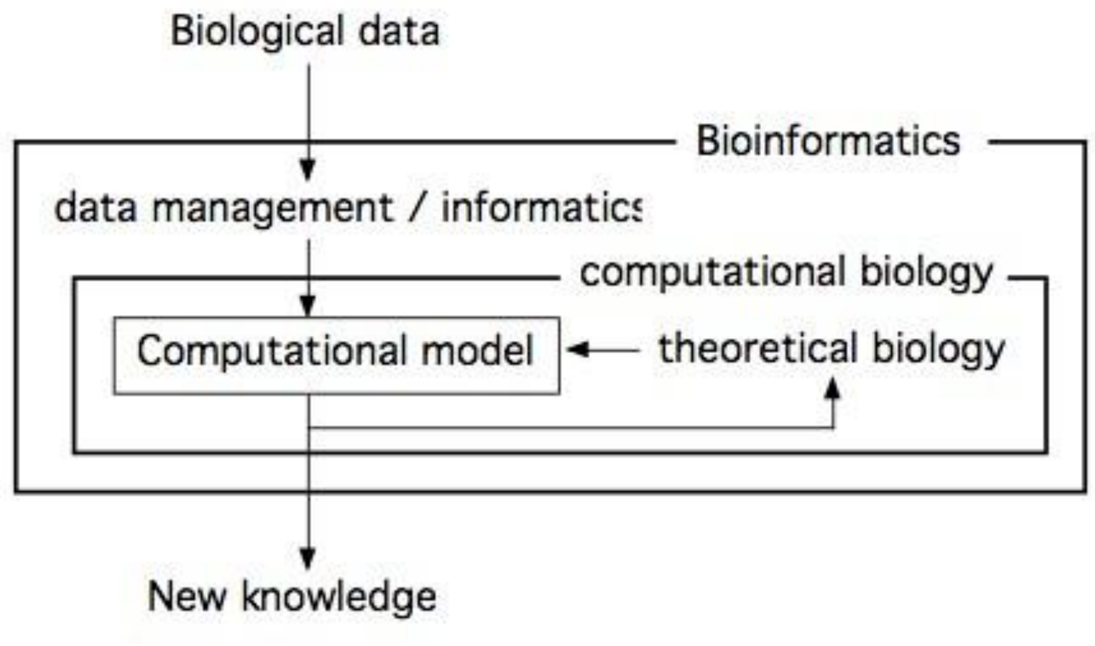


**Figure 1.1 Data Flow Diagram in Bioinformatics**[6]

### 1.3.1 GROUPING ANALYSIS

Succession examination is the most primitive operation in computational science. This operation involves finding which part of the natural courses of action are comparable and which part fluctuates in the midst of restorative examination and genome mapping frames[7]. The progression examination surmises subjecting a DNA or peptide course of action to gathering plan, progression databases, repeated progression looks or changed BioInformatics methodologies on a PC.

### 1.3.2 GENOME ANNOTATION

with regards to genomics, the comment is the way toward denoting the qualities and other natural elements in a DNA grouping. The essential genome clarification programming system was created in 1995 by Dr. Owen White.

### 1.3.3 INVESTIGATION OF GENE EXPRESSION

The disclosure of different qualities can be controlled by measuring mRNA levels with different frameworks, for example, microarrays, conferred cDNA strategy name EST sequencing, serial examination of significant worth expression SAGE tag sequencing, exceptionally parallel stamp sequencing MPSS, or particular uses of multiplexed in-situ hybridization and whatnot[8]. These methods are to a mind-boggling degree tumult inclined and subject to slant in the normal estimation. Here the certified research zone joins making quantifiable instruments to separate flag from the commotion in high-throughput quality expression contemplates.

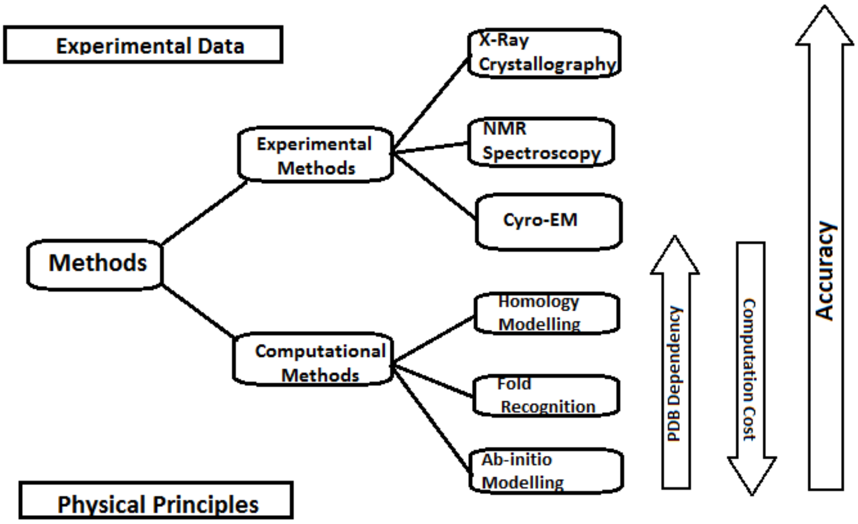


**Figure 1.2 Methods for Protein Structure Prediction**[9]

### 1.3.4 EXAMINATION OF PROTEIN EXPRESSION

Quality expression is measured from different points of view including mRNA and protein expression; however protein expression is one of the best signs of true blue quality advancement since proteins are regularly last main thrusts of cell action. Protein microarrays and high throughput HT mass spectrometry MS can give an audit of the proteins show up in a trademark case. BioInformatics is particularly required in acknowledging protein microarray and HT MS information.

### 1.3.5 EXAMINATION OF MUTATIONS IN CANCER

In tumor, the genomes of influenced cells are balanced in disease or even abnormal ways. Tremendous sequencing attempts are used to see suitable cloud point changes in a grouping of traits in harm[10]. Bioinformaticians continue making particular automated systems to manage the sheer volume of get-together data passed on, and they make new counts and programming to isolate the sequencing happens with the making get-together of human genome developments and germline polymorphisms. New physical ID developments are used, for instance, oligonucleotide microarrays to see chromosomal growth and hardships and single-nucleotide polymorphism groups to see known point changes. Another kind of data that requires novel informatics development is the examination of wounds saw to be unusual among various tumors.

### 1.3.6 PROTEIN STRUCTURE PREDICTION

The amino damaging strategy of a protein assembled, essential structure can be sufficiently looked over the movement on the quality that codes for it. In by a long shot the vast majority of the cases, this fundamental structure exceptionally picks a structure in its close-by environment. Information of this structure is enter in discernment the point of confinement of the protein. For nonappearance of better terms, major data is regularly named aide, tertiary and quaternary structure. Protein structure craving is a champion among the most basic for medication mastermind and the system of novel chemicals. A general reaction for such gauges remains an open issue for the geniuses.

### 1.3.7 SIMILAR GENOMICS

Similar genomics is the examination of the relationship of genome structure and breaking point crosswise over various regular species. Quality finding is a key use of close genomics, as is divulgence of new, non-coding utilitarian fragments of the genome. Relative genomics maul both similarities and complexities in the proteins, RNA, and legitimate locale of various creatures. Computational ways to deal with oversee genome examination have beginning late changed into an average ask about theme in programming building.

### 1.3.8 DISPLAYING BIOLOGICAL SYSTEMS

Demonstrating trademark structures is a basic errand of structures science and numerical science[11]. Computational systems science intends to make and use beneficial figurings, data structures, and representation and specific mechanical assemblies for the compromise of enormous measures of natural data with the goal of PC showing. It incorporates the usage of PC entertainments of common structures, as cell subsystems, for instance, the systems of metabolites and blends, hail transduction pathways and quality administrative structures to both break down and picture the mind-boggling relationship of these cell shapes. Fake life is an attempt to grasp transformative strategies through the PC diversion of direct life outlines

### 1.3.9 HIGH-THROUGHPUT IMAGE ANALYSIS

Computational advances used to enliven or totally motorize the taking care of, assessment and examination of a considerable measure of high-information content biomedical pictures. Frontline picture examination systems grow an onlooker's ability to make estimations from an inconceivable or complex game plan of pictures. A totally made examination system may absolutely supplant the onlooker. Biomedical imaging is ending up being more fundamental for both diagnostics and research. A portion of the occurrences of research around there are: clinical picture examination and representation, gathering clone covers in DNA mapping, Bioimage informatics, and so on.

### 1.3.10 PROTEIN-PROTEIN DOCKING

In the most recent two decades, incalculable three-dimensional structures have been controlled by X-shaft crystallography and Protein atomic engaging reverberation spectroscopy (protein NMR). One focal question for the trademark pro is whether it is utilitarian to foresee conceivable protein-protein joint efforts just in light of these 3D shapes, without doing protein-protein correspondence tests. A gathering of methodology have been made to handle the Protein-protein docking issue, regardless it gives there is still much work to be done in this field.

## 1.4 PROCESSING OF DATA

The 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 the computer so as to convert it into machine readable form and thus process it and convert it back to human readable form[12]. Earlier in the 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 much independent equipments's was started to carry out data processing. Individuals that were appointed tr 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 a large extent and the accuracy and efficiency of' data processing have increased tremendously. Due to the innovation of computers and supercomputers. 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

Handling implies gathering and dominant knowledge to make a significant data. Within the wake of extricating the knowledge data from the various information sources refinement and breaking down of knowledge is finished. There square measure various strategies for breaking down the various types of data like grouping, characterization so on. Monumental data implies data that cannot be handled and handled clearly. Previous data square measure spared, then stacked to some plate and showing intelligence perform a minimum of one examination on data. In any case, currently within the today's universe of giant data the intuitiveness, handling of knowledge is fast Machine learning:

As statistical methods have some disadvantages. Statistical method faces difficulty incorporating subjective information in their models and it also faces the 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

## 1.5 FUZZ INFERENCE SYSTEM

Fuzzy rules are the new approach that is used to mine quantitative data frequently present in the databases. Fuzzy rules use the fuzzy logic to change numerical type attributes to fuzzy type attributes[13]. The fuzzy data mining technique that is Fuzzy Repeated Pattern growth (FRP-growth) algorithm can be used to mine the large datasets. It treats each data item as variable and its portioned is based on the linguistics values.

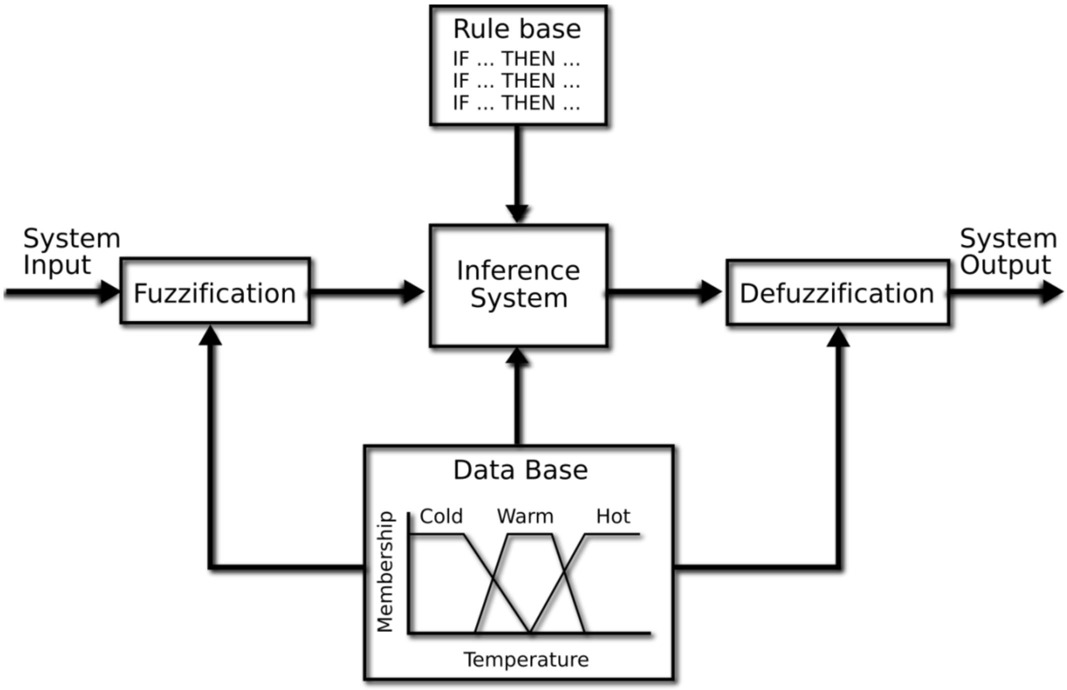


**Figure 1.3 Fuzzy Inference System**[14]

Various blocks of fuzzy reasoning system and various inferences that are drawn based on the If Then kind of System are performed only in FIS and those are:

1. First of all Comparison is done between inputs variable along with membership functions that are the previous antecedent to get the new membership values corresponding to every linguistic values and this process is termed as Fuzzification.

2. Then Combination which are mostly multiplication which are used to get membership values in premises which gives the degree to which extent the rule is followed.

3. Thirdly Generation of qualification is done which can be either fuzzy or crisp values and also every rule is dependent on the firing strength.

4. Finally Aggregation of qualified consequents corresponding to antecedent is done to give a crisp value as output. This process is called defuzzification.

