Spring 5-15-2015

A Balanced Secondary Structure Predictor

Md Nasrul Islam
University of New Orleans, nasrul.shohan@gmail.com

A Balanced Secondary Structure Predictor

A Thesis

Submitted to the Graduate Faculty of the
University of New Orleans
In partial fulfillment of the
Requirements for the degree of

Master of Science
in
Computer Science

by

Md Nasrul Islam

B.Sc. Bangladesh University of Engineering and Technology, 2008

May, 2015

Acknowledgement

First of all, I would like to humbly express my profound gratitude to my supervisor Dr. Md Tamjidul Hoque for being so kind, enduring and at the same time vigilant in every aspect of my research and academic progress during the whole time I have been here at University of New Orleans. He has been tireless to explain every nuances again and again till I felt little uncomfortable to understand any concept. I must appreciate his continuous guidance, critical and insightful advice, helpful and inspiring criticism, valuable suggestions and commendable support and quick feedbacks to my queries throughout the way towards completion of my thesis.

Secondly, I would like thank Dr. Christopher M. Summa and Dr. N. Adlai A. DePano for their kind consent to be a board member of my thesis committee, despite their hectic schedule and important other priorities.

I would also like to thank University of New Orleans for providing me an excellent environment for research and financial support in the form of graduate research assistantship.

I must mention my adorable wife and also my lab partner, Sumaiya Iqbal, for her continuous encouragement and support, each and every day in every aspect of my life. And I would like to express my sincere gratitude to my mother in law, Salma Iqbal, for taking the full responsibility of looking after my new born baby during last two hectic semesters.

Table of Contents

List of Figures

List of Tables

Abstract

Secondary structure (SS) refers to the local spatial organization of the polypeptide backbone atoms of a protein. Accurate prediction of SS is a vital clue to resolve the 3D structure of protein. SS has three different components- helix (H), beta (E) and coil (C). Most SS predictors are imbalanced as their accuracy in predicting helix and coil are high, however significantly low in the beta. The objective of this thesis is to develop a balanced SS predictor which achieves good accuracies in all three SS components. We proposed a novel approach to solve this problem by combining a genetic algorithm (GA) with a support vector machine. We prepared two test datasets (CB471 and N295) to compare the performance of our predictors with SPINE X. Overall accuracy of our predictor was 76.4% and 77.2% respectively on CB471 and N295 datasets, while SPINE X gave 76.5% overall accuracy on both test datasets.

Protein, Secondary structure, MetaSSPred, Support vector machine, Genetic algorithm, balanced prediction

1. Introduction

1.1 Introduction

In the modern scientific world bioinformatics has attained a very crucial position as a research discipline, promising the potential of benefitting human endeavor to understand and analyze biological phenomena. It is a multidisciplinary field where computer scientists can provide biologists critical tools to study genomics, proteomics, medicine and many more. Proteomics, as a discipline, studies the function and structure of proteins and requires the processing and analysis of enormous amounts of data. The number of different structures and functions of proteins in a single organism is staggering [1]. Considering either from a quantitative or a functional perspective, proteins are arguably the most important macromolecules in all living organisms. They make possible all of the chemical reactions in living cells [1]. More than half of the dry weight of cells are constituted by proteins of various shapes and sizes and they play a significant role in the functions of cells. Chemical organization of proteins is relatively simple. They are linear chains of amino acids connected through covalent bonds commonly known as peptide bonds [2]. The main constituents of proteins, amino acids, are small molecules with a common backbone consisting of several C, H, O, and N atoms and a side chain with up to 30 more atoms. This apparently simple linear chain of amino acids adopts a specific folded three-dimensional (3D) shape, which enables proteins to perform various tasks. This 3D shape of a particular protein may not remain fixed, rather, the protein may explore a wide array of kinetically accessible conformations. The spatial arrangement of atoms in a protein macromolecule is called its conformation. Sets of possible conformations of a protein include any structural state that can be formed without breaking covalent bonds. However, all points in the protein's conformational space

are not equally probable. The conformation that exists under a given set of conditions is usually the one that is thermodynamically most stable and has the lowest Gibbs free energy. This structural dynamism also yields additional functionality to proteins. Some examples of tasks carried out by proteins are transportation of small molecules such as haemoglobin that transports oxygen in the bloodstream, storage, as done by ferritin for iron storage and release, catalyzing biological functions, providing structure to collagen and skin, controlling sense, regulating hormones, processing emotion, etc. [3]. Without the catalyzing effect of proteins, many chemical reactions in the cell of living organism would happen in a rate which can be deemed as negligible [1]. Knowledge about the structure of a protein reveals important information about the location of probable functional or interaction sites, distantly related protein identification, and detection of the important regions of the protein sequence which are crucial in maintaining the structure of the protein, and so on [4]. A widely accepted fundamental principle in protein science is that *protein structure leads to protein function* [5]. For these reasons, prediction of protein structure and function has become one of the most important problems in molecular biology. The first successful discovery of protein structure is credited to Max Perutz and John Kendrew of Cambridge University [6]. Perutz and John Kendrew discovered the high resolution structure of myoglobin and hemoglobin respectively through X-ray crystallography. Since then, X-ray crystallography has been the most widely used experimental method to determine protein structure. Over 80% of the three-dimensional macromolecular structure data in the Protein Data Bank (PDB) were obtained by X-ray crystallography [7]. Nuclear magnetic resonance (NMR) spectroscopy is the second most widely used experimental method for obtaining three dimensional structure of proteins at atomic level resolution [8]. More than 10,000 structures, about 16% of the total structures in PDB, have been solved by NMR [7]. However experimental methods for protein structure determination are

very time consuming as well as expensive. For example, one experiment cost around $250,000 in 2000 and $65,000 in 2008 [9]. One structure determination through such experimental methods may take months or even years [10]. The drawbacks of experimental methods are not limited to time and cost only. For some proteins it is not possible to apply such experimental methods. For example, X-ray crystallography is extraordinarily difficult for membrane protein structure determination. Protein molecules with highly hydrophobic portions generally cannot be crystallized [11]. NMR techniques cannot be applied for larger proteins of size more than 100 KDa. On the other hand, computational approaches have the potential to overcome the previously mentioned difficulties or disadvantages associated with the experimental approaches by utilizing the correlation between the primary sequence information and the final 3D structure. During the last several decades the amount of biological data has rapidly increased. Genome sequences of a number of species, including human, have been completely mapped thanks to the world-wide genome sequencing project. The gap between the number of known sequence and the number of known structures is widening rapidly [12]. A detailed understanding of the biological role of the majority of proteins is not possible through genome sequencing alone, rather we need structural and functional information for that [13]. With increasing research to discover new biological processes and to master existing known processes, the development of well performing and new computational tools becomes increasingly necessary and useful. If the path of protein folding from sequence to 3D structure can be properly modeled, it will bring about radical benefits in combating many diseases by solving either fully or partially various currently existing crucial medical, agricultural and biotechnological bottlenecks. Therefore, high throughput computational models for the prediction of protein structure from sequence is an immediate need. For this reason prediction of protein structure and function from sequences has been referred to as the second half

of genetics [14]. The principle of Anfinsen asserts that all information required to specify the structure of a protein is encoded in its amino acid sequence [15, 16]. However this task of predicting the 3D structure of protein from sequence information is not straight forward as how to read this information off of the sequence so as to reconstruct 3D structure remains unclear [17]. Mostly because proteins exhibit some general patterns and a degree of regularity in their folding, it is possible to apply computational techniques to investigate this challenging problem. Although prediction of protein 3D structures is the ultimate target, the structure yet cannot be accurately predicted directly from sequences [12]. However this final 3D structure prediction problem is usually approached by solving course-grained intermediate problems such as secondary structure prediction (SSP) [12, 18, 19]. Secondary structure (SS) refers to the local spatial organization of a polypeptide backbone atoms of a protein [20]. It may be deemed as a notion of residue-level local sub-structure. Secondary structures are determined by examining the pattern of hydrogen bonds between side chains and amino acid residues in a protein. As proposed by Pauling and his colleagues, there are mainly two types of secondary structure- alpha helix (α) or helix (H) and beta (β) strand or beta sheet (E) [21]. These are all regular polypeptide folding patterns. Dominant hydrogen bonding patterns are turn and bridge. Repeated turns give rise to helixes while repeated bridges generate strands [22]. However another type, turn or coil (C), is also considered as a kind of secondary structure. The third type is generally referred to the structure of those residues which are neither helix nor beta sheet. Prediction of SS greatly simplifies the ultimate 3D structure prediction problem [12].

