

**Georgia State University**  
**ScholarWorks @ Georgia State University**

---

Computer Science Dissertations

Department of Computer Science

---

Spring 3-21-2012

# Protein Tertiary Model Assessment Using Granular Machine Learning Techniques

Anjum A. Chida

Follow this and additional works at: [https://scholarworks.gsu.edu/cs\\_diss](https://scholarworks.gsu.edu/cs_diss)

---

## Recommended Citation

Chida, Anjum A., "Protein Tertiary Model Assessment Using Granular Machine Learning Techniques." Dissertation, Georgia State University, 2012.  
[https://scholarworks.gsu.edu/cs\\_diss/65](https://scholarworks.gsu.edu/cs_diss/65)

This Dissertation is brought to you for free and open access by the Department of Computer Science at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Computer Science Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact [scholarworks@gsu.edu](mailto:scholarworks@gsu.edu).

# PROTEIN TERTIARY MODEL ASSESSMENT USING GRANULAR MACHINE LEARNING TECHNIQUES

by

ANJUM A. CHIDA

Under the Direction of Dr. Robert W. Harrison and Dr. Yan-Qing Zhang

## ABSTRACT

The automatic prediction of protein three dimensional structures from its amino acid sequence has become one of the most important and researched fields in bioinformatics. As models are not experimental structures determined with known accuracy but rather with prediction it's vital to determine estimates of models quality. We attempt to solve this problem using machine learning techniques and information from both the sequence and structure of the protein. The goal is to generate a machine that understands structures from PDB and when given a new model, predicts whether it belongs to the same class as the PDB structures (correct or incorrect protein models). Different subsets of PDB (protein data bank) are considered for evaluating the prediction potential of the machine learning methods. Here we show two such machines, one using SVM (support vector machines) and another using fuzzy decision trees (FDT). First using a preliminary encoding style SVM could get around 70% in protein model quality assessment accuracy, and improved Fuzzy Decision Tree (IFDT) could reach above 80% accuracy. For the purpose

of reducing computational overhead multiprocessor environment and basic feature selection method is used in machine learning algorithm using SVM.

Next an enhanced scheme is introduced using new encoding style. In the new style, information like amino acid substitution matrix, polarity, secondary structure information and relative distance between alpha carbon atoms etc is collected through spatial traversing of the 3D structure to form training vectors. This guarantees that the properties of alpha carbon atoms that are close together in 3D space and thus interacting are used in vector formation. With the use of fuzzy decision tree, we obtained a training accuracy around 90%. There is significant improvement compared to previous encoding technique in prediction accuracy and execution time. This outcome motivates to continue to explore effective machine learning algorithms for accurate protein model quality assessment.

Finally these machines are tested using CASP8 and CASP9 templates and compared with other CASP competitors, with promising results. We further discuss the importance of model quality assessment and other information from proteins that could be considered for the same.

**INDEX WORDS:** Protein 3D Structures, Protein model assessment, Feature selection, Support vector machines, Decision tree, Fuzzy ID3 and Machine learning.

PROTEIN TERTIARY MODEL ASSESSMENT USING GRANULAR MACHINE LEARNING  
TECHNIQUES

by

ANJUM A. CHIDA

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2012

Copyright by  
Anjum Ashraf Chida  
2012

PROTEIN TERTIARY MODEL ASSESSMENT USING GRANULAR MACHINE LEARNING  
TECHNIQUES

by

ANJUM A. CHIDA

Committee Chair: Robert Harrison  
Yan-Qing Zhang

Committee: Raj Sunderraman  
Yichuan Zhao

Electronic Version Approved:

Office of Graduate Studies  
College of Arts and Sciences  
Georgia State University  
May 2012

## **DEDICATION**

*To my daughter Ayra*

## ACKNOWLEDGEMENTS

I wish to take this opportunity to thank many people without whom this dissertation would not have been accomplished. First and foremost, I would like to thank my advisor, Dr. Yanqing Zhang. I was able to achieve this task, with his help, guidance, encouragement, and the time that he has spent on directing my dissertation. I wish to thank Dr. Harrison for sharing his expertise in field of biochemistry and biotechnology and also for accepting to be my dissertation committee chair. I also wish to thank my committee members, Dr. Sunderraman and Dr. Zhao, for taking time to evaluate my simulation results and to review my dissertation document. I would like to express my gratitude to Dr. Nael Abu-halaweh for providing the improved fuzzy decision tree algorithm for this research work.

Last but not least, I wish to express my gratitude to my parents and my brother who have pushed me this far. I would certainly like to thank my husband Ashraf, who supported me and encouraged me all these years of my education.



## TABLE OF CONTENTS

ACKNOWLEDGEMENTS .....	v
LIST OF TABLES.....	ix
LIST OF FIGURES.....	xi
LIST OF ABBREVIATIONS. ....	xii
CHAPTER 1. INTRODUCTION.....	1
CHAPTER 2. PROTEIN STRUCTURE PREDICTION AND MODEL ASSESSMENT A REVIEW .....	8
2.1 Protein Structure Prediction -- Overview .....	10
2.1.1 Homology Modeling .....	11
2.1.2 Protein threading .....	13
2.1.3 Ab initio methods .....	14
2.1.4 Prediction Errors .....	15
2.1.5 Application of Structure Prediction .....	17
2.2 Literary Review of Current Model Assessment Techniques.....	19
CHAPTER 3. MACHINE LEARNING TECHNIQUES .....	26
3.1 Support Vector Machines.....	27
3.1.1 Nonlinear SVM - Kernel Method .....	31
3.1.2 SVM Software .....	32
3.2 Decision Trees .....	33
3.2.1 Improved Fuzzy ID3 Algorithm .....	35

3.2.2	<i>Extended Improved Fuzzy ID3 Algorithm</i>	38
<b>CHAPTER 4. PRELIMINARY ENCODING SCHEME</b>		<b>41</b>
4.1	Enumeration of Protein Structure	42
4.2	Vector Formation	43
4.3	Implementation Using SVM	44
4.3.1	<i>Simulation I</i>	45
4.3.2	<i>Simulation II</i>	47
4.3.3	<i>Feature Selection Algorithm</i>	48
4.4	Implementation Using IFID3	50
<b>CHAPTER 5. ENHANCED ENCODING SCHEME</b>		<b>53</b>
5.1	Methodology	57
5.2	Enumeration of Features	58
5.2.1	<i>Principle of Polarity</i>	58
5.2.2	<i>Secondary Structure of Proteins</i>	61
5.3	Vector Formation	64
5.4	Implementation Results	65
5.5	Meaningful Rules and Inference	68
<b>CHAPTER 6. TESTING USING CASP TEMPLATES</b>		<b>72</b>
6.1	Synopsis on CASP Experiments	73
6.2	Template Based Modeling in light of CASP	75
6.3	Testing EE_IFDT Algorithm with CASP Dataset	77

6.3.1	<i>Testing Using CASP8 and CASP9 Templates –Testing Phase I</i>	78
6.3.2	<i>Comparison with other Model Assessment Techniques -Testing Phase II and III</i>	80
<b>CHAPTER 7. FUTURE RESEARCH AVENUES</b>		<b>91</b>
7.1	<b>Scoring Technique</b>	<b>91</b>
7.2	<b>Future enhancements in current methodology</b>	<b>92</b>
7.2.1	<i>Type-2 Fuzzy Decision Tree Algorithm</i>	93
7.3	<b>Extracting features distinguishing good protein models from bad ones</b>	<b>93</b>
7.3.1	<i>Feature Pool</i>	94
7.3.2	<i>New Random Fuzzy Forest</i>	94
<b>CHAPTER 8. CONCLUSIONS</b>		<b>96</b>
<b>BIBLIOGRAPHY</b>		<b>101</b>
<b>APPENDIX</b>		<b>112</b>

## LIST OF TABLES

<b>TABLE 2.1 OVERVIEW OF CURRENT MODEL ASSESSMENT METHODS.....</b>	<b>23</b>
<b>TABLE 4.1 ENCODING SCHEME: PROFILE + DISTANCE MATRIX .....</b>	<b>46</b>
<b>TABLE 4.2 ENCODING SCHEME: BLOSUM MATRIX + DISTANCE MATRIX.....</b>	<b>46</b>
<b>TABLE 4.3 ACCURACY BEFORE FEATURE SELECTION .....</b>	<b>47</b>
<b>TABLE 4.4 ACCURACY AFTER FEATURE SELECTION.....</b>	<b>49</b>
<b>TABLE 4.5 COMPARISON OF ACCURACIES BEFORE AND AFTER FEATURE SELECTION .....</b>	<b>50</b>
<b>TABLE 4.6 SEVEN FOLD RESULTS USING IFID3.....</b>	<b>51</b>
<b>TABLE 5.1 8-TO-3 STATE REDUCTION METHOD IN SECONDARY STRUCTURE ASSIGNMENT.....</b>	<b>63</b>
<b>TABLE 5.2 PRELIMINARY ENCODING SEVEN FOLD RESULTS .....</b>	<b>66</b>
<b>TABLE 5.3 ENHANCED SPATIAL ENCODING SEVEN FOLD RESULTS .....</b>	<b>67</b>
<b>TABLE 5.4 COMPARISON OF TWO ENCODING SCHEMES .....</b>	<b>67</b>
<b>TABLE 5.5 AVERAGE NUMBER OF RULES .....</b>	<b>68</b>
<b>TABLE 6.1 RESULTS USING CASP TEMPLATES AS TEST DATA .....</b>	<b>79</b>
<b>TABLE 6.2 MQA METHODS CLASSIFICATION .....</b>	<b>82</b>
<b>TABLE 6.3 POSITIVE TEMPLATE RESULTS .....</b>	<b>84</b>

<b>TABLE 6.4 NEGATIVE TEMPLATE RESULTS .....</b>	<b>85</b>
<b>TABLE 6.5 PEARSON'S CORRELATION FOR TARGET T0635 .....</b>	<b>88</b>
<b>TABLE 6.6 PEARSON'S CORRELATION FOR TARGET T0578 .....</b>	<b>88</b>
<b>TABLE 6.7 PEARSON'S CORRELATION FOR TARGET T0635 .....</b>	<b>89</b>

## LIST OF FIGURES

<b>FIGURE 2.1 GROWTH OF BIOLOGICAL DATABASE .....</b>	<b>9</b>
<b>FIGURE 2.2 FLOWCHART OF PROTEIN STRUCTURE PREDICTION (PICTURE ADOPTED FROM [31]) .</b>	<b>16</b>
<b>FIGURE 2.3 PROTEIN STRUCTURE PREDICTION SOFTWARE TRENDS .....</b>	<b>20</b>
<b>FIGURE 3.1 A SEPARATING HYPER PLANE FOR A 2-D TRAINING SET [82]. .....</b>	<b>29</b>
<b>FIGURE 3.2 LINEAR SEPARATING HYPER PLANE FOR NON-SEPARABLE CASE .....</b>	<b>31</b>
<b>FIGURE 3.3 TRANSFORMATION TO HIGHER DIMENSIONAL SPACE.....</b>	<b>32</b>
<b>FIGURE 4.1 FEATURE VECTOR FORMATION .....</b>	<b>42</b>
<b>FIGURE 5.1 DATA STRUCTURE OF ALPHA CARBON ATOM .....</b>	<b>54</b>
<b>FIGURE 5.2 ENHANCED ENCODING SCHEME ILLUSTRATION I.....</b>	<b>55</b>
<b>FIGURE 5.3 ENHANCED ENCODING ILLUSTRATION II.....</b>	<b>55</b>
<b>FIGURE 5.4 VECTOR FORMATION.....</b>	<b>65</b>
<b>FIGURE 5.5 FUZZY DECISION TREE-I .....</b>	<b>69</b>
<b>FIGURE 5.6 FUZZY DECISION TREE-II.....</b>	<b>70</b>
<b>FIGURE 7.1 FEATURE POOL.....</b>	<b>94</b>

## LIST OF ABBREVIATIONS

BLOSUM	– Block Substitution Matrix
CASP	– Critical Assessment of Structure Prediction
DNA	– Deoxyribonucleic acid
DSSP	– Dictionary of Protein Secondary Structure
EE_IFDT	– Enhanced Encoding with Improved Fuzzy Decision Tree
GDT_TS	– Global Distance Test Total Score
IFID3	– Improved Fuzzy ID3
NMR	– Nuclear Magnetic Resonance
PDB	– Protein Data Bank
PSSM	– Position Specific Scoring Matrix
RBF	– Radial Basis Function
RMSD	– Root Mean Square Deviation
SOV	– Segment Overlap Measure
SRM	– Structural Risk Minimization
SVM	– Support Vector Machines
TBM	– Template Based Modeling

## CHAPTER 1. INTRODUCTION

Proteins are large polypeptides constructed from same set of twenty different amino acids. The primary structure is the specific sequence of amino acids specified by the genes. This linear string folds into an intricate three-dimensional structure that is unique to each protein. It is this three-dimensional structure that allows proteins to function. Thus in order to understand the details of protein function at a molecular level, one must understand protein structure and hence it is necessary to determine the three-dimensional structure. In structure biology protein structures are often determined by techniques like X-ray crystallography, NMR spectroscopy and electron microscope. A repository of these experimentally determined structures is organized as a data bank called Protein Data Bank (PDB). This data bank is freely accessible on the internet [1].

Around 90% of protein structures available in PDB have been determined by X-ray crystallography. X-ray crystallography can provide very detailed atomic information, showing every atom in a protein or nucleic acid along with atomic details of ligands, inhibitors, ions, and other molecules that are incorporated into the crystal. However, the process of crystallization is difficult and can impose limitations on the types of proteins that may be studied by this method. For example, X-ray crystallography is an excellent method for determining the structures of rigid proteins that form nice, ordered crystals. Flexible proteins, on the other hand, are far more difficult to study by this method because crystallography relies on having many, many molecules aligned in exactly the same orientation, like a repeated pattern in wallpaper. NMR spectroscopy



may also be used to determine the structure of proteins. The protein is purified, placed in a strong magnetic field, and then probed with radio waves. The major advantage of NMR spectroscopy is that it provides information on proteins in solution, as opposed to those locked in a crystal or bound to a microscope grid, and thus, NMR spectroscopy is the premier method for studying the atomic structures of flexible proteins. The technique is currently limited to small or medium proteins, since large proteins present problems with overlapping peaks in the NMR spectra. Electron microscopy is also used to determine structures of large macromolecular complexes. A beam of electrons is used to image the molecule directly. For a few particularly well-behaved systems, electron diffraction produces atomic-level data, but typically, electron micrographic experiments do not allow the researcher to see each atom. Overall predicting protein structure by experiment alone is not feasible in every case [2].

The success of genome sequencing program resulted in massive amounts of protein sequence data (that are produced by DNA sequencing) [HUMAN GENOME PROJECT]. The generation of a protein sequence is much easier than the determination of protein structure. The structure of the protein gives much more insight about its function than its sequence. Therefore computational methods for the prediction of protein structure from its sequence have been developed. These methods aim to predict protein three-dimensional structure from its primary sequence. *Ab initio* prediction methods use just the sequence of the protein based on the physical principles governing any molecular structure. These techniques usually require vast computational capabilities and have been tried on only small proteins sequence. Threading and Homology Modeling methods can build a 3D model for a protein of unknown structure from experimental structures of evolutionary related proteins. The theory behind this being, even though the

number of proteins is vast, there are limited number of tertiary structure motifs to which most proteins belong. Homology modeling is based on the assumption that two homologous proteins will share similar structure. Sequence alignment algorithms are used to search for homologous protein, the structure hence predicted is more accurate if the alignment between target and template proteins is good. In protein threading algorithms scans amino acid sequence of unknown structure against database of solved structure. As long as a detailed physicochemical description of protein folding principles does not exist, structure prediction is the only method available to see the structure of some proteins. Experts agree it is possible to construct high quality full length models for almost all single domain proteins by using best possible template structure in PDB and state-of-the-art modeling algorithm [1] [2] [3]. This suggests that the current PDB structure universe may be approaching completion. So it all comes down to selecting that model in a pool of models.

Protein structure prediction has been an important conundrum in field of bioinformatics and theoretical chemistry due to its importance in medicine, drug design, biotechnology, etc. Critical Assessment of Techniques for Protein Structure Prediction (CASP) is a worldwide competition for protein structure prediction conducted every two years [4]. The primary goal of CASP is to establish the capabilities and limitation of current methods of modeling protein structure from sequence. Methods are assessed on the basis of the analysis of a large number of blind predictions of protein structure. The targets for this competition are proteins where the experimental structure is not yet public, but will be available shortly. The accuracy of three-dimensional structure models are primarily evaluated using two metric. One is GDT\_TS [5], a multi threshold measure related to difference in position of main chain C $\alpha$  atoms between a

model and corresponding experimental structure. The other is the alignment accuracy AL0 [6] [7], showing how well the assigned amino acid positions accord with those in the experimental structure. Other user methods are also considered. Model quality prediction has always been included in CASP, but has received much attention only recently. If structure modeling field is to be taken seriously, it is critical that we develop for reliably informing users how accurate these models are or are not [3] [4] [5].

How to assess a complex protein model quality is a long-term problem. Usually model quality assessment (considered in CASP and other user algorithms) is done by comparison of the model to true structure of protein. Recently there are methods that aim at determining a scoring function with no knowledge of the true structure. Over the past two decades, several approaches have been developed to analyze correctness of protein structures and models. These methods use techniques like stereochemistry checks, molecular mechanics energy based functions, statistical potentials and machine-learning approaches to tackle the problem. Typically, the features taken into account are the molecular environment, hydrogen bonding, secondary structure, solvent exposure, pair-wise residue interactions and molecular packing. A rapid development of new methods in model quality assessment is taking place, the necessity of an unbiased evaluation of these methods led to their inclusion as a separate category in CASP. Very few methods and techniques aim at determining the model quality without the experimental structure [2] [3].

As models are not experimental structures determined with known accuracy but predictions its vital to present the user with corresponding estimates of model quality. We aim to obtain a learning algorithm that studies known structures from PDB and when given a protein model predicts whether it belongs to the same class as PDB structures. Since using a whole primary

protein sequence to determine a 3D protein structure is very difficult, it is necessary to design new intelligent algorithms to find key features from a large pool of relevant features in biology and geometry to effectively evaluate 3D protein models. The central focus of this study is to develop and implement new granular decision machines to find biologically meaningful features for assessing 3D protein structures accurately and efficiently. The methodology for the new granular decision machines is that multiple intelligent decision making methods are systematically organized in an integrated algorithm with accurate performance and high efficiency. Some traditional machine learning methods are black-box methods (such as neural networks, and support vector machines). Biologist and chemists really want to know how a decision is made (i.e., meaning, reasons). To solve the black-box problem, we aim to develop the granular decision machines that can perform meaningful knowledge discovery of protein structures. The important innovation is that the granular knowledge discovery algorithms can automatically generate protein structure assessment rules with both key sequence features and important geometrical features. This effort will lead to a better understanding of internal mechanism governing 3D protein structures such as how and why the key biological features and geometrical features can dominate a 3D protein structure, and what these critical sequence features and geometrical features are.

The amount and nature of information given to the machine learning system will have an impact on the final output regarding the quality measure of given 3D structure. There are various ways of representing a protein three dimensional structure, like backbone sketch of the protein, the entire distance matrix of alpha carbon atoms, a fractal dimension of the structure, 3D information with its sequence data etc. These methods of representing protein structure are most-

ly used in comparison and classification problems and they are well studied and researched fields. In order to be useful the decision model will have to be both accurate and efficient, it will be required to rank as many as 100 models in reasonable amounts of time. While the fuzzy decision tree algorithms developed will likely achieve the accuracy required, they may be too slow when run in a single process. Therefore it will be a necessity to develop parallel or granular algorithms that can scale effectively for high performance.

Next area of exploration will be in active learning. Machine learning requires training data, and active learning is a strategy for choosing training data to give the most accurate decision model. Current approaches use sequence identity to try to form a set of non-redundant proteins. Active learning would identify a non-redundant set by eliminating those proteins that do not add information to the decision model, and therefore should be much more rigorous.

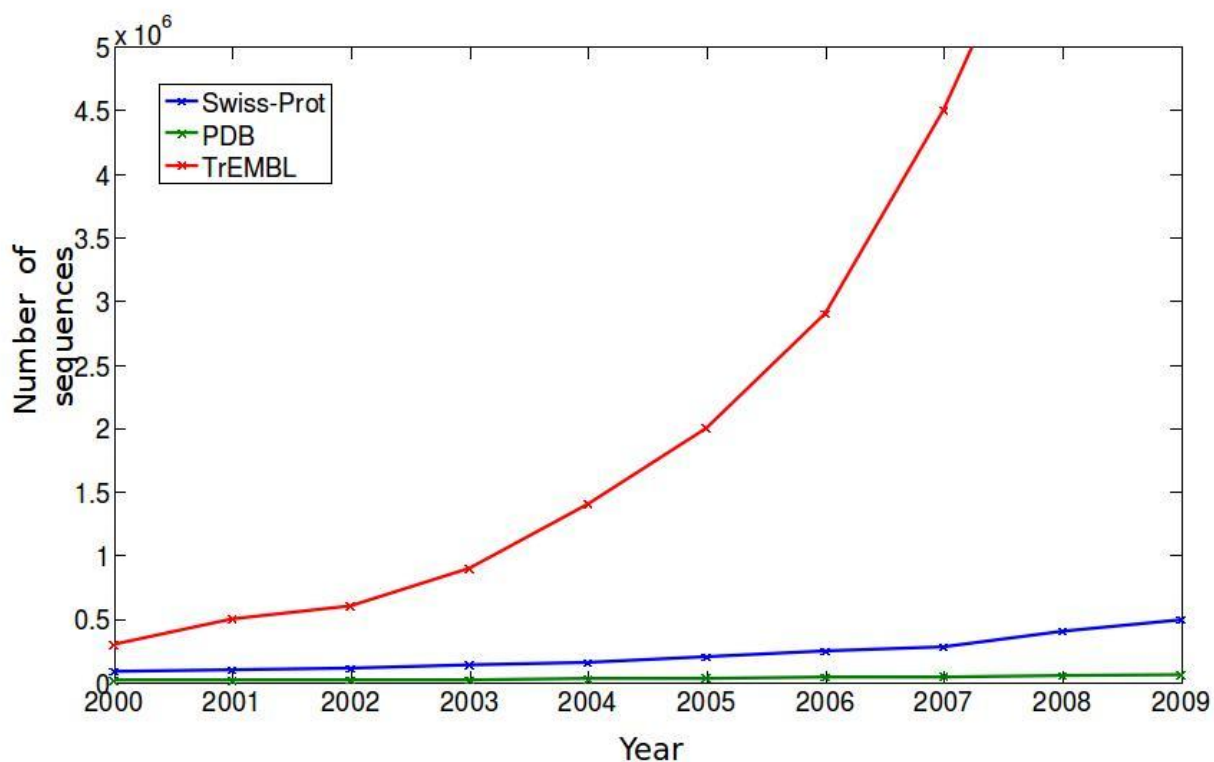
Solving this problem will help study problems from different domains that have the same intensity of information. There are many complex and important application problems with huge geometric factors. A short list of common problems where geometry is a critical factor would include: social and computer network structure, traffic analysis, computer vision. Biomedical examples include 3D structural features of a protein that are directly related to basic functionality and are crucial for drug design.