# CHAPTER 2

# REVIEW OF LITERATURE

**Stefano Cresci** *et al(2016).* In this creator proposed a strikingly novel, basic, and powerful way to deal with model online client conduct, it separate and dissect computerized DNA successions from client online activities and utilize Twitter as a benchmark to test the proposal[15]. It additionally acquires a sharp and minimized DNA-roused portrayal of client activities. At that point, it applies standard DNA investigation strategies to separate amongst veritable and spambot accounts on Twitter. A test battle underpins this proposition, demonstrating its adequacy and reasonability. To the best of information, they guaranteed to be first ones to recognize and adjust DNA-propelled systems to online client behavioral displaying. While Twitter spambot discovery is a particular utilize case on a particular web-based social networking, they proposed approach is stage and innovation rationalist, thus making ready for various behavioral portrayal errands.

**M. Masseroli** *et al(2015).* In this examination paper noteworthy advances in biotechnology and structure science is done which is making an astounding measure of bimolecular data and semantic remarks, moreover which increases in number and quality, yet it is scattered and just for the most part related. Compromise and mining of these scattered and making information and data have the high limit of finding canvassed biomedical learning obliging in appreciation complex natural miracles, run of the mill or hypochondriac, and finally of enhancing investigation, foresight, and treatment, however such fuse stances gigantic challenges[16]. Also, it has endeavored to handle them by working up a novel and summed up way to deal with describe and successfully keep up overhauled and intensify a compromise of many progressing and heterogeneous data sources. Maker's approach showed handiness to expel biomedical data about complex natural methodology and diseases. Understanding complex normal marvels incorporates taking note of complex biomedical request on different biomolecular information at the same time, which are passed on through different genomic and proteomic semantic comments Scattered in many appropriated and heterogeneous information sources, such heterogeneity and scattering hamper the master's capacity of asking general request and performing general assessments. To beat this issue, producer has made programming planning to make and keep up a Genomic and Proteomic Knowledge Base, which combines two or three the most significant wellsprings of such scattered data

**Alfredo Cuzzocrea** *et al(2015).* The author found that Provenance of Big Data is an interesting issue in the database and information mining research groups[17]. Fundamentally, provenance is the way toward identifying the genealogy and the inference of information and information items, and it assumes a noteworthy part in database administration frameworks and also in work process administration frameworks and conveyed frameworks. Regardless of this, the provenance of huge information research is still in its embryonic stage, and a considerable measure of endeavors should at present be done in this area. Motivated by these contemplations, this paper gives an outline of pertinent issues and difficulties with regards to huge information provenance examine, by likewise highlighting conceivable future endeavors inside these examination headings.

**Eric P. Xing** *et al(2015).*If there is any precise approach to productively apply a large range of cutting edge ML projects to modern scale issues then that is, utilizing Huge Models ranging up to hundreds of billions of different parameters used in on Big Data which reaches up to terabytes or pet bytes in size[18]. Cutting edge parallelization procedures utilize fine-grained operations and booking past the great mass synchronous preparing worldview promoted by MapReduce, or even specific chart construct execution that depends with respect to a diagram which indicates the use of ML programs. The assortment within methodologies tries to describe frameworks and calculations outline in various headings, and it stays hard to locate an all inclusive stage pertinent to an extensive variety of ML projects at scale. A universally useful structure, Petuum ,that methodically addresses information and model-parallel difficulties in expansive scale ML, by watching that numerous ML projects are in a general sense enhancement driven and concede blunder tolerant, iterative-joined algorithmic arrangements. Creator exhibited the viability of these framework plans versus surely understood the usage of present day ML calculations, demonstrating that Petuum permits ML projects to keep running in substantially minimal time and with the impressively bigger representation sizes, or we can say even on unobtrusively measured process groups.

**Suvarna** *et al(*2015) In this examination paper critical advances in biotechnology and structure science is done which is making an amazing measure of biomolecular information and semantic comments, besides which increments in number and quality, yet it is scattered and only generally related[19]. The trade off and mining of these scattered and making data and information have the high furthest reaches of discovering peddled biomedical learning obliging in gratefulness complex regular supernatural occurrences, common or despondent person, lastly of upgrading examination, premonition, and treatment, however such circuit positions monstrous challenges. Likewise, it has tried to handle them by working up a novel and summed up approach to managing to depict and effectively keep up upgraded and strengthen a bargain of many advancing and heterogeneous information sources. Producer's approach demonstrated handiness to oust biomedical information about complex normal system and illnesses. Understanding complex typical wonders consolidate observing complex biomedical demand on various biomolecular data in the meantime, which are gone on through various genomic and proteomic semantic remarks Scattered in many appropriated and heterogeneous data sources, such heterogeneity and dispersing hamper the ace's ability to ask general demand and performing general appraisals. To beat this issue, maker has made programming wanting to make and keep up a Genomic and Proteomic Knowledge Base, which consolidates a few the most noteworthy wellsprings of such scattered information

**Sanjima Manocha** *et al(2014).* In this, the creator has investigated the essential components of information mining strategies in distributed computing and securing the information utilizing edge identification strategy. What's more, tries to incorporate information mining strategies into distributed computing and picture handling making it a mutt approach[20]. The use of data mining methodology through disseminated figuring inclinations the customers to separate important concealed prescient data from for all intents and purposes incorporated information stockroom that decreases the expenses of capacity and foundation. Unify administration of programming and information stockpiling, with an affirmation of effective, solid also, secure organizations for their customers through Edge area based approach for picture steganography. As the learning to be removed from the information is put away over the distributed storage, likewise, the trouble to get every one of the information to a brought together capacity is discussed. Subsequently, the significance for an information mining calculation that works over non incorporated information stockpiling over dispersed environment is examined.

**Sanjima Manocha** *et al(2014).* In this, creator has investigated the essential components of information mining strategies in distributed computing and securing the information utilizing edge identification strategy. What's more, tries to incorporate information mining strategies into distributed computing and picture handling making it a half-breed approach[20]. The usage of information mining procedures through distributed computing urges the clients to separate important concealed prescient data from for all intents and purposes incorporated information stockroom that decreases the expenses of capacity and foundation. Unify administration of programming and information stockpiling, with an affirmation of effective, solid and secure administrations for their clients through Edge location-based approach for picture steganography. As the learning to be removed from the information is put away over the distributed storage, likewise, the trouble to get every one of the information to a brought together capacity is discussed. Subsequently, the significance for an information mining calculation that works over non-incorporated information stockpiling overdispersed environment is examined.

**Avita Katal** *et al(2013)* portrayed Big information as a significant measure of information which requires new improvements and models with the target that it finds the opportunity to be unmistakably conceivable to think respect from it by getting and examination handle. Because of such interminable size of information, it winds up being enormously hard to perform productive examination utilizing the present routine techniques[21]. Immense information because of its unmistakable properties like volume, speed, gathering, fluctuation, respect and multifaceted nature set forth various inconveniences. Since Big information is a late pending improvement in the market which can pass on immense ideal conditions to the business affiliations, it persuades the chance to be especially important that various inconveniences and issues related to passing on and changing as per this progression are gotten into light this paper. This paper likewise presented the Big information headway near to its significance in the present world and existing attempts which are persuading and essential in changing the likelihood of science into goliath science and society as well. The unmistakable difficulties and issues in adjusting and persevering Big information improvement, its mechanical congregations like Hadoop are in addition broke down in detail near to the issues Hadoop is going up against.

**Xiaohua** *et al(2011).* Creator has given a semantic-based approach for multi-source BioInformatics information incorporation[22]. In this approach, a metamodel is used to speak to the ace hunt composition, and a compelling interface extraction calculation in view of the progressive structure of the web and example is created to catch the rich semantic connections of the online BioInformatics information sources. The objective was to build up a meta-scan interface for scholars as a solitary purpose of access to various online BioInformatics databases. In content mining, a portion of the testing issues in mining and looking the biomedical writing are tended to, and writer exhibits abound together design i.e. Biomedical Literature Searching, Extraction and Text Data Mining, examine some novel calculations, for example, semantic-based dialect demonstrate for writing recovery, semi-managed design learning for Information extraction of organic connections from biomedical writing. In another part, diagram based information mining, the attention is on chart based mining in organic systems are talked about additionally appropriateness of chart based mining strategies and calculations in the examination of measured and progressive structure of natural systems, how to recognize and assess the subsystems from confused organic systems, and present the exploratory outcomes are found is talked about. To assemble these pieces, a brought together structure is acquainted with coordinate the three sections in the BioInformatics information mining methodology.

**Dennis Wegener** *et al(2011)* This paper centers Bioinformatics and information mining techniques are working together to execute and assess apparatuses and systems for the expectation of ailment repeat and movement, reaction to treatment, and also new bits of knowledge into different oncogenic pathways by considering the client requirements with their heterogeneity[23]. In view of given advancement, medication gives the experience an insurgency that is notwithstanding changing the way of medicinal services from receptive to proactive. The p-solution group is making a biomedical phase to encourage the interpretation from given present observe to a prescient, customized, defensive, actively taking part and psycho-subjective drug in coordinating VPH modeling, scientific work on, image and other kind of omics dataset. This author shows and introduces the difficulties for information mining dependent examination in biology and therapeutic informatics which shows our approach towards an information mining conditions tending to given necessities in the solution stage.

**Shane Dixon** *et al(2010).* In this paper creator portrayed about BioInformatics, which is an information escalated field of innovative work. The reason for BioInformatics information mining is to find the connections and examples in extensive databases to give valuable data to biomedical examination and diagnosis[24]. In this paper, calculations in view of counterfeit safe frameworks with fake nervous systems are utilized in BioInformatics information mining. There are discussed three unique varieties of genuine esteemed pessimistic choice calculation and with multiple layers bolster forward neural system models given are examined, tried and thought about by means of PC reproductions. It additionally demonstrates ANN has shown results that yields finest general outcome while the calculation is invaluable in conditions where just the "typical" information is accessible.

**Gowtham Atluri** *et.al* (2009) proposed the particular sorts of association illustrations and some of their applications in Bioinformatics. Challenges which are ought to have been directed to make alliance examination based strategies round fragment more material to different intriguing issues in Bioinformatics[25]. In a general sense, IWo sorts of cases were used that were viewed Limited thing set illustration and Affiliation represent plan. A part of the sorts of illustrations is Customary Regular cases. 1-lyper group Examples. Bumble Tolerant Examples and Discriminative Example Missing. Association examination has wound up being a powerful approach for dismembering standard market bushel data and has even been found profitable for a couple issues in Bioinformatics in two or three cases. Regardless, there are different other crucial issues in Bioinformatics. for instance, finding biomarkers using thick data like SNP data and real regarded data like quality expression data, where Limited thing set and Affiliation control outline methodology could exhibit to him especially supportive, bungalow channel starting at now he easily and effectively associated. A basic instance of illustrations which are not effectively gotten by the traditional connection examination framework and its present expansions is a get-together of characteristics that are co-imparted together over a subset of conditions in a quality expression enlightening gathering

**Chanchal Kumar**, *et.al* (2009) reviewed the limitations of biochemical methods also, other related advances that are recently fitting Of single sort proteins[26]. Mass spectrometry when joined with other imaginative methods and advanced test and computational methodologies engage to finish a tremendous scale consider Of proteins of cell level. Another methodology name SILAC was used when no less than two common states were ought to have been breaking down. Programming named MaxQuant was used for the parallel treatment of complex enlightening records. These datasets were the eventual outcome of a blend of contemporary mass spectrometry and advanced spectrometry. As needs are. MaxQuant makes a multidimensional lattice that would contain data concerning proteomes. There curve moves one needs to confront, for instance, mapping while in the meantime finishing such kind of examination. There are certain structures like BioMari that help with settling mapping issues, lodge still it has slip-ups and abnormalities while making occurs

**N. Jacq** *et aI.* (2003) developed a bioinformatics instrument which helps in securing the normal data. Framework applications were used for adding and separating the natural data[27]. The system mechanical assembly was used to store and institutionalize the data. An interface was used to predict the sub-nuclear limits and to perceive the plans molded in science. Affect computation was in like manner used which set away progressions and besides consider each datum. Gadget still prompts troubles in managing all data, as it can't cover the duplication of the data

**Nir Friedman** *et al (2000)* proposed a Bayesian network based model for analysis of biological pattern. Quality illustrations were bankrupt down using the Bayesian thought of frameworks[28]. The two procedures that were used as a piece of this examination work are a novel chase figuring and a creamer approach. The guideline concern was to think components of the quality expression data and to associate between the qualities. A couple issues were Factual parts of decoding the results, algorithmic versatile quality issues in picking up from the data, and the choice of close-by probability models. Techniques for expression examination would he have the capacity to upgraded by working up the speculation for learning neighborhood probability models which are sensible for the kind of associations that appear in expression data. Combining natural learning as prior data to the examination upgrades the request heuristics.

# CHAPTER 3

# PRESENT WORK

## 3.1 PROBLEM FORMULATION

Proteins are the inadequately comprehended and anticipated in view of proteomic and genomic information sets accessible. The expectation for the most part incorporates information serious computational mining. Robotized comment foreseeing protein work exercises are extremely mind boggling. With the expansion in the quantity of sequenced genomes quickly, all the protein items must be commented on numerically. Conditions upon computational forecasts then it is vital that the estimation of these techniques be high. Different discoveries have appeared now and again that:

(i) Nowadays best protein work forecast calculations unmistakably beat all the more oftentimes utilized more established era techniques, with significant advantages on a wide range of target information and

(ii) Even the highest strategies have indicated alright learning to guide tests.

The right expectation of protein capacity is primary component to comprehend and see life at the point by point sub-atomic level and has high biomedical and pharmaceutical ramifications. Be that as it may, with this much-inbuilt intricacy and expense, the speculative portrayal of capacity can't extend to hold the tremendous measure of arrangement information effectively accessible. The computational clarification of protein capacity has subsequently risen as a difficulty at the cutting edge of computational and atomic science.