From a machine learning point of view, SSP is a classification problem, where based on relevant features we have to decide on each amino acid in a protein belongs to protein secondary structures, namely helix, beta or coil. More specifically it is a three class classification problem.

SSP problem has been approached using various machine learning algorithms including neural networks (NN), hidden Markov model (HMM), support vector machine (SVM), etc. Such machine learning approaches used so far for SSP vary in basic algorithm used, and/or in feature sets employed. Despite numerous effort the accuracy of SSP stuck at around 80% for last five decades. This accuracy is also very much dataset dependent. Needless to say, the increase in accuracy of SSP, is crucial for biological and medical development.

In this thesis, we have employed SVM with a radial basis function (RBF) kernel, along with several novel features such as disorder probability, bigram and monogram. SVM is a well performing algorithm in biological application compared with other machine learning algorithms as they are effective in controlling the classifier's capacity and the associated potential for over fitting ensuring maximum margin of the decision boundary separating two classes [18]. Instead of directly trying three class classification, we have employed three binary classifiers *viz*- H/~H, E/~E and C/~C separately and then to combine. This provides us the opportunity to efficiently attack the problem in simpler form and then to optimally combine them. We consolidated the predictions from these three binary classifiers to come up with a final three class prediction with optimal weighting using heuristics obtained from a genetic algorithm. We also developed a meta-predictor by combining the prediction of our combined SVM predictor and SPINE X [23], a state-of-the-art secondary structure predictor. Our meta-predictor comes up with a highly balanced overall prediction as well as the prediction of three secondary structure components with higher accuracy and improving the accuracy of beta structure prediction significantly in particular.

1.2 Motivation

Since the 3D structure of protein is a pivotal clue to the study of a protein's function, a good number of prominent researchers have devoted a significant amount of effort to find methods to

predict protein 3D structure. Despite numerous experimental and analytical endeavors, protein 3D structure prediction is still an unresolved problem. This issue is commonly known as the protein folding problem. Notable pioneers, Pauling and Corey devoted decades to find way to predict the accurate structure of amino acids, peptides and other substances that construct the structure of protein. They suggested that interatomic distances, bond angles and other configurational parameters might aid such prediction [21]. They used such information to develop models of two different hydrogen bonded helical conformations, keeping in mind that such things are likely to develop significant part of the structure of both globular and fibrous proteins. With thorough analysis of different bond length, hydrogen bond distances and neighboring atoms' influences, they proposed the idea of helices with different non-integer number of residues per helix. They also proposed the idea of a planar peptide bond that drastically simplifies the study and the understanding of protein structure. This successful prediction of alpha helix is a significant contribution of Linus Pauling. He achieved it due to his assumptions of planar peptide bonds, equivalency of amino acids with respect to backbone conformation and hydrogen bonds between amide protein and the O atom of adjacent residue with an N–O bond distance of 2.72 Å [24]. In this regard, a significant contribution is Anfinsen's work [25, 26].

Anfinsen and his colleagues established the "Thermodynamic hypothesis" that the three dimensional structure of a protein in its native environment is the one that minimizes Gibbs free energy of the whole system. This notion ultimately translates into that the native structure is determined by inter atomic interaction, hence by the sequence of amino acids of protein. This finding brought Anfinsen Nobel prize in 1972. Guzzo in 1965 suggested significant influence of amino acid sequence on the location of helical and non-helical part of a protein structure based on known sequences and structures of myoglobin, and alpha and beta hemoglobin [27]. Guzzo also

emphasized on the notion that without considering other components of the cell, the enzymatically active secondary and tertiary structure of protein may be resolved solely from the interaction between and amino acids and solvent in a solution. For instance, Guzzo said that presence of proline, aspartic acid, glutamic acid, or histidine are important to form a helical disruption.

Further, the famously known *Levinthal Paradox* suggests that proteins fold into their specific 3D conformations in a time-span way [28] shorter than it would be possible for protein molecules to actually search the entire conformational space for the lowest energy state. Therefore, hierarchical approaches are very suitable for this critical problem solving. Secondary structure prediction is a critical building block towards protein fold recognition. One reason is that secondary structure gives local structural preferences which limits the possible number of configurations to each part of a polypeptide chain. In the amino and carboxyl termini of alpha helices, often very strong sequence-structure correlations are observed [29]. Therefore, secondary structure information significantly reduces the conformational search space for fold recognition. Accurately computing protein structure is also important for crucial biological applications such as virtual ligand screening [9, 30], structure based protein function prediction [31] and structure based drug design [32]. Therefore, every single advancement towards solving the protein folding problem is vitally important for human kind. For all these reasons we have taken the challenge to enhance the accuracy of SSP problem.

2. Background and Related Works

2.1 Fundamentals of Protein

Proteins are the most versatile macromolecules in living organisms and play significant roles essentially in all biological processes [33]. Proteins are large biological polymers composed of single or multiple chain of amino acid residues. An amino acid is an organic compound that has a central carbon atom, usually known as alpha carbon (Cα) or chiral carbon that uses its four valences to create bond with a carboxylic group, an amino group, a hydrogen atom and a side chain. A simple amino acid molecule along with its ionic condition is shown in the Figure 1.

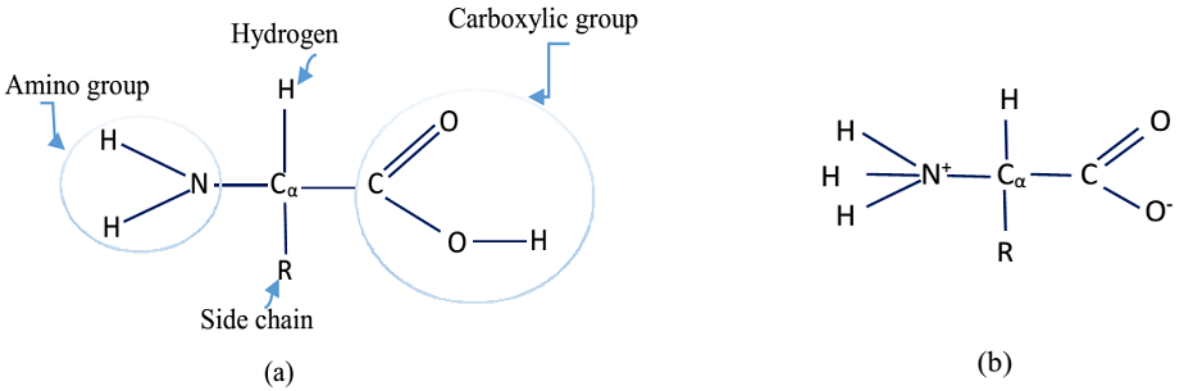

Figure 1: (a) An amino acid with its bonds, (b) An amino acid at pH 7.0.