The next chapters (chapter 2) gives back ground information on protein three dimensional structure prediction using experimental techniques and computational techniques. It also has a brief discussion on prediction errors and application of protein models and it gives overview on current model assessment technique as well. Chapter 3 gives back ground on machine learning techniques, in particular about Support Vector Machines and Fuzzy Decision Tress. Chapter 4

introduces a preliminary encoding scheme, its methodology, implementation and results using SVM and IFDT are discussed. Chapter 5 introduces enhanced encoding scheme with its description and results. Chapter 6 has a brief introduction on CASP, on conduction of competition, on current trends in CASP, on model quality assessment and few prominent competitors in CASP. Finally the chapter shows results of using CASP templates as testing data. Chapter 7 discusses future development in this research study and chapter 8 has the final concluding remarks.

## CHAPTER 2. PROTEIN STRUCTURE PREDICTION AND MODEL ASSESSMENT -- A REVIEW

Modern experimental methods for determining protein structure through X-ray crystallography or NMR spectroscopy can solve only a small fraction of proteins sequenced by the large-scale genome sequencing projects, because of technology limitations and time constraints. Currently, there are more than seven million protein sequences accumulated in the nonredundant protein sequence database (NR; accessible through the national Center for Biotechnology Information: <ftp://ftp.ncbi.nlm.nih.gov/blast/db/>) and fewer than eighty thousand protein structures in the Protein Data Bank (PDB; <http://www.rcsb.org.pdb/>). With these numbers at hand, it seems that the only way to bridge the ever growing gap between protein sequence and structure is computational structure modeling. The Figure 2.1 shows the growth of biological databases like Swiss-Prot, TrEMBL and PDB. The number of genes whose function we are determining using experiments is but a drop in the ocean compared to number of genes we have sequences and whose function is not known. The red line is the growth of protein sequences deposited in TrEMBL, a comprehensive protein sequence database. The blue line illustrates the growth proteins in TrEMBL whose function is known, or at least can be predicted with some reasonable accuracy. The green line is the growth in the proteins whose 3D structure has been solved. Note the logarithmically increasing gap between what we know (blue) and what we do not know (red). (Image courtesy of Predrag Radivojac ). Swiss-Prot stands for Swiss Institute of Bioinformatics containing non-redundant protein sequence database. These data base is manually annotated. TrEMBL contains high quality computationally analyzed records.



**FIGURE 2.1 GROWTH OF BIOLOGICAL DATABASE**

With improvement of prediction technique, protein models are becoming genuinely useful in biology and medicine, including drug design. There are numerous examples where in silico models were used to infer protein functions, hint at protein interaction partners and binding site locations, design or improve novel enzymes or antibodies, create mutants to alter specific phenotypes or explain phenotypes of existing mutants. Current available template-based methods can reliably generate accurate high-resolution models, comparable in quality to the structure solved by low resolution X-ray crystallography, when sequence similarity of a homolog to an already solved structure is high( 50 % or greater). As alignment problems are rare in these cases, the



main focus shifts to accurate modeling of structurally variable regions (insertions and deletions relative to known homology) and side chains, as well as to structure refinement. The high-quality comparative models often present a level of detail that is sufficient for drug design, detecting sites of protein-protein interactions, understanding enzyme reaction mechanisms, interpretation of disease-causing mutations and molecular replacement in solvent crystal structures [8] [9] [10] [11].

## 2.1 Protein Structure Prediction -- Overview

According to Anfinsen's (1973) [12] thermodynamic hypothesis, proteins are not assembled into their native structures by a biological process. Protein folding is a purely physical process that depends only on the specific amino acid sequence of the protein and the surrounding solvent [12]. This would suggest that one should be able to predict, at least theoretically, the three-dimensional (3D) conformation of a protein from its sequence alone. Since then, many efforts have been devoted to this fascinating and challenging problem, attempting to tackle this problem from different angles including biophysics, chemistry, and biological evolution. Solving the problem of predicting a protein's 3D structure from its amino acid sequence has been called the "holy grail of molecular biology" and is considered as equivalent to deciphering "the second half of the genetic code" [13]. The study of the principles that dictate the 3D structure of natural proteins can be approached either through the laws of physics or the theory of evolution. Each of these approaches provides the foundation for a class of protein structure prediction methods [14]. Accordingly, theoretical structure prediction can be divided into two extreme camps: homology modeling and *ab initio* methods [15]. The boundaries between these two extreme classes of pre-

diction techniques have started to become blurred as scientists have started to integrate the strengths of different methods to make their prediction methods more effective and more generally applicable. Also, a third class of protein structure prediction methods has appeared: protein threading. Homology modeling makes structure predictions based primarily on its sequence similarity to one or more proteins of known structures. *Ab initio* methods predict the three-dimensional structure of a given protein sequence without using any structural information of previously solved protein structures; instead, methods belonging to this group are entirely based on the first principles of physics [16]. Protein threading, sometimes referred as fold recognition (FR) is an approach between the two extremes which uses both sequence similarity information when it exists, and structural fitness information between the query protein and the template structure [17]. Below is a brief discussion of each of these methods, which emphasizes their advantages and disadvantages from a user's point of view.

### **2.1.1 Homology Modeling**

Homology modeling, also referred to as comparative modeling (CM), is a class of methods based on the fact that proteins with similar sequences adopt similar structures, as most protein pairs with more than 30 out of 100 identical residues were found to be structurally similar [18]. Homology modeling is facilitated by the fact that the 3D structure of proteins from the same family is more conserved than their amino acid sequences [19]. When the structure of one protein in a family has been determined by experimentation, other members of the same family can be modeled based on their alignment to the known structure. This high robustness of structures with respect to residue exchanges explains partly the robustness of organisms with respect to gene-replication errors, and it allows for the variety in evolution. Comparative modeling con-

sists of five main stages: (a) identification of evolutionary related sequences of known structure; (b) aligning of the target sequence to the template structures; (c) modeling of structurally conserved regions using known templates; (d) modeling side chains and loops which are different than the templates; (e) refining and evaluating the quality of the model through conformational sampling [20]. The accuracy of predictions by homology modeling depends on the degree of sequence similarity between the target sequence and the template structures. When the sequence identity is above 40%, the alignment is straightforward, there are not many gaps, and 90% of main-chain atoms could be modeled with an RMSD (root-mean-square distance) error of about 1 Å [15]. In this range of sequence identity, predictions are of very good to high quality, and have been shown to be as accurate as low-resolution X-ray predictions [11]. When the sequence identity is about 30-40%, obtaining correct alignment becomes difficult where insertions and deletions are frequent. For sequence similarity in this range, 80% of main-chain backbone atoms can be predicted to RMSD 3.5 Å, while the rest of the residues are modeled with larger errors [15].

When the sequence identity is below 30%, the main problem becomes the identification of the homolog structures, and alignment becomes questionable, thereby giving rise to the name of the 20 -30 % zone – the twilight zone of protein sequence alignments [18]. From a user point of view, the main difficulty in homology modeling is finding the target sequence to be used as a template. Approximately 57% of all known sequences have at least one domain that is related to at least one protein of known structure [21]. The probability of finding a related known structure for a randomly selected sequence from a genome ranges from 30% to 65% [15]. The percentage is steadily increasing because projects like Protein Structure Initiative promise to fulfill within

the next decade [22] the task of experimentally determining the 16 000 optimally selected new structures needed so that homology modeling can cover 90% of protein domains [23].

### **2.1.2 Protein threading**

Also known as fold recognition (FR), protein threading is a class of methods that aims at fitting a target sequence to a known structure in a library of folds. Generally, similar sequence implies similar structure but the converse is not true: similar structures are often found for proteins for which no sequence similarity to any known structure can be detected [24]. This means that the actual number of different folded protein structures is significantly smaller than the number of different sequences generated by the large scale genome projects [20]. An optimistic view is that the number of existing folds is a few orders of magnitudes smaller than the number of different sequences, possibly ranging from a few hundred to a few thousand. The basic idea of protein threading is to literally “thread” the amino acids of a query protein, following their sequential order and allowing for insertions and gaps, into the structural positions of a template structure in an optimal way measured by a scoring function. This procedure is repeated for each template structure in a database of protein folds. The quality of a sequence-structure alignment is typically assessed using statistical-based energy and the “best” sequence-structure alignment provides a prediction of the backbone atoms of the query protein. The main drawback of this class of methods is the fact that it is very demanding on the computing power and also, that there is still a need for target identification. Currently, the Protein Data Bank contains enough structures to cover small singleton main protein structures up to a length of about 100 residues, so the method has the best chances of success with proteins within this limit [25] [26].

### 2.1.3 *Ab initio* methods

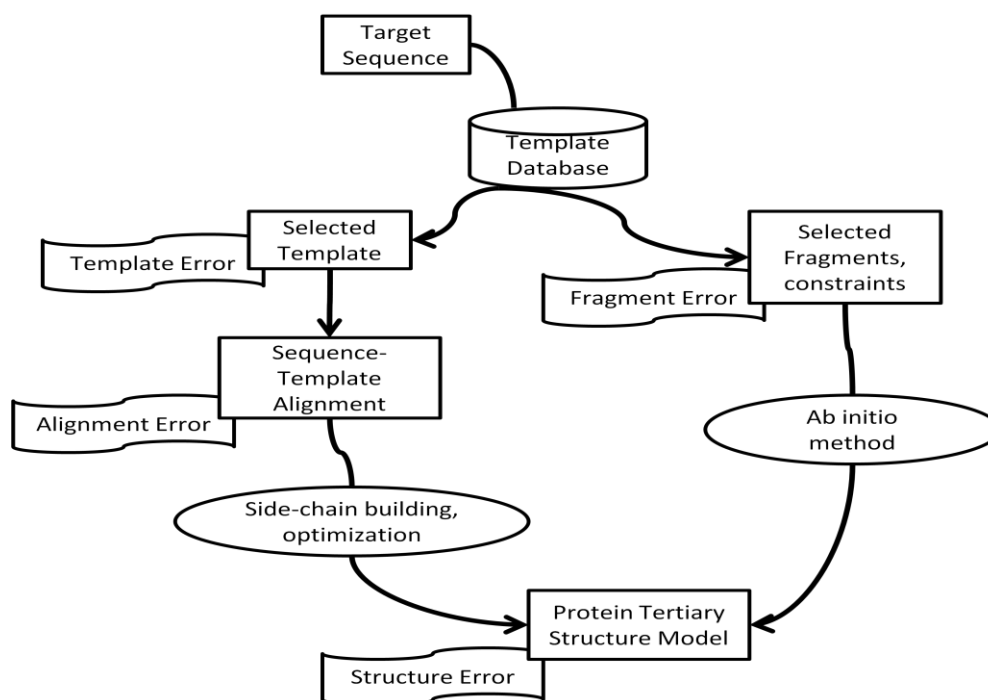
Also known as *de novo* methods, “first principle” methods or “free modeling” [27], these methods assume that the native structure corresponds to the global free energy minimum accessible during the lifespan of the protein, and attempt to find this minimum by an exploration of many conceivable protein conformations [14]. The term *ab initio* methods referred initially to methods for structure prediction that do not use experimentally known structures [24]. Lately, this term has become vaguer since the introduction of novel fragment based methods. These methods primarily utilize the fact that, although we are far from observing all folds used in biology [28], we probably have seen nearly all substructures [29]. Structure fragments are chosen on the basis of the compatibility of the substructure with the local target sequence and assembled into one new structure. The field of *ab initio* prediction methods is thereby divided into two main classes: *ab initio* methods with database information and *ab initio* methods without database information [24]. Even though the methods from this last class are computationally very demanding and still lack accuracy [14], they are continuously used and developed for several reasons. Firstly, in some cases, even a remotely related structural homolog may not be available. In these cases, *ab initio* methods are the only alternative. Secondly, new structures continue to be discovered which could not have been identified by methods which rely on comparison to known structures. Thirdly, knowledge-based methods have been criticized for predicting protein structures without having to obtain a fundamental understanding of the mechanisms and driving forces of structure formation. First principle structure prediction methods, in contrast, base their predictions on physical models for these mechanisms. As such, they can therefore help to deepen the understanding of the mechanisms of protein folding [24]. From a user point of view the main

bottlenecks of *ab initio* methods are the resolution of generated models and the computing power required to generate these models. The low resolution of *ab initio* generated models resides in our limited understanding of the protein folding problem and despite significant progress in this direction [30], it remains applicable to a limited number of sequences of less than approximately 100 residues [14].

#### **2.1.4 Prediction Errors**

Flowchart of protein structure prediction methods is shown in Figure 2.2, the figure also highlights the occurrence of error in each phase. If an appropriate template structure for a target sequence is found in a template database by a threading method, a structure model will be built on the template structure (the left branch of the chart). If not, an *ab initio* method can be employed (the right branch). Most of the current *ab initio* methods use fragment structure taken from template database. Errors can occur at each step of this procedure. In template recognition step, wrong templates with a different fold but in correct fold class are often recognized in threading (template recognition level error). A severe template level error can occur when the template database does not contain exactly correct structures. In that case, a threading program still ranks templates in the database according to their scores, and the top ranking structure which has a similar, but not exactly correct fold, may gain a statistically significant score. In template-based structure prediction, it is almost impossible to fix a template level error if the template is considerable different from the correct one. When a recognized template does not share sufficient sequence similarity to the target sequence, it is not easy to align the template and the target correctly (alignment level error). Finally, each procedure in the full-atom model construction i.e.

loop modeling and side-chain building, and the refinement step will cause errors (tertiary structure level error) [2] [31].



**FIGURE 2.2 FLOWCHART OF PROTEIN STRUCTURE PREDICTION (PICTURE ADOPTED FROM [31])**

It is desired that model quality assessment programs that predict the real quality value of a model and give the output as a single score be developed. These tools will then become the key for bridging computational and experimental biology, bringing the structure prediction tools into experimental biology labs. These tools could also significantly broaden the applicability of protein structure models of a moderate resolution. For protein structure prediction methods to enable

fruitful research exploration together with experimental methods, error estimation and quality assessment are indispensable research focus [27] [31] [32] .

### 2.1.5 Application of Structure Prediction

A 3-D model does not have to be absolutely perfect to be helpful in biology, but the type of question that can be addressed with a particular model does depend on its accuracy. Depending on the prediction approach applied [14] the accuracy of a model differs. Comparative modeling generates structures that have a root mean square deviation (RMSD) of 1–2 Å from the experimental structure, achieving the accuracy of medium resolution NMR or low-resolution X-ray structures [33] . Threading provides models with an RMSD of 2–6 Å, with errors mainly occurring in the loop regions [34]. For target proteins without solved template structures, *ab initio* methods are limited to small proteins (<120 residues) with an accuracy in the range of 4–8 Å. For low accuracy models (RMSD >3 Å) RMSD is no longer a meaningful measure of modeling quality [14] and TMscore is preferred. By definition, TM-score lies in a 0.1 interval. A TM value of 1 indicates a very accurate model (equivalent of RMSD 0 Å), a value >0.5 indicates a model with a roughly correct topology, and a value 0.17 indicates a random prediction regardless of the protein size [22]. High-resolution models obtained by homology modeling at more than 50% sequence identity can usually meet the highest structural requirements in the case of single-domain proteins and have been used in a wide range of applications, as docking, designing and improving ligands for a given binding site [35], designing mutants to test hypotheses about a protein's function [36] [37], identifying active and binding sites [38], simulating protein protein docking [39], facilitating molecular replacement in X-ray structure determination [40], refining models based on NMR constraints [41] and rationalizing known experimental observations [42]. For models of



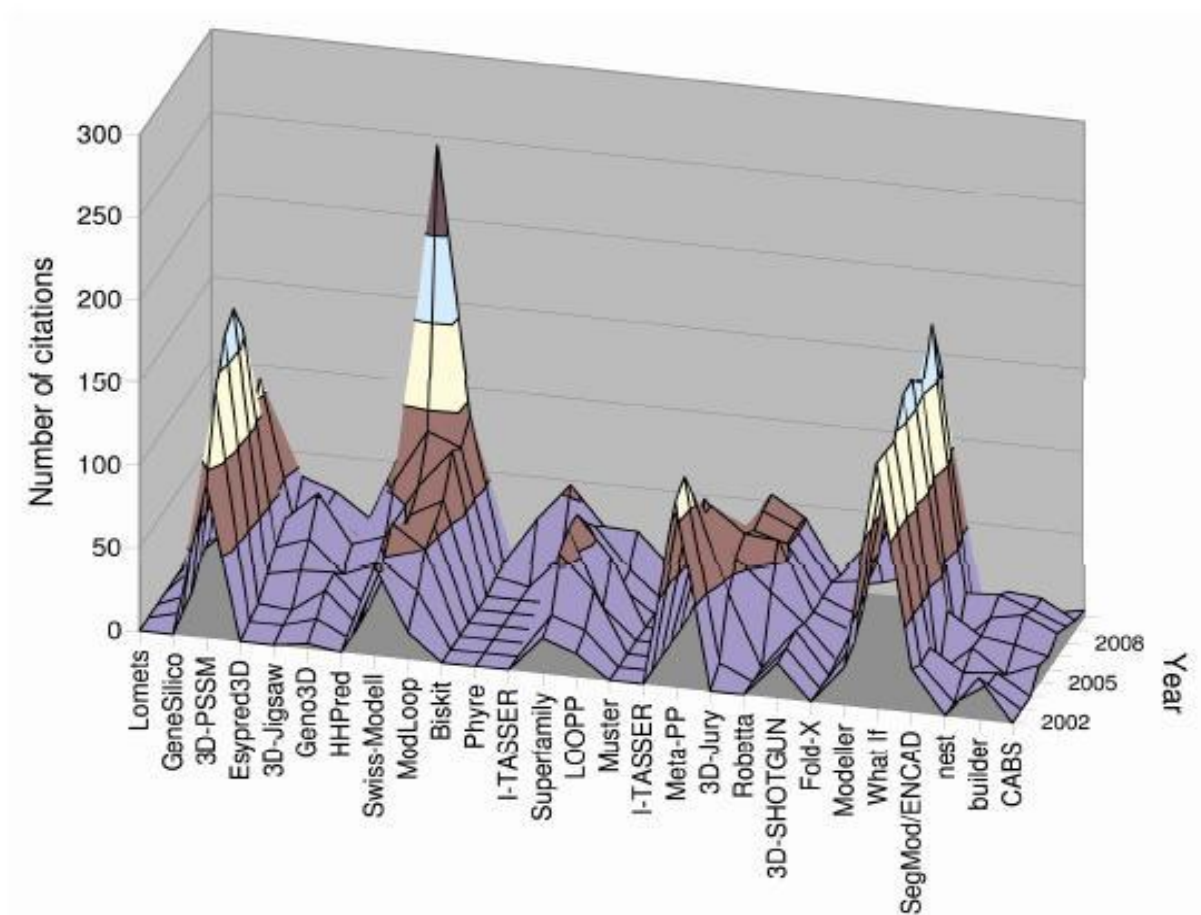
medium-resolution, with an RMSD between 2.5-5 Å, typically generated by comparative modeling from distantly homologous templates or by fold recognition, the structural predictions are useful for identification of the spatial locations of functionally important residues, such as active sites and the sites of disease-associated mutations. Arakaki et al. [43] assessed the possibility of assigning the biological function of enzyme proteins by matching the structural patterns (or descriptors) of the active sites with structure decoys of various resolutions. Boyd et al. [44] used structural models generated by the automated I-TASSER server to help interpret mutagenesis experiments with the Sec1/Munc18 (SM) proteins on the basis of the spatial clustering of the mutated residues. Models with the lowest resolution from free modeling approaches or based on weak hits from threading, have a number of uses including protein domain boundary identification [45], topology recognition, or family/superfamily assignment. For example, the TASSER structural predictions placed the RDC1 receptor in the family of chemokine receptors because the predicted RDC1 structure is closest to the predicted structure of the CXCR4 chemokine receptor [46]. This finding was later confirmed by binding experiments [47].

Protein structure prediction has been thought of as a “grand challenge” for some time now. As more and more researchers need and use the protein prediction tools, rapid progress has been made in recent years in this field. The massive amounts of sequence and structural data becoming available and the low cost and accessibility of computing power has led to an explosion of available tools and methods for protein prediction. The choice of one or another method still depends on the protein sequence, as well as the expected quality of the result. The rapid growth of automated servers means that protein prediction is no longer only for only a handful of researchers, but is available for the masses. The process is not completely automated, the feedback

of the user is still required when deciding on the most trustful method and the usefulness of the result.

## 2.2 Literary Review of Current Model Assessment Techniques

An important task in both structure prediction and application is to evaluate the quality of a structure model. Half a century has passed since it was shown that an amino acid sequence of a protein determines its shape, but a method to translate reliably into the 3D structure still remains to be developed. So it is important to develop methods that determine the quality of a model. Over the past two decades, a number of approaches have been developed to analyze correctness of protein structures and models. Traditional model evaluation methods use stereochemistry checks, molecular mechanics energy-based functions and statistical potentials to tackle the problem. As a server or a program will almost always return an answer, using two or more of such tools means that one will get more than just one computer generated model. This becomes hard to know which model to choose. As opposed to experimental structure evaluation, there are not too many reliable procedures to assess the quality of a computer-generated model [48]. Before tackling with any *in silico* protein prediction problem, a non-bioinformatician has to check the CASP website. Choosing a tool from most highly ranked in the latest CASP experiment will assure the best possible start in terms of reliability of the results. Beside the CASP rank, another important factor in choosing the right tool is the protein to be modeled. Figure 2.3: Protein structure prediction software – trends in the number of citations per year for some of the most common docking programs and servers, analyzed from the ISI Web of Science (2009). The figure has been adopted from reference [49].