Various arrangements have been foreseen in the most recent four decades, however, the errand of computational utilitarian derivation as often as possibly depends on customary methodologies, for instance, recognizing spaces or discovering basic kept Alignment Search which is a principle in the midst of proteins with tentatively firm capacity.

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. Further, we have seen the error rate and have tried to minimize the error and therefore 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.

Lately, the openness of genomic-level succession data for a large number of animal groups, tied with huge high-throughput investigational information, has made new open doors for the capacity forecast. An incredible number of strategies have been foreseen to use this information, including capacity expectation from amino corrosive grouping, protein-protein communication systems, and protein structure information.

An unbiased appraisal of these distinctive strategies can give understanding into their ability to separate proteins practically and can control natural investigations. In this way, however, a total assessment consolidating a substantial and divergent arrangement of target groupings has not been led due to down to earth troubles and issues in giving an effectively explained target set.

Data analysis along with data mining of all the data is a very staggering task due o the inconsistency and vagueness in the data generated  to uncover the uncertainties in the data fuzzy computations can be considered by effective evaluation of the proteomic data

## 3.2 OBJECTIVES OF THE STUDY

The key objective should be together with the prediction of average error rate in protein function prediction and implement it by using Fuzzy inference system. Learning of different objectives which define the exact aims for the learning should be visibly declared. Integration of the  Proteomic and genomic data from heterogeneous data sets should be done to focus on function prediction. This study will evaluate the outcomes of new improved results which are better than previous existing results derived from Protein Ontology. So the basic Objectives can be stated as:

1. To predict the behavior of protein functions successfully from biological information available.
2. To apply Classification and Predictions on hidden patterns obtained.
3. To represent the predicted results through Fuzzy Inference System (FSI) using Fuzzy Classification.
4. To enhance the predicted results by reducing the average error rate and therefore improving the Protein prediction

## 3.3 RESEARCH METHODOLOGY

### 3.3.1 STEPS TO BE FOLLOWED FOR EXECUTION

**Step 1: Browsing Dataset** Yeast dataset is browsed loaded into the Matlab. It contains output names of classes which are total 10 in number which is that are in non-numeric form like MIT, ME1, VAC, POX, ERL etc. These class names are replaced with numeric values from 1 to 10. The Class names and corresponding values and biological short forms are shown in the table no. 3.1 and 3.2 as shown below.

**Table 3.1 Class name corresponding to biological short forms**

| | |
|---|---|
| Cytosolic or Cytoskeletal | CYT |
| Nuclear | NUC |
| Mitochondrial | MIT |
| Membrane Protein without N-terminal signal | ME3 |
| Membrane Protein with Uncleaved signal | ME2 |
| Membrane Protein with cleaved signal | ME1 |
| ExtraCellular | EXC |
| Vacuolar | VAC |
| Peroxisomal | POX |
| Endoplasmic Reticulum Lumen | ERL |

**Table 2.2 Mapping of class names to numerical values**

| | |
|---|---|
| MIT | 1 |
| NUC | 2 |
| CYT | 3 |
| ME1 | 4 |
| EXC | 5 |
| ME2 | 6 |
| ME3 | 7 |

|  |  |
|---|---|
| VAC | 8 |
| POX | 9 |
| ERL | 10 |

**Step 2**: **Attribute Description** The data set has 1484 instances of yeast data with total 9 attributes out of which 8 attributes are predictive and the last one is Class name. The name of all the attributes with their corresponding information is shown in the table no 3.3 as shown below:

**Table 3.3 Attribute Description**

| **Sequence Name** | **Description** |
|---|---|
| mcg | McGeoch's method for signal sequence recognition. |
| gvh | von Heijne's method for signal sequence recognition. |
| alm | Score of the ALOM membrane spanning region prediction program. |
| mit | Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins. |
| erl | Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Binary attribute. |
| pox | Peroxisomal targeting signal in the C-terminus. |
| vac | Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins. |
| nuc | Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins. |

**Step 3: Applying Algorithm** On the available dataset the feature selection is applied to do preprocessing to identify the subsets of data that mainly compute the results

from a large amount of data and provides us with the features or attributes that have minimal and maximal contribution to computing results.

For this purpose, we have used the Relief algorithm which helps to improve the accuracy of classification and also reflects the attributes that need to be taken care of to improve the results.

### 3.3.2 RELIEF ALGORITHM

Relief algorithm is the easy, high-speed, and efficient approach for attribute weighting. The end result of Relief algorithm is an assignment of weights which lie between the range of −1 to 1 for every single attribute, and with more number of positive weights shows or indicate more attributes that can be predicted[29]. Relief Algorithm has various variants which further depend on the category or nature of data also on attributes and characteristics of data. This Algorithm workings are based on the principles as follows:

In this, the weight of a given attribute is updated in different iterations in the procedure mentioned below. A random sample of data is selected, after that the sample is identified which lie in the same zone or belongs to the same given class and neighbouring nearest sample is called (nearest hit) and on the other hand those samples which do not lie nearby and belong to some other different class are identified and are said to be belonging to opposite class which is termed as (nearest miss). Relief algorithm basically works on this principle of Nearest Hit and Nearest Miss which are building blocks of this algorithm. The next thing we calculate is the distance between Nearest Hit and Chosen Sample and same with Nearest Miss and chosen Sample which is calculated on the basis of Manhattan distance between two points. Then there is a change in feature weights which are taken into consideration for feature selection in the classification of the given target class. These features found are then given extra weights for the process of classification. So we can say that weights obtained of feature play very important role for locating the accurate or correct class.

The Relief algorithm used is shown as under which follows the steps as shown[30]:

Input: For every instance training data there are vector values of Attribute and Class which is called Class label.
Output: Estimated vector *W* with corresponding weights of different components.

```
1. Set all the weights W[A]:=0.0;
2. for every i:=1 to m do begin
3. Randomly select an instance $r_i$ ;
4.Then find k-nearest hits $h_j$;
5. for every class C class($r_i$) do
6. from class C find k-nearest misses $m_j$ (c);
7. for A:=1 to a
8. W[A]=W[A]- $\sum_{j=1}^{k} \frac{diff\left(A,r_i,h_j\right)}{m.k}$ +    (1)
9. $\sum_{C\neq class\ r_i} \frac{\left[\frac{p(c)}{1-p(class\ (r_i}\sum_{j=1}^{k} diff\ (a,r_i,h_j)\right]}{m.k}$    (2)
10. End
```

**Step 4: Ranking of Attributes** After applying relief algorithm and assigning weights, the ranking of the attributes changes. The most important one are named on the top of the list and thus all are ranked in decreasing order of their importance and contribution towards class attribute.

Fuzzy Inference System (FIS) which have 8 input values and 1 output value is implemented with the help of MATLAB 2013b. After removing the first or initial attribute, the remaining dataset will contain 9 attributes only. Out of that first 8 attributes are given as input values, and the remaining last one will give output values to the FIS.

Ranking of all the attributes with their weight are shown in table 4 given below:

**Table 3.4 Attributes with corresponding ranks and weights obtained**

| | | |
|---|---|---|
| / mcg/ aln | 3 | 0.0061607 |
| / gvh/ mcg | 1 | 0.0047937 |

|  |  |  |
|---|---|---|
| / aln/ gvh | 2 | 0.0078266 |
| / mit/ mit | 4 | 0.0042892 |
| / pox/ pox | 6 | 6.5501e-05 |
| ME3 | 7 | 0.0038163 |
| EXC | 5 | 0.00014708 |
| VAC | 8 | -0.0028697 |

**Step5: Selection of Attributes** The attributes which are solely responsible for maximizing the accuracy of classification and help to reduce the error rate are selected. Here we are selecting 5 attributes out of 8 attributes based on the weights obtained in table no 3.4

**Table 3.5 Selected Attributes**

|  |  |  |
|---|---|---|
| CYT/ mcg/ aln | 3 | 0.0061607 |
| MIT/ gvh/ mcg | 1 | 0.0047937 |
| NUC/ aln/ gvh | 2 | 0.0078266 |
| ME1/ mit/ mit | 4 | 0.0042892 |
| ME2/ pox/ pox | 6 | 6.5501e-05 |

**Step 6: Range Decomposition** The selected 5 attributes are provided to fuzzy inference system as input and 1 attribute as output in MATLAB 2013b. Now the range of all 6 attributes(5 input and 1 output) is obtained from the dataset and then decomposed to a fuzzy set of values as shown in 6,7,8,9,10 tables below. It is to note that there are 5 attributes mcg, gvh, aln, mit and pox for which decomposition of values is done.

Table 3.6 Decomposition Range of 1st Attribute (mcg)

| Decomposition Range of 1st Attribute (mcg) | | |
|---|---|---|
| | **Range Value** | **Fuzzy Set Values** |
| 1 | 0.42 to 0.64 | Lowl |
| 2 | 0.33 to 0.61 | Low2 |
| 3 | 0.40 to 0.73 | Low3 |
| 4 | 0.91 to 0.70 | Medium 1 |
| 5 | 0.49 to 0.89 | Medium 2 |
| 6 | 0.54 to 0.94 | Medium 3 |
| 7 | 0.28 to0.54 | High 1 |
| 8 | 0.28 to0.80 | High 2 |
| 9 | 0.32 to 0.68 | High 3 |
| 10 | 0.7 to 0.86 | Very high |

Decomposition of 1st attribute mcg into fuzzy set of variables which are divided into low, medium, high and very high.

Table 3.7 Decomposition Range of 2nd Attribute (gvh)

| Decomposition Range of 2nd Attribute (gvh) | | |
|---|---|---|
| | **Range Value** | **Fuzzy Set Values** |
| 1 | 0.40 to 0.67 | Low l |
| 2 | 0.31 too.60 | Low 2 |
| 3 | 0.39to0.63 | Low 3 |
| 4 | 0.66 to 0.88 | Medium l |
| 5 | 0.39 to 0.87 | Medium 2 |
| 6 | 0.42 to 0.75 | Medium 3 |
| 7 | 0.24 to 0.58 | High 1 |
| 8 | 0.32 to 0.82 | High 2 |
| 9 | 0.27 to 0.68 | High 3 |
| 10 | 0.56 to 0.92 | Very high |

Decomposition of 2nd attribute gvh into fuzzy set of variables which are divided into low, medium, high and very high.

**Table 3.8 Decomposition Range of 3rd Attribute (aln)**

| Decomposition Range of 3rd Attribute (aln) | | |
|---|---|---|
| | **Range Value** | **Fuzzy Set Values** |
| 1 | 0.45 to 0.66 | Low l |
| 2 | 0.43 to 0.69 | Low 2 |
| 3 | 0.42 to 0.60 | Low 3 |
| 4 | 0.30 to 0.47 | Medium 1 |
| 5 | 0.36 to 0.58 | Medium 2 |
| 6 | 0.33 to 0.58 | Medium 3 |
| 7 | 0.21 to 0.42 | High 1 |
| 8 | 0.26to0.57 | High 2 |
| 9 | 0.43 to0.59 | High 3 |
| 10 | 0.38 to 0.58 | Very high |

Decomposition of 3rd attribute aln into fuzzy set of variables which are divided into low, medium, high and very high.

**Table 3.9 Decomposition Range of 4th Attribute (mit)**

| Decomposition Range of 4th Attribute (mit) | | |
|---|---|---|
| | **Range Value** | **Fuzzy Set Values** |
| 1 | 0.13 to 0.65 | Low l |
| 2 | 0.13 to 0.43 | Low 2 |
| 3 | 0.11 to 0.35 | Low 3 |
| 4 | 0.23 to 0.78 | Medium 1 |
| 5 | 0.23 to 0.37 | Medium 2 |
| 6 | 0.4 to 0.49 | Medium 3 |
| 7 | 0.l2 to 0.31 | High 1 |
| 8 | 0.08 to 0.28 | High 2 |
| 9 | 0.10 to 0.49 | High 3 |
| 10 | 0.25 to 0.40 | Very high |

Decomposition of 4th attribute mit into fuzzy set of variables which are divided into low, medium, high and very high.

**Table 3.10 Decomposition Range of 5th Attribute (pox)**

| **Decomposition Range of 5th Attribute (pox)** | | |
|---|---|---|
| | **Range Value** | **Fuzzy Set Values** |
| 1 | 0.46 to 0.53 | Low l |
| 2 | 0.47 to 0.68 | Low 2 |
| 3 | 0.49 to 0.58 | Low 3 |
| 4 | 0.43 to 0.58 | Medium 1 |
| 5 | 0.39 to 0.56 | Medium 2 |
| 6 | 0.40 to 0.59 | Medium 3 |
| 7 | 0.43 to0.55 | High 1 |
| 8 | 0.39 to 0.60 | High 2 |
| 9 | 0.40 to 0.54 | High 3 |
| 10 | 0.53 to 0.58 | Very high |

Decomposition of 5th attribute pox into fuzzy set of variables which are divided into low, medium, high and very high.