Side chains are unique features of amino acids. Side chains distinguish one amino acid from other amino acids. Their interaction in protein with the surroundings depends on this side chain. Depending on the nature of this side chain, the properties of different amino acids also vary. They can be hydrophilic, hydrophobic, acidic or basic, etc. The simplest side chain may be a hydrogen atom (H). In general 20 different amino acids are found in protein molecules. These amino acids are monomeric building block of protein. Protein length is usually expressed in terms of amino acids in its structure. A protein structure may contain different number of amino acids ranging

from fewer than 20 to more than 5000, however on an average a protein has 350 amino acid residues. The possible space for variation of protein could be as large as 20^{4500} or 10^{5850} [1]. Protein structure and function are mainly determined by the sequence of amino acids in a particular polymer. Two adjacent amino acids are connected through linear peptide bond. Peptide bonds, a type of covalent bond, are created through a condensation reaction between the carboxylic group of one amino acids and the amino group of another adjacent amino acid. The general reaction that forms the peptide bond backbone of proteins is shown in Figure 2 below:

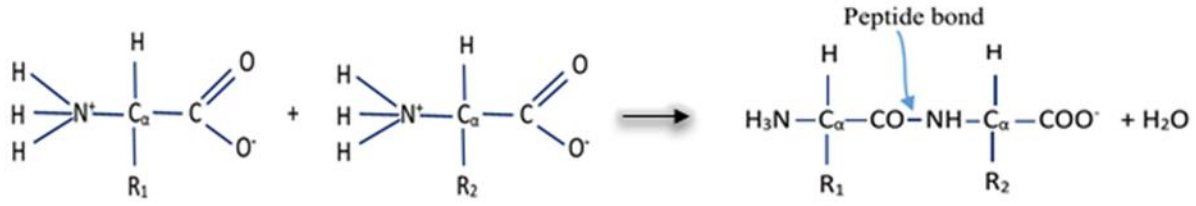

Figure 2: Condensation reaction that forms protein chain by developing peptide bond between two amino acids. Here R_1 and R_2 stand for side chains.

Compound created through peptide bonds are known as polypeptide. In this sense proteins are also polypeptides, however only short chains of amino acids are usually known as polypeptides. The polypeptide chain begins with the amino group and ends with the carboxyl group. Terminal with amino groups is also known as N- terminus while the terminal with carboxyl group is known as C- terminus. Every protein has a unique amino acid sequence. This sequence is based on the codons in the encoding gene [34]. Three nucleotides constitute a codon which determines the particular amino acid to be added in a particular position of the protein chain. There exist 64 different codons to specify the set of possible 20 amino acids.

From chemical point of view, proteins are one of the most complex and functionally important molecules [35]. Important properties that enable proteins for a wide variety of functions [33] are discussed below:

- Polymer type: Proteins are made up of only 20 different amino acids, the combination of those amino acids greatly varies. Such variation of combining amino acids make possible formation of myriad number of different proteins with different functionalities.
- Functional groups: Protein molecules may contain a wide variety of functional groups such as alcohols, thiols, thioethers, carboxylic acids, carboxamides, etc. These functional groups are oriented in protein molecules in numerous fashion, which give rise to a broad spectrum of protein functionality. They also interact in such a way that the chemical reactivity of amino acid side chains enhances [35].
- Formation of complex assembles: Proteins are capable of interacting with one another and with other biological macromolecules. Such interactions enable proteins to form complex assemblies. These assemblies may act as macro-molecular machines which are capable of precisely replicating DNA, transmitting signals within cells, and also helping in many other essential biological processes.
- Mix of structural rigidity and flexibility: Some parts of the protein structure may be very rigid which may act as the skeleton of the macro-molecule while other parts may be flexible. These flexible parts with their limited flexibility may work as hinges, springs, and levers and may assist in assembling of proteins with one another and with other molecules into complex units, and in transmitting information within and between cells.

- Diversity in size, shape and chemical properties: Protein size varies; as we have noted earlier that the number of amino acid residues in a protein may vary from only 20 to many thousands. Protein may be hydrophobic, hydrophilic, fibrous, globular etc. Proteins may also have affinity to binding to a variety of different compounds, atoms or molecules commonly known as *prosthetic groups* [1]. Prosthetic groups may be organic such as vitamin or inorganic such as metal ions that may bind to a specific site of a protein and are important for different functionality of proteins. All these different features add to the spectrum of diverse functionality of protein. A well-known example of prosthetic group is *heme* which binds oxygen to protein hemoglobin.

2.2 Structure of a Protein

A fundamental principle in protein science is that the *protein structure leads to protein function* [5]. If we want to know how proteins function, we must know the structure of the proteins accurately. Therefore, the study of protein function is inseparable from the study of protein structure.

It has been though for a long time that proteins are random colloids of structures until it was shown by Bernal and Crowfoot that if a crystal of pepsin yields a discrete diffraction pattern if placed in a beam of X-ray [36]. This finding is a pioneering evidence that protein structures are not random colloid rather a large structured molecule consists of ordered array of atoms. Now through studies on a large number of proteins it has been established that protein shows significant degree of structural regularities in terms of repetitive structural patterns, which may be classified into distinct categories [20].

Following discussion focuses on the levels of structure of protein. Protein structures are usually categorized into the following four levels' of complexities:

- Primary structure
- Secondary structure
- Tertiary structure and
- Quaternary structure

2.2.1 Primary structure

Primary structure is defined as the linear sequence of amino acids. Sequence of amino acids in a particular protein is not random, rather fixed which was first discovered by Frederick Sanger [37]. He established this idea for protein insulin and for the first time determining the complete amino acid sequence of a protein, the B chain of insulin. B chain is one of the two polypeptide chains that form the insulin. Before this discovery, the predominant notion was that the proteins are random molecules with a kind of center of gravity as well as with appreciable micro-heterogeneity [38]. Therefore, this work of Sanger brought about paradigm shift in the knowledge of scientists in this field.

Although proteins are linear sequence of amino acids, detail and specific mapping of its structure from the sequence is not straightforward, rather it has remained as a widely studied yet to solve critical problem of molecular biology.

2.2.2 Secondary structure

Secondary structure (SS) refers to the local spatial organization of a polypeptide's backbone atoms of a protein [20]. Secondary structures are determined by examining the pattern of hydrogen bonds between side chains and amino acid residues in a protein. It may be deemed as a notion of local sub-structure. As proposed by Pauling and his colleagues, there are mainly two types of secondary

structure- alpha helix (α) or helix (H) and beta (β) strand or beta sheet (E) [21]. These are all regular polypeptide folding patterns. Dominant hydrogen bonding patterns are turn and bridge. Repeated turns give rise to helixes while repeated bridges generate strands [22]. However another type, turn or coil (C), is also considered as a kind of secondary structure. The third type is generally referred to the structure of those residues which are neither helix nor beta sheet. In Figure 3 a protein 3D structure with secondary components is shown. To classify the secondary structure, we investigate the geometric properties of peptides to verify their formations.

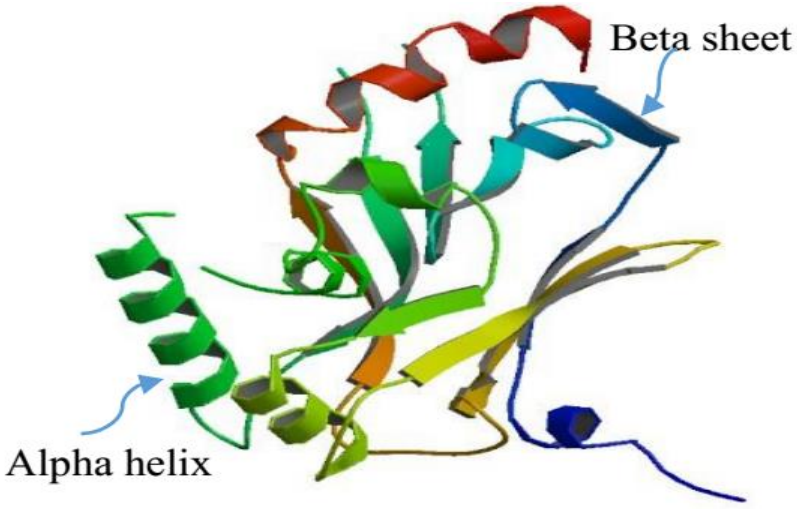

Figure 3: Tertiary structure of protein with secondary structure component [39].