**FIGURE 2.3 PROTEIN STRUCTURE PREDICTION SOFTWARE TRENDS**

There is a basic rule to follow. If your protein has at least 40% similarity with a known structure, comparative modeling is the method to use. For lower similarities, protein threading is preferred. When the target sequence has no similarities with known structure, *ab initio* methods are the last resort. Two types of evaluation of the computer-generated models can be carried out. Internal evaluation of self consistency checks whether or not a model satisfies the restraints used to calculate it. Generally, each of the tools used in the construction of a model, template selec-

tion, alignment, model building, and refinement has its own internal measures of quality [48]. Nevertheless, assessment of the stereochemistry of a model (e.g., bonds, bond angles, dihedral angles and nonbonded atom-atom distances) can be additionally checked with programs such as PROCHECK [50], WHAT-IF [51] and WHAT-CHECK [52]. External evaluation relies on information that was not used in the calculation of the model, like the calculation of the pseudo energy profile of a model performed by tools like PROSA [53], Verify3D [54] and QMEAN [55]. Finally, a model should be consistent with any existing experimental observations, such as site directed mutagenesis, cross-linking data and ligand binding [14].

Recently machine learning methods using algorithms like neural networks and support vector machines that are trained on structure models to predict model quality are introduced [33]. There are various techniques available to determine the quality of a predicted model, either by comparing it with the native structure or with no knowledge of known structure. In the following paragraphs, the most familiar assessment techniques are categorized and importance of each category is discussed.

These techniques can be divided into several categories based on their scoring strategy – local vs. global, absolute vs. relative or single vs. multiple (consensus or ranking methods). There are methods that predict the quality of local regions such as distance between the position of a residue in a protein model and its native structure as suppose to predicting an overall score of a model. Some methods predict both local and global quality like Pcons [56]. Wallner and Elofsson's Pcons is a consensus-based method capable of a quite reliable ranking of model sets for both easy and hard targets. Pcons uses a meta-server approach (i.e. combines results from several available well-established QA methods) to calculate a quality score reflecting the average

similarity of a model to the model ensemble, under the assumption that recurring structural patterns are more likely to be correct than those observed only rarely. It should be underscored that, while the consensus-based methods are useful in model ranking, they can be biased by the composition of the set and, in principle, are incapable of assessing the quality of a single model. This brings us to another category based on scoring that is absolute score vs. relative score. Relative scoring methods discriminate near-native structure from decoys; these methods are different from methods which produce absolute score. A relative score can only select or rank models but does not tell how good a model is, which is critical for using the model. The techniques could also be grouped according to the information needed to make judgment. In prominent assessment approaches 3D co-ordinates, sequence information, sequence alignment, alignment information, template, secondary structure information, etc are generally used to make judgment on quality. Model evaluation methods can be classified into single-model approach such as ProQ, Proq-LG, ProQ-MX and MODCHECK and multiple model approaches such as clustering methods whose output depends on number of input models. We can also group based on prediction techniques, machine learning tools like neural networks and SVM, clustering and consensus approach, etc. Some of these methods are used in CASP (Critical Assessment of Structure Prediction), as one of many analysis involved in assessment phase [5] [6] [57]. There are many methods that aim at finding the model quality but very few come up with an absolute score using a single model and information from only its primary sequence and 3D co-ordinates [57] [58] [59].

In the two years following CASP7, a considerable increase in method development in the area of model quality assessment can be observed. More than a dozen papers have been published on the subject, and 45 quality assessment methods, almost double the CASP7 number,

have been submitted for evaluation to CASP 8 [34]. CASP evaluation is based on comparison of each model with the corresponding experimental structure. GDT\_TS [5] score is used in several CASP competitions, which is defined as average coverage of the target sequence of the substructure with four different distance threshold.

**TABLE 2.1 OVERVIEW OF CURRENT MODEL ASSESSMENT METHODS**

<b>Method</b>	<b>Year</b>	<b>Scoring</b>	<b>Remark</b>
GDT_TS [5]	1999	Single Score	Compares to native struc-
ProQ [56]	2003	Single Score	Uses Neural networks
3D-Jury [64]	2003	Ranking	Consensus method
SPICKER [62]	2004	Ranking	By clustering
MODCHECK [60]	2005	Single Score	Classical threading poten-
Undertaker [63]	2005	Single Score	Uses full 3D information
ProSa- Web [73]	2007	Single Score	Uses evolutionary infor-
ModFOLD [65]	2008	Single Score	TM score is used
ModelEvaluators [59]	2008	Single Score	Support Vector Regression
Tasser [66]	2008	Ranking	Structure feature and statis-

There are other similar techniques that obtain an absolute scoring by comparing the model to its experimental structure. A strikingly different domain is assessing the models with no known structure. There are several methods proposed in recent years to solve this problem. Single-model approaches like ProQ [56], ProQ-LG, ProQ-MX [58], and MODCHECK [60] assign a

score to a single model, whereas, multiple-model approaches, such as clustering and consensus methods, require a large pool of models as inputs to rank them. These methods cannot be used to assess the quality of a single model. They may not reliably evaluate the quality of a small number of input models [32] [31]. Machine learning methods such as neural networks and SVM that are trained on structure models predict model quality [59] Zhou and Skolnick [61] [62] differ by including the consensus-based features (i.e. incorporating in the analysis information from multiple models on the same target).

MODCHECK [60] places emphasis on benchmarking individual methods and also offers neural network-based meta techniques that combine them. Modfold merges four original approaches in a program. Some of the recent methods that make use of single model, 3D coordinate information and primary sequence to evaluate an absolute score for model quality assessment are ModelEvaluator [59] and Undertaker [63]. In ModelEvaluator [59] they use normalized GDT\_TS score with SVM regression to train SVM to learn a function that accurately maps input features. To get a general overview of available techniques please refer TABLE 2.1.

Since models are not experimental structures determined with known accuracy but predictions, it is vital to present the user with the corresponding estimates of model quality. Much is being done in this area but further development of tools to assess model quality reliably is needed. Our approach is quite different from any recent study; we aim to classify the models into two classes, a protein or not a protein. With thousands of protein structures available in Protein Data Bank is it possible to train a machine learning algorithm to study protein structure and predict when given a model whether it closely resembles these structures or not. From initial results we can say with some assurance that it is possible to achieve such a learning curve.

The amount and nature of information given to the machine learning system will have an impact on the final output regarding the quality measure of given 3D structure. There are various ways of representing a protein three dimensional structure, like backbone sketch of the protein, the entire distance matrix of alpha carbon atoms, a fractal dimension of the structure, 3D information with its sequence data etc. These methods of representing protein structure are mostly used in comparison and classification problems and they are well studied and researched fields [75] [76] [79].



### CHAPTER 3. MACHINE LEARNING TECHNIQUES

Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge. This capacity to learn from experience, analytical observation, and other means, results in a system that can improve its own speed or performance, i.e., its efficiency and/or effectiveness. Machine-learning techniques have been used to create self-improving software for decades, but recent advances are bringing these tools into the mainstream. The exponential growth of the amount of biological data available raises two problems: on one hand, efficient information storage and management and, on the other hand, the extraction of useful information from these data. The second problem is one of the main challenges in computational biology, which requires the development of tools and methods capable of transforming all these heterogeneous data into biological knowledge about the underlying mechanism. These tools and methods should allow us to go beyond a mere description of the data and provide knowledge in the form of testable models. By this simplifying abstraction that constitutes a model, we will be able to obtain predictions of the system. Machine learning algorithms are widely used in many biological fields to name a few – genomics, proteomics, microarrays, system biology, evolution and text mining. The main purpose of a machine learning algorithm is to make intelligent decisions based on available knowledge from some database. For this research we have considered the following algorithms [76] [79].

### 3.1 Support Vector Machines

Support Vector Machines (SVM) are learning systems that use a hypothesis space of linear function in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. This learning strategy introduced by Vapnik and co-workers is a principled and very powerful method that in the recent years since its introduction has already outperformed most other systems in a wide variety of applications. In supervised learning the learning machine is given a training set of examples (or inputs) with associated labels (or output values). Once the attributes vectors are available, a number of sets of hypotheses could be chosen for the problem. Among these, linear functions are best understood and simplest to apply. The development of learning algorithm became an important sub field of artificial intelligence, eventually forming the separate subject area of machine learning [80].

Kernel representations offer an alternative solution by projecting the data into a high dimensional feature space to increase the computational power of the linear learning machines. Another attraction of kernel methods is that the learning algorithms as theory can largely be decoupled from the specifics of the application area, which must simply be encoded into the design of an appropriate kernel function. Hence the problem of choosing architecture for a neural network application is replaced by the problem of choosing a suitable kernel for a Support Vector Machines. The introduction of kernel greatly increases the expressive power of the learning machines while retraining the underlying linearity that will ensure that learning remains tractable. The increased flexibility however, increases the risk of over fitting as the choice of separating hyperplane becomes increasingly ill-posed due to the number of degrees of freedom. Success-

fully controlling the increased flexibility of kernel-induced feature spaces requires a sophisticated theory of generalization, which is able to precisely describe which factors have to be controlled in the learning machines in order to guarantee good generalization. There is a remarkable family of bounds governing the relation between the capacity of a learning machines and its performance. The theory grew out of consideration of under what circumstances and how quickly, the mean of some empirical quantities converges uniformly, as the number of data points increases, to the true mean [75] [80] [81].

Since SVM approach has a number of superior values such as effective avoidance of over fitting, the ability to handle large feature spaces, information condensing of the given data set etc. It has been successfully applied to a wide range of pattern recognition problems, including isolated handwritten digit recognition, objective recognition, speaker identification, and text categorization, etc [82].

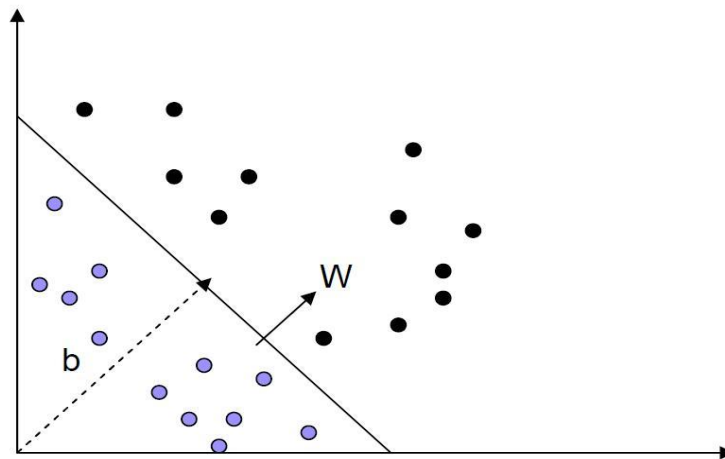
Binary classifier is frequently implemented by using a real-valued function  $f: X \subseteq \mathcal{R}^n \rightarrow \mathcal{R}$  in the following way: the input  $x = (x_1, \dots, x_n)'$  is assigned to the positive class, if  $f(x) \geq 0$ , and otherwise to the negative class. If we consider the case where  $f(x)$  is a linear function of  $x \in X$ , so that it can be written as

$$f(x) = w \bullet x + b \quad (2.1)$$

$$= \sum_{i=1}^n w_i x_i + b \quad (2.2)$$

Where,  $(w, b) \in \mathbb{R}^n \times \mathbb{R}$  are the parameters that control the function and the decision rule given by  $\text{sgn}(f(x))$ . And these parameters must be learned from the data.

If we interpret this hypothesis geometrically, input space  $X$  is split into two parts by the hyperplane defined by the equation  $w \cdot x + b = 0$ . For example, in Figure 3.1, the hyper plane is the dark line, with the positive region above and the negative region below. The vector  $w$  defines a direction perpendicular to the hyperplane, while varying the value of  $b$  moves the hyperplane parallel to itself. And these quantities are referred as the weight vector and bias which are the terms borrowed from the neural networks literature [82].



**FIGURE 3.1 A SEPARATING HYPER PLANE FOR A 2-D TRAINING SET [82].**

The above algorithm for separable data, when applied to non-separable data, will find no feasible solution: this will be evidenced by the objective function) i.e. the dual Lagrangian) growing arbitrarily large. To extend these ideas to handle non-separable data, the constraints (2.1) and (2.2) are relaxed, but only when necessary, that is, a further cost (i.e. an increase in the

primal objective function) is introduced. This can be done by introducing positive slack variables  $\xi_i, i=1, \dots, l$  in the constraints, which then become:

$$x_i \bullet w + b \geq +1 - \xi_i \quad \text{for } y_i = +1 \quad (2.3)$$

$$x_i \bullet w + b \leq -1 + \xi_i \quad \text{for } y_i = -1 \quad (2.4)$$

Thus, for an error to occur the corresponding  $\xi_i$  must exceed unity, so  $\sum_i \xi_i$  is an upper bound on the number of training errors. Hence a natural way to assign an extra cost for error is to change the objective function to be minimized from  $\|w\|^2/2$  to  $\|w\|^2/2 + C(\sum_i \xi_i)^k$ , where  $C$  is a parameter to be chosen by the user, a larger  $C$  corresponding to assigning a higher penalty to error [81] [82].

The soft margin classifier is an extension of linear SVM. The kernel method is a scheme to find the nonlinear boundaries. The concept of the kernel method is transformation of the vector space to a higher dimensional space. By transforming the vector space from two-dimensional to three-dimensional space, the non-separable vectors can be separated.

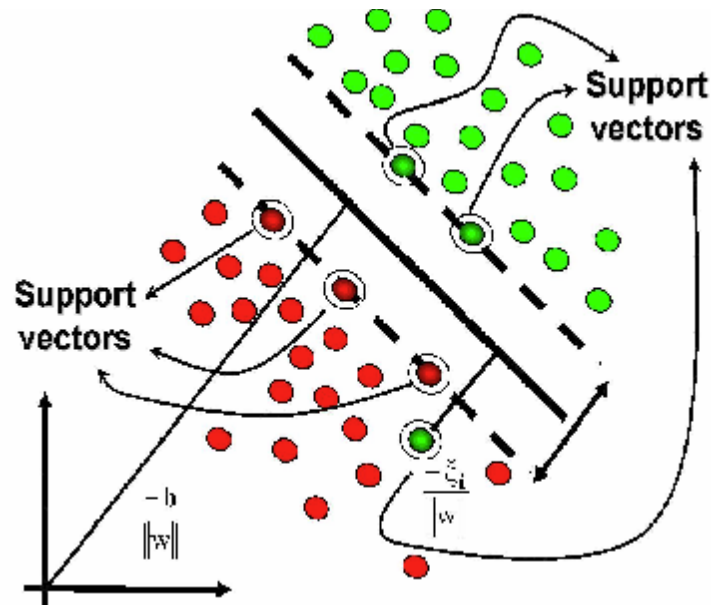
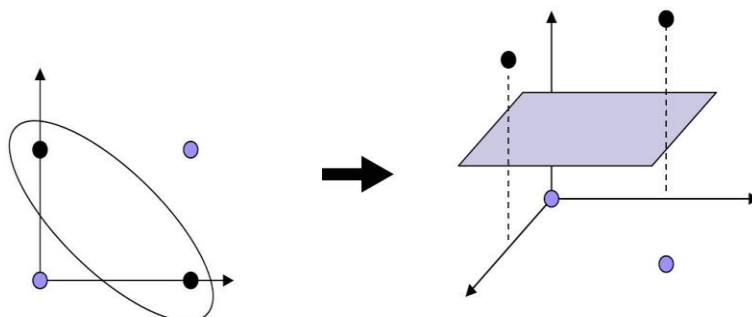


FIGURE 3.2 LINEAR SEPARATING HYPER PLANE FOR NON-SEPARABLE CASE

### 3.1.1 Nonlinear SVM - Kernel Method

The soft margin classifier is an extension of linear SVM. The kernel method is a scheme to find the nonlinear boundaries. The concept of the kernel method is transformation of the vector space to a higher dimensional space. As can be seen from Figure 3.3, by transforming the vector space from two-dimensional to three-dimensional space, the non-separable vectors can be separated [80] [82] .



**FIGURE 3.3 TRANSFORMATION TO HIGHER DIMENSIONAL SPACE**

### 3.1.2 SVM Software

The prediction of protein secondary structure is done using  $SVM^{light}$  software.  $SVM^{light}$  software is the implementation of Vapnik's Support Vector Machine (Vapnik 1995) for the problem of pattern recognition, regression and ranking function.  $SVM^{light}$  software consists of two parts, the first part i.e. is the `svm_learn` part takes care of the learning module and the second part `svm_classify` part does the classification of the data after training.

The input data to both the parts should be given in the following format

`<line> .=. <target> <feature>:<value> <feature>:<value> ...`

`<target> .=. +1 | -1 | 0 | <float>`

`<feature> .=. <integer> | "qid"`

`<value> .=. <float>`

The target value and each of the feature/value pairs are separated by space character. Feature/value pairs must be ordered by increasing feature number. Features with value zero can be

skipped. For classification, the target value denotes the class of the example +1 as the target value marks a positive example, -1 negative example respectively. So, for example, the line

```
-1 1:0.43 3:0.12 2345:0.9
```

Specifies a negative example for which feature number 1 has value 0.43 feature number 3 has value 0.12 feature number 2345 has the value 0.9 and all the other features have value 0. The order of the predictions is the same as in the training data [81].

### 3.2 Decision Trees

Decision trees are one of the most popular machine-learning techniques. They are known for their ability to represent the decision support information in a human comprehensible form, however, they are recognized as a highly unstable classifier with respect to small changes in training data [86] [87]. One of the most popular algorithms for building decision trees is Interactive Dichotomizer3 (ID3) algorithm proposed by Quinlan in 1979 [88] [89]. Generally trees produced by ID3 known as crisp decision trees are sensitive to small changes in feature values, cannot handle data uncertainties caused by noise and/or measurement errors [86] [87]. To overcome these problems, several ID3 extensions were proposed to handle continuous and multi-valued data [86] [87]. Some statistical approaches and tree pruning are used to overcome the problem of over fitting the training data and to improve the generalization of the decision model produced by decision trees. Trees produced by ID3 – known as crisp decision trees- cannot handle data uncertainties and spurious precision in the data. Fuzzy ID3 (FID3) is an extension of ID3 algorithm. It integrates ID3 and fuzzy set theory [89] to overcome some of the deficiencies in ID3. Trees produced by the FID3 algorithm are known as fuzzy decision trees ( FDTs); and



they are more immune to data uncertainties caused by measurement errors, noise, missing and/or inconsistent information.

The fuzzy decision tree building procedure is very similar to that of ID3 ; the fuzzified training dataset is partitioned recursively based on the value of a selected splitting (also called branching or test) feature (attribute). The splitting feature is selected such that a certain information measure of separating data belonging to different classes is maximized. Several information measures exist in the literature, typical information measures used include: Information gain (IG) [90], classification ambiguity (CA) [91] and gini-index [92][92]. Existing fuzzy ID3 algorithms use either information gain or classification ambiguity to select a branching feature. A gini-index based fuzzy decision tree algorithm was proposed recently in [89]. A node in the tree is considered a leaf node, when all the objects at the node belong to the same class, the number of objects in the node is less than a certain threshold, the ratio between objects' memberships in different classes is greater than a given threshold, or no more features are available. Tuning these thresholds is crucial to the performance and the quality of the produced fuzzy decision trees. A feature can appear only once on any path from the root node to a leaf node in the tree. Unlike crisp decision trees where an object can propagate only to one child node, fuzzy decision trees allow a data item or an object to propagate to more than a child node.

Nael Abu-halaweh and team proposed an improved FID3 algorithm (IFID3) [93]. The IFID3 integrates classification ambiguity and fuzzy information gain to select the branching attribute. The IFID3 algorithm outperformed the existing FID3 algorithm on a wide range of datasets. They also introduce an extended version of the IFID3 algorithm (EIFD3). EIFID3 extends IFID3 by introducing a new threshold on the membership value of a data instance to propagate

down the decision tree from a parent node to any of its children nodes during the tree construction phase. Using the new threshold significant reduction in the number of rules produced, an improved accuracy and a huge reduction in execution time is achieved. They automate the generation of the membership functions, by two simple approaches, in the first approach the ranges of all numerical features in a dataset are divided evenly into an equal number of fuzzy sets. In the second one, the dataset is clustered and the resulted cluster centers are used to generate fuzzy membership functions. These fuzzy decision trees were applied to the microRNA prediction problem, their results showed that fuzzy decision trees achieved a better accuracy than other machine learning approaches such as Support Vector Machines (SVM) and Random Forest (RF) [93].

### ***3.2.1 Improved Fuzzy ID3 Algorithm***

Improved fuzzy ID3 (IFID3) uses the same fuzzy decision tree building procedure as that of ID3 and Fuzzy ID3 algorithm. It uses the classification ambiguity measure introduced in to extend the Fuzzy ID3 algorithm presented in [87]. As mentioned earlier, IFID3 uses attribute classification ambiguity to select the branching attribute at the root node, and fuzzy information gain elsewhere. Given a data set  $D$ , with attributes  $A_1, A_2, \dots, A_L$  and a classified class  $C = \{C_1, C_2, \dots, C_n\}$  and fuzzy sets  $F_{i1}, F_{i2}, \dots, F_{im}$  for the attribute  $A_i$  (each attribute may have a different value of  $m$ ). Let  $D^{C_k}$  be a fuzzy subset in  $D$  with class  $C_k$ , and let  $|D|$  be the sum of membership values in a fuzzy set of data  $D$ . IFID3 works as follows:

1. Generate the root node with a fuzzy set of all training data with membership value of 1 (This not necessary the case, membership values can be initialized manually, for example instances in the training dataset can be associated with weights) [93].
2. The node is a leaf node, if the fuzzy set of the data at that node satisfies any of the following conditions:

- a. The number of objects in the node data set  $D$  is less than a given threshold; this threshold is cold leaf control threshold  $\theta_n$ . that is:

$$|D| < \theta_n.$$

- b. The proportion of a data set of any class  $C_k$   $|D^{C_k}|$  in the node data set  $D$  is greater than or equal to a given threshold. This threshold is called fuzziness control threshold  $\theta_r$ .