**Step 7: Creating Fuzzy** Rules On the basis of input valued given to fuzzy inference system the rule base is created and few of the rules are shown in the given table 10

**Table 3.11 Rule Base for FIS**

| Rule No. | Rules |
|---|---|
| 1 | If (mcg is low1) and (gvh is low1) and (aln is low1) and (mit is low1) and (pox is low1) then (output1 is MIT) (1) |
| 2 | If (mcg is low2) and (gvh is low2) and (aln is low2) and (mit is low2) and (pox is low1) then (output1 is NUC) (1) |
| 3 | If (mcg is low3) and (gvh is low3) and (aln is low3) and (mit is low3) and (pox is low1) then (output1 is CYT) (1) |
| 4 | If (mcg is medium1) and (gvh is medium1) and (aln is medium1) and (mit is medium1) and (pox is medium1) then (output1 is ME1) (1) |
| 5 | If (mcg is medium2) and (gvh is medium2) and (aln is medium2) and (mit is |

| | |
|---|---|
| | medium2) and (pox is medium1) then (output1 is EXC) (1) |
| 6 | If (mcg is medium3) and (gvh is medium3) and (aln is medium3) and (mit is medium3) and (pox is medium1) then (output1 is ME2) (1) |
| 7 | If (mcg is high1) and (gvh is high1) and (aln is high1) and (mit is high1) and (pox is high1) then (output1 is ME3) (1) |
| 8 | If (mcg is high1) and (gvh is high2) and (aln is high2) and (mit is high2) and (pox is high1) then (output1 is VAC) (1) |
| 9 | If (mcg is high1) and (gvh is high3) and (aln is high3) and (mit is high3) and (pox is high1) then (output1 is POX) (1) |
| 10 | If (mcg is high1) and (gvh is Veryhigh) and (aln is veryhigh) and (mit is veryhigh) and (pox is high1) then (output1 is ERL) (1) |

**Step8: Various Combinations to generate Output** After this membership function to all the input and output has been applied. There are many combination member functions. Four Commonly used combinations are listed in the table given below. It is here to note that if one input function belongs to one particular membership class then the rest of them also belong to that membership class.

**Table 3.12 Input- output membership function combination**

| Sl No. | Membership function for Input variable | Membership function for Output variable |
|---|---|---|
| 1 | Gaussiaii2 | Gaussian2 |
| 2 | Gaussian2 | Triangular |
| 3 | Trapezoidal | Trapezoidal |
| 4 | Trapezoidal | Triangular |
| **i.** | Bell shaped | Bell shaped |
| 6. | Sigmoid | Sigmoid |
| 7. | Pie shaped | Pie shaped |

Out of all these combinations, the least error rate was shown in the combination of Trapezoidal Triangular means that whenever we give all the inputs as Trapezoidal and the output is in the form of Triangular.

**Table 3.13 Average error rate in different combinations**

| S. No. | Membership function for Input variable | Membership function for Output variable | Average Error |
|---|---|---|---|
| **1.** | **Trapezoidal** | **Triangular** | **0.36806** |
| 2. | Trapezoidal | Trapezoidal | 0.3751 |
| 3. | Gaussian2 | Gaussian2 | 0.92 |
| 4. | Ganssian2 | Triangular | 2.59 |
| 5. | Bell shaped | Bell shaped | 1.008709762 |
| 6. | sigmoid | sigmoid | 0.92496 |
| 7. | Pie shaped | Pie shaped | 1.021738 |

Therefore we have tried to improve the error rate by taking into consideration the membership combination of Trapezoidal to Triangular and the average error rate which was earlier 0.36806 is improved.

### Trapezoidal Membership Function

It is shown by a lower limit A, an upper limit d, a lower support limit B, and an upper support limit C, where A < B < C < D.

$$\mu_A(x) = \left\{ \begin{array}{c} 0, \text{if}(\text{x} < A) or (x > D) \\ \frac{x-A}{B-A}\ if\ A \ \leq \ \mathbf{x} \ \leq \ B \\ 1, if\ \text{B} \ \leq \ x \ \leq \ D \\ \frac{D-x}{D-C} if\ C \leq x \leq D \end{array} \right\} \qquad (3)$$

### Triangular Membership function

It is shown by a lower limit A, an upper limit B, and a value M, where A<M<B.

$$\mu_A(x) = \left\{ \begin{array}{c} 0, \text{if}(\text{x} \leq \text{A}) \\ \frac{x-A}{M-A}\ if\ A \ \leq \ \text{x} \ \leq \ M \\ \frac{B-x}{B-M} if\ \text{M} < \ x < \ B \\ 0\ if\ x \leq B \end{array} \right\} \qquad (4)$$

### 3.3.3 ERROR CALCULATION

The error between of the two methods of the classification can been evaluated via calculating estimated error along with average error. Estimated error is $E_i$ of an individual instance i and it is given by equation

$$E_i = \frac{(|P_i - T_i|)}{T_i} \tag{5}$$

Here, $P_i$ is the output value of class estimated for given instance. $T_i$ is shown as actual output value of class for that given instance Average Error can be derived by using equation

$$\mathrm{A} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}-1}^{\mathrm{n}} \mathrm{E_i} \tag{6}$$

## FLOW CHARTS

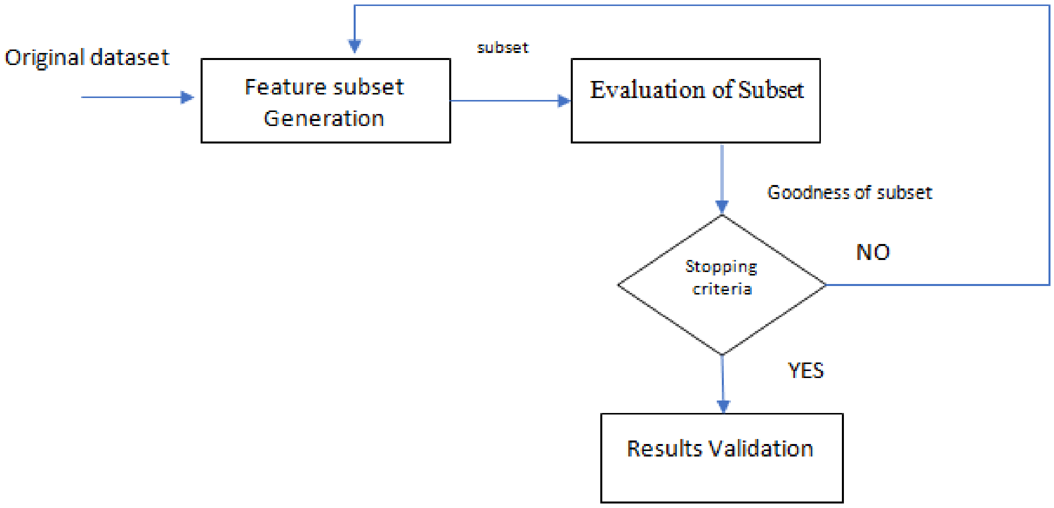
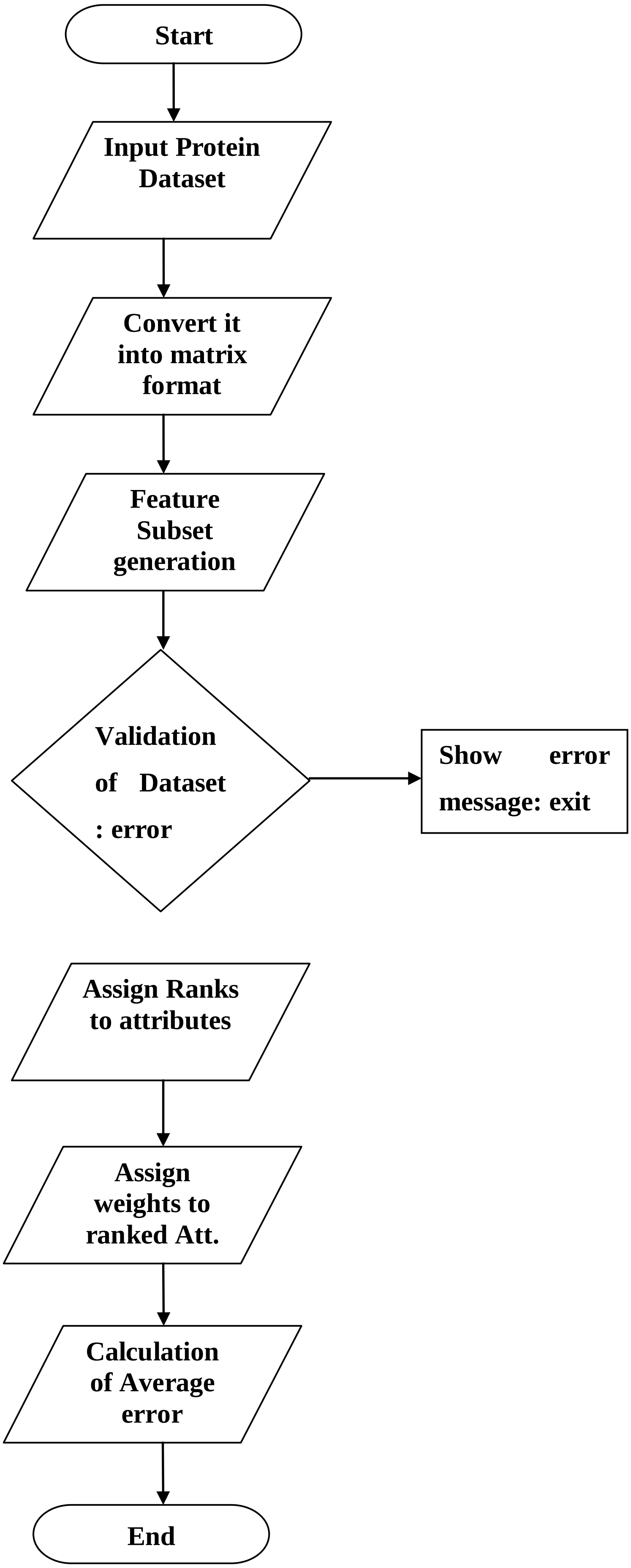


**Figure 3.1 Various Steps followed during execution process**

First of all the yeast dataset is browsed and loaded into the Matlab upon which processing is to be done. Then in feature subset Generation we apply relief algorithm on the dataset. In this we create the subsets with different ranks and assigning Weights to different ranks. Next step is to evaluate the Subset in which we check if the subset obtained is good enough which can help in reducing the error rate or not, if not then we backtrack to second step and again apply relief algorithm to generate ranks and corresponding weights. And if the Subset is good enough then we calculate the error rate by comparing the results of fuzzy values and already available data.


Start
Input Protein Dataset
Convert it into matrix format
Feature Subset generation
Validation of Dataset : error
Show error message: exit
Assign Ranks to attributes
Assign weights to ranked Att.
Calculation of Average error
End

Initilly we start the project by opening the Matlab. Then we input the protein dataset from the GUI interface through browse option which will load the yeast dataset into Matlab and then Popup window is shown giving success message.Then everything is converted into Matrix format for processing which is done internally by Matlab.Then on the dataset loaded we apply relief algorithm for feature selection from all the 8 attributes available. Then validation of data is done. If the data is valid then assignment of Ranks to all attributes is done by relief algorithm. Then weights are allocated based on the ranks.Then fuzzy based data mining is done in which rule base is created and and fuzzy output is created. Finally we calculate the average error rate by comparing the fuzzy output obtained with the output of the yeast dataset which is already with is in the form of Class attribute.

# CHAPTER 4

# RESULTS & DISCUSSION

Various authors have tried to predict the protein structure and protein functional activates, so we have the different techniques and methodologies available. But as with the passage of time the data required to be processed has increased manifold so we need to focus on the accuracy of the results predicted. For that the error rate that occurred in prediction need to be minimized. Because not only in research areas proteins are quite helpful in other aspects also like to invent new drugs, diagnose the patients with diseases. So in this key areas not only prediction will solve our purpose, we need to reduce or minimize the average error rate as well so that we can be more sure about the predicted end results.

As discussed earlier that the error rate which was given by other authors on the yeast dataset is 0.36806 which is minimum as per their study. On the other hand we have tried and successfully reduced the average error rate on the same dataset and the new improved results show that we have 0.35945 average error rate.

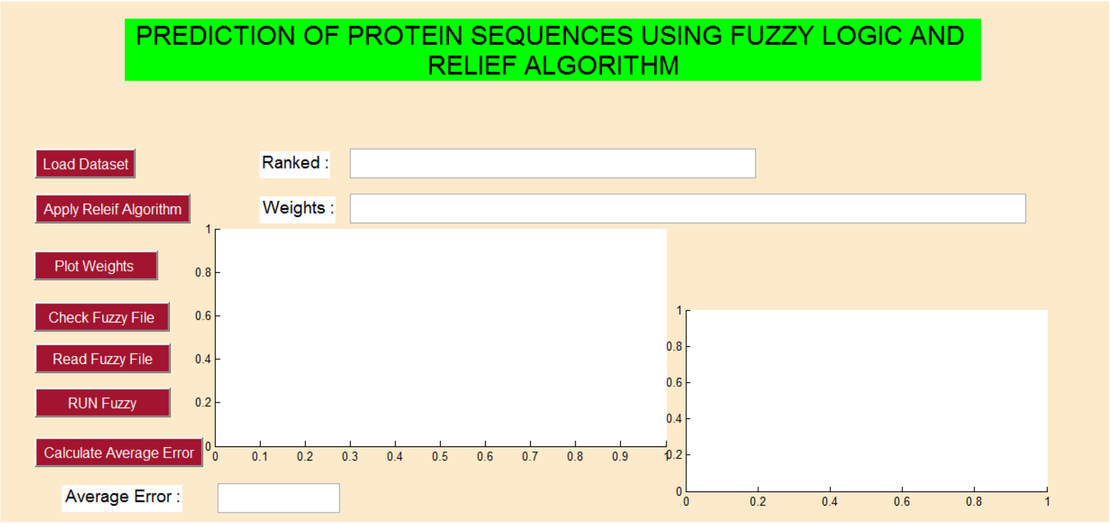


**Figure 4.1 Home Page of GUI interface**

This is the home page of the implemented thesis work. All the Buttons and Text boxes are placed on the home page. From this page itself we can give commands by simply pressing the buttons and therefore getting the results.