However, when the structure or, the geometric properties are not yet discovered, we rely on prediction from the primary sequences alone. The task of secondary structure prediction basically is to predict to which of these three types of structures (α, β or turn) each residue in a particular sequence belongs. For example, if we consider **L**LATGCLL**K**NK*G*KSEHTFTIKKLGID*V*VVESG…. – a primary sequence of a protein, a secondary structure prediction method may suggest, say, from first **L** to the first **K** are in helix, after that from *G* to first *V* are in beta sheet, and the rest are in turns and so on. However this

secondary structure prediction does not give the complete 3D structure of a protein, which is the famous folding problem. While complete understanding of protein folding is excruciatingly complicated, however, the problem is usually tried through simpler steps and secondary structure prediction is one such very prominent step [40]. Predicting secondary structure is also important because there is a strong coupling between secondary and tertiary structure of proteins [41]. Accurate prediction of SS is urgent, since predicted SS is an essential input feature for other important predictors such as tertiary protein structure predictor, disorder predictor, binding and non-binding predictor, statistical energy function, etc. Successful SSP can also help us to step forward to answer the reasons behind many critical diseases such as Cancer, Cardiovascular diseases, Alzheimer's disease, type two diabetes, Parkinson's disease and many more.

2.2.3 Tertiary structure

Tertiary structure refers to the folding of its secondary structure element by specifying the position of each atom in the protein in a three dimensional structure. Scope of secondary structure is within the spatial arrangement of adjacent amino acid residues. On the other hand, tertiary structure encompasses longer-aspects of amino acid sequence by capturing the interactions of amino acids in the polypeptide sequence that are far apart and belongs to different types of secondary structures. These distant amino acids with respect to their positions in the primary sequence may come closer when the protein folds and interacts within the completely folded structure of a protein. Therefore, we might say that while secondary structure is all about local sub-structure patterns, tertiary structure is defined as the global structural conformation of proteins. Although it is well established that the sequence of amino acids determines the three dimensional structure of proteins, precise mapping of three dimensional structure and amino acid sequence remains a big challenge [1]. Moreover, protein three dimensional structure is not always fixed, rather it may move and flex

within certain geometric constraints. These different three dimensional structures of same protein are known as conformations. This conformational variability, however difficult to capture, is very important for protein functioning [1, 42, 43].

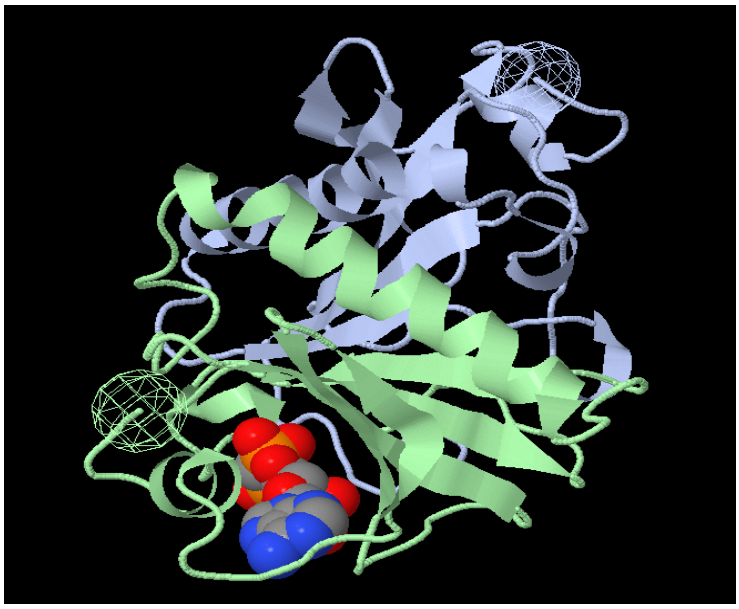

Figure 4: Tertiary structure of protein. This is a cartoon image of a protein (PDB ID: 1AV5) generated by Jmol, an open source Java based viewer for chemical structures.

2.2.4 Quaternary structure

We know that proteins are polypeptide chains. In some instances, two or more polypeptide chains known as subunits may combine and form a structure different from regular tertiary structures, whereas in contrast the tertiary structure consist of a single polypeptide chain. Spatial arrangement of these subunits is known as protein quaternary structure [20]. The subunits may be identical or different. The simplest type of quaternary structure consists two identical subunits. This type of structure is known as homo-dimer [44]. An example structure of homo-dimer is shown in Figure 5. In a quaternary structure, there may exist more than 2 subunits as well. Two or more subunits may bind together by means of hydrogen bonds, disulfide bonds or salt bridges.

Figure 5: A simple example of protein quaternary structure is the structure of Quinone reductase. This is a cartoon image collected from PDB (PDB ID: 1QRD) generated by Jmol, an open source Java based viewer for chemical structures.

Quaternary structures are sometimes crucial for some important functions. Interactions among subunits in quaternary structure play a vital role in biochemical reaction regulation and catalysis [45]. For example, the quaternary structure of quinone reductase contains enzyme that catalyzes the reaction of reducing obligatory NAD(P)H-dependent two-electron from quinones and protects cells against the toxic and neoplastic effects of free radicals and reactive oxygen species that arise from one electron reduction [46]. Such reduction of two-electron helps the process of reductive bioactivation of cancer chemotherapeutic agents such as mitomycin C in tumor cells.

2.3 Functions of Proteins

We already have discussed the reasons for which proteins are involved in a wide array of functionality. Here we discuss major functions of proteins:

- Antibody: Some specialized proteins participate in the defense mechanism of the living body to identify and defend the attack of bacteria, viruses, and other foreign intruders. These proteins travel through the blood stream and helps the white blood cells to destroy antigens.
- Movement: Some proteins help in cell movement and muscle contraction. Example of such proteins are actin and myosin [47]. Myosin acts as a molecular motor which converts the chemical energy to mechanical energy and thus generates force and movement.
- Enzyme: A very important and fundamental task of protein is that it acts as enzyme. Enzyme proteins are catalyst that significantly increase the chemical reactions within a cell. These enzyme proteins are important because they catalyze most of the biological reactions [47]. Lactase and pepsin are the two important examples of enzyme protein. Lactase helps to break down the sugar lactose of milk while pepsin helps in digestion of proteins in food.
- Transportation: Proteins are important agents in various transportations within living organisms. Protein like hemoglobin transports oxygen in the blood [48]. Cytochrome *bc* complexes help in electron transportation as well as proton translocation across the membranes of bacteria, mitochondria, and chloroplasts [49].
- Hormonal functions: Some proteins known as hormones coordinate different function in the body. For example, insulin is responsible for regulating glucose metabolism by controlling the blood-sugar concentration. Somatotropin is well known as growth hormone [50] which stimulates protein production in muscle cells.
- Structural support: Some proteins, usually fibrous give structural support. For example, Keratine is one of the widely known structural protein [51]. It strengthen the covering part of the body such as hair, feather, horn, beak, etc.

- Storage: Some proteins help in storing amino acids. For example, ovalbumin is found in the white part of egg. Amino acids stored by proteins help in the embryonic development of animals or plants.

2.4 Approaches to Secondary Structure Prediction

Protein secondary structure prediction (SSP) approaches may be categorized into two broad areas:

- Experimental approach
- Computational approach

2.4.1 Experimental approach

Two most widely used experimental approaches for protein secondary structure prediction are X-ray crystallography and Nuclear Magnetic Resonance (NMR) [52]. The other experimental methods used are: fiber diffraction, electron microscopy, and so on [53]. Most structural solution at atomic level resolution is solved either by X-ray crystallography or NMR [54].