That is:

$$\frac{|D^{C_k}|}{|D|} \geq \theta_r$$

- c. No more attributes are available for classification.
3. The class name assigned to a leaf node depends on the inference method and is either the name of the class with the greatest membership value, or the node is assigned all class names along with membership values.
4. If the node does not satisfy any of the above conditions then do the following:
  - a. If this node is the root node of the decision tree, then calculate the Attribute classification Ambiguity for all attributes in the dataset and select the attribute with the minimum attribute classification ambiguity as the test attribute.

- b. Else, the node is not the root node of the decision tree, calculate the fuzzy information gain for all attributes; and select the attribute that has greatest fuzzy information gain as the test attribute.
  - c. Divide the fuzzy data set at the node into fuzzy subsets using the selected test attribute, with the membership value of an object in a subset set to the product of its' membership value in parent node dataset and the value of the selected attribute fuzzy term.
  - d. For each of the subsets, generate new node with the branch labeled with the fuzzy term of the selected attribute.
5. For each new generated node repeat recursively from step 2.

This modified fuzzy ID3 algorithm presents a simple approach to integrate two information measures. It is shown that such integration can lead to better performance by experimental results using several standard datasets [93]. This algorithm makes use of attribute classification ambiguity to select the branching attribute at the root node of the decision tree, and fuzzy information gain to select the branching attribute at all the other non-leaf nodes in the tree. One of the drawbacks of our method is the huge number of generated fuzzy rules. Reducing the number of rules required for a decision is very important, both because it increases the computational performance of the fuzzy decision tree induction process and for the more fundamental reason that it improves the falsifiability of decision model and improves its interpretability and its applicability to real time applications.

### 3.2.2 *Extended Improved Fuzzy ID3 Algorithm*

Rule set reduction refers to the process of generating a smaller set of rules from a larger set of rules. This is achieved mainly by removing redundant rules and/or by merging adjacent rules leading to same decision. The main purpose of rule reduction is improving the human interpretability of the decision model by reducing its complexity. In addition, rule set reduction makes the process of validating the resulting decision model easier and improves the applicability of the resulting models to real time applications and can improve the system accuracy. To reduce the number of generated fuzzy rules, they introduce a modified version of improved Fuzzy ID3 algorithm [93]. This modified version introduces a new threshold on the membership value of a given data item to propagate down a fuzzy decision tree from parent node to any of its child nodes during the fuzzy decision tree generation step. The propagation threshold is not used as data-partitioning-stopping criteria, but is used as a filter that prevents data instances with membership values less than this threshold to propagate down to child nodes. The algorithm for Extended Fuzzy ID3 method is same as the one in the previous section except step 4 is as follows

4. If the node does not satisfy any of the above conditions then do the following:
  - a. If this node is the root node of the decision tree, then calculate the Attribute classification Ambiguity for all attributes in the dataset and select the attribute with the minimum attribute classification ambiguity as the test attribute.
  - b. Else, the node is not the root node of the decision tree, calculate the fuzzy information gain for all attributes; and select the attribute that has greatest fuzzy information gain as the test attribute.

- c. Divide the fuzzy data set at the node into fuzzy subsets using the selected test attribute, with the membership value of an object in a subset set to the product of its' membership value in parent node dataset and the value of the selected attribute fuzzy term.
- d. For each of the fuzzy subsets produced in c do:
  - i) For each data instance in the fuzzy subset If the membership of this data instance  $<$  object propagation threshold then remove this instance from the fuzzy set.
- e. For each of the subsets, generate a new node with the branch labeled with the fuzzy term of the selected attribute.

With experimental results they show that the modified version of IFID3 produces better accuracy and achieves significant reduction in the number of resulting fuzzy rules [94] [95]. Overall with their new fuzzy decision tree they have improved the accuracy and execution time of induction algorithms by integrating fuzzy information gain and classification ambiguity to select the branching feature. By introducing a new threshold on the membership value of a data object to propagate down the decision tree from parent node to any of its child nodes they have significantly reduced number of fuzzy rules generated [94] [95].

Both these features appeal to our data set and objective of assessing models. The rules generated will also help us understand the concept/rules governing protein structure formation. The main reason for using this machine learning technique is the rule set it generates. Using these rules, we could be able to map out the path for correct protein structure models.

In the initial part of the chapter, we discussed many methods available in literature to validate protein models. Our approach is quite different from any recent study; we aim to classify the models into two classes, a protein or not a protein. With thousands of protein structures available in Protein Data Bank is it possible to train a machine learning algorithm to study protein structure and predict when given a model whether it closely resembles these structures or not. From initial results we can say with some assurance that it is possible to achieve such a learning curve.

The amount and nature of information given to the machine learning system will have an impact on the final output regarding the quality measure of given 3D structure. There are various ways of representing a protein three dimensional structure, like backbone sketch of the protein, the entire distance matrix of alpha carbon atoms, a fractal dimension of the structure, 3D information with its sequence data etc. These methods of representing protein structure are mostly used in comparison and classification problems and they are well studied and researched fields. Our proposal is to study these various structure representations and machine learning tools and eventually produce a well trained machine that is capable of determining the correctness of any protein structure model.

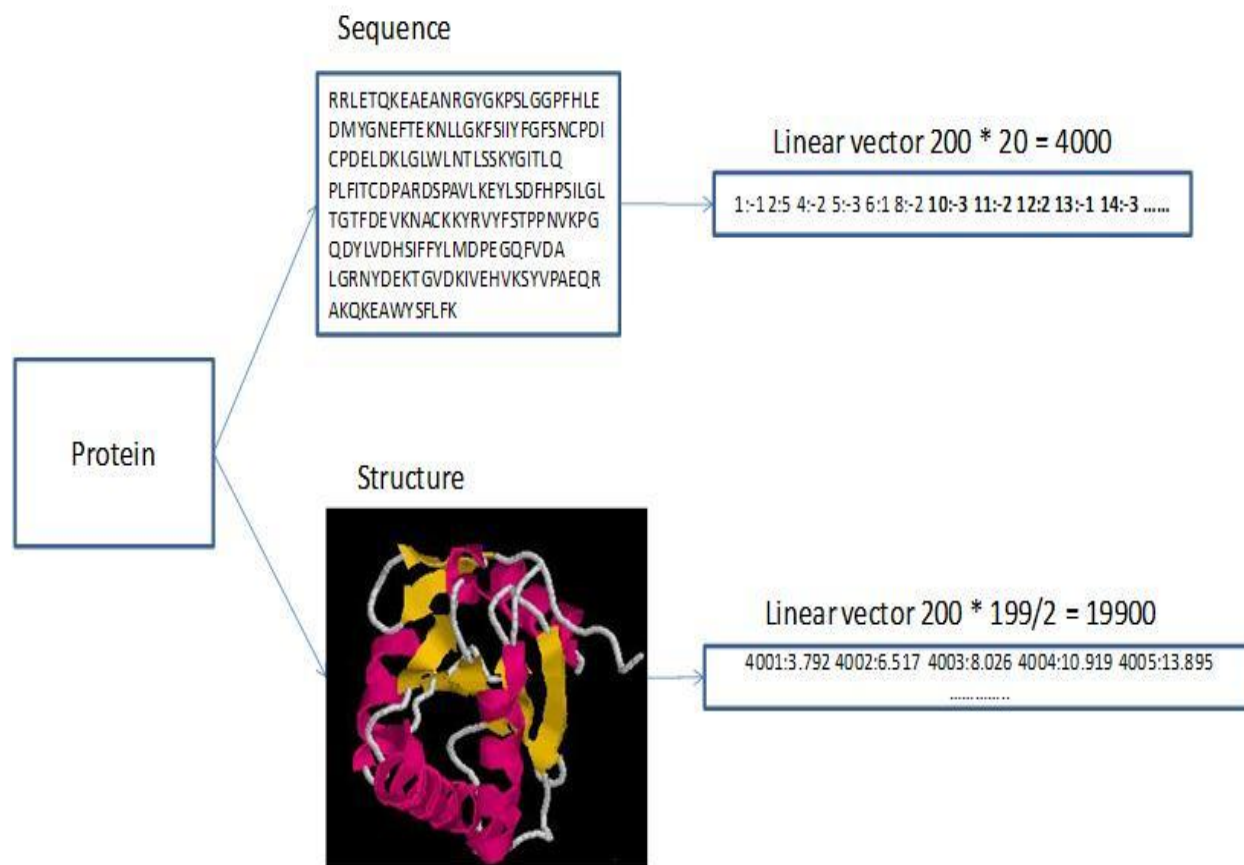
## CHAPTER 4. PRELIMINARY ENCODING SCHEME

In order to classify whether a three dimensional object is a protein structure or not, the structure should be represented in machine understandable format. In this methodology we represent each protein as one data vector. Each data vector should contain both structural information as well as sequence information of that protein structure. For training and testing cycles with machine learning algorithm, both positive and negative data vectors are needed. Positive vectors can be structures from PDB (protein data bank) data base, as these structures are experimentally determined ones. Negative vectors are generated by misaligning the sequence and structure information, so that we have wrong structure for a particular sequence. Different kernel methods and encoding schemes are used to observe their effectiveness in classifying proteins as correct or wrong. The goal here is to encode protein information in numerical form understandable by machine learning technique. The encoding is done in the following order.

### **Sequence Information + Structure Information**

In the following section, different methods for representing sequence and structure information are discussed. These are not the only methods for encoding protein information but they are made popular in other research domains like structure alignment, protein function classification, protein secondary structure classification etc [75] [76] [77].





**FIGURE 4.1 FEATURE VECTOR FORMATION**

#### **4.1 Enumeration of Protein Structure**

Protein sequence is a one dimensional string of 20 different amino acids. To represent each amino acid we can use one of the two very popular matrices. BLOSUM (BLOCKS of AminoAcid Substitution Matrix) is a substitution matrix. The scores measure the logarithm for the ratio of the likelihood of two amino acids appearing with a biological sense and the likelihood of the same amino acids appearing by chance. A positive score is given to the more likely substitutions while a negative score is given to the less likely substitutions. The elements in this

matrix are used as features for data vectors. For each amino acid there are 20 features to be considered. Profile is a table that lists the frequencies of each amino acid in each position of protein sequence. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. Similar to BLOSUM matrix we have 20 features per amino acid. In preliminary studies both methods were used to encode sequence information. For IFID3 only BLOSUM matrix is used. This method is further illustrated in the Figure 4.1.

The distance matrix containing all pair wise distances between C $\alpha$  atoms, is one commonly used rotationally and translationally invariant representation of protein structure. This technique is used in DALI [77] for protein structure comparison by detecting spatial similarities. The major difficulty with distance matrices is that they cannot distinguish between right-handed and left-handed structures. Other evident problem with this method is computability, as there too many parameters or attributes to optimize in the case of feature selection or optimization, which are important steps in machine learning process. Two different simulations are done for implementation purpose of the design. The first simulation is the direct implementation of a larger dataset done using support vectors technique. Due to some short comings in computational domain and poor prediction accuracy the second simulation is considered. The second simulation uses fuzzy decision tree to obtain better prediction accuracy.

## 4.2 Vector Formation

Single Positive Vector is a single protein with its sequence information followed by its own structure information. Single Negative Vectors has one protein's sequence information followed by another protein's structure information. For initial implementation we have considered

two kernels linear and Gaussian. Sequence information is represented using both BLOSUM as well as profiles to observe their individual performance. In case of structural information only distance matrix is considered to represent protein 3D structure. For example to encode protein chain 1M56D of sequence length 51 using profile + distance matrix encoding scheme, its complete sequence and structure information has to be included. For every amino acid in the sequence we need to input 20 features corresponding to its position specific score (PSSM or profiles), so the example protein will have 1020 ( $51 * 20$ ) features to represent its sequence. For structure information we have to consider (upper or lower) half of distance matrix, this will result in 1275 ( $51*50/2$ ) features to represent its structure. Total of 2295 features are used to represent the protein chain. (Note the entire sequence and structure details are considered .i.e. for a protein of length 200 we will 4000 features for sequence information and 19900 for structure.) This will be a positive vector as it's from the PDB data base. Negative vectors are generated by choosing sequence and structure information of two random proteins in similar manner. For BLOSUM matrix + Distance matrix encoding scheme BLOSUM matrix is used instead of profiles.

The PDB entries are culled based on their sequence length and relative homology using PISCES server [32] [78]. The culled list has no more than 25% homology among different protein sequences. This will ensure that negative vectors are not false negative with highest probability.

### **4.3 Implementation Using SVM**

In this chapter, implementation of the proposed method using support vector machines is shown. Different simulations are done to study the effectiveness of the selected machine learning

technique in this data scenario. A specific set of PDB entries are culled based on sequence length (200) and homology. Positive and negative data are obtained from the same set as discussed in the previous chapter.

#### **4.3.1 *Simulation I***

For this simulation about two thousand PDB entries are culled from the entire PDB data bank. These entries form the positive vectors for the learning system. The same number of negative vectors is generated by randomly selecting two PDB entries, one for sequence and other for structure information. Number of training vectors hence obtained is 4670 and number of testing vectors is 780. Results after the implementation are shown in Table 4.1 and Table 4.2, results show seven fold testing.

From the tables we see BLOSUM matrix to be better in accuracy than profiles. Thought with BLOSUM matrix for encoding we see Gaussian kernels unable to give comparable results to that of liner kernel. This might be due to incorrect optimization parameters. The above results are not sufficient to agree upon any one encoding scheme or kernel, more simulations are required. Simulation two is done to determine the effect of using all features as suppose to some randomly selected ones.

**TABLE 4.1 ENCODING SCHEME: PROFILE + DISTANCE MATRIX**

Linear Kernel		RBF Kernel	
Test Case	Accuracy	Test Case	Accuracy
1	70.52%	1	50.58%
2	64.74%	2	50.43%
3	63.29%	3	50.58%
4	64.45%	4	50.00%
5	66.18%	5	50.72%
6	65.32%	6	50.87%
7	71.82%	7	50.29%
<b>Average</b>	<b>66.62%</b>	<b>Average</b>	<b>50.50%</b>

**TABLE 4.2 ENCODING SCHEME: BLOSUM MATRIX + DISTANCE MATRIX**

Linear Kernel		RBF Kernel	
Test Case	Accuracy	Test Case	Accuracy
1	71.11%	1	53.08%
2	63.59%	2	52.95%
3	70.77%	3	53.95%
4	72.31%	4	52.18%
5	73.21%	5	52.44%
6	71.41%	6	52.82%
7	75.51%	7	53.08%
<b>Average</b>	<b>71.13%</b>	<b>Average</b>	<b>52.93%</b>

### 4.3.2 Simulation II

From the simulation one results it is noted that the machine learning algorithm suffers from curse of dimensionality which affects computational efficiency and final accuracy. This could be due to the huge number of features considered in representing protein 3D structure. Feature reduction and selection techniques like redesigning the feature, selecting appropriate subset of features or combining features could be considered to solve this problem. The training and testing datasets are constructed similar to simulation one.

**TABLE 4.3 ACCURACY BEFORE FEATURE SELECTION**

<b>Test Case</b>	<b>BLOSUM Matrix Encoding</b>	<b>Profile Encoding</b>
1	58.24%	59.72%
2	65.29%	66.67%
3	59.41%	63.89%
4	60.59%	63.19%
5	60.00%	52.78%
6	62.94%	66.67%
7	64.12%	58.33%
<b>Average</b>	<b>61.51%</b>	<b>61.60%</b>

To obtain this we will adopt a static scheduling algorithm that will schedule each training vector in a different processor. This scheduling will continue until the desired accuracy (equal or greater than linear kernel accuracy with all features) or maximum number of tries is attained.

For effective analysis of the feature selection procedure a different dataset was culled similar to one shown in section 4.1 but with only 600 PDB entries. Number of training vectors hence obtained is 1030 and number of testing vectors is 170. The results show the effectiveness of feature selection. The accuracies have also increased after feature selection.

Since dataset considered here is different from simulation one, we have calculated the seven fold accuracy for this dataset. These results are shown in Table 4.3. There is a drop in average percentage accuracies; this might be due to lesser number of training vectors.

#### **4.3.3 Feature Selection Algorithm**

A simple algorithm is devised for feature selection purpose. On number of features to be selected several percentage of features were tested to compare their performance with vectors that have all the features. After several trails 2 % was proved to be sufficient to obtain the same accuracy. For improving the speed of the algorithm we use multiple processors. Each processor is scheduled to perform first the feature selection then SVM training then SVM testing. Once completed it will do the same task over until the desired accuracy is obtained or for maximum number of tries. The algorithm has the following steps

1. Select 2 % of features from the training set
2. Schedule a processor to train this training set using SVM light software
3. Repeat step 1 & 2 for generating 10 such training sets
4. Schedule the processors that have completed the training with testing
5. Check testing accuracy of each set with the testing accuracy of set with 100 % features (previous result from Table 4.3).

- if the testing accuracy is better than previous results then record the feature numbers used and quit
- Else repeat step 1 to 5 until desired results is obtained or a maximum number of tries is reached.

**TABLE 4.4 ACCURACY AFTER FEATURE SELECTION**

<b>Test Case</b>	<b>BLOSUM Matrix Encoding</b>	<b>Profile Encoding</b>
1	60.50%	64.58%
2	66.47%	68.75%
3	62.55%	66.67%
4	65.29%	65.28%
5	63.53%	58.33%
6	65.29%	69.44%
7	67.06%	63.89%
<b>Average</b>	<b>64.37%</b>	<b>64.69%</b>

Table 4.4 shows the results obtained by using the above algorithm. The average has improved in both encoding schemes. Table 4.5 clearly shows the average accuracies of both encoding schemes before and after feature selection. This emphasizes the fact that all that features are not needed to make the binary decision. This leads us to look into other encoding schemes and representations of protein sequence and structure information. Other novel kernels should also be considered as so far only linear kernel has shown any real prediction ability [40].



**TABLE 4.5 COMPARISON OF ACCURACIES BEFORE AND AFTER FEATURE SELECTION**

<b>Encoding Scheme</b>	<b>Before Feature Selection</b>	<b>After Feature Selection</b>
BLOSUM Matrix	61.51%	64.37%
Profile	61.60%	64.69%

#### 4.4 Implementation Using IFID3

Three different datasets are considered. Each is a subset of the same dataset with different number of proteins. Proteins from PDB are culled as discussed in section III. The proteins within a specific length range (150-200) are considered. Since there is a length variation among different proteins, smaller proteins have “?” as attribute value for positions where is no amino acid. This kind of attribute definition is acceptable with IFID3 algorithm. This representation just means the attribute has no value or no meaning in our case.

As mentioned earlier, IFID3 uses attribute classification ambiguity to select the branching attribute at the root node, and fuzzy information gain elsewhere. Given a data set  $D$  with attributes  $A_1, A_2, \dots, A_n$ , based on the given parameter values different fuzzy sets are formed for each attribute. Fuzzy ID3 algorithm requires the given dataset to be in a fuzzy form. In our case the data set is not in fuzzy form but in continuous numerical form, so it needs to be fuzzified first. To obtain the optimal number of fuzzy sets, the number of fuzzy sets can be given as a parameter that needs to be tuned. The tree is build according to the given conditional parameters and each node becomes a leaf node if number of dataset is less than a given threshold, if proportion of any

class in the node is greater than the given fuzziness condition threshold or no more attributes are available for classification [93] [94].

**TABLE 4.6 SEVEN FOLD RESULTS USING IFID3**

<b>Test Case</b>	<b>100 Proteins</b>	<b>200 Proteins</b>	<b>700 Proteins</b>
1	82.14%	87.72%	80.50%
2	78.57%	78.95%	82.50%
3	85.71%	80.70%	77.50%
4	75.43%	78.95%	81.50%
5	82.14%	85.96%	85.00%
6	78.57%	85.96%	80.50%
7	78.12%	84.48%	79.50%
<b>Average</b>	<b>80.10%</b>	<b>83.25%</b>	<b>81.00%</b>

For our dataset we considered ten fuzzy sets and triangular membership function for each attribute. The performance of the tests is shown in Table 4.6. About 20 rules are generated for these data sets. There are three datasets, each with different number of proteins. The first one has 100 proteins and hence 100 positive vectors and 100 negative vectors. Similarly data set 200 and 700 proteins has 200/700 positive vectors and negative vectors respectively. For sequence information only BLOSUM matrix encoding used. The table shows seven fold test results and average of these seven folds [85].

The prediction accuracy of IFID3 is much better when compared to SVM results. More simulations can be done with increased number of proteins to check if there is any improvement to prediction results.

## CHAPTER 5. ENHANCED ENCODING SCHEME

This chapter introduces enhanced encoding of protein structure data to effectively distinguish a well formed model from poorly formed model. The initial encoding scheme (explained in the previous chapter) used in this study for protein structure assessment has numerous concerns. The number one issue encountered while using the preliminary encoding scheme is the number of features required in representing the model. This results in enormous computational overload on the algorithm. The sequence length becomes primary factor in coding the protein because of this, proteins of varying sequence lengths transformed into vectors of varying features. This requires a necessary tuning step while implementing with fuzzy decision tree algorithm. The tuning of the vectors is done in order to get uniform lengths along the entire data set. This makes rule understanding and inference very hard and cumbersome, despite excellent prediction accuracy. For the above reason and to increase the prediction accuracy more enhanced spatial encoding technique is considered. This encoding scheme is explained elaborately in this chapter

As in previous studies each protein three dimensional structure is considered as a single vector. For the training phase we generate both positive and negative vectors from PDB structures. For positive vectors both sequence and structure details are collected from the same structure. For negative vectors sequence information comes from one structure and structural information from another (similar to previous work). We follow a different methodology in generating these vectors when compared to previous studies.