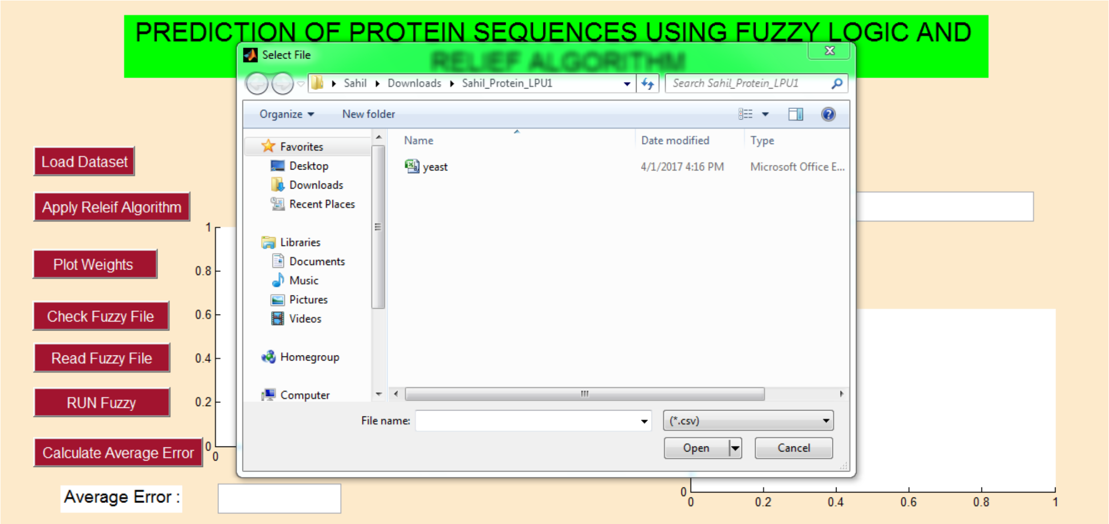
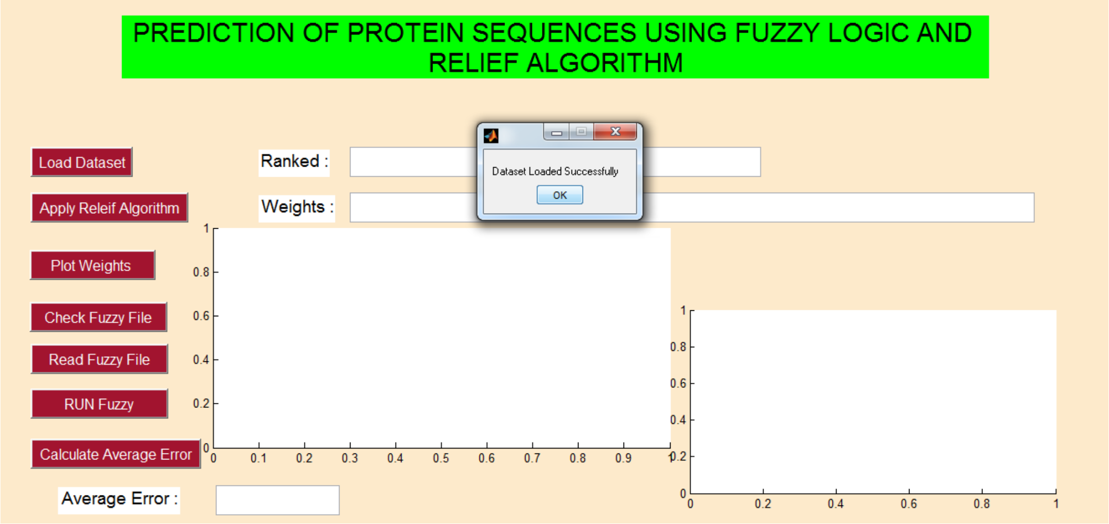


**Figure 4.2 Loading Dataset in Matlab**

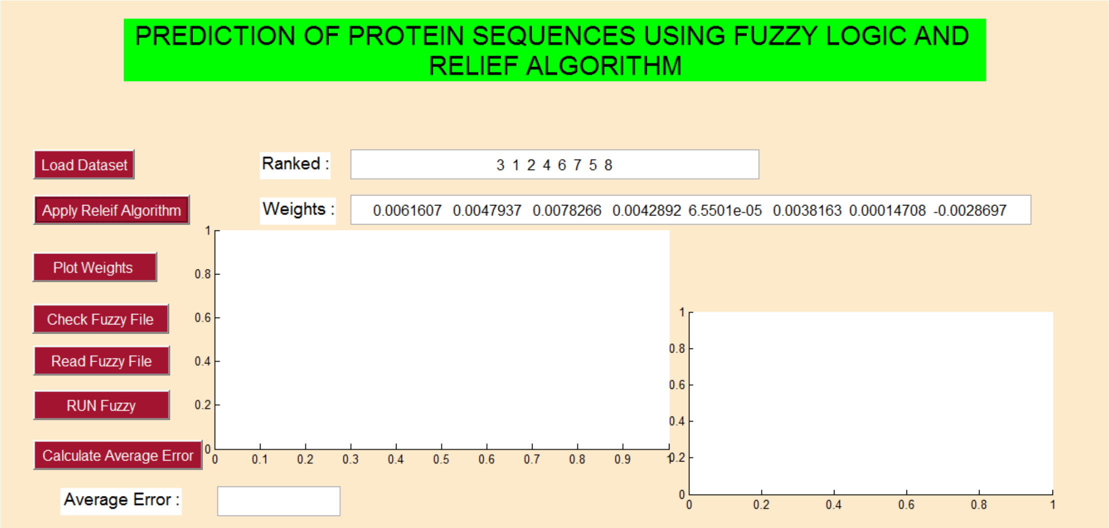
In the beginning we will browse the excel file which contain the dataset from ‘Load Dataset’ option which will prompt a message window that dataset is successfully loaded in the Matlab.



Dataset Loaded Successfully

**Figure 4.3 Popup showing dataset successfully loaded**

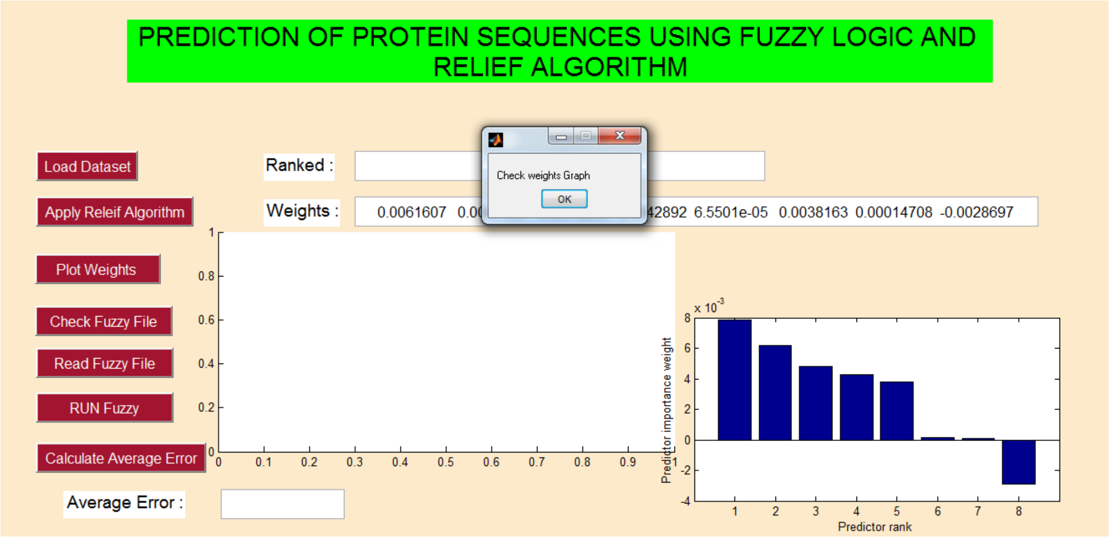
In the next step we will click on the second option which is ‘Apply Relief Algorithm’. In this, on the dataset we have loaded Feature selection is done inside the code and corresponding to that weights are assigned to each attribute according to the ranking.



Ranked : 3 1 2 4 6 7 5 8

Weights : 0.0061607 0.0047937 0.0078266 0.0042892 6.5501e-05 0.0038163 0.00014708 -0.0028697

**Figure 4.4 Applying Relief Algorithm**

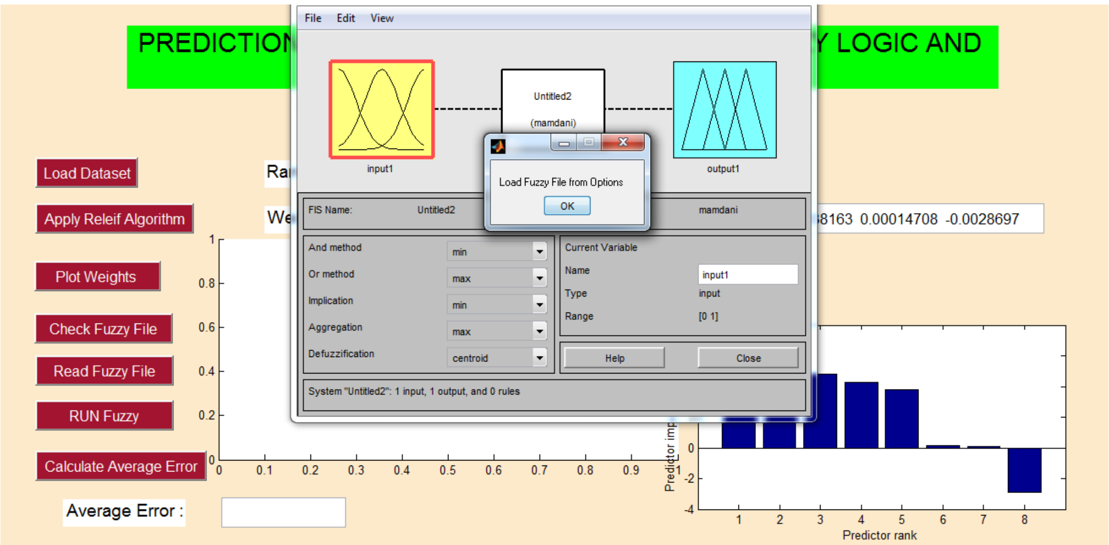


Check weights Graph

Weights : 0.0061607 0.0( 42892 6.5501e-05 0.0038163 0.00014708 -0.0028697

**Figure 4.5 Weight assignment to corresponding ranks of attributes**

Next comes the plotting of Weight graph, which shows Predictor’s rank and graphically shows that which attribute is contributing maximum and minimum in Protein Prediction.



8163 0.00014708 -0.0028697

Load Fuzzy File from Options

FIS Name: Untitled2 mamdani

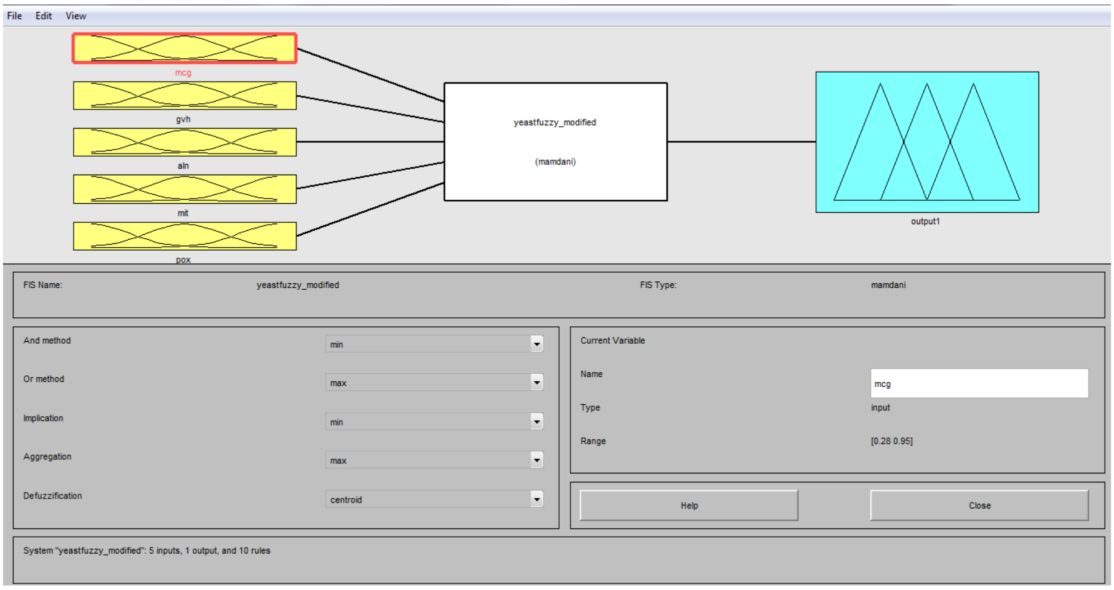
And method min
Or method max
Implication min
Aggregation max
Defuzzification centroid

Current Variable
Name input1
Type input
Range [0 1]

System "Untitled2": 1 input, 1 output, and 0 rules

**Figure 4.6 Popup to show fuzzy window**

After this we need to read fuzzy file from options. Fuzzy file will show all the membership functions and output class variable. It also shows various If-Then rules based on which Prediction is done.