2.4.1.1 X-ray crystallography:

In 1895, Wolhelm Röntgen discovered X-ray [55], an epoch-making event in the history of science. The first successful deployment of X-ray crystallography in determining protein structure is the credit of two Cambridge scientists, Max Perutz and John Kendrew because of their discovery of the structures of hemoglobin and myoglobin respectively [6]. Since then, X-ray crystallography is widely used for protein structure determination. Over 80% of the three-dimensional macromolecular structure data in the Protein Data Bank (PDB) were obtained by X-ray crystallography [7]. In the Figure 6, we can see the growth of structures in PDB solved by X-ray crystallography. This graph also affirms that the method is very widely used., However, the method requires long and careful steps for crystallization and then to go through very tedious and complex computational to retrieve true structural image from the orthogonal image generated by

X-ray [53]. In the following paragraph, a brief discussion of how X-ray crystallography works in determining macromolecular structure of protein is presented.

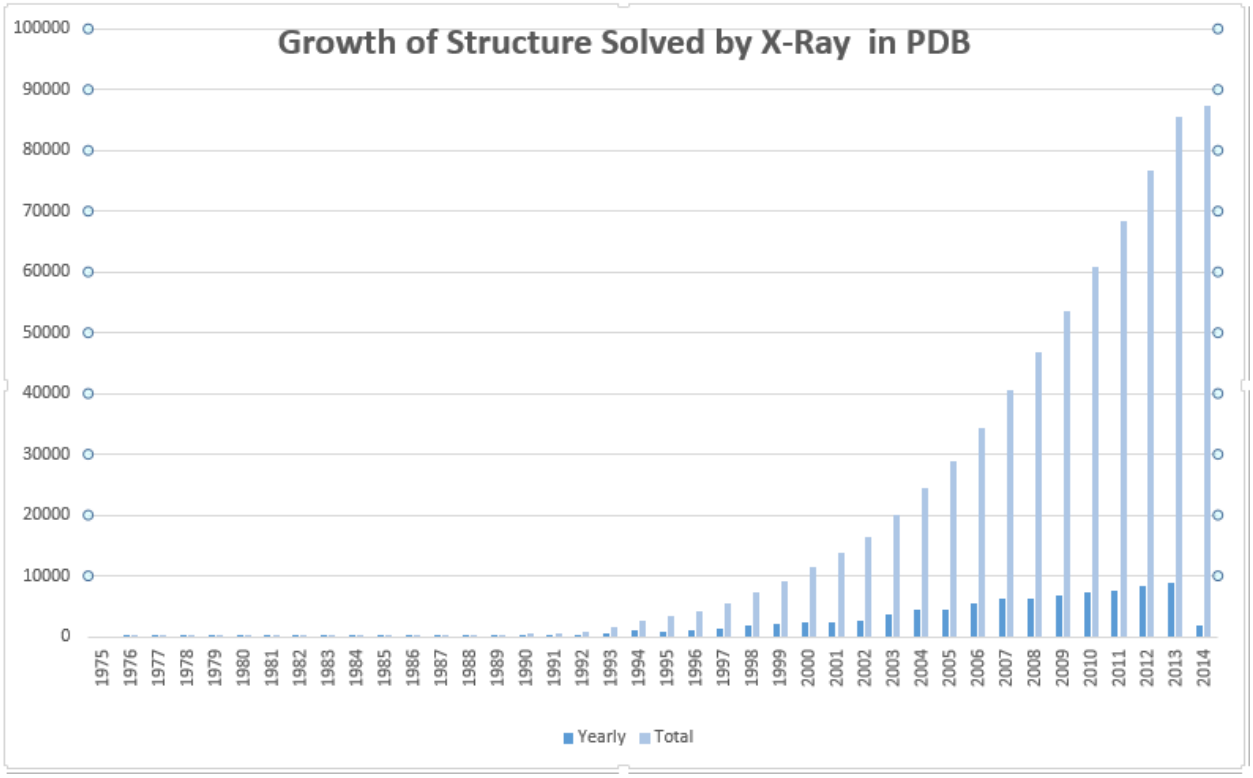

Figure 6: Yearly growth of X-ray crystallographic structure. Source: PDB [36, 37]

X-ray crystallography in protein is basically a form of very high resolution microscopy, which facilitates us to visualize atomic level protein structure. It works on the principle of well-known optical phenomena- interference and diffraction. Superimposed light waves from any source enhanced each other in one direction, while destruct each other in another direction. The enhancement process is known as constructive interference while the later one is called destructive interference. After agitating a surface with light of certain wave-length, we may visualize its structure by analyzing the diffraction or interference pattern of the light waves diffracted from that surface. To simulate the atomic structure, the wave-length of the incident light has to be of the order of magnitude of inter atomic distances of the substance under investigation. Binding

distances between atoms of a protein vary between 1-3 Å(approx.). X-ray may be found with wavelength range of 0.05-100 Å. If X-ray beam of proper wave-length is imposed on a particular protein crystal, which is actually a regularly spaced molecules and atoms, it will create certain diffraction pattern specific to that particular protein [6]. Such diffraction pattern was first described by Bragg in 1913 [56]. However, we have to keep it in mind that diffraction signal from a single protein molecule is very weak. Therefore, to have suitable diffraction pattern, we have to use ordered three- dimensional array of protein molecules, which we call crystal [57]. If the molecules are not properly ordered in the crystal, the diffraction pattern will not yield an adequately high resolution structure with subtle detail. A crystal of similar protein may be considered as a 3D diffraction grating as unit cells of highly similar structural motifs are repeated through the entire crystal in a periodic style. The larger unit cells, the more diffraction pattern may be observed obtaining, thereby, a more discernable signal [53]. Therefore, more specifically, if a crystal of particular protein is exposed to X-rays, a diffraction pattern is found consisting of a series of reflected light rays with varying intensities because of the scattering of X-rays by the electrons of atoms in the crystal. If we know the geometry or symmetry of the crystal well, we may obtain the diffraction spots for every ordered atom in the molecule by rotating the crystal through some defined angle as determined by its symmetry. Each diffraction spot actually represents the diffracted beam which is defined by three well known parameters- amplitude, wavelength and phase. We must know all of these three parameters to correctly obtain the location of each atom in three dimensional space. Amplitude can easily be determined from the intensity of the spot. Wavelength is actually dependent on the selection of used X-ray. Phase is the most critical parameter among these three as it is lost during X-ray measurement. Therefore, it remains as an important challenge to reasonably estimate the phases for all diffracted beams using indirect

methods. Many different methods exist for determining phases in protein crystallography. Most of them typically start with an initial approximate electron-density distribution in crystal which is iteratively improved until a reliable model is attained [53]. This is very important to keep in mind that when the crystal of protein is exposed to X-ray, diffraction occurs simultaneously from all the molecules in the crystal lattice. Intensities of reflections by any single atom is influenced by the reflections of many other atoms in the same crystal. Therefore, partial derivation of structure of any part of the crystal is not possible without modeling the whole [53]. This factor is taken into account as the final three dimensional structure is obtained through a time-average of all the pictures of entire lattice [43].

Although X-ray crystallography has widely been used by scientific community to obtain 3D structure of protein, the method is not completely flawless, owing particularly to the limited resolution and un-precise phase information, among many other reasons. There is no hard and fast methodology for it, rather is subject to experience, individual preference and expectations. Therefore, errors in X-ray crystallography is almost unavoidable [58]. DePristo, Bakker and Blundell found some errors in several structures obtained through X-ray crystallography [42]. They also opined that accuracy of X-ray crystal structures has been widely overstated and that the analyses depending on small changes in atom position may be flawed. In the following discussion, some loop-holes of X-ray crystallography will be discussed.