In previous study, only structure information taken into consideration is the distance matrix along with substitution matrix to represent amino acids. In this research, different factors of

alpha carbon atoms like polarity, secondary structure information are also considered as features along with distance matrix and substitution matrix see Figure 5.1.

```

Alpha carbon atom CA data structure has the following features

{
    //Sequence Info
        amino acid
        polarity
    //Structure Info
        x-coordinate
        y-coordinate
        z-coordinate
        Secondary structure information
        Distance to the center of the structure
}

```

**FIGURE 5.1 DATA STRUCTURE OF ALPHA CARBON ATOM**

To obtain uniform vector length throughout the data set containing proteins of varying sequence lengths, fixed number of alpha carbon atoms is selected from different local areas of the model. The local areas are generated as spheres of fixed diameter. Throughout the data set number of spheres per model is fixed as well. These constants are generated as functions of certain geometric factors of the model and are explained further below. Alpha carbon atom features are selected within a fixed diameter by traversing the protein back born and after finishing with one diameter we move to the next. This is illustrated diagrammatically in the Figure 5.2 and Figure 5.3

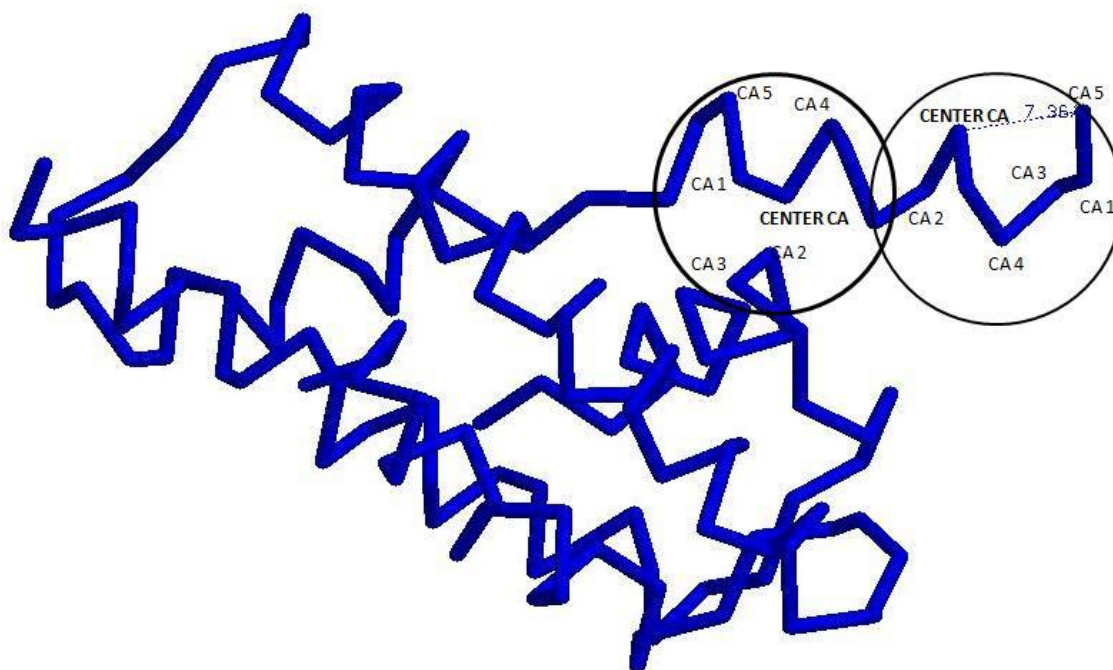


FIGURE 5.2 ENHANCED ENCODING SCHEME ILLUSTRATION I

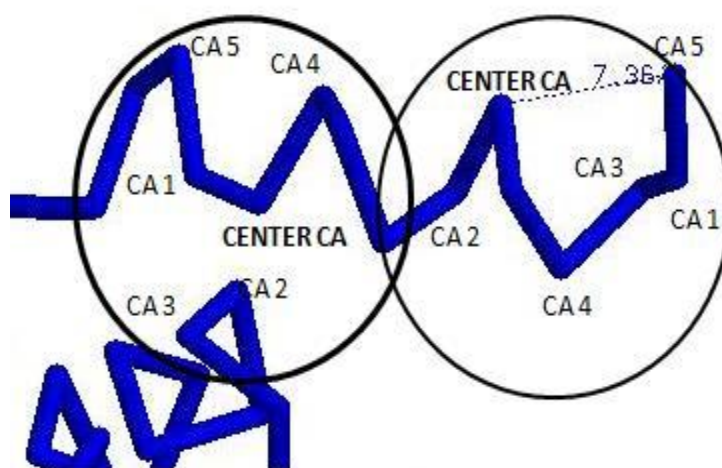


FIGURE 5.3 ENHANCED ENCODING ILLUSTRATION II

Before we begin building protein vectors, we need to effectively describe each alpha carbon atom and then calculate two constants that are based on the geometric structure of the protein model. The first constant is a point called Structure center, this is calculated to further estimate distance of every alpha carbon atom to this point and relative to each other. The second constant is called delta ( $\delta$ ) this is the fixed distance that determines the groups.

Once we extract all information in the data format shown in Figure 5.1, we calculate the two constants as follows

*Structure center* is calculated as

$$\left( \frac{\sum_{i=1 \text{ to } N} CA[i].x}{N}, \quad \frac{\sum_{i=1 \text{ to } N} CA[i].y}{N}, \quad \frac{\sum_{i=1 \text{ to } N} CA[i].z}{N} \right)$$

Distance Delta is calculated as

$$\delta = \frac{\sum_{i=1 \text{ to } N} \text{distance between } CA[i] \text{ and center}}{c * N}$$

Where N is number of amino acid in the structure and c is a fixed constant

## 5.1 Methodology

1. All alpha carbon atom information is extracted in the described data format by traversing the primary structure.
2. The features are extracted in groups. Each group has one alpha carbon atom considered as the center CA. This atom is selected to be the center, if it satisfy the following two conditions
  - i. No other center CA within a fixed distance  $\delta$
  - ii. At least 5 alpha carbon atoms that are within distance  $\delta$  to the center CA , that do not belong to any other group.
2. Once the center CA is selected for the group then group is formed with 5 other alpha carbon atom that are within distance  $\delta$  and don't belong to any other group.
3. This process is repeated for 10 groups.

{Selecting five neighboring atoms and ten groups is for pure information extraction purpose. More groups could be selected if it enhances the overall machine performance. For the particular data set with the sequence length in the range of 150 and 200, it is determined ten groups and six alpha carbon atoms per group to be optimal after several simulations. We need to select atoms from different structural zones, this is enforced by not selecting other group members and



group centers. Also after number of iterations number five is considered to be an optimum number of group members so we do not have scenarios like

Two group centers too close ( $<$  than delta distance)

Partially filled groups ( $<$  than 5 atoms within delta distance).

Over lapping groups (once an atom is selected to be in one group, it can't be in any other group))}

The picture shows the method figuratively.

## 5.2 Enumeration of Features

Before going into details of vector formation, the following two sections explain briefly the fundamental concepts of polarity and secondary structures in amino acids. This is of significance as these features of amino acid are used in the feature space to provide maximum information about the considered alpha carbon atom.

### 5.2.1 *Principle of Polarity*

Amino acids are classified into different ways based on polarity, structure, nutritional requirement, metabolic fate, etc. Generally used classification is based on polarity. Amino acid polarity chart in APPENDIX section shows the polarity of amino acids.

Based on polarity, amino acids are classified into four groups as follows,

- Non-polar amino acids
- Polar amino acids with no charge

- Polar amino acids with positive charge
- Polar amino acids with negative charge

Each amino acid has at least one amine and one acid functional group as the name implies. The different properties result from variations in the structures of different R groups. The R group is often referred to as the amino acid side chain. The greater the electro negativity differences between atoms in a bond, the more polar the bond. Partial negative charges are found on the most electronegative atoms, the others are partially positive.

### **Non-Polar Side Chains:**

Side chains which have pure hydrocarbon alkyl groups (alkane branches) or aromatic (benzene rings) are non-polar. Examples include valine, alanine, leucine, isoleucine, phenylalanine. The number of alkyl groups also influences the polarity. The more alkyl groups present, the more non-polar the amino acid will be. This effect makes valine more non-polar than alanine; leucine is more non-polar than valine.

### **Polar Side Chains:**

Side chains which have various functional groups such as acids, amides, alcohols, and amines will impart a more polar character to the amino acid. The ranking of polarity will depend on the relative ranking of polarity for various functional groups as determined in functional groups. In addition, the number of carbon-hydrogens in the alkane or aromatic portion of the side chain should be considered along with the functional group.

*i. Acidic Side Chains:*

If the side chain contains an acid functional group, the whole amino acid produces an acidic solution. Normally, an amino acid produces a nearly neutral solution since the acid group and the basic amine group on the root amino acid neutralize each other in the zwitterion. If the amino acid structure contains two acid groups and one amine group, there is a net acid producing effect. The two acidic amino acids are aspartic and glutamic.

*ii. Basic Side Chains:*

If the side chain contains an amine functional group, the amino acid produces a basic solution because the extra **amine** group is not neutralized by the acid group. Amino acids which have basic side chains include: lysine, arginine, and histidine.

*iii. Neutral Side Chains:*

Since an amino acid has both an amine and acid group which have been neutralized in the zwitterion, the amino acid is neutral unless there is an extra acid or base on the side chain. If neither is present then the whole amino acid is neutral.

Based on fact that how much polarity of an amino acid matters on where the amino acid is placed in protein three dimensional structure, we incorporate this crucial information into our feature space. The information is enumerated as follows

1 --> Neutral Non-Polar

2 --> Neutral Polar

3 --> Basic Polar

4 --> Acidic Polar

The complete list of amino acids and their three letters and single letter abbreviations, their structure and their polarity information is given in the table in the appendix.

### 5.2.2 *Secondary Structure of Proteins*

In biochemistry and structural biology, secondary structure is the general three-dimensional form of *local segments* of biopolymers such as proteins and nucleic acids (DNA/RNA). It does not, however, describe specific atomic positions in three-dimensional space, which are considered to be tertiary structure.

Secondary structure can be formally defined by the hydrogen bonds of the biopolymer, as observed in an atomic-resolution structure. In proteins, the secondary structure is defined by the patterns of hydrogen bonds between backbone amide and carboxyl groups. In nucleic acids, the secondary structure is defined by the hydrogen bonding between the nitrogenous bases. The hydrogen bonding patterns may be significantly distorted, which makes an automatic determination of secondary structure difficult [75].

The secondary structure may be also defined based on the regular pattern of backbone dihedral angles in a particular region of the Ramachandran plot; thus, a segment of residues with such dihedral angles may be called a helix, regardless of whether it has the correct hydrogen bonds. The secondary structure may be also provided by crystallographers in the corresponding PDB file [75].

The rough secondary-structure content of a biopolymer (e.g., "this protein is 40%  $\alpha$ -helix and 20%  $\beta$ -sheet.") can often be estimated spectroscopically. For proteins, a common method is far-ultraviolet (far-UV, 170-250 nm) circular dichroism. A pronounced double minimum at 208 and 222 nm indicate  $\alpha$ -helical structure, whereas a single minimum at 204 nm or 217 nm reflects random-coil or  $\beta$ -sheet structure, respectively. A less common method is infrared spectroscopy, which detects differences in the bond oscillations of amide groups due to hydrogen-bonding. Finally, secondary-structure contents may be estimated accurately using the chemical shifts of an unassigned NMR spectrum [75].

The **DSSP** algorithm is the standard method for assigning secondary structure to the amino acids of a protein, given the atomic-resolution coordinates of the protein. DSSP begins by identifying the hydrogen bonds of the protein using a purely electrostatic definition, assuming partial charges of  $-0.42 e$  and  $+0.20 e$  to the carbonyl oxygen and amide hydrogen respectively, their opposites assigned to the carbonyl carbon and amide nitrogen. A hydrogen bond is identified if  $E$  in the following equation is less than  $-0.5$  kcal/mol:

$$E = 0.084 \left\{ \frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right\} \cdot 332 \text{ kcal/mol}$$

Based on this, eight types of secondary structure are assigned. The  $3_{10}$  helix,  $\alpha$  helix and  $\pi$  helix have symbols **G**, **H** and **I** and are recognized by having a repetitive sequence of hydrogen bonds in which the residues are three, four, or five residues apart respectively. Two types of beta sheet structures exist; a beta bridge has symbol **B** while longer sets of hydrogen bonds and beta bulges have symbol **E**. **T** is used for turns, featuring hydrogen bonds typical of helices, **S** is used

for regions of high curvature (where the angle between  $\overrightarrow{C_i^\alpha C_{i+2}^\alpha}$  and  $\overrightarrow{C_{i-2}^\alpha C_i^\alpha}$  is less than  $70^\circ$ ), and a blank (or space) is used if no other rule applies, referring to loops. These eight types are usually grouped into three larger classes: helix (**G**, **H** and **I**), strand (**E** and **B**) and loop (all others). This classification is shown in the Table 5.1 (partially adopted from [75]).

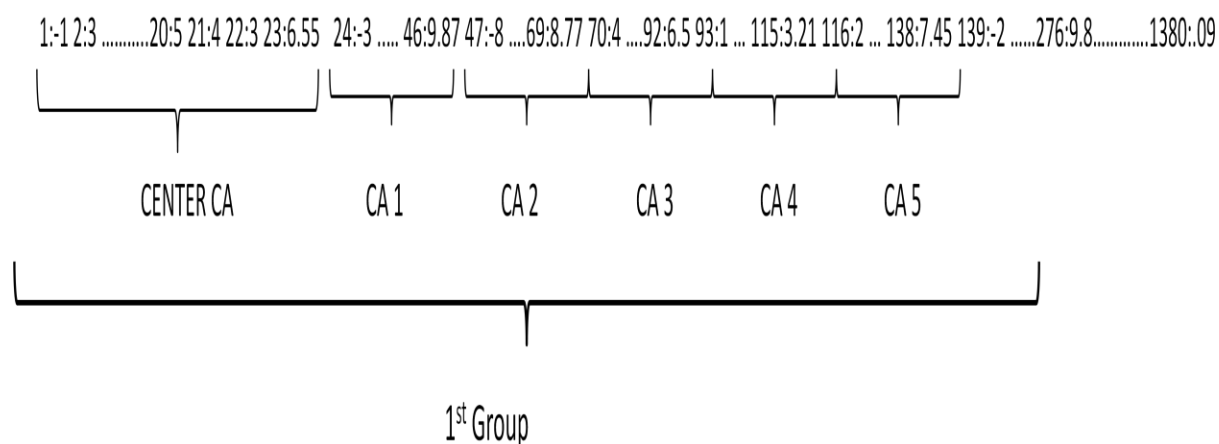
**TABLE 5.1 8-TO-3 STATE REDUCTION METHOD IN SECONDARY STRUCTURE ASSIGNMENT**

DSSP Class	8-state Symbol	3 – state Symbol	Class Name	Enumeration
$3_{10}$ helix	G	H	Helix	1
$\alpha$ helix	H			
$\pi$ helix	I			
$\beta$ strand	E	E	Sheet	2
Isolated $\beta$ - bridge	B	C	Loop	3
Bend	S			
Turn	T			
Rest	-			

For showing each amino acid to belong to one of the three classes in the algorithm, these values are enumerated as 1 for helix, 2 for Sheets and 3 for coil. In most cases the PDB files downloaded from protein data bank website, contains the secondary structure information. In certain case if that is not available, the proper enumeration is done by calculating the secondary structure as explained in the above paragraphs. For CASP templates only coordinate information is provided and secondary structure is calculated in every case using the DSSP algorithm,

### 5.3 Vector Formation

This section describes the vector formation procedure. For each group first the *center CA* details are included followed by group members. There are 23 features to describe each alpha carbon atom. First 20 represents the BLOSUM matrix sequence corresponding to the amino acid residue 21<sup>st</sup> is to represent polarity 22<sup>nd</sup> for secondary structure 23<sup>rd</sup> for distance (to *structure center* for *center CA* and to *center CA* for other group members). There are 23 features to represent one alpha carbon atom, 6 members in a group, so 138 to represent a group and with ten groups we have a total of 1380 features to represent a single protein. This vector formation methodology enforces that proteins of different sequence length will all have same number of features, unlike previous method which ended with different number features for very vector and needed certain tweaking to manage the flaw. The concept is diagrammatically illustrated in the Figure 5.4. Due to this format of representing the features, each feature number has a special significance, like feature numbers 21, 44, 67 etc represent the polarity of amino acids that are within delta distance, correspondingly feature numbers 22, 45, 68 etc represent secondary structures of the same amino acids and feature number 23, 46, 69 etc represent distance either to *center CA* or the *structure center*. This way of representation helps during understanding of rules in the process of knowledge inference from the rules.



**FIGURE 5.4 VECTOR FORMATION**

The PDB entries are culled based on their sequence length and relative homology. The culled list has no more than 25% homology among different protein sequences. This will ensure that negative vectors are not false negative with highest probability. Only structures obtained by X-ray crystallography are selected as they offer better resolutions. We sub selected these proteins to form idle training phase.

## **5.4 Implemeantation Results**

For implementation purpose five subsets of protiens are selected, to determine the optimal number of training data required for obtaining the highest accuracy. These five subsets have 100, 350, 500, 600 and 700 proteins respectively. The table shows the seven fold cross validation results of these subsets. There is a significant improvement in accuracy compared to previous encoding technique. Overall subset suggestion containing 500 proteins is seen to



perform better compared to other ones. Different dataset are considered to analyzise the rule tree generated by fuzzy decision tree algorithm.

**TABLE 5.2 PRELIMINARY ENCODING SEVEN FOLD RESULTS**

<b>Test Case</b>	<b>100 Proteins</b>	<b>350 Proteins</b>	<b>500 Proteins</b>	<b>600 Proteins</b>	<b>700 Proteins</b>
1	82.14%	82.35%	88.00%	82.56%	80.50%
2	78.57%	75.49%	78.67%	81.39%	82.50%
3	85.71%	86.27%	80.67%	84.30%	77.50%
4	75.43%	79.41%	79.33%	79.65%	81.50%
5	82.14%	78.43%	85.96%	80.23%	85.00%
6	78.57%	78.43%	86.00%	77.32%	80.50%
7	78.12%	85.29%	84.66%	79.65%	79.50%
<b>Average</b>	<b>80.10%</b>	<b>80.81%</b>	<b>83.33%</b>	<b>80.73%</b>	<b>81.00%</b>

This camparison is made in Table 5.4. There are also other improvements like considerable reduction in terms of computational time and number of features considered in spatial encoding. Different dataset are considered to analyzise the rule tree generated by fuzzy decision tree algorithm.

**TABLE 5.3 ENHANCED SPATIAL ENCODING SEVEN FOLD RESULTS**

<b>Test Case</b>	<b>100 Proteins</b>	<b>350 Proteins</b>	<b>500 Proteins</b>	<b>600 Proteins</b>	<b>700 Proteins</b>
1	86.67%	88.24%	98.72%	89.53%	89.00%
2	80.00%	85.30%	97.44%	97.10%	95.00%
3	83.33%	86.27%	99.36%	94.77%	94.50%
4	86.66%	89.22%	92.31%	96.51%	88.00%
5	90.00%	88.24%	91.66%	90.11%	96.00%
6	86.67%	87.25%	95.51%	95.35%	89.00%
7	86.66%	83.33%	100.00%	92.44%	97.00%
<b>Average</b>	<b>85.71%</b>	<b>86.84%</b>	<b>96.43%</b>	<b>93.69%</b>	<b>92.64%</b>

**TABLE 5.4 COMPARISON OF TWO ENCODING SCHEMES**

<b>Subset</b>	<b>Preliminary Encoding</b>	<b>Enhanced Encoding</b>
<b>100 Proteins</b>	80.10%	85.71%
<b>350 Proteins</b>	80.81%	86.84%
<b>500 Proteins</b>	83.33%	96.43%
<b>600 Proteins</b>	80.73%	93.69%
<b>700 Proteins</b>	81.00%	92.64%
<b>Average</b>	<b>81.2%</b>	<b>91.1%</b>

Implementation using SVM was considered but the results are not shown as they were sub optimal compared to the above results.

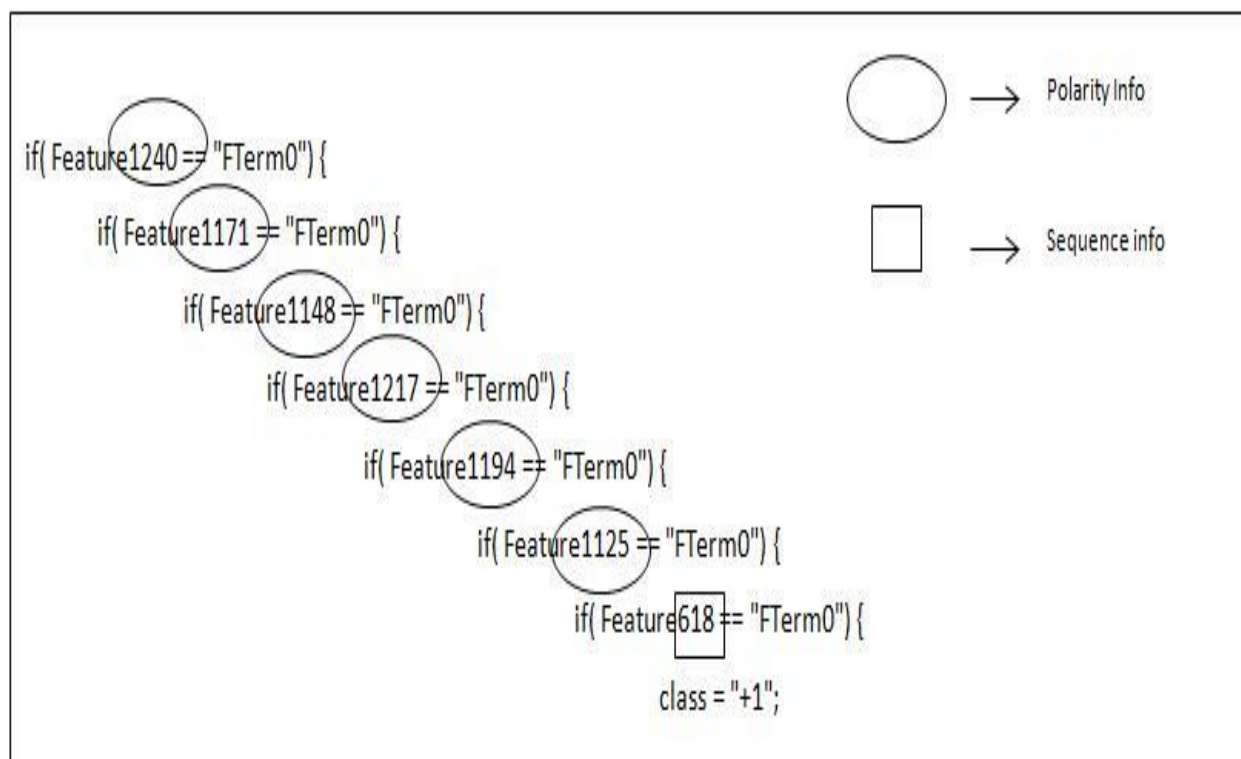
## 5.5 Meaningful Rules and Inference

The rules are analyzed to deduce any association to known knowledge. The number of rules for subset 1 and 2 with 100 and 350 proteins were around 20, but in the third subset the number of rules spiked to be around 60. The accuracy in the subset 3 has also increased. In the subsets 4 and 5 the number of rules continued to increase but the accuracy dropped slightly. The average numbers of rules in all subsets are shown in Table 5.5. These rules are further studied to underscore any meaningful information, as well as to see if they correlate to already known knowledge about these structure.