FIS Name: yeastfuzzy_modified
FIS Type: mamdani

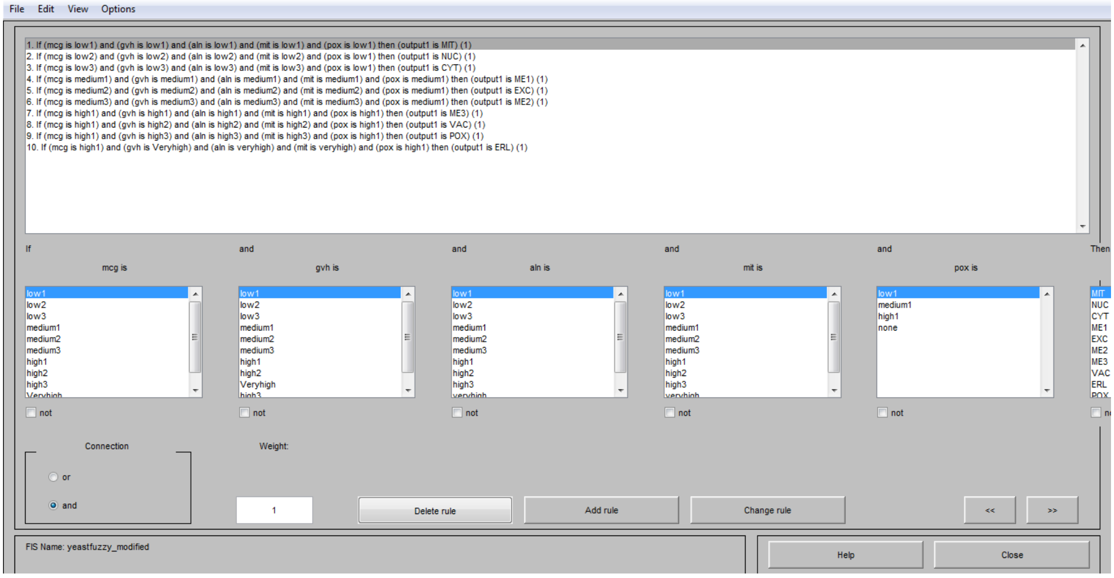
And method min
Or method max
Implication min
Aggregation max
Defuzzification centroid

Current Variable
Name mcg
Type input
Range [0.28 0.95]

System "yeastfuzzy_modified": 5 inputs, 1 output, and 10 rules

**Figure 4.7 Fuzzy window to show Input-Output Linkage**

Here we can see that out of 8 attributes we have selected the 5 attributes which have shown maximum contribution in Prediction. Also we can see that how Input variables and Output variables are connected.



1. If (mcg is low1) and (gvh is low1) and (aln is low1) and (mit is low1) and (pox is low1) then (output1 is MIT) (1)
2. If (mcg is low2) and (gvh is low2) and (aln is low2) and (mit is low2) and (pox is low1) then (output1 is NUC) (1)
3. If (mcg is low3) and (gvh is low3) and (aln is low3) and (mit is low3) and (pox is low1) then (output1 is CYT) (1)
4. If (mcg is medium1) and (gvh is medium1) and (aln is medium1) and (mit is medium1) and (pox is medium1) then (output1 is ME1) (1)
5. If (mcg is medium2) and (gvh is medium2) and (aln is medium2) and (mit is medium2) and (pox is medium1) then (output1 is EXC) (1)
6. If (mcg is medium3) and (gvh is medium3) and (aln is medium3) and (mit is medium3) and (pox is medium1) then (output1 is ME2) (1)
7. If (mcg is high1) and (gvh is high1) and (aln is high1) and (mit is high1) and (pox is high1) then (output1 is ME3) (1)
8. If (mcg is high1) and (gvh is high2) and (aln is high2) and (mit is high2) and (pox is high1) then (output1 is VAC) (1)
9. If (mcg is high1) and (gvh is high3) and (aln is high3) and (mit is high3) and (pox is high1) then (output1 is POX) (1)
10. If (mcg is high1) and (gvh is Veryhigh) and (aln is veryhigh) and (mit is veryhigh) and (pox is high1) then (output1 is ERL) (1)

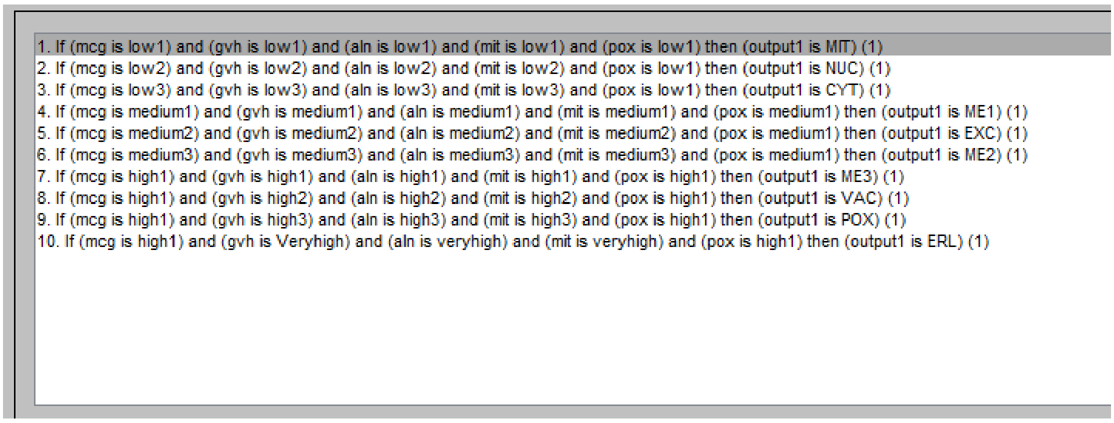
Connection
○ or
● and

Weight: 1

FIS Name: yeastfuzzy_modified

**Figure 4.8 Rule base showing If-Then rules**

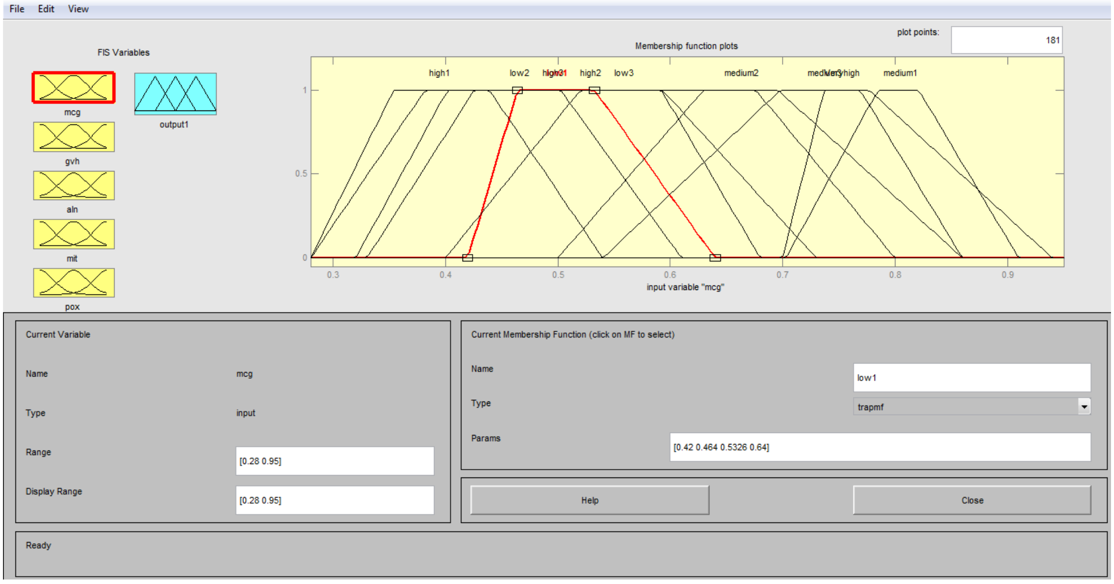
From View Option we can go to the Rules, on the basis of which input membership function changes and corresponding change Is observed in the Output class Variable.



1. If (mcg is low1) and (gvh is low1) and (aln is low1) and (mit is low1) and (pox is low1) then (output1 is MIT) (1)
2. If (mcg is low2) and (gvh is low2) and (aln is low2) and (mit is low2) and (pox is low1) then (output1 is NUC) (1)
3. If (mcg is low3) and (gvh is low3) and (aln is low3) and (mit is low3) and (pox is low1) then (output1 is CYT) (1)
4. If (mcg is medium1) and (gvh is medium1) and (aln is medium1) and (mit is medium1) and (pox is medium1) then (output1 is ME1) (1)
5. If (mcg is medium2) and (gvh is medium2) and (aln is medium2) and (mit is medium2) and (pox is medium1) then (output1 is EXC) (1)
6. If (mcg is medium3) and (gvh is medium3) and (aln is medium3) and (mit is medium3) and (pox is medium1) then (output1 is ME2) (1)
7. If (mcg is high1) and (gvh is high1) and (aln is high1) and (mit is high1) and (pox is high1) then (output1 is ME3) (1)
8. If (mcg is high1) and (gvh is high2) and (aln is high2) and (mit is high2) and (pox is high1) then (output1 is VAC) (1)
9. If (mcg is high1) and (gvh is high3) and (aln is high3) and (mit is high3) and (pox is high1) then (output1 is POX) (1)
10. If (mcg is high1) and (gvh is Veryhigh) and (aln is veryhigh) and (mit is veryhigh) and (pox is high1) then (output1 is ERL) (1)

**Figure 4.9 Fuzzy Rules**

Here we can see the individual Input Membership function and their Trapezoidal graph for all the rules written in the Rule Base.



plot points: 181

Current Variable

Name mcg

Type input

Range [0.28 0.95]

Display Range [0.28 0.95]

Current Membership Function (click on MF to select)

Name low1

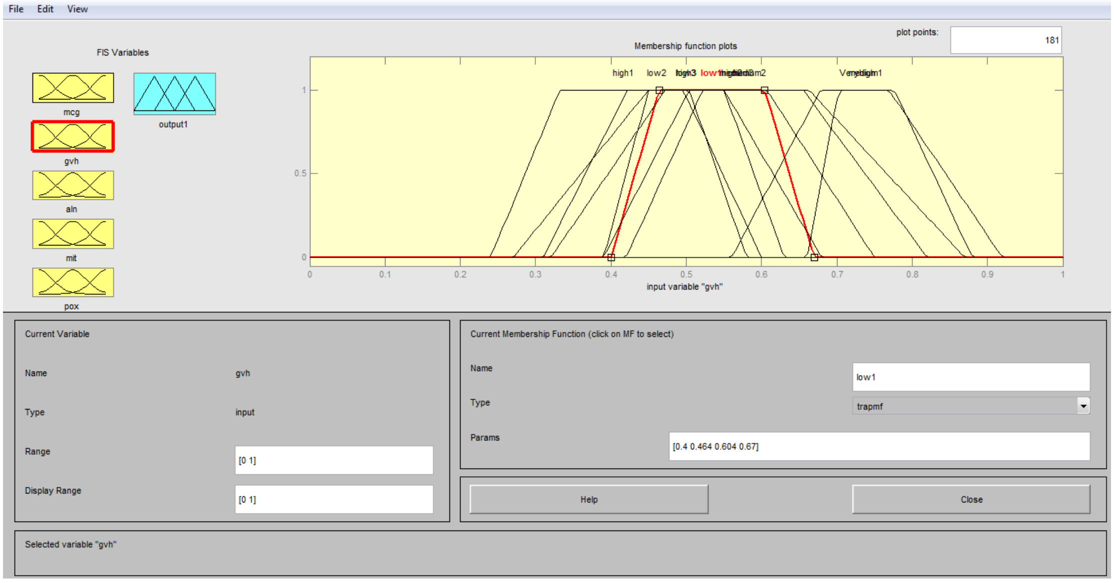
Type trapmf

Params [0.42 0.464 0.5326 0.64]

Ready

**Figure 4.10 Input 1 mcg Membership function Graph**

Different Membership function behaves in different ways so we have different graphs for all the 5 input functions.



plot points: 181

Current Variable

Name gvh

Type input

Range [0 1]

Display Range [0 1]

Current Membership Function (click on MF to select)

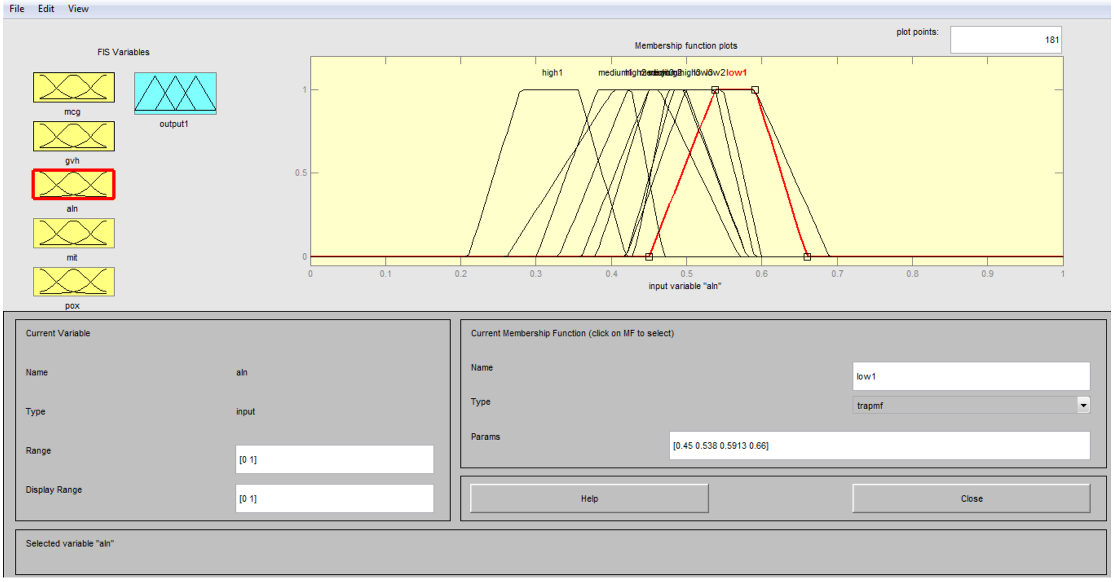
Name low1

Type trapmf

Params [0.4 0.464 0.604 0.67]