To determine a precise structure of a protein through X-ray crystallography, first we need a proper crystal to be formed which will produce quality diffraction. In practical situation, having the crystal with desired accuracy/quality is not always an easy task [59]. For new proteins, the right way of making the crystal may not be known. If the crystal obtained for crystallography is found to be not up to the mark, we may not always find the next right course of action. Even we

may never be sure whether a suitable crystal may be obtained or not. Therefore, although, rapid invention of modern computer software and algorithm, development of high quality X-ray sources and synchrotron radiation [60] have eased the challenges of crystallography, finding the right crystal remains as a major bottle-neck for this method. However, still it takes month or even year to find a single structure of protein from this method [10] .

Proteins are dynamic and heterogeneous macro molecules [52, 61, 62]. Dynamic in a sense that structure of protein molecules are actually not stationary rather they evolve among different possible conformation. This concept also refers to the translation and rotation of the entire protein molecule, domain reorientation, conformation exchange, side chain rotation, bond vibration, etc. [63]. Scientists also found that proteins show individual anisotropic motion as well as collective large-scale motion over time [61]. Because of the complex energy issues of protein, they show large population of significantly different conformations distinguished by high energy differences [64-66]. However these dynamism and heterogeneity are largely responsible for different functions of proteins [65, 67]. In X-ray crystallography this molecular dynamics is restrained, whereas it reports larger expected uncertainties of around 0.5 Å which yields less accurate structure compared to that obtained through theoretical calculations [68, 69].

We have seen from the above discussion that for X-ray crystallography, protein molecule have to be crystallized. However, some regions of the protein in crystalline form may have highly different conformation of structure than the structure of that region in solution or native environment [70]. Because in reality proteins are not crystalline, rather they work in a highly concentrated aqueous environment, widely known as native environment of protein. For this reason, the obtained structure from crystal may not represent the native situation structure. This may sometimes be misleading for the analysis of functions of the protein if the crystal is not formed

maintaining all required characteristics such as purity, singularity (not stuck with one or more other proteins), etc. [71].

Since large amount of solution (around 30-70%) is present in protein crystals, these are highly likely to damage if exposed to X-ray. Such event may disorder the molecules within the crystal lattice [43]. This event is known as radiation damage, which occurs mainly because of the primary interactions between the molecules that forms the crystal and the X-ray beam [72]. Such reaction generates heat leading to vibration of the molecules and also provides sufficient energy to break the bonds between atoms in a molecule. This is another limitation of X-ray crystallography. Radiation damage may be reduced if data is collected at liquid nitrogen temperature. This technique of using liquid nitrogen temperature has become common practice.

2.4.1.2 Nuclear magnetic resonance:

Nuclear magnetic resonance (NMR) spectroscopy is the second available method for obtaining three dimensional structure of protein in atomic level resolution [8]. Data that we obtain through NMR is complementary to X-ray crystallography in many aspects. The works of Adelinda *et. al.* found that X-ray crystallography and NMR have different advantages and disadvantages in terms of sample preparation, data collection and analysis. They showed a comparison of 263 unique proteins screened by both NMR spectroscopy and X-ray crystallography in their structural proteomics pipeline. They found only 21 targets (8%) were deemed amenable to structure determination by both methods. However, when applied both methods in their pipeline the amenable target increased to 107, where only 43 were amenable to NMR and 43 were amenable to X-ray crystallographic methods [73]. Therefore, NMR creates opportunities to further delve into the structure and function of a greater varieties of proteins. So far more than 10,000 structures,

about 16% of the total structures in PDB, are solved by NMR [7]. The growth of NMR resolved structure in PDB is shown in the Figure 7 below:

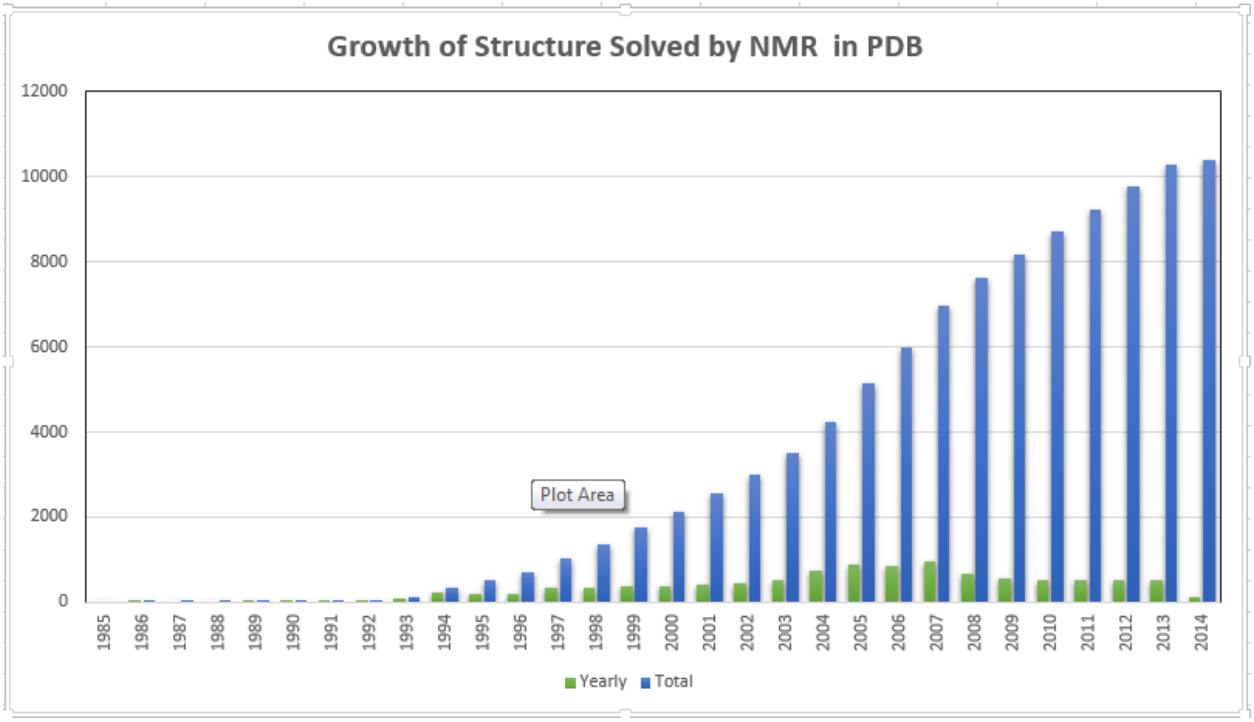

Figure 7: Yearly growth of NMR structure. Source: PDB [36].

In 1957, the first NMR spectrum for a protein (ribonuclease) was reported [74, 75]. Importance of NMR is immense in this field, because its output includes not only structural data but also important information of molecular dynamics, conformational equilibriums as well as intra or intermolecular interactions [76-78]. Instead of producing a direct image of protein, NMR produces huge amount of indirect data, from which we may find the three dimensional structure of macromolecules after complex computation, such as Fourier transformation and analysis [54].

NMR is based on the a quantum mechanical property of nucleus known as spin [54]. Spin refers to a small atomic level magnetic dipole with two different states – up and down. These two

states are separated by a small energy barrier. Jump from one state to another is accompanied by a small electromagnetic radiation or absorption. Two main components of a NMR experiment are:

- a strong field of super conductive magnet capable of producing highly homogeneous strong static magnetic field
- a console which can produce electromagnetic waves in an expected combination

A concentrated solution of molecule of interest is kept in a bore of the super conductive magnet in room temperature. A slight imbalance in the nuclear magnetic moment oriented parallel and anti-parallel creates small polarization of the nuclear spin in the sample. This magnetization can be manipulated to the desired level of the analyzer applying suitable electromagnetic irradiation [54, 79, 80]. This electromagnetic radiation provides the required energy to shift the spin from one phase to another [33]. Every nuclei in NMR spectrum is detected by its characteristic resonance frequency as different nuclei's resonance frequencies widely varies [54]. For example, resonance frequency of a proton (^{1}H) is 4 times higher than that of a carbon (^{13}C) nucleus. Although the resonance frequencies of the nuclei of same atoms are usually within a very narrow range, the frequencies vary at different locations of a molecule for various local interactions among nuclei. NMR signals are observed after disturbing the spin equilibrium with suitable radio frequencies. The system usually returns to equilibrium within 100 miliseconds through *free induction decay* (FID). Meanwhile, the FID and NMR signals are recorded. Afterwards, NMR frequency spectrum is measured through Fourier transformation of these data. Detail of the protein spectra is analyzed based on bond and space correlation. Bond correlations group individual spins into overall spin system to analyze the spectra while space correlations form the basis for geometric information which ultimately determines the final three dimensional structure of macro-molecule within conformational constraint.