**TABLE 5.5 AVERAGE NUMBER OF RULES**

<b>Subset</b>	<b>Average No. Of Rules</b>
100 Proteins	15
350 Proteins	18
500 Proteins	67
600 Proteins	81
700 Proteins	84

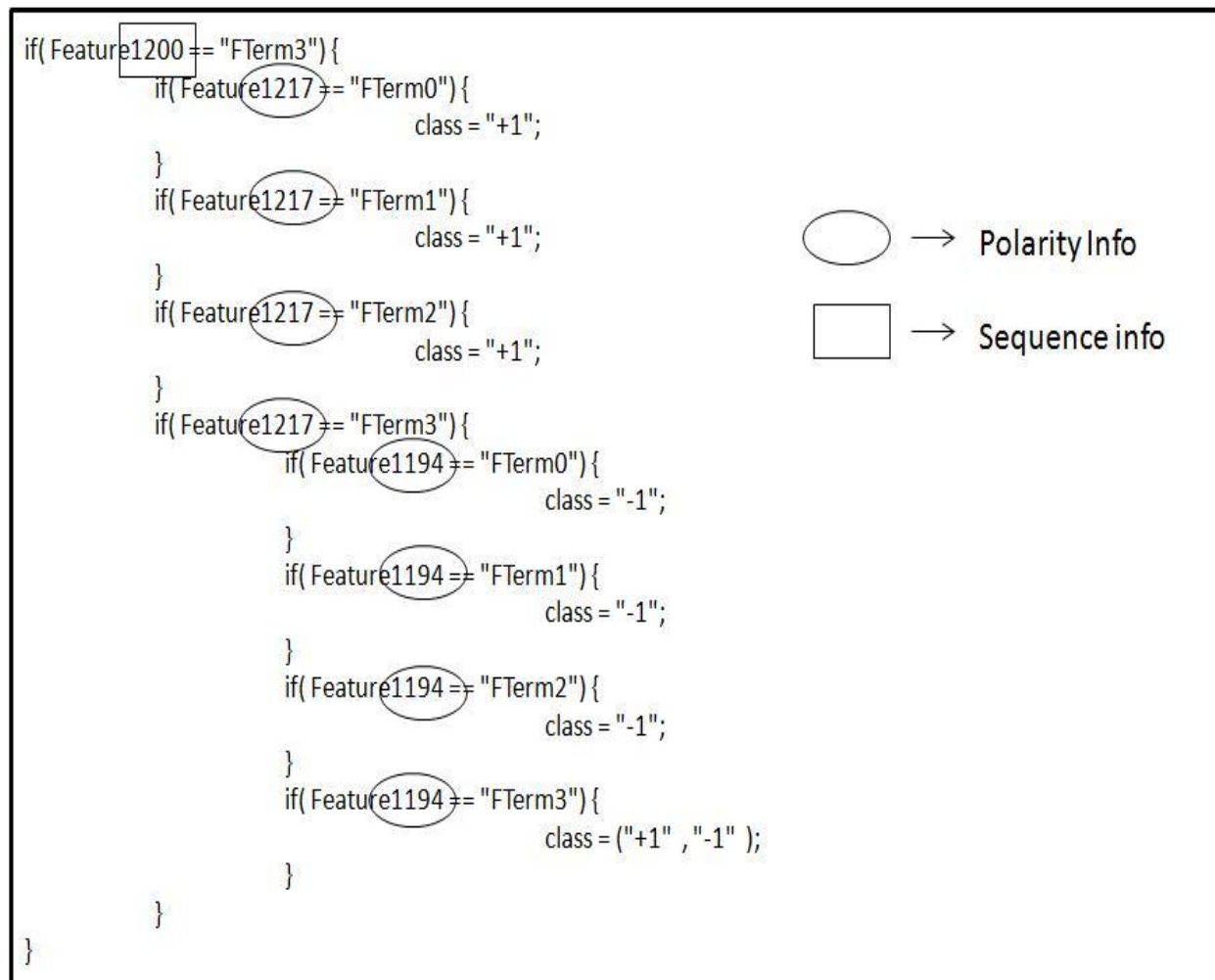
Most of these rules are based on checking the polarity of nearby atoms. The next degree of checks is performed on their amino acid sequence information followed by the distance between atoms. Mostly secondary structure information is least considered in forming the judgment. This correlates directly with well known factor that amino acid properties like charge, hydrophilicity or hydrophobicity, are important to protein 3D structure formation.



**FIGURE 5.5 FUZZY DECISION TREE-I**

Figure 5.5 shows one such fuzzy decision tree partially. It is noticed that polarity features are used more often than any other features. The outline of the rules seems to follow a particular style, always checking polarity of atoms belonging to same spatial group. Those are the atoms that are placed within delta ( $\delta$ ) distance in the three dimensional structure. In different trees different spatial groups are seen to determine the outcome, but the overall style remains the same in trees of all test cases.

Figure 5.6 shows a similar tree but the structure is slightly different compared to first one. Here the first feature considered to classify is a one containing sequence information and then polarity of several atoms all within the same group are considered to make the final decision.



**FIGURE 5.6 FUZZY DECISION TREE-II**

Overall the results look promising and we can further test the system against CASP data set. This is done and explained in the next chapter. By using enhanced encoding results show that, we can distinguish a badly formed model that has no association with structure in PDB from the good ones that belong to PDB. In next few chapters, this novel technique is called using the abbreviation EE\_IFDT. It stands for Enhanced Encoding with Improved Fuzzy Decision Tree.

## CHAPTER 6. TESTING USING CASP TEMPLATES

Every other year since 1994, protein structure modelers from around the world dedicate their late spring and summer to testing their methods in the worldwide modeling experiment called CASP (Critical Assessment of Structure Prediction) [100]. CASP meetings have become one of the most influential venues for assessing protein structure modeling methods. Predictors with expertise in applied mathematics, computer science, biology, physics and chemistry in well over 100 scientific centers around the world work for approximately three months to generate structure models for the set of several tens of protein sequences selected by the experiment organizers. The proteins suggested for prediction (in the CASP jargon – ‘targets’) are either structures soon-to-be solved or structures already solved and deposited at the PDB but kept inaccessible to the general public until the end of the CASP season. The prediction center has been organized to provide the means of objective testing of these methods via the process of blind prediction. These experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused. There have been nine previous CASP experiments. The tenth experiment is planned to start in April 2012. Description of these experiments and the full data (targets, predictions, interactive tables with numerical evaluation results, dynamic graphs and prediction visualization tools) can be accessed through CASP website <http://predictioncenter.org>. Details of the experiments have been published in a scientific journal *Proteins: Structure, Function and Bioinformatics*. These journals include papers describing the structure and conduct of the experiments, the numerical evaluation measures, reports from the assessment teams highlighting state of the art in different prediction categories, methods from some of the most successful prediction

teams, and progress in various aspects of the modeling. Prediction methods are assessed on the basis of the analysis of a large number of blind predictions of protein structure. Summary of numerical evaluation of the methods tested in the latest CASP experiment can be found on their website. Some of the best performing methods are implemented as fully automated servers and therefore can be used by public for protein structure modeling [100] [101].

This chapter contains a detailed overview on CASP experiments, with explanation on procedures followed during the course of the competition. There is also a section on template based modeling and methods that perform near perfection using this technique. Then, the chapter gives a comprehensive description on model quality assessment part of CASP and the methods of evaluating the MQA programs. Finally the chapter shows the Enhances encoding technique's (EE\_IFDT) performance with CASP templates in comparison with other prominent techniques that have been shown to give good results in CASP9 and CASP8.

## **6.1 Synopsis on CASP Experiments**

The main goal of CASP is to obtain an in-depth and objective assessment of our current abilities and inabilities in the area of protein structure prediction. To this end, participants will predict as much as possible about a set of soon to be known structures. These will be true predictions, not 'post-dictions' made on already known structures.

The results of CASP (Critical Assessment of Structure Prediction) are published as articles in special issue of journal PROTEINS. The experiment is dedicated to assess the state of the art in protein structure modeling. The following paragraphs explain the conduct of experiment, the categories of prediction and assessment procedure. There have been nine previous CASP ex-



periments, at two year intervals from 1994 through 2010, and these were reported in respective special issues of PROTEINS [100][101]. The specific challenges in constructing the best possible model of a particular protein depend on a number of factors. To reflect these considerations, CASP modeling targets are divided into categories. The categories have evolved over the course of the experiments, as the capabilities of the methods have changed. This time, as in CASP8, targets were divided into two primary categories – TBM, where a relationship to one or more experimentally determined structure could be identified, providing at least one modeling template and often more [100]. The second category - template-free modeling, where there are either no usefully related structures or the relationship is so distant that it cannot be detected. In addition to evaluating the overall accuracy of three dimensional structure models, CASP also examines other key aspects of structure modeling. There are four different articles published in PROTEINS journal, assessing major aspects in the following areas: prediction of the accuracy of a model, critical to determining whether it is suitable for a particular purpose; prediction of the presence of structural disorder, important since some parts of proteins do not exhibit a single three-dimensional structure under all circumstances; intramolecular contact identification, a source of auxiliary information for template-free modeling; and the identification of ligand binding sites, a central application of models.

The structure of the experiment has three main steps [100]:

1. Participants are required to register for the experiment in one or both of two categories: as human teams, where a combination of computational methods and investigator expertise may be used, and as servers, where methods are only computational and fully automated, so that a target sequence is sent directly to a machine.

2. Information about “soon to be solved” structures is collected from the experimental community and passed on to the modeling community. The trend in recent CASPs has been, nearly all targets were obtained from the structural genomics community, particularly the NIH Protein Structure Initiative Centers (the PSI, <http://www.nigms.nih.gov/Research/FeaturedPrograms/PSI>).

3. Models are collected in accordance with predefined deadlines. Groups are limited to a maximum of five models per target and are instructed that most emphasis in assessment would be placed on the model they designated as the most accurate (referred to as “model 1”). This self-ranking of model quality is also used as part of the evaluation of the state of the art in assigning relative accuracy to models. The models were compared with experiment, using numerical evaluation techniques and expert assessment and a meeting is held to discuss the significance of the results [100] [101].

Early CASP experiments saw dramatic improvements from round to round. Recently, progress has been more gradual, but nevertheless, steady and cumulatively very significant [100]. The CASP web site (<http://predictioncenter.org>) provides extensive details of the targets, the predictions, and the numerical analyses. A CASP10 experiment is planned, beginning in the spring of 2012 and culminating in a meeting in December of that year. The meeting is planned to take place in Italy.

## **6.2 Template Based Modeling in light of CASP**

Assessment of the template-based predictions in the recently completed CASP (CASP 9 [100]) identified the top two teams achieving particularly promising results: groups of Yang Zhang (University of Kansas) [102] and David Baker (University of Washington) [103]. Both

groups used highly automated computational approaches and, while Baker's group utilized hundreds of thousands of CPUs distributed worldwide to build the optimal model (<http://boinc.bakerlab.org/rosetta/>), Zhang's methodology necessitated considerably less CPU time. Zhang's approach is based on the improved I-TASSER methodology [102] that threads targets through the PDB library structures, uses continuous fragments in the alignments to assemble the global structure, fills in the unaligned regions by means of *ab initio* simulations and finally refines the assembly by an iterative lowest energy conformational search. Baker's TBM approach uses three different strategies depending on the target length and target–template sequence similarity [103], and in general relies on computationally demanding sampling of conformational space coupled with an iterative all-atom refinement. The predictions from both groups improved over the best existing templates for the majority of template-based targets in CASP9.

Analyzing the progress of server performance in successive CASPs, it is evident that the gap between the best servers and the best human-expert groups is narrowing over time [104]. Especially in the case of easy TBM, the progress of automated servers is impressive, with the fraction of targets where at least one server model is among the best six submitted models – increasing from 35% in CASP5 to 65% in CASP6, and to over 90% in CASP7. This also confirms the notion that the impact of human expertise on modeling of easy comparative targets is now marginal. In general, in CASP7 servers were at least on par with humans (three or more models in the best six) for about 20% of targets; and significantly worse than the best human model for only very few targets. In CASP7 special attention was dedicated to the assessment of model details in high accuracy TBM. Conceptually, the TBM procedure starts from identifying and selecting the appropriate templates and follows with the target– template sequence alignment. And, after

years of development, the level of target–template structural conservation and the accuracy of the alignment still remain the two issues having the major impact on the quality of resulting models.

Summarizing, currently available template-based methods can reliably generate accurate high resolution models, comparable in quality to the structures solved by low resolution X-ray crystallography, when sequence similarity of a homolog to an already solved structure is high (50% or greater) [104] [105]. As alignment problems are rare in these cases, the main focus shifts to accurate modeling of structurally variable regions (insertions and deletions relative to known homology) and side chains, as well as to structure refinement. The high-quality comparative models often present a level of detail that is sufficient for drug design, detecting sites of protein–protein interactions, understanding enzyme reaction mechanisms, interpretation of disease-causing mutations and molecular replacement in solving crystal structures [8] [9] [10].

### **6.3 Testing EE\_IFDT Algorithm with CASP Dataset**

Human predictors and assessors are not likely to be able to handle many more test sequences than at the past CASP meetings. Predictors only have a few months to generate their models, and an assessor only has about 2 months to examine approximately 1000 models calculated by 100 methods; a rigorous examination that goes beyond the use of a single model quality criterion must depend on consideration of tens of quantitative assessment criteria and visual inspection of each model. It appears that testing with hundreds of sequences can be achieved only by automating both the modeling and assessment methods. Although the CAFASP section of CASP [104], already evaluates automated prediction methods, this assessment is the same as that of the other models and is thus exposed to the same problems. Two research groups, Pcons

(Sweden) [106] and LEE (Korea) [104], outperformed other CASP7 participants in a statistically significant manner based on the results of the paired t-test assessment. Wallner and Elofsson's Pcons [106] is a consensus-based method shown in CASP7 that is capable of a quite reliable ranking of model sets for both easy and hard targets. Pcons uses a meta-server approach (i.e. combines results from several available well-established QA methods) to calculate a quality score reflecting the average similarity of a model to the model ensemble, under the assumption that recurring structural patterns are more likely to be correct than those observed only rarely. The LEE group, by contrast, based their technique on a comparison of query models with their own, and assigned rank in accordance with the distance between the models. There are several other groups that follow the same method for ranking the models. Although both methods could provide a ranking significantly correlated with the one derived from CASP data, they were not able to select the best model consistently from the entire collection of models, indicating that considerable additional effort is needed in this area.

### ***6.3.1 Testing Using CASP8 and CASP9 Templates –Testing Phase I***

We aim to establish effectiveness of our assessment methods by testing our fuzzy decision trees using CASP data. The templates selected from CASP are used only in the testing phase, training data comes from PDB (Protein Data Bank). The vector formation is explained in detail in the previous chapter (Chapter 5). In the model template files are similar to the PDB files with few missing details like secondary structure information and other meta data.

Secondary structures are explicitly given in most PDB files. Templates files only contain the coordinates of atoms and the corresponding secondary structure is calculated using popular DSSP algorithm.

Templates are selected from CASP9 and CASP8 conducted in years 2010 and 2008 respectively. Other CASP data are not selected due to incomplete data in the CASP official website (like missing results summary table etc.). Our algorithm does not rank model, it classifies them in to two classes, similar to protein structure and not similar to protein structure found in PDB. In order to pick only models that are either good or bad and not to pick the ones that fall on the middle grey area, templates are selected based on their GDT\_TS scores. Only good and bad scored templates are selected average scores are avoided.

**TABLE 6.1 RESULTS USING CASP TEMPLATES AS TEST DATA**

	<b>CASP 9</b>	<b>CASP8</b>
<b>Number of Positive Templates</b>	243	187
<b>Number of Negative Templates</b>	85	83
<b>Protein 500</b>	79.13 %	72.22 %
<b>Protein 700</b>	79.61 %	67.78 %
<b>Protein 1000</b>	77.18 %	69.63 %

First only templates that have residue range of 150 to 200 are chosen since the training proteins fall in same range. Among these models, positive and negative data are separated based on their GDT\_TS scores. A score of above 90 is considered as positive data point and a score of less than 10 is selected as negative point, other templates are voided. The data is formatted similar to the training data using spatial feature extraction technique (explained in previous chapter). Among thousands of template, only few hundred satisfy these constraints, the results of these test data are shown in the Table 6.1. Different subsets of the initial culled structures comprising 1000 proteins are used and their individual results are recorded. The subsets are the same that

was used for training phase. In the training phase they were divided into seven folds to check their performance using extended fuzzy decision trees. Now they are not divided into any fold the entire set is considered to check their performance against CASP templates.

The results are used to further emphasis how effective machine learning techniques are for assessing protein models. To obtain a direct comparison of our model assessment technique with other CASP MQAPs , we have to design a scoring method for each test case. Since the fuzzy decision tree used here only classifies each test case and does not explicitly score them (but it does give the rule with maximum weight and the distance between the rule and the data point), we might either deduce a unique scoring mechanism or use entirely a different machine learning algorithm like SVM to score as well as classify the test cases.

### ***6.3.2 Comparison with other Model Assessment Techniques -Testing Phase II and III***

Our technique is different from most of the methods shown in CASP competitions. First most of these techniques are structure alignment techniques that measure the accuracy of the model by comparing it with the native structure. Some techniques do not use the native structure to assess the model, but they are consensus or clustering methods that need a whole lot of models to rank the models. These methods do not give reliable results when number of models is less. Very few methods that use only the single model and its own structure to give assessment on the model are available. These techniques most definitely underperform the consensus methods. In CASP9, targets that were difficult for structure prediction also appeared to be difficult for model quality prediction [107]. This fact can be explained, in part, by the observation that the best performing methods are consensus methods, which work better for the TBM targets for which the cluster center is dominated by the presence of structurally similar templates, while for hard mod-

eling cases, there is usually no consensus or, in some cases, a wrong one. As results from structure comparison programs become less meaningful below some cut-off (e.g., a model with a GDT\_TS score of 20 does not superimpose with a target significantly better than a model with a GDT\_TS score of 15), the relationship between model quality estimates and structure similarity scores for difficult targets can be misleading. In CASP, model quality predictions are evaluated by comparing submitted estimates of global reliability and per-residue accuracy of structural models with the values obtained from the sequence-dependent LGA superpositions of models with experimental structures. There two prediction techniques in CASP are called QA1 and QA2. In QA1, the model assessor are required to give a single score for the entire model and rank the models (global) and in QA2 they are required to provide per –residue scores(local) In both prediction modes, estimated and observed data are compared on a target-by-target basis and by pooling all models together. The first approach rewards methods that are able to correctly rank models regardless of their absolute GDT-TS values, while the second accentuates how well the method is able to assign different scores to models of different quality regardless of their ranking within the set of models for the specific target [107]. Correlation between the predicted accuracy scores and the corresponding GDT\_TS values for the submitted server models are used as a main evaluation measure for assessing the QA1 results.



TABLE 6.2 MQA METHODS CLASSIFICATION

Method	Consensus /Single Model	Local/Global Score	Scoring function
QMEANclust [108]	Consensus	Local + Global	QMEAN-weighted mean GDT_TS deviation of the model to all models in the subset
Multicom-cluster [110]	Consensus	Local + Global	Average GDT-TS between the model and all other models in a decoy set (similar to the NAÏVE_CONSENSUS method - see Materials)
Distill_NNPIF [107]	Single	Local + Global	An artificial neural network based on $\text{Ca-Ca}$ contact interactions
ProQ2 [109]	Single	Local + Global	A successor of ProQ54; combines evolutionary information, multiple sequence alignment and structural features of a model using an SVM

Several top performing groups obtained very similar results. This visual conclusion is confirmed by the results of the statistical significance tests on the common set of predicted targets. According to the paired Student's *t*-test, the top-ranked eight predictors (MuFOLD-WQA, MuFOLD-QA, QMEANClust , United3D, Multicom-cluster, Mufold, MetaMQAPclust, and MQAPmulti—all using clustering techniques) [107] appear to be indistinguishable from each

other and perform better than the rest of the groups at the  $P = 0.01$  significance level. It should be noted that not all groups submitted quality estimates for all models, and therefore correlation coefficients for different groups on a specific target might be calculated on slightly different subsets of models. This may raise a question of reliability of direct comparisons of the scores for different groups.

Comparison of the QA1 results from the latest CASPs points to clear though modest progress in the area: all assessment scores have improved since CASP8 and correlation coefficients for the best groups have nearly reached saturation (0.97), and so it may seem that the QA1 problem has been solved. But a closer look reveals hidden problems and issues that need attention. As in two previous CASPs, all top performing methods in CASP9 relied on a consensus technique to assess model quality. However, for real-life applications, researches may want to obtain estimates for single models downloaded from one of the many widely used model databases. Therefore, there is an urgent need for methods that can assign a quality score to a single model without requiring the availability of tens of models from diverse servers. Unfortunately, these methods lag behind the best consensus-based techniques: the best quasi-single model method in CASP9 was ranked 18th, while the best “pure” single-model method was ranked only 28<sup>th</sup> [107].

So far our technique does not explicitly score the models, so it poses a problem of comparing our technique with other best performing methods submitted in CASP. We still need to make some comparison to make a reliable statement about our technique. So to do this we selected few methods among the ones submitted for CASP. They are QMeanClust [108], ProQ2 [109], Multicom\_Cluster [110] and Distill\_NNPIF [107]. Classification and a brief description about these methods can be found in the

Table 6.2. These particular methods are selected because of their performance in CASP9. QMeanClust and Multicom\_Cluster are consensus techniques that give both local and global score for the model they among the top performing methods in CASP9. ProQ2 and Distill\_NNPIF are techniques that give both local and global score by using only a single model and they are among top performers in this category (single model category). Also Distill\_NNPIF uses neural networks and ProQ2 uses SVM, these can also be categorized under machine learning algorithms, similar to our approach to the problem.

### 6.3.2.1 Results using classification – Testing Phase II

To implement our technique specific targets are selected from CASP9 data archives. Targets T0635 and T0578 are selected because they have residues in the range 150-200 and all the above selected methods have submitted their score files for the templates.