Selected variable "gvh"

**Figure 4.11 Input 2 gvh Membership function Graph**



plot points: 181

Current Variable

Name aln

Type input

Range [0 1]

Display Range [0 1]

Current Membership Function (click on MF to select)

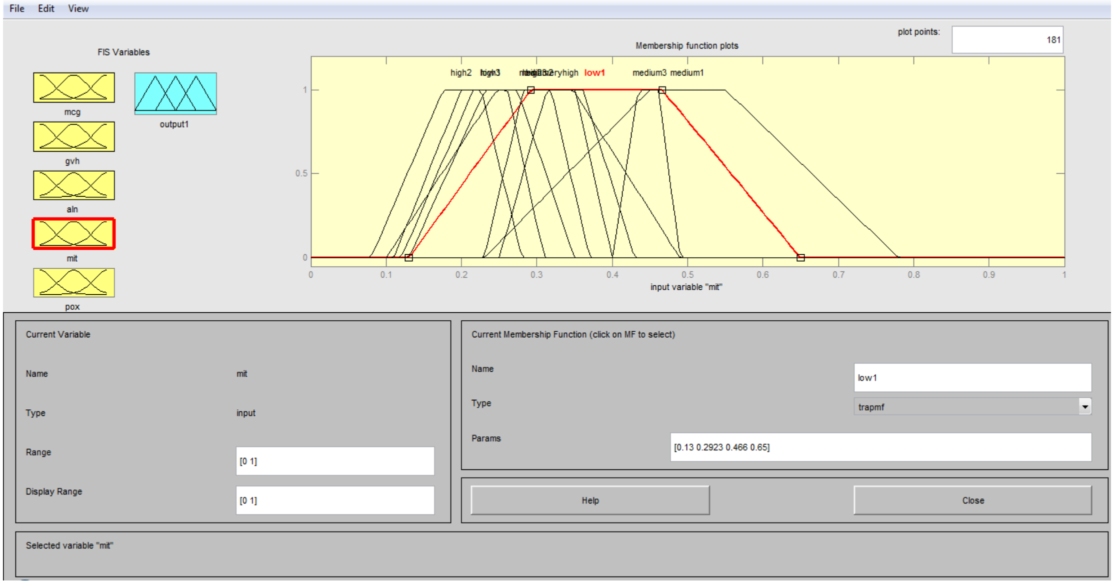
Name low1

Type trapmf

Params [0.45 0.538 0.5913 0.66]

Selected variable "aln"

Figure 4.12 Input 3 aln Membership function Graph



plot points: 181

Current Variable

Name mit

Type input

Range [0 1]

Display Range [0 1]

Current Membership Function (click on MF to select)

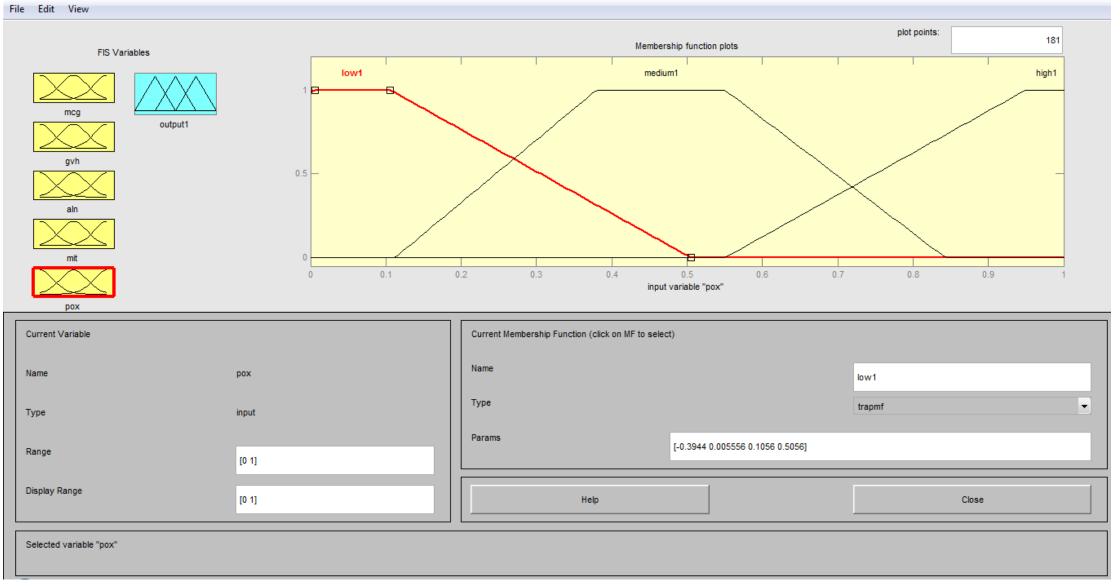
Name low1

Type trapmf

Params [0.13 0.2923 0.466 0.65]

Selected variable "mit"

Figure 4.13 Input 4 mit Membership function Graph



plot points: 181

Current Variable

Name pox

Type input

Range [0 1]

Display Range [0 1]

Current Membership Function (click on MF to select)

Name low1

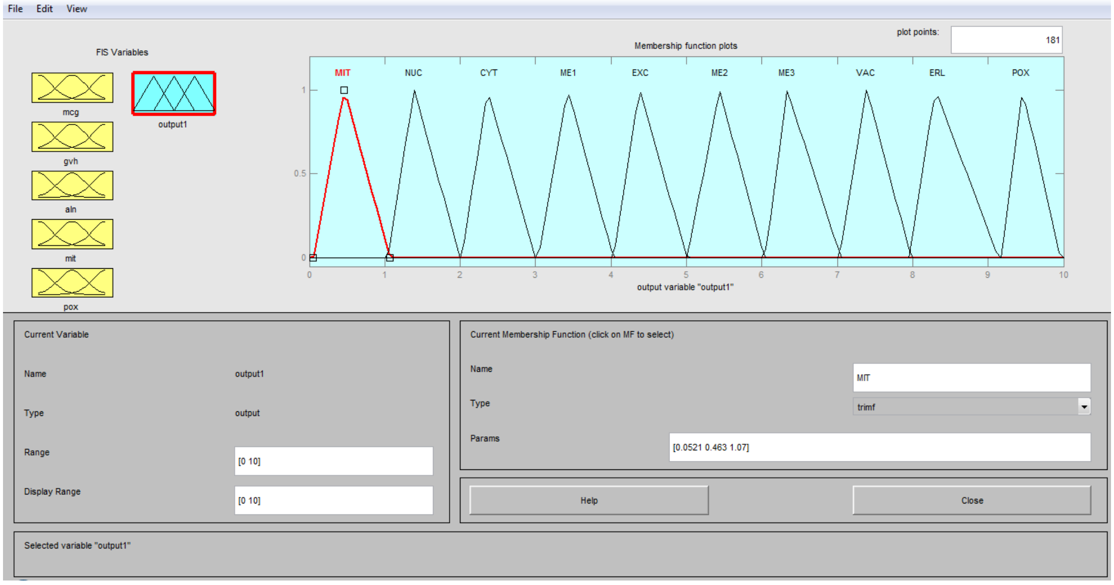
Type trapmf

Params [-0.3944 0.005556 0.1056 0.5056]

Selected variable "pox"

**Figure 4.14 Input 5 pox Membership function Graph**

This as we can see is the final graph which shows results in the form of triangular graph corresponding to all 5 input membership functions.



plot points: 181

Current Variable

Name output1

Type output

Range [0 10]

Display Range [0 10]

Current Membership Function (click on MF to select)

Name MIT

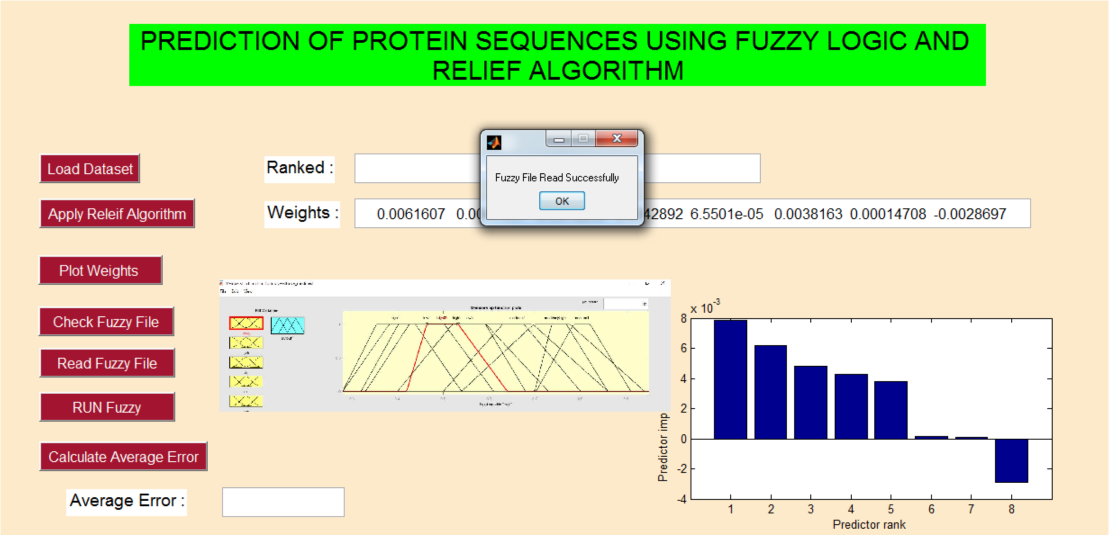
Type trimf

Params [0.0521 0.463 1.07]

Selected variable "output1"

**Figure 4.15 Output Class Membership Function**

The next step is to execute this fuzzy file, for that we need to read fuzzy file.



PREDICTION OF PROTEIN SEQUENCES USING FUZZY LOGIC AND RELIEF ALGORITHM

Ranked :

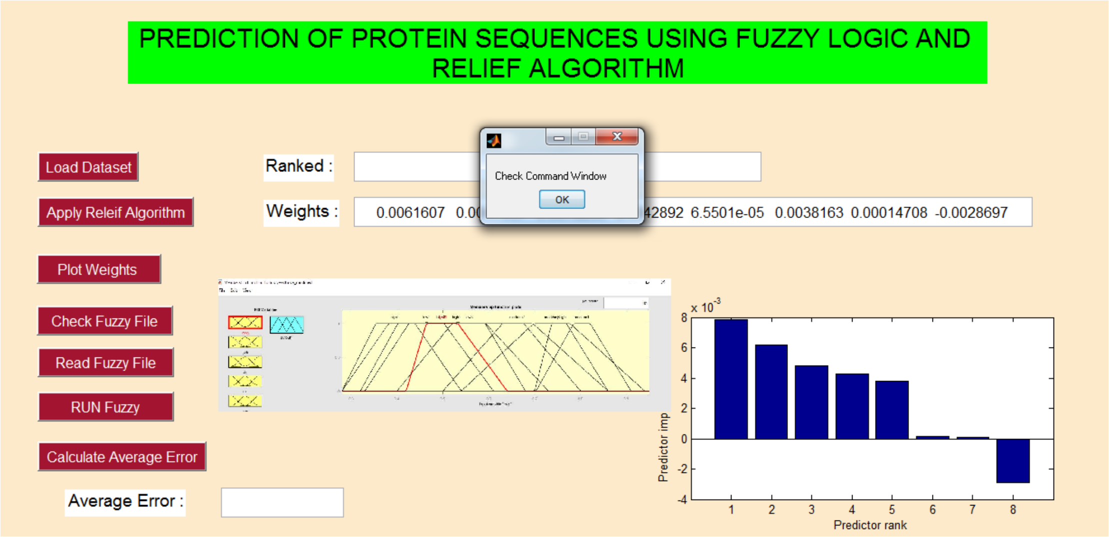
Weights : 0.0061607 0.00 42892 6.5501e-05 0.0038163 0.00014708 -0.0028697

Fuzzy File Read Successfully

Average Error :

**Figure 4.16 Execution of Fuzzy File**

After Execution of Fuzzy file we get the output fuzzy values to all the membership function which can be seen by minimizing this window and going to Command window.