A unique feature of NMR spectroscopy is that it can be conducted in highly concentrated solution to obtain atomic resolution structure [8, 33] whereas X-ray crystallography is done on crystal of protein. Therefore NMR environment better mimics the native environment of protein compared to that X-ray crystallography does as in general protein remains in concentrated aqueous solution. Structure obtained through NMR can capture the dynamic nature of protein structure by producing not a single structure but an ensemble of different conformation [81]. Because of this ensemble derivation NMR has been a popular method for the structural studies of disordered protein where a definite single structure is very unlikely [82]. Because of the feasibility of NMR in solution, it may also be conducted in real living cell [63, 83]. We know that protein functions in solution and the concentration of the solution can reach as high as 400 g/l [63]. Therefore, we can say that NMR is conducted in an experimental environment which is more like native environment for protein functionality, because most biological reaction occurs in a concentrated solution environment [84]. However, most NMR experiments are done in a single protein solution with a concentration much lower than that in cell [85, 86] because it is a big challenge to keep protein mono-dispersed in a solution having concentration higher than 0.5mM [87]. Although NMR spectroscopy is uniquely identified for its capacity to determine structure of protein in solution of atomic level resolution [8], NMR can be conducted in solid state as well [88] which is suitable for determining the structures of insoluble macro-molecules. In PDB there are 38 unique protein structure determined through solid state NMR as on 19 May 2014 [89].

From the above discussion, it is clear that NMR has been playing a vital role in the research of obtaining protein 3D structure. However, the method has some significant limitations as well, which are briefly discussed in the following paragraphs.

A pivotal element in NMR spectroscopy is the energy difference between two spin states, which when irradiated with suitable radiation, emits or absorbs certain nuclei specific electromagnetic radiation. However, the emission depends mainly on the net number of parallel or anti-parallel spins, which is usually very small in room temperature to generate a good NMR signal [54]. For example difference of the number of spins oriented parallel or anti-parallel for ^{1}H is only 60 per million in room temperature and in the maximum magnetic field strengths available for NMR. For this reason, NMR is usually considered as an insensitive technique.

Structure from NMR is estimated by analyzing the NMR spectrum generated by individual active atom nuclei. However, this analysis becomes almost impossible with reasonable accuracy when the protein size is very large, especially when molecular weights of protein exceeds 50-60 KDa (kilo-Dalton) [54, 90]. This limitation of size is mainly because of two factors: first, larger molecules exhibits shorter NMR signal relaxation time and slower tumbling rate. Second, increased number of active nuclei in larger molecules increases the local interaction and complexity of NMR spectrum [91]. There have been significant advancement of the NMR technology during past few decades [92]. Because of this progress, the size limit of protein molecule that can be studied with NMR has been reported as high as 100 KDa with the same detail that was found for smaller proteins previously [93].

However, insightful, this multiple structure derivations, known as the ensemble, from NMR has a short coming in homology modelling, where we have to select one single structure [81]. Because in the ensembles all possible structures derived under structural constraints can differ widely. In such case, protein crystallography is a better choice.

Experimental methods such as X-ray crystallography and NMR have so far helped determine the three dimensional structure of a large number of proteins. However, one common

limitations of both the methods is that those are highly time consuming and expensive task [9]. Therefore, development and advancement of computational methods is a significant requirement in this field, given the higher and increasing volume of genome sequence data becoming available and waiting to be analyzed for 3D structure to know its function.

2.4.2 Computational Approaches

To analyze the massive amount of protein sequences generated by genome project highly efficient theoretical methods for predicting SS is of immense importance as experimental methods are highly time consuming, costly and in some cases inefficient [9, 23]. Scientists have been attempting to solve SSP problem with a wide variety of computational models for last five decades, however the accuracy stuck around 80%. Scientist have been attempting to solve SSP problem applying a wide variety of theoretical models or machine learning approaches such as artificial neural network (ANN), hidden Markov model (HMM), support vector machine (SVM), etc. In this chapter we will briefly discuss the theory and applications of these machine learning approaches.

2.4.2.1 Hidden Markov Model:

Concept of Hidden Markov Model (HMM) stemmed from the concept of Markov process. Therefore, for the sake of clarity, after a short discussion of Markov process we will ultimately focus on HMM. A Markov process is a stochastic process that satisfies Markov property and Markov property is defined as the property that in any stochastic process the next state depends only the present state with some conditional probability but not on the state or series of states preceding the present one. We can express this relation as:

$$\mathrm{P}(q_t = s_i) = \mathrm{P}(q_t = s_i \mid q_{t-1} = s_j) \qquad (1)$$

where, s_i and s_j are two consecutive states in a given sequence, i and j may or may not be the same and q_t stands for any state at position $t > 0$.

Since the next state depends only on the present state and not on any past states, Markov property is also called *memorylessness* of a stochastic process. A stochastic process is said to satisfy Markov property, therefore, is qualified to be a Markov process, if the prediction of a discrete state at any point of time or position in the process using only the immediate preceding state information is identical to the prediction of that state using the full information of the process. We may express this relation by extending (1) as:

$$\mathrm{P}(q_t = s_i) = \mathrm{P}(q_t = s_i \mid q_{t-1} = s_j) = \mathrm{P}(q_t = s_i \mid q_{t-1} = s_j, q_{t-2} = s_k, \ldots q_0 = s_x) \quad (2)$$

where, *s* subscripted with *i, j, k… x* indicates any possible state.

This means, that any discrete state in Markov process depends only on the previous state information and is independent of any other observation prior to the previous one [94]. More specifically, this type of Markov process is known as *first order* Markov process. Markov process can be of any order [95], however, to introduce Markov process here we will confine ourselves within first order Markov process only. A Markov process with finite number of states is known as Markov chain. In a Markov chain, transition from one state to another is governed by a transition probability matrix. It is to be noted that, a state may allow self-transition as well, i.e., the next state may be the same as the present one in a Markov process or chain. So far we have not focused on the probability of finding any state at the beginning of the sequence. This phenomena is governed by another set of probabilities. Therefore, a Markov chain is a single stochastic process which can fully be defined with:

- a set of states
- a set of probabilities that indicates the probability of finding any possible state at the beginning of the sequence
- a transition probability matrix that contains the probability of transitioning form one state to another

If we have all of these three set of information mentioned above, we may estimate how probable a given sequence of states is using Bayesian formula given the Markov process. We may also stochastically develop different possible sequence of states using this Markov process.

Let us consider a situation, where we have a sequence of stochastic observations or outcomes. Each observation is generated from one of a set of states according to some probabilities. We do not know the specific state that has generated any particular observation in the sequence. Therefore, the states of the observation sequence are hidden. However, we know the corresponding probabilities of generating all of the observations by all of the states. We may go from one hidden state to another or to itself according to another set of probabilities known to us. Now we have to estimate the sequence of hidden states that might have generated the given stochastic sequence of observations. This sequence of hidden state is usually known as path. We may apply HMM to solve this problem.