**TABLE 6.3 POSITIVE TEMPLATE RESULTS**

<b>Method</b>	<b>Total number of Positive Templates</b>	<b>Templates with available results</b>	<b>True Positive</b>	<b>False Negative</b>
QMEANclust	162	115	115	0
Multicom-Cluster	162	115	52	0
Distill_NNPIF	162	115	13	102
ProQ2	162	111	107	4
<b>EE_IFDT</b>	<b>162</b>	<b>162</b>	<b>162</b>	<b>0</b>

Among the templates only positive and negative templates are selected based on their GDT\_TS scores. For a GDT\_TS score of above 90 is considered as positive and GDT\_TS below 10 is considered as negative data value.

**TABLE 6.4 NEGATIVE TEMPLATE RESULTS**

<b>Method</b>	<b>Total number of Negative Templates</b>	<b>Templates with available results</b>	<b>True Negative</b>	<b>False Positive</b>
QMEANclust	45	44	44	0
Multicom-Cluster	45	45	45	0
Distill_NNPIF	45	45	45	0
ProQ2	45	40	40	0
<b>EE_IFDT</b>	<b>45</b>	<b>45</b>	<b>11</b>	<b>34</b>

In the target T0635 all score were above 10 GDT\_TS, so this target is used to generate only positive values and T0578 which had no templates with score above 90 is used to generate only negative values. Their performances are shown in the table and table.

The consensus methods that have similar score to that of GDT\_TS and they have flawless performance. The single model techniques are not perfect in their performance. Our technique is good in finding the true positives and not that well in finding true negatives. This could be due to the fact we judge the structure on its sequence and how close it is to other pdb structures. Most of

models are derived from PDB structures and hence might have some of the same features as the pdb structure. We do not penalize the models for missing residues and other obvious mistakes in template based modeling technique. The other major improvement in our technique should be in deducing a scoring function. This will enable us to check its performance among all models to GDT\_TS as well as other model quality assessment techniques in the literature. These results shows how effective our technique is when compared to others. It is noted that how insignificant our technique is when used to classify negative values.

### **6.3.2.2 Results Using Scoring Methodology – Testing Phase III**

For comparing EE\_IFDT technique with other CASP competitors, deducing a scoring mechanism is necessary. Fuzzy decision tree algorithm does not explicitly score each data point. It does however; give the rules that are fired with their individual weights. The rule with maximum weight is used for classification of the data point. It is possible to calculate the distance between the rule with maximum weight and the data point. This distance can be used as a parameter for providing a confidence on the classification itself. This gives us a neat gave of scoring each model using IFDT.

Each data point is scored as

$$score = \frac{\text{maximum weight}}{\text{distance between rule and data point}}$$

After normalizing this score between 0 and 1, the final score is given for positive and negative data points as.

For positive data points, the final score (FS) is calculated as

$$FS = 50 + \text{normalized score} * 50$$

For negative data points, the final score (FS) is calculated as

$$FS = 50 - \text{normalized score} * 50$$

The final score will always be between 0 and 100. For positive vector it will be more than 50 and for negative it will be less than 50.

To compare this score of each data point with GDT\_TS score given in CASP, Pearson's correlation is calculated for both scores. In statistics, the Pearson product-moment correlation coefficient (sometimes referred to as the PPMCC or PCCs, and typically denoted by  $r$ ) is a measure of the correlation (linear dependence) between two variables  $X$  and  $Y$ , giving a value between +1 and -1 inclusive. The correlation coefficient ranges from -1 to 1. A value of 1 implies that a linear equation describes the relationship between  $X$  and  $Y$  perfectly, with all data points lying on a line for which  $Y$  increases as  $X$  increases. A value of -1 implies that all data points lie on a line for which  $Y$  decreases as  $X$  increases. A value of 0 implies that there is no linear correlation between the variables. Suppose we have two variables  $X$  and  $Y$ , with means  $\bar{X}$  and  $\bar{Y}$  respectively and standard deviations  $\sigma_x$  and  $\sigma_y$  respectively. The correlation is computed as

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n - 1)\sigma_x\sigma_y}$$

This correlation is calculated for every method submitted to CASP per target. In following tables the correlation between the few selected techniques from CASP 9 and GDT\_TS score is given along with EE\_IFDT's correlation with GDT\_TS.

**TABLE 6.5 PEARSON'S CORRELATION FOR TARGET T0635**

<b>Method</b>	<b>Total number of Templates</b>	<b>Templates with available results</b>	<b>Pearson's Correlation</b>
QMEANclust	387	324	0.998
Multicom-Cluster	387	325	0.998
Distill_NNPIF	387	325	0.937
ProQ2	387	314	0.728
<b>EE_IFDT</b>	<b>387</b>	<b>387</b>	<b>0.678</b>

**TABLE 6.6 PEARSON'S CORRELATION FOR TARGET T0578**

<b>Method</b>	<b>Total number of Templates</b>	<b>Templates with available results</b>	<b>Pearson's Correlation</b>
QMEANclust	625	321	0.814
Multicom-Cluster	625	322	0.794
Distill_NNPIF	625	322	0.498
ProQ2	625	314	0.572
<b>EE_IFDT</b>	<b>625</b>	<b>625</b>	<b>0.050</b>

From the tables it is seen that Enhanced encoding scheme using improved fuzzy decision trees has the least correlation with GDT\_TS. On closer look, it is seen that EE\_IFDT provides results for all models whereas other give only for selected few. The scoring technique used with

EE\_IFDT certainly needs more refinement as it does not consider parameters unique to model prediction techniques. Even though direct comparison results in a conclusion that EE\_IFDT is not a good classifier or evaluator when it comes to CASP templates, we still have to look at the methodology used here to make any judgment. EE\_IFDT does not use parameters from other scoring techniques. It uses only the coordinate information from the 3Dstructure itself to make the decision. Even though the EE\_IFDT performance is not good in above tables, in case of Target T0635 it does perform better than some of submitted methods in CASP like (Baltymas, Splicer, Splice\_QA, PconsR, PconsD etc.)

**TABLE 6.7 PEARSON'S CORRELATION FOR TARGET T0635**

<b>Method</b>	<b>Total number of Templates</b>	<b>Templates with available results</b>	<b>Pearson's Correlation</b>
Baltymas	387	325	0.000
Splicer_QA	387	313	0.000
PconsR	387	137	0.338
PconsD	377	325	0.040
<b>EE_IFDT</b>	<b>387</b>	<b>387</b>	<b>0.678</b>

In this chapter we use EE\_IFDT to examine the templates submitted to recent CASP competitions. First we selected only templates that are classified as either good or bad and used it as test data set to evaluate (around 10 % of all templates in CASP 8 and 9 since we only chose model 1 of all groups). The training data set is the same ones from training phase of EE\_IFDT algorithm (results given in previous chapter). The prediction accuracy using CASP



template is around 70%, (refer Table 6.1) in testing phase I. To get a better understanding and to compare EE\_IFDT with other methods two targets from CASP 9 are selected for further investigation in testing phases II and III. All templates in this target are used not just the model number 1 of every group, as in testing phase I. These templates are classified as positive and negative based on their GDT\_TS scores similar to phase I of testing. These results are recorded in Table 6.3 and Table 6.4. So far no scoring scheme has been used, since the only way to compare EE\_IFDT with other CASP competitors is to score each model. For this purpose a rudimentary scoring scheme is introduced and its performance in comparison to other prominent CASP competitors is recorded in Table 6.5 and

Table 6.6. In tables EE\_IFDT is seen to have less correlation score compared to others but it is also noted that EE\_IFDT scores all models whereas other have not given scores to few templates. In CASP competitors are not required to score every model and this makes any comparison difficult. Even then, it is noted that the performance of EE\_IFDT is not up to the mark. In the final table, Table 6.7, EE\_IFDT is shown to be better compared to some other competitors of CASP.

## **CHAPTER 7. FUTURE RESEARCH AVENUES**

The main purpose of this research is to identify factors that distinguish between a well formed protein structure and a poorly shaped structure. The first challenge in this scenario is to design a fast and accurate system that classifies the protein models. Secondly we would like to know the features/factors in the learning system that distinguishes a well defined model from low scoring models. Thirdly we would want to study common protein folds that span the protein three dimensional structure space.

### **7.1 Scoring Technique**

For our technique to compete in CASP, we need to deduce a more refined scoring scheme along with classification of templates. The scoring scheme should consider the structure correctness with respect to the sequence as well as other flaws like missing residues etc while evaluating the overall score of the model or CASP templates. This will enable our technique to participate in future CASP competitions. The goal is to design a scoring mechanism that will have a better performance in comparison with other techniques. Along with global score given to the model, a local, per-residue scoring methodology will further enhance the usefulness of the technique. This will also be useful to compare its performance with other techniques that give local score to the models. This will require more knowledge on other features governing the protein space. We try to accomplish something that most other model assessment programs in literature don't. We try to judge a protein model using only information from its own structure and se-

quence and without considering other scoring techniques or consensus methods. The Fuzzy ID3 used could be a phase in a pipeline designed to evaluate models.

## 7.2 Future enhancements in current methodology

The following algorithm explains new methodologies discussed in previous sections to be undertaken in each step of the design.

### Algorithm

#### *Step 1: Data Selection:*

*Currently data is culled using DunBrack lab  
New methodology would include active learning, other algorithm to  
include proteins from different families and classifications*

#### *Step2: Data Encoding:*

##### *Step 2.1: Encoding of Protein Sequence*

*Current Method: BLOSUM Matrix  
Profiles*

*New Method: String Kernel*

##### *Step 2.2: Encoding of Protein Structure*

*Current Method: Distance Matrix*

*New Method: Atom interactions, Hydrogen bonds and other features  
from mentioned feature pool*

#### *Step 3: Learning System:*

*Current Technique: Support Vector Machines*

*Fuzzy Decision Trees*

*New Techniques: Type-2 Fuzzy Decision Tree Algorithm*

*New Random Fuzzy Forest*

Some of these methods like type-2 fuzzy decision trees, fuzzy random forest, use of string kernel for sequence information encoding, use of structure related features for structure information encoding etc. are included to enhance the accuracy of the learning system and others like active learning, ball traverse of the structure etc. are available to improve the speed.

It has been noted that initial encoding method used is not effective especially for the sequence information. Use of string kernels to encode information regarding the protein sequence will result in more efficient data vectors. This will require one more stage of kernel method usage before decision trees. Methods like type-2 fuzzy decision trees, granular neural networks, data clustering etc could be explored. Features other than geometry could also affect the learning capabilities.

### ***7.2.1 Type-2 Fuzzy Decision Tree Algorithm***

This technique would be employed to improve accuracy. Type-1 FDT algorithm got above 80% protein model quality assessment accuracy. Since Type-2 fuzzy sets have more parameters to be optimized, we will design a new Type-2 FDT algorithm and hope to improve protein model quality assessment accuracy by optimizing relevant parameters.

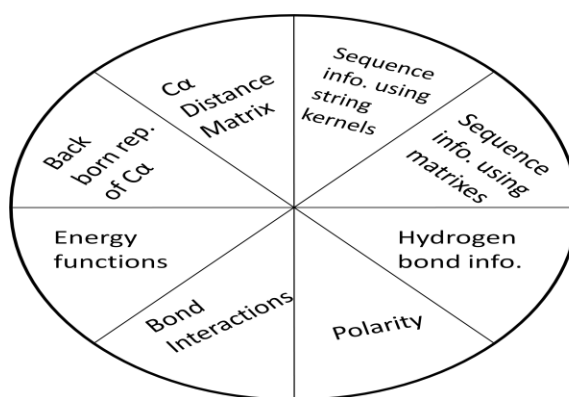
## **7.3 Extracting features distinguishing good protein models from bad ones**

It is expected for the decision tree to produce large amount of rules depending on the data vectors. Rules generated by the decision tree should make sense in terms of biology, chemistry and structural dynamics of the protein. This might not be true for all rules but there should be

few rules that relate to formerly deduced information. This will ensure the process used for rule formation tally with protein structure formation.

### 7.3.1 *Feature Pool*

We try to expand our feature pool to extract as much information about the protein as possible. Any combination of attributes shown in Figure 7.1 could be combined to generate unique representation of protein structures. In our preliminary and enhanced encoding schemes we do make use of some of the feature combinations like sequence information using matrixes, distance matrix, secondary structure information and polarity.



**Figure 7.1 Feature Pool**

### 7.3.2 *New Random Fuzzy Forest*

Many rule-based machine learning methods explicitly use features to build classifiers, so positions and frequencies of using these features in classifiers can be considered as ranking criteria. The key is to build numerous accurate rule-based classifiers from original datasets. Random forest can be altered for our feature selection purpose.

Random forest for imbalanced data learning has already been studied. However, most researchers only focus on prediction performance and ignore the features that have been used in trees. Our approach is to calculate the frequencies and the positions of each biomarker used in these decision trees. Based on its importance in the random forest, each biomarker can be ranked. More importantly, we can find coupling gene markers (i.e., two markers together correlate well with the class label).

## CHAPTER 8. CONCLUSIONS

Most aspects of experimental protein structure predictions process are difficult, time consuming, expensive, labor intensive and problematic. Scientists have agreed it is an impossible task to determine a complete set of all protein structures found in nature, since the number of proteins is much larger than the number of genes in an organism. On the other hand, scientists also believe that there is but a limited number of single domain topologies such that at some point the library of solved protein structure in PDB would be sufficiently complete that the likelihood of finding a new fold is minimal. Earlier even though there were several thousand structures in PDB, most of these structures were not unique instead they were many variant of same structure and sequence. So these did not significantly expand our knowledge of protein structure space. Now experts believe we have sufficient knowledge of protein structure space. This information is critical because it suggest that PDB structures provide a set from which other proteins can be modeled using computational techniques [99]. These fact leads to important task of estimating correctness of the prediction techniques and quality of protein models.

The role of protein structure modeling in biomedical research is steadily increasing. Models are routinely used to address various problems in biology and medicine. Contrary to experimentally derived structures, where accuracy can be deduced from experimental data and typically falls within a narrow range, theoretical models are usually unannotated with quality estimates and can span a broad range of the accuracy spectrum. Thus, reliable a priori estimates of global and local accuracy of models are critical in determining the usefulness of a model to address a specific problem. For example, high-resolution models ( $GDT\_TS > 80$ ) often are sufficiently accurate for detecting sites of protein-ligand interactions, understanding enzyme reaction

mechanisms, interpreting the molecular basis of disease-causing mutations, solving crystal structures by molecular replacement and even for drug discovery. A model of medium accuracy ( $GDT\_TS > 50$ ) can still be useful for detecting putative active sites in proteins, virtual screening, or predicting the effect of disease-related mutations. Low resolution models can be useful for providing structural characterization of macromolecular ensembles, recognizing approximate domain boundaries, helping choose residues for mutation experiments, or formulating hypotheses on the protein molecular function. In response to these needs, the computational biology community has focused on the model quality assessment (MQA) problem, that is, on the possibility of predicting the accuracy of structural models when experimental structural data are not available [107].

Evaluating the accuracy of predicted models is critical for assessing structure prediction methods. This problem is not trivial, a large number of assessment measures have been proposed by various groups and has already become an active subfield of research. Most of these methods are normalized scoring functions that compare the given model to experimental structure. In this research we aim to obtain a binary classifier that studies structures from protein data bank and classifies models as good or bad.

The sheer volume of known structures available makes it possible to develop a machine learning system that studies protein structures and eventually predicts the quality of any new structure model. The most important task in this approach is representing the protein 3D structure in best possible manner and using appropriate machine learning algorithm to get good assessment accuracy. The machine learning techniques considered in this paper are support vector machines and fuzzy decision tree. To solve protein model assessment problem we employed attributes from both protein's sequence and structure.



First a preliminary encoding scheme is employed to achieve this task. The encoding scheme attempts to incorporate number of features to represent the structure and sequence effectively. Two different machine learning technique namely SVM (Support vector machines) and Fuzzy Decision Trees are used to show their individual performance with this selected dataset. SVM did not show any good performance and suffered due to huge number of features involved. To reduce the computational complexity of the program we employ feature selection and parallel processing to be used with SVM. We have implemented kernels to understand a complex 3D object and to judge if the object represents a protein structure and got accuracy at nearly 70%. By the use of improved fuzzy decision tree (IFID3) we could get prediction accuracy above 80%. The accuracy is borderline satisfactory but the main drawback is the uneven feature space and the inconvenience in rule inference. Numerous other concerns arise from this encoding scheme, primarily the number of features required in representing the model and consequent computational overload on the algorithm. Also since sequence length becomes primary factor in coding, proteins of varying sequence length resulted in vectors of varying features. This required further tuning of the vectors to get uniform lengths. This made rule understanding and inference very hard and cumbersome. For the above reason and to increase the prediction accuracy more enhanced spatial encoding technique is considered.

As stated the preliminary encoding scheme is seen to be less efficient in representing the protein structure and due to heavy computational overhead, both machine learning algorithms were limited in their overall performance. This led to an enhanced encoding scheme, which gives prediction accuracy above 95% using certain subsets. The features considered in this scheme are more refined and also contains additional information like polarity of amino acid and secondary structure. This further enhances the overall performance. The vector space is uniform due to

novel spatial feature extraction technique and good data representation. The most important benefit is the inference from the rule tree obtained after classification. The classification is done based on rules that show some correlation to how the protein structure folds. The rules could be further studied to deduce more information on folding of protein structures.

At the end this method is shown to judge the results of previous CASP competitions. The templates from CASP are classified as good or bad (positive or negative) based on their individual GDT\_TS scores. These templates are used only in the testing phase and they are not included in any form in the training data. The tests are performed in several styles to get overall estimate about the newly introduced protein model evaluation technique. First only good and bad models from CASP8 and CASP9 are selected based on GDT\_TS scores. The prediction accuracy for this test is recorded around 70% using fuzzy decision tree and enhanced encoding scheme (EE\_IFDT) for CASP8 and CASP9 templates respectively. Since the methodology studied so far in this research only classifies the data and does not score them, it was hard to compare it with other MQA programs. Still to get some estimate, two targets from CASP9 and few MQA programs that have shown to perform well in CASPs are selected to chart out their performance. The performance of fuzzy decision tree algorithm with enhanced encoding was shown to perform comparable to other MQAP. A rudimentary scoring method is introduced and its correlation with GDT\_TS scores is compared with other CASP competitors. The main drawback is not having the scoring mechanism that is tailored to score models especially the ones predicted using template based modeling technique. Template based modeling is a technique in which PDB structures are used to construct as many parts of the model as possible and then few unaligned residues positions are predicted based on some energy function. Since this methodology will

mostly result in a structure that loosely resembles any PDB structure, the classifier using fuzzy decision tree (EE\_IFDT) is not that sensitive to correct and incorrect models from CASP. Free modeling technique in which no template is used are much harder to evaluate and EE\_IFDT method could be used in this category. Also other important evaluation requirement is to provide score for every residue position and in future this should also be made possible. Both local and global scoring technique and implementation will be the biggest part of future enhancements.

Overall the results look promising, but improvements in data set and parameters could further improve the accuracy of prediction. Improvements like making use of graph kernels, string kernels and kernel fusion methods, decision fusion methods could further enhance the learning system. Deducing a novel scoring technique to effectively score models will be a major focus of future additions.

## BIBLIOGRAPHY

- [1] Berman H.M, Westbrook J, Feng Z, Gilliland G, Bhat T.N, Weissig H, Shindyalov I.N, Bourne P.E. The Protein Data Bank. *Nucleic Acids Research* 28, 2000, pp. 235-242.
- [2] Berman H.M., The Protein Data Bank: a historical perspective. *Acta Crystallographica*. 2008 A 64 : 88-95
- [3] Zhang Y, Skolnick J. The protein structure prediction problem could be solved using current PDB library. *Proc Natl. Acad. Sci. U S A* 2005, 102, pp – 1029-1034.
- [4] Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Science Direct* 2005; 15: 285-289.
- [5] Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure prediction. *Proteins* 1999; Suppl 3: 22- 29.
- [6] Siew N, Elofsson A, Rychlewski L, Fisher D. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* 2000; 16(9): 776-785.
- [7] Ortiz A.R, Strauss C.E., OlmeaO. MAMMOTH (matching models obtained from theory) an automated method or model comparison. *Protein Sci* 2002; 11(11) 2606-2621
- [8] Moult, J. Comparative modeling in structural genomics. *Structure* 2008; 16, 14–16
- [9] Nayeem, A. A comparative study of available software for high-accuracy homology modeling: from sequence alignments to structural models. *Protein Sci*. 2006; 15, 808–824
- [10] Raimondo, D. Automatic procedure for using models of proteins in molecular replacement. *Proteins* 2007; 66, 689–696

- [11] Kopp J., and T. Schwede. Automated protein structure homology modeling: a progress report. *Pharmacogenomics*, 2004; 5(4), 405-416.
- [12] Anfinsen C.B. Principles that govern the folding of protein chains. *Science* 1973; 181(96), 223-230.
- [13] Kolata G. Trying to crack the second half of the genetic code. *Science* 1986; 233(4768), 1037-1039.
- [14] Fiser A. Protein structure modeling in the proteomics era. *Expert Rev Proteomics* 2004; 1(1), 97-110.
- [15] Xiang Z. Advances in homology protein structure modeling. *Curr Protein Pept Sci* 2006; 7(3), 217-227.
- [16] Pillardy J., Czaplewski C., Liwo A., Lee J., Ripoll R., Kaźmierkiewicz R., Ołdziej S., Wedemeyer J., Gibson D., Arnautova A., Saunders J., Ye Y., and A. Scheraga. Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences of the United States of America* 2001; 98(5), 2329-2333.
- [17] Jun-Tao, G., Kyle, E. and X. Ying. A Historical Perspective of Template-Based Protein Structure Prediction in Z. Mohammed and B. Christopher (Eds.), *Protein Structure Prediction* 2008; (Vol. 4, 3-42). Humana Press.
- [18] Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999, 12(2), 85-94.
- [19] Lesk A M., and C. Chothia. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Bio* 1980; 136(3), 225-270.
- [20] Floudas C A. Computational methods in protein structure prediction. *Biotechnol Bioeng* 97(2), 207-213, 2007.
- [21] Pieper U., Eswar N., Stuart C., Ilyin A., and A. Sali. MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res*, 30(1), 255-259, 2002.