PREDICTION OF PROTEIN SEQUENCES USING FUZZY LOGIC AND RELIEF ALGORITHM

Ranked :

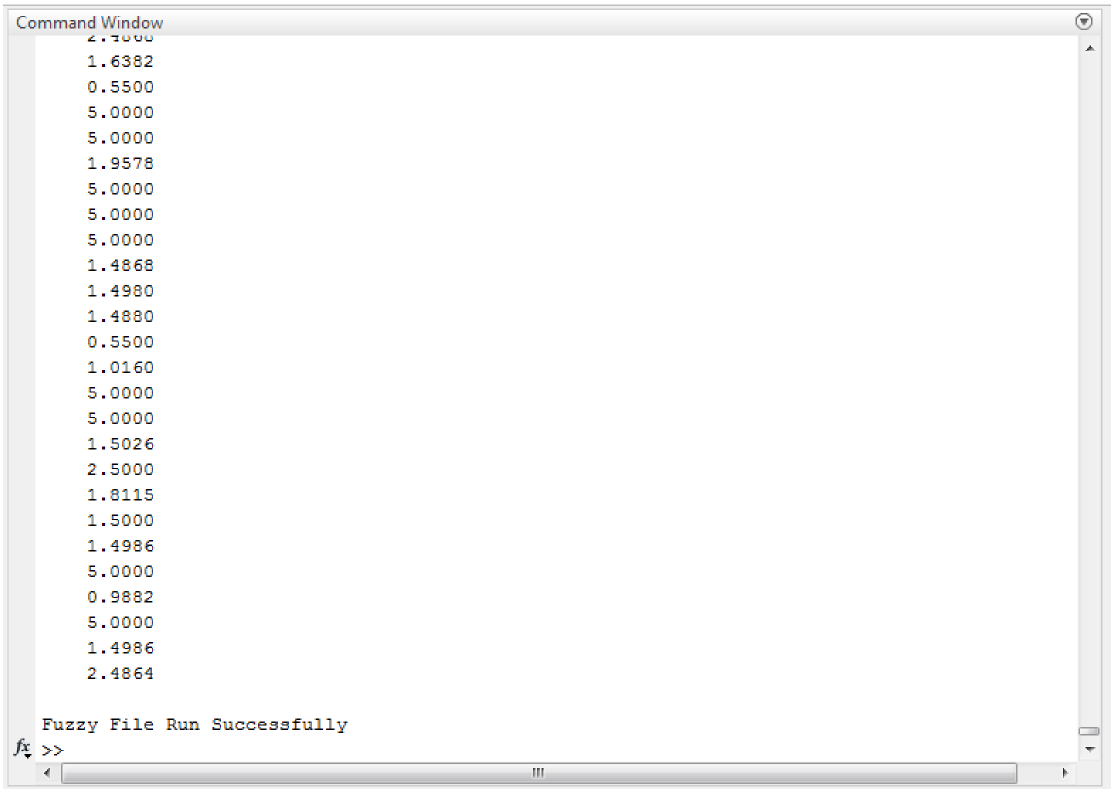
Weights : 0.0061607 0.00 42892 6.5501e-05 0.0038163 0.00014708 -0.0028697

Check Command Window

Average Error :

**Figure 4.17 Results of Fuzzy File on Command Window**

This is how Command window looks. It contain all the fuzzy values which are required for further calculations.

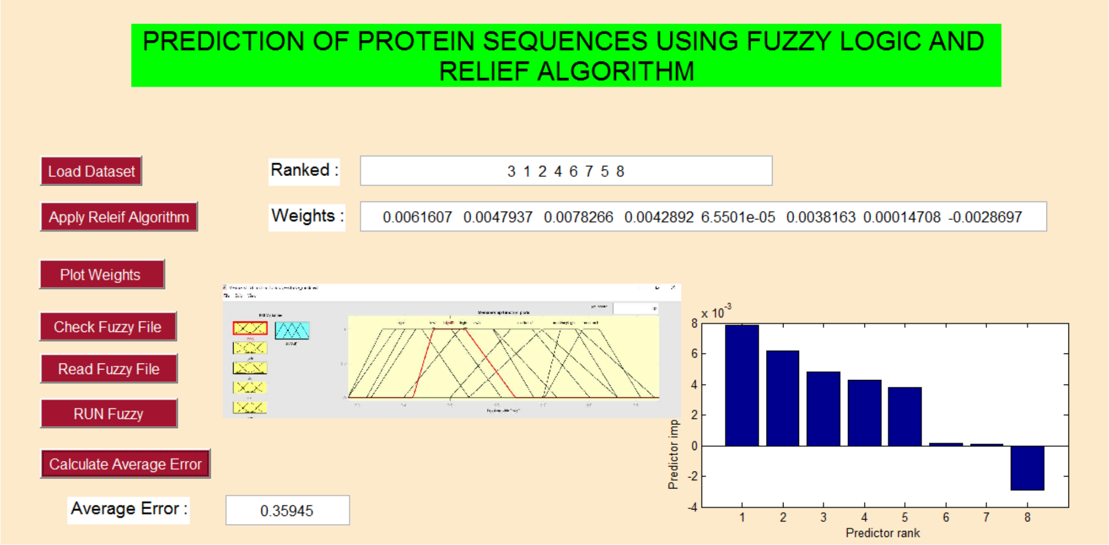


```
    1.6382
    0.5500
    5.0000
    5.0000
    1.9578
    5.0000
    5.0000
    5.0000
    1.4868
    1.4980
    1.4880
    0.5500
    1.0160
    5.0000
    5.0000
    1.5026
    2.5000
    1.8115
    1.5000
    1.4986
    5.0000
    0.9882
    5.0000
    1.4986
    2.4864

Fuzzy File Run Successfully
>>
```

**Figure 4.18 Output on Command Window**

After this step the final step is to calculate the average error rate which is calculated by given formula of Error calculation from fuzzy values obtained and the values already available in the dataset.


PREDICTION OF PROTEIN SEQUENCES USING FUZZY LOGIC AND RELIEF ALGORITHM

Ranked : 3 1 2 4 6 7 5 8

Weights : 0.0061607 0.0047937 0.0078266 0.0042892 6.5501e-05 0.0038163 0.00014708 -0.0028697

Average Error : 0.35945

**Figure 4.19 Calculated Average Error Rate**

# CHAPTER 5

# CONCLUSION & FUTURE SCOPE

Fuzzy method can be used to efficiently analyze the Bioinformatics. Some of the fuzzy applications are intelligent control and pattern classification which can be incorporated for structuring the proteomic data. This method help different user to prepare their knowledge and to simplify the procedure to recover and explore 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.

So the improvement in the average error rate using Relief algorithm over the previous existing technique is concluded in the given table below:

**Table 5.1 Result Comparison**

| Description | Input Membership Function | Output Membership Function | Results |
|---|---|---|---|
| Existing error rate | Trapezoidal | Triangular | 0.36806 |
| New Improved error rate | Trapezoidal | Triangular | 0.35945 |

Various extraction and analysis techniques are also discussed for mining and analyzing 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

Usages of data mining to Bioinformatics fuse quality finding, protein work space acknowledgment, work subject distinguishing proof, protein work inference, disease investigation, contamination reckoning, affliction treatment headway, protein and quality affiliation sort out proliferation, data purification, and protein cells region estimate.

The approach of the moderately new teach, data innovation, has helped in the improvement of Bioinformatics as an establishment of biotechnology. Its primary concentration is on the natural data administration, free of the birthplace or representation of the biotechnological information. The computational and numerical methodologies in dissecting different natural information supplemented with different systems of laser mass spectrometry and X-beam procedures, creating information about the structure and capacity between connections of bimolecular would additionally fortify Bioinformatics[17]. This train is empowering life sciences to imagine novel medication revelation and in addition tranquilize conveyance frameworks to gain biotechnological ground much quicker. Such creations accomplish significance in the present situation of licenses and WTO administration. There is most likely the appearance of Bioinformatics will revolutionalize biotechnology. The achievement of biosciences would rely on the databases. The contribution of industry has set Bioinformatics in a post-genomic age and now it has framed its own particular society, the International Society for Computational Biology20. It is normal that soon a hypothetical researcher of the post genomic period with a comprehension of systems, pathways and falls, flag transduction, digestion system and hereditary direction would have the capacity to manage rDNA look into in a way like 'turn around transcriptase'. The quickening improvements in data innovation would make basic the accessibility of perfect equipment as well, which can adapt to the accessible Bioinformatics instruments. There is doubtlessly Bioinformatics has gone to an age inside recent years to wind up distinctly a bonafide teach'. The coming of the web and WWW has created Bioinformatics massively, making it competent to shape another general public of biotechnology and natural researchers, who may call themselves Bioinformaticians and the train itself might be authored bio-data innovation at standard with data innovation.

# REFERENCES/BIBLIOGRAPHY

[1] J. L. C. Ramos, R. E. D. Silva, R. L. Rodrigues, J. C. S. Silva, and A. S. Gomes, “A Comparative Study between Clustering Methods in Educational Data Mining,” vol. 14, no. 8, 2016.

[2] B. E. V. Comendador, L. W. Rabago, and B. T. Tanguilig, “An educational model based on Knowledge Discovery in Databases (KDD) to predict learner’s behavior using classification techniques,” *2016 IEEE Int. Conf. Signal Process. Commun. Comput.*, pp. 1–6, 2016.

[3] J. P. A. Ioannidis, “Why most published research findings are false: Author’s reply to Goodman and Greenland [7],” *PLoS Medicine*, vol. 4, no. 6. pp. 1132–1133, 2007.

[4] “D. M. Creasy and B. T. Tanguilig Bioinformatics, Data ming tasks vol. 17, no. 24, pp. 3481–3571, 2014” .

[5] “Guest Editorial : Data Mining in Bioinformatics , Biomedicine , and Healthcare Informatics,” vol. 18, no. 2, p. 2306988, 2014.

[6] “2148d7aa1af45d7b4655ccb39013487cc8adb461 @ sciblogs.co.nz.” .

[7] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell, “Probability-based protein identification by searching sequence databases using mass spectrometry data,” *Electrophoresis*, vol. 20, no. 18, pp. 3551–3567, 1999.

[8] “164d5730f2fef68cf287acb5971386339bea474f @ www.ncbi.nlm.nih.gov.” .

[9] “protein-structure-prediction-using-coarse-grain-force-fields-6530787 @ www.slideshare.net.” .

[10] J.-K. Yu, Y.-D. Chen, and S. Zheng, “An integrated approach to the detection of colorectal cancer utilizing proteomics and bioinformatics.,” *World J. Gastroenterol.*, vol. 10, no. 21, pp. 3127–31, 2004.

[11] S. Mission, S. Mission, H. Topics, and H. Topics, “The Eighth Annual Bio-

Ontologies Meeting," *PLoS Comput. Biol.*, vol. 1, pp. 2005–2006, 2005.

[12] A. Rai, "Concepts of Bioinformatics," *Online Content Creat. Manag. an eLearning Environ.*, pp. 333–352, 2001.

[13] N. H. Reyes, "Fuzzy Inference Systems," 2012.

[14] "node16 @ www.cs.bris.ac.uk." .

[15] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "DNA-inspired online behavioral modeling and its application to spambot detection," 2016.

[16] A. Cuzzocrea, "Provenance Research Issues and Challenges in the Big Data Era," pp. 684–686, 2015.

[17] alfredo Mann. M Nachman, 1., & Pe'er, D. (2000). Issues , Challenges , Tools and Good Practices in " vol. 5963, no. c, 7, pp. 7–11, 2013

[18] E. P. Xing *et al.*, "Petuum : A New Platform for Distributed Machine Learning on Big Data," vol. 1, no. 2, pp. 49–67, 2015.

[19] S. V. Koneru and D. B. S, "Divide and Conquer Approach to Contact Map Overlap Problem using 2D-Pattern Mining of Protein Contact Networks," vol. 5963, no. c, pp. 1–9, 2015.

[20] S. Manocha, "A Novel Hybrid Approach for Secure Cloud Mining using Lossless Image Format," vol. 98, no. 7, pp. 7–11, 2014.

[21] A. Katal, "Big Data : Issues , Challenges , Tools and Good Practices," pp. 404–409, 2013.

[22] X. Hu, "Data Mining and Its Applications in Bioinformatics : Techniques and Methods," p. 4577, 2011.

[23] D. Wegener, S. Rossi, F. Buffa, M. Delorenzi, and R. Stefan, "Towards an Environment for Data Mining based Analysis Processes in Bioinformatics & Personalized Medicine," pp. 570–577, 2011.

[24] S. Dixon and X. Yu, "Bioinformatics Data Mining Using Artificial Immune

Systems and Neural Networks,” pp. 440–445, 2010.

[25] Jacq. N., Blanchet, C., Combct. C., Cornillot, E.. Durct. L., Kurata. K. I., ... & Breton, V. (2004). Grid as a hioinforrnatic tool. Parallel Computing, 30(9), 1093—I 107.

[26] Agrawal, R., Imieliñski, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In ALM SIGMOD Record(Vol. 22, No. 2, pp. 207-2 16). ACM. Atluri, G.. Gupta. R.. Fang. G.. Pandey. G.. Steinhach. M.. & Kumar. V. (2009). Association analysis techniques for biointrniatics problems. Inl)u)u1ormaucs and (‘omputailoital Biology (pp. 1-13). Springer Berlin Heidelberg

[27] Friedman, N., Linial, M., Nachman, 1., & Pe’er, D. (2000). Using Bayesian networks to analyze expression data. Journal of computational biology, 7(3-4). 601-620.

[28] Kumar, C., & Mann. M. (2009). Bioinforrnatics analysis of mass spectromeiry-based proleomics data sets. FEBS letters, 583(11), 1703-1712.

[29] S. F. Rosario and K. Thangadurai, “RELIEF : Feature Selection Approach,” *Int. J. Innov. Res. Dev.*, vol. 4, no. 11, pp. 218–224, 2015.

[30] L. Gao, T. Li, L. Yao, and F. Wen, “Research and application of data mining feature selection based on relief algorithm,” *J. Softw.*, vol. 9, no. 2, pp. 515–522, 2014.