- [22] Zhang Y. . Protein structure prediction: when is it useful?. *Current Opinion in Structural Biology*, 19(2), 145-155, 2009.
- [23] Vitkup D., Melamud E., Moult J., and C. Sander . Completeness in structural genomics. *Nat Struct Mol Biol*, 8(6), 559-566, 2001.
- [24] Floudas CA., Fung HK., McAllister SR., Monnigmann M., and R. Rajgaria . Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, 61(3), 966-988, 2006.
- [25] Kihara Daisuke., and J. Skolnick. The PDB is a Covering Set of Small Protein Structures. *Journal of Molecular Biology*, 334(4), 793-802, 2003.
- [26] Zhang Y., Devries E., and J. Skolnick. Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS Comput Biol*, 2(2), e13, 2006.
- [27] Zhang Y. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3), 342-348, 2008.
- [28] Coulson F W., and J. Moult . A unfold, mesofold, and superfold model of protein fold use. *Proteins* 2002; 46(1), 61-71.
- [29] Du P., Andrec M., and M. Levy. Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Eng*, 16(6), 407-414, 2003.
- [30] Bonneau R., and D. Baker . Ab initio protein structure prediction: progress and prospects. *Annu Rev Biophys Biomol Struct*, 30, 173-189, 2001.
- [31] Kihara D, Chen H and Yang Y D. Quality assessment of protein Structure Models. *Current Protein and Peptide Science*, 10, pp - 216-228, 2009.
- [32] Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. *Proteins*, 69( S8): 175-183, 2007.

- [33] Read J., and G. Chavali . Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins*, 69 Suppl 8, 27-37, 2007.
- [34] Jauch R., Yeo Hock C., Kolatkar R., and D. Clarke. Assessment of CASP7 structure predictions for template free targets. *Proteins*, 69 Suppl 8(), 57-67, 2007.
- [35] Ring C S., Sun E., McKerrow J H., Lee G K., Rosenthal P J., Kuntz I D., and F E Cohen. Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc Natl Acad Sci U S A*, 90(8), 3583-3587, 1993.
- [36] Vernal J., Fiser A., Sali A., Muller M., Cazzulo J., and C. Nowicki. Probing the specificity of a trypanosomal aromatic alpha-hydroxy acid dehydrogenase by site-directed mutagenesis. *Biochem Biophys Res Commun*, 293(1), 633-639, 2002.
- [37] Wu G., Fiser A., ter Kuile B., Sali A., and M. Muller. Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proc Natl Acad Sci U S A*, 96(11), 6285-6290, 1999.
- [38] Sheng Y., Sali A., Herzog H., Lahnstein J., and S A Krilis: Site-directed mutagenesis of recombinant human beta 2-glycoprotein I identifies a cluster of lysine residues that are critical for phospholipid binding and anti-cardiolipin antibody activity. *J Immunol*, 157(8), 3744-3751, 1996.
- [39] Vakser I A. Protein docking for low-resolution structures. *Protein Eng*, 8(4), 371-377, 1995.
- [40] Howell P L., Almo S C., Parsons M R., Hajdu J., and G A Petsko. Structure determination of turkey egg-white lysozyme using Laue diffraction data. *Acta Crystallogr B*, 48 ( Pt 2)(), 200-207, 1992.
- [41] Modi S., Paine M J., Sutcliffe M J., Lian L Y., Primrose W U., Wolf C R., and G C Roberts. A model for human cytochrome P450 2D6 based on homology modeling and NMR studies of substrate binding. *Biochemistry*, 35(14), 4540-4550, 1996.

- [42] Eswar N., Webb B., Marti-Renom A., Madhusudhan M S., Eramian D., Shen M., Pieper U., and A. Sali. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci*, Chapter 2, Unit 2.9, 2007.
- [43] Arakaki Adrian K, Zhang Y., and J. Skolnick. Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics*, 20(7), 1087-1096, 2004.
- [44] Boyd A., Ciufo F., Barclay W., Graham E., Haynes P., Doherty K., Riesen M., Burgoyne D., and A. Morgan. A random mutagenesis approach to isolate dominant-negative yeast sec1 mutants reveals a functional role for domain 3a in yeast and mammalian Sec1/Munc18 proteins. *Genetics*, 180(1), 165-178, 2008.
- [45] Tress M., Cheng J., Baldi P., Joo K., Lee J., Seo J., Lee J., Baker D., Chivian D., Kim D., and I. Ezkurdia. Assessment of predictions submitted for the CASP7 domain prediction category. *Proteins*, 69 Suppl 8, 137-151, 2007.
- [46] Zhang Y., Hubner I.A., Arakaki A.K., Shakhrovich E., Skolnich J., “On the origin and highly likely completeness of single-domain protein structures”, *Proceedings of National Academy of Science*, vol. 103 no. 81, pp. 2605-2610, Feb. 21 2006.
- [47] Miao Zhenhua., Luker E., Summers C., Berahovich R., Bhojani S., Rehemtulla A., Kleer G., Essner J., Nasevicius A., Luker D., Howard C., and J. Schall . CXCR7 (RDC1) promotes breast and lung tumor growth in vivo and is expressed on tumor-associated vasculature. *Proc Natl Acad Sci U S A*, 104(40), 15735-1574, 2007.
- [48] Petrey D., Xiang Z., Tang L., Xie L., Gimpelev M., Mitros T., Soto S., Goldsmith-Fischman S., Kernytsky A., Schlessinger A., Koh Y Y., Alexov E., and B. Honig . Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, 53 Suppl 6, 430-435, 2003.



- [49] Mihasan M. Basic protein structure prediction for the biologist: a review. Arch. Biol. Sci., Belgrade, 62 (4), 857-871, 2010
- [50] Laskowski RA., Macarthur MW., Moss DS., and JM Thornton. {PROCHECK}: a program to check the stereochemical quality of protein structures. J. Appl. Cryst, 26(), 283-291, 1993.
- [51] Vriend G. WHAT IF - a molecular modeling and drug design program. Journal of Molecular Graphics, 8(1), 52, 1990.
- [52] Hooft R W., Vriend G., Sander C., and E E Abola. Errors in protein structures. Nature, 381(6580), 272, 1996.
- [53] Sippl M J. Knowledge-based potentials for proteins. Curr Opin Struct Biol, 5(2), 229-235. Soding J., Biegert A., and AN Lupas. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research, 33 (Suppl.2), W244-W248, 2005.
- [54] Eisenberg D., Luthy R., and J U Bowie. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol, 277(), 396-404, 1997.
- [55] Benkert P., Tosatto C E., and D. Schomburg. QMEAN: A comprehensive scoring function for model quality assessment. Proteins, 71(1), 261-277, 2008.
- [56] Wallner B, Fang H, Elofsson A. Automatic consensus-based fold recognition using pcons, proq, and pmodeller. Proteins 2003; 53: 534-541.
- [57] Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci 2003; 12: 1073-1086.
- [58] Wallner B, Elofsson A. Can correct regions in protein models be identified? Protein Sci 2006; 15: 900-913.
- [59] Wang , Tegge and Cheng. Evaluating the absolute quality of a single protein model using structural features and support vector machines. Proteins: Structure, Function, and Bioinformatics Volume 75 Issue 3, pp 638 – 647. Sep 2008

- [60] Pettitt CS, et al. Improving sequence-based fold recognition by using 3D model quality assessment. *Bioinformatics* 2005;21(17):3509–3515. [PubMed: 15955780]
- [61] Zhou H, Skolnick J. Protein model quality assessment prediction by combining fragment comparison and a consensus contact potential. *Proteins* 2007; 71: 1211-1218.
- [62] Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem* 2004; 25: 865-871.
- [63] Karplus K, Katzman S, Shackleford G, Koeva M., Draper J, Barnes B, Soriano M., Hughey R. SAM-T04: what is new in protein structure prediction for CASP6. *Proteins* 2005; 61: 135-142.
- [64] Kajan L, Rychlewski L. Evaluation of 3D-jury on casp7 models. *BMC Bioinformatics* 2007; 8: 304.
- [65] McGuffin LJ. The ModFOLD server for the quality assessment of protein structural models. *Bioinformatics* 2008; 24: 586-587.
- [66] Zhou H, Skolnick J. Ab initio protein structure prediction using chunk-TASSER. *Biophys J* 2007;93 (5):1510–1518. [PubMed: 17496016]
- [67] Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins* 2003; 51: 434-441.
- [68] Qiu J, Sheffler W, Baker D, Noble WS. Ranking predicted protein structures with support vector regression. *Proteins* 2008; 71: 1175-1182.
- [69] Luthy R, Bowie J.U, Eisenberg D. Assessment of protein models with three-dimensional profiles. *Nature* 1992; 356, 83-85
- [70] Tress M, Jones D, Valencia A. Predicting reliable regions in protein alignments from sequence profiles. *J Mol Biol* 2003; 330: 705-718.

- [71] McGuffin L, Bryson K, Jones D. What are the baselines for protein fold recognition? *Bioinformatics* 2001; 17: 63-72.
- [72] McGuffin L. Benchmarking consensus model quality assessment for protein fold recognition. *BMC Bioinformatics* 2007; 8: 345.
- [73] Wiederstein M, Sippl M. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* 2007; 35: W407-410.
- [74] Wu Y, Lu M, Chen M, Li J, Ma J. OPUS-Ca: a knowledge-based potential function requiring only Ca positions. *Protein Sci* 2007; 16: 1449-1463.
- [75] Reyaz-Ahmed A, Zhang Y.-Q. and Harrison R, "Evolutionary Neural SVM and Complete SVM Decision Tree for Protein Secondary Structure Prediction," Special Issue on Computational Intelligence in Knowledge Technology, the International Journal of Computational Intelligence Systems, 2009
- [76] Chen X.J, Harrison R. and Zhang Y.-Q., "Genetic Fuzzy Classification Fusion of Multiple SVMs for Biomedical Data," Special Issue on Evolutionary Computing in Bioinformatics, *Journal of Intelligent and Fuzzy Systems*, vol. 18, no. 6, pp. 527-541, 2007.
- [77] Holm L, Sander C. Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol* 1993; 233: 123-138
- [78] Wang G, Dunbrack R.L., Jr. PISCES: a protein sequence culling server. *Bioinformatics* 2003;19:1589-1591.
- [79] Jin B, Zhang Y.-Q. and Wang B.H, "Granular Kernel Trees with Parallel Genetic Algorithms for Drug Activity Comparisons," *International Journal of Data Mining and Bioinformatics*, vol. 1, no. 3, pp. 270-285, 2007.

- [80] Shawe-Taylor J, Cristianini N. Kernel Method for Pattern Analysis. Cambridge University Press 2004.
- [81] Joachims T, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
- [82] Vapnik V, Corter C. Support Vector Networks. Machine Learning 1995; 20 : 273-293.
- [83] Tang Y.C, Zhang Y.-Q. and Huang Z, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, no. 3, pp. 365-381, July-September 2007.
- [84] Zhi D, Krishna S, Cao H, Pevzner P, Godzik A. Representing and comparing protein structures as paths in three-dimensional space. BMC Bioinformatics 2006; 7:460.
- [85] Reyaz-Ahmed A., Harrison R. and Zhang Y.-Q., "3D Protein Model Assessment Using Geometric and Biological Features," Proceedings of SEDM 2010, Chengdu, June 23-25, 2010.
- [86] Lee K., Lee K., Lee J., Lee-Kwang H., " A Fuzzy Decision Tree Introduction Method for Fuzzy Data", Proc. IEEE Conf. on Fuzzy Systems, FUZZ-IEEE 99, Seoul, Vol. 1, pages 16-25, 1999
- [87] Umamo M., Okamoto H., Hatono I., Tamura H., Kawachi F., Umedzu S., Kinoshita J., "Fuzzy Decision Trees by Fuzzy ID3 Algorithm and its Application to Diagnosis Systems", in Proc. 3rd IEEE Conf. on Fuzzy Systems, Orlando, Vol. 3, pages 2113-2118
- [88] Quinlan J. R., "Discovering Rules by Induction from Large Collection of Examples", in D. Michi(ed): Expert Systems inMicro Electronics Age, Edinburgh University Press, 1979.
- [89] Chandra B., Varghese P. P., "Fuzzifying Gini Index Based Decision Trees", Expert Systems with Application, Vol. 36 Issue4, pages 8549-8559, 2009
- [90] Quinlan J.R., "Introduction of Decision Trees", Machine Learning, Vol. 1 pages 81-106, 1986

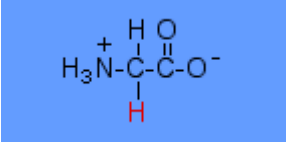
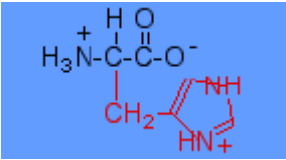
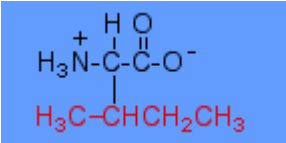
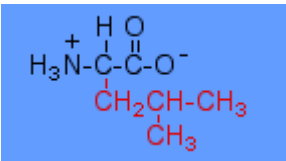
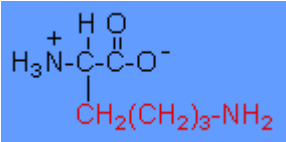
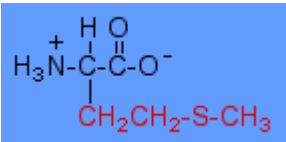
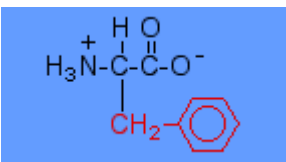
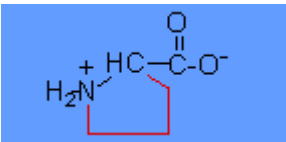
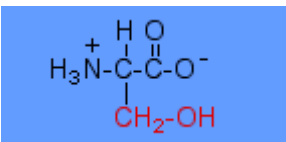
- [91] Yuan Y., Shaw M.J., "Introduction of Fuzzy Decision Trees", Fuzzy Sets and Systems, Vol. 69, Issue 2, pages 125-139, 1995.
- [92] Breiman L., Friedman J. H., Olshen J. A., & Stone . J, "Classification and regression trees. Belmont", CA: Wadsworth International Group. 1984
- [93] Abu-halaweh N, Harrison R.W, "Prediction and Classification of Real and Pseudo MicroRNA Precursors via Data Fuzzification and Fuzzy Decision Tree", Proceedings of ISBRA, pages 323-334, 2009.
- [94] Abu-halaweh N, Harrison R.W, "Rule Set Reduction in Fuzzy Decision Trees", Proceedings of NAFIPS, 2009, p.p. 1-4.
- [95] Abu-halaweh N, Harrison R.W, "Practical Fuzzy Decision Trees", Proceeding of IEEE Symposium on Computational Intelligence and Data Mining (ICTAI), pages 203-206, 2000.
- [96] Peng H., Long F., Ding C.. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8):1226-1238, 2005.
- [97] Tang Y.C., Zhang Y.-Q., Huang Z., Hu X.H. T., and Zhao Y., "Recursive Fuzzy Granulation for Gene Subsets Extraction and Cancer Classification," IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 6, pp. 723-730, Nov. 2008.
- [98] Reyaz-Ahmed A., Abu-halaweh N., Harrison R. and Zhang Y.-Q., "Protein Model Assessment via Improved Fuzzy Decision Tree," Proc. of BIOCOMP 2010, Las Vegas, July 12-15, 2010.
- [99] Zhang Y., Hubner I.A., Arakaki A.K., Shakhrovich E., Skolnich J., "On the origin and highly likely completeness of single-domain protein structures", Proceedings of National Academy of Science, vol. 103 no. 81, pp. 2605-2610, Feb. 21 2006.

- [100] Moult J, Fidelis K, Kryshtafovych A, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)- Round IX. *Proteins* ; 79(Suppl 10):1–5, 2011.
- [101] Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction- Round VII. *Proteins*;69(Suppl 8):3–9, 2007.
- [102] Zhang, Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* 69 (Suppl. 8), 108–117, 2007
- [103] Baker, D. and Sali, A. Protein structure prediction and structural genomics *Science* 294, 93–96, 2001
- [104] Kryshtafovych A. and Fidelis K. Protein structure prediction and model quality assessment. *Drug Discovery Today* \_ Volume 14, Numbers 7/8 \_ April, 2009.
- [105] Ginalski, K. Comparative modeling for protein structure prediction. *Curr. Opin. Struct. Biol.* 16, 172–177, 2006
- [106] Wallner, B. and Elofsson, A. Prediction of global and local model quality in CASP7 using Pcons and ProQ. *Proteins* 69 (Suppl. 8), 184–193, 2007
- [107] Kryshtafovych A, Fidelis K, Tramontano A. Evaluation of model quality predictions in CASP9. *Proteins*;79(Suppl 10):91–106, 2011.
- [108] Benkert P, Tosatto SC, Schwede T. Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust. *Proteins* ;77 (Suppl 9):173–180, 2009.
- [109] Larsson P, Skwark MJ, Wallner B, Elofsson A. Assessment of global and local model quality in CASP8 using Pcons and ProQ. *Proteins*;77:167–172, 2009.
- [110] Wang Z, Eickholt J, Checg J. MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. *Bioinformatics* Vol. 26 no. 7, pages 882–888, 2010

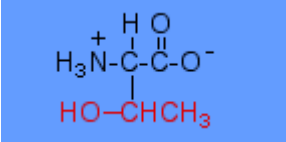
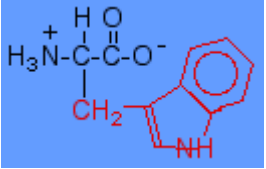
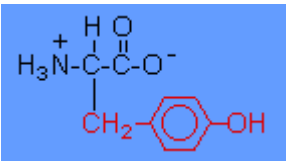
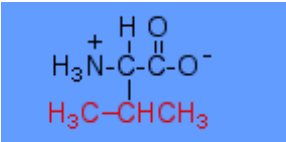
## APPENDIX

## Structure of Amino Acids

Amino Acid Name	Abbrev.	Abbrev.	Structure of R group (red)	Polarity Information
Alanine	ala	A	$\begin{array}{c} \text{H} \quad \text{O} \\   \quad    \\ \text{H}_3\text{N}^+-\text{C}-\text{C}-\text{O}^- \\   \\ \text{CH}_3 \end{array}$	Neutral Non-polar
Arginine	arg	R	$\begin{array}{c} \text{H} \quad \text{O} \quad \quad \text{H} \\   \quad    \quad \quad   \\ \text{H}_3\text{N}^+-\text{C}-\text{C}-\text{O}^- \quad \text{NH}_2^+ \\   \quad \quad \quad   \\ (\text{CH}_2)_3-\text{N}-\text{C}-\text{NH}_2 \end{array}$	Basic Polar
Asparagine	asn	N	$\begin{array}{c} \text{H} \quad \text{O} \\   \quad    \\ \text{H}_3\text{N}^+-\text{C}-\text{C}-\text{O}^- \\   \quad \quad \quad    \\ \text{CH}_2-\text{C}-\text{NH}_2 \end{array}$	Neutral Polar
Aspartic Acid	asp	D	$\begin{array}{c} \text{H} \quad \text{O} \\   \quad    \\ \text{H}_3\text{N}^+-\text{C}-\text{C}-\text{O}^- \\   \quad \quad \quad    \\ \text{CH}_2-\text{C}-\text{OH} \end{array}$	Acidic Polar
Cysteine	cys	C	$\begin{array}{c} \text{H} \quad \text{O} \\   \quad    \\ \text{H}_3\text{N}^+-\text{C}-\text{C}-\text{O}^- \\   \\ \text{CH}_2-\text{SH} \end{array}$	Neutral Slightly Polar
Glutamic Acid	glu	E	$\begin{array}{c} \text{H} \quad \text{O} \\   \quad    \\ \text{H}_3\text{N}^+-\text{C}-\text{C}-\text{O}^- \\   \quad \quad \quad    \\ \text{CH}_2\text{CH}_2-\text{C}-\text{OH} \end{array}$	Acidic Polar
Glutamine	gln	Q	$\begin{array}{c} \text{H} \quad \text{O} \\   \quad    \\ \text{H}_3\text{N}^+-\text{C}-\text{C}-\text{O}^- \\   \quad \quad \quad    \\ \text{CH}_2\text{CH}_2-\text{C}-\text{NH}_2 \end{array}$	Neutral Polar

Glycine	<b>gly</b>	G		Neutral Non-polar
Histidine	<b>his</b>	H		Basic Polar
Isoleucine	<b>ile</b>	I		Neutral Non-polar
Leucine	<b>leu</b>	L		Neutral Non-polar
Lysine	<b>lys</b>	K		Basic Polar
Methionine	<b>met</b>	M		Neutral Non-polar
Phenyl-alanine	<b>phe</b>	F		Neutral Non-polar
Proline	<b>pro</b>	P		Neutral Non-polar
Serine	<b>ser</b>	S		Neutral Polar



Threonine	<b>thr</b>	T	 <p>The chemical structure of Threonine is shown. It features a central alpha-carbon bonded to a hydrogen atom (H), an amino group (H<sub>3</sub>N<sup>+</sup>), a carboxylate group (COO<sup>-</sup>), and a side chain (CH(OH)CH<sub>3</sub>). The side chain is highlighted in red.</p>	Neutral Polar
Tryptophan	<b>trp</b>	W	 <p>The chemical structure of Tryptophan is shown. It features a central alpha-carbon bonded to a hydrogen atom (H), an amino group (H<sub>3</sub>N<sup>+</sup>), a carboxylate group (COO<sup>-</sup>), and a side chain (CH<sub>2</sub> attached to an indole ring). The side chain is highlighted in red.</p>	Neutral Slightly polar
Tyrosine	<b>tyr</b>	Y	 <p>The chemical structure of Tyrosine is shown. It features a central alpha-carbon bonded to a hydrogen atom (H), an amino group (H<sub>3</sub>N<sup>+</sup>), a carboxylate group (COO<sup>-</sup>), and a side chain (CH<sub>2</sub> attached to a para-hydroxybenzene ring). The side chain is highlighted in red.</p>	Neutral Polar
Valine	<b>Val</b>	V	 <p>The chemical structure of Valine is shown. It features a central alpha-carbon bonded to a hydrogen atom (H), an amino group (H<sub>3</sub>N<sup>+</sup>), a carboxylate group (COO<sup>-</sup>), and a side chain (CH(CH<sub>3</sub>)<sub>2</sub>). The side chain is highlighted in red.</p>	Neutral Non-polar

## BLOSUM62 MATRIX

[illegible]