An HMM is an extension of Markov process consisting of finite number of hidden states which are capable of self-transitioning as well as transitioning to other states according to some probabilities, and from each state we may observe a visible outcome from a possible set of outcomes according to another set of probabilities. Therefore, unlike a Markov chain which has a single stochastic process of state transition, HMM is "*a doubly embedded stochastic process that is not observable (hidden), but can only be observed through another set of stochastic processes that produce the sequence of observations*" [96]. The set of visible outcomes may be represented by a set of symbols. For example, in the context of our secondary structure prediction, 20 different amino acids are the visible outcomes, where each amino acid is represented by a unique letter, the symbol.

From now on, we will also refer to visible outcome or observation from any state as emission of symbol as a generalized expression. However in secondary structure context emission of symbols and emission of amino acids are equivalent. Therefore, in the context of secondary structure, the emission of amino acids and the emission of symbols or observations may interchangeably be used. Here in the secondary structure prediction context, state means different type of secondary structures: helix, turn or sheet. In HMM, the state, from where the symbol is emitted, remains invisible or hidden [97] similar to a given protein sequence of unknown structure where we do not know the secondary structure to which every amino acid belongs. The secondary structure information of every amino acid in a protein sequence remains unknown or hidden until and unless the structure is revealed through experiments such as X-ray crystallography, NMR or through computational approaches. If we model protein sequences with necessary parameters required to define a HMM, the secondary structures will represent the hidden states, whereas the set of letters that represents corresponding set of amino acids will represent the visible set of symbols. Every state will have a set of probabilities to emit any of the 20 different symbols representing 20 different amino acids or residues. Transition between connected states is governed by another set of probabilities, known as transition probabilities. For example, probability of finding the next symbol in helix after a sheet structure in a protein sequence is the transition probability of sheet state to go to a helix state. Using the HMM we can predict the path, the sequence of states, for a given observation sequence of symbols when all the parameters necessary to fully describe a HMM is available.

An HMM can fully be described using the following factors [96]:

- A set of hidden states, $S = \{s_1, s_2, s_3 \ldots . s_n\}$, where n = number of hidden states in the HMM

- A transition probability matrix *A*, for transitioning from one state to another. Here any element of *A*, $a_{ij} = P(q_t = s_j | q_{t-1} = s_i)$, $1 \leq i \leq n$, $1 \leq j \leq n$ and q_t represents any state at position *t* of a sequence.
- A set of emitted symbols, $O = \{o_1, o_2, o_3 \ldots\ldots o_m\}$ where, *m* = number of symbols required to represent all the possible visible outcomes.
- An emission probability matrix *E*, that represents the probability of emitting any possible symbol from any state. Here any element of *E*, $e_{ij} = P(o_j|s_i)$, $1 \leq i \leq n$ and $1 \leq j \leq m$, $o_j \epsilon O$, *and* $s_i \epsilon S$
- A set of probabilities that represents the probability of finding any state at the beginning, $B = \{b_1, b_2, b_3 \ldots\ldots\ldots b_n\}$, where any element $b_i = P(q_{t=0} = s_i)$, $1 \leq i \leq n$

In Figure 8, an example of a simple HMM with 3 states and *m* symbols is shown. Every arrow that shows state-transition is associated with a probability value in the transition probability matrix *A* and every arrow that shows emission of symbol is associated with a probability value in the emission probability matrix *E*. Every Arrow from start state is associated with a probability value in the beginning probability matrix *B*. Here we see that every state is self-connected as well as both way connected to all other states and every state may emit all possible symbols. In reality it may or may not be the case. If it is not possible to transit from one state to another, probability of such transition will be 0. If it is not possible for any state to emit any of the symbols, associated emission probability will be 0. In other words, any impossible transition or emission has an associated probability of 0 in corresponding matrix.

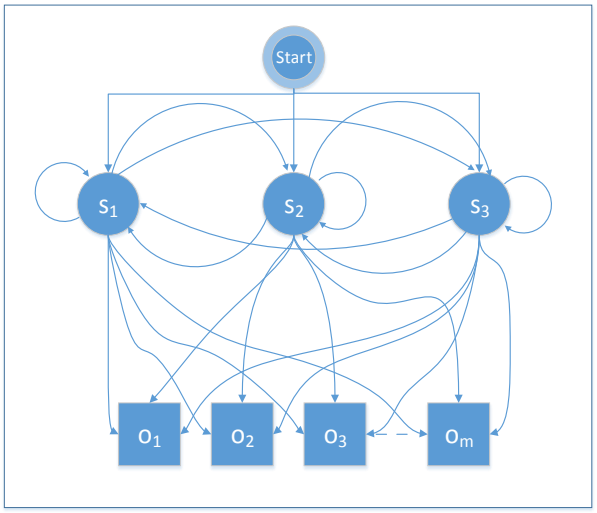

Figure 8: A hidden Markov model with 3 hidden states– s_1, s_2, and s_3 and m number of observations denoted by o_i that may be emitted from any of these states. Here, $1 \leq i \leq m$. We also see a start state here. This start state is actually a pseudo state which emits no symbol, rather just indicates the start of the sequence. Here the arrows from one state to another or to itself represent the transitions from one state to another or self-transition, whereas the arrows from states to symbols represent the emission of symbols from corresponding states.

At any given position $t > 0$ of a sequence, the probability of finding any symbol o_m ($m \geq 1$) from any state $q_t = s_j$ depends on the probability of transitioning from state $q_{t-1} = s_i$ to $q_t = s_j$ and the probability of emitting the symbol o_m from the state s_j. If $t = 0$, which indicates the beginning of the sequence, the probability of finding any symbol o_m from any state s_j depends on the probability of finding s_j at the beginning and the probability of emitting o_m from the state s_j. Therefore, if we have all the five information - *S*, *A*, *O*, *E* and *B*, we may find the sequence of hidden states for any related query sequence of symbols using Bayesian formula. Finding the hidden path is known as decoding [98]. For any query sequence, only *S* and *O* sets are known. *A*, *E* and *B* are called HMM parameters. We can estimate these parameters by training our HMM with a large number of training sequences for which we know some relevant attributes and the hidden paths. Among different available criteria for training HMM and estimating HMM parameters, mostly applied one

is the *Maximum Likelihood Estimation* (MLE) [99]. It maximizes the probability of training sample with respect to the model. Once the HMM is trained, i.e., the parameters are estimated, we can stochastically estimate the unknown path of states for any query sequence of nature similar to the training sequences. However there could be as high as n^l different paths for a sequence of length l with n number of possible hidden states, given all states are connected to each other. Therefore, finding the most likely path is important. Recursive Viterbi algorithm is the mostly used one to find the path with maximum probability [98]. This algorithm offers an effective means to find the most likely state sequences of a finite state discrete time Markov process in terms of maximum posteriori probability.

HMM may be a good choice for sequence analysis because of two reasons mainly. First, the model is based on mathematical structure, therefore theoretically sound for a wide range of application. Second, it works very well in several practical applications, such as speech recognition, temperature measurement, biological sequence analysis and similar application [96]. The protein secondary structure prediction process is a biological sequence analysis problem, and the solution is to assign a right label of structure on every residue or amino acid of the sequence [97]. The labels are usually helix, sheet or turn, to indicate the structure of each amino acid. We can easily obtain the transition and emission probability matrixes from a large number of training sequences of known structure. We may also obtain the initial state probabilities for every state. Therefore, we may estimate the unknown sequence of hidden states (the secondary structures) for a query sequence of protein utilizing HMM.

Asai, Hayamizu and Handa implemented HMM based protein secondary structure prediction model for the first time [100, 101]. They trained only four HMMs for helix, sheet, turn and others. Each HMM is capable of predicting one of these four types of structure only. Once a

test sequence is passed, each HMM gives a path of structure. The HMM that gives the highest probability is accepted as the predicted structure. They used only 120 sequence from Brookhaven PDB for training and testing purpose [100] and their accuracy was not very good. They achieved a Q3 score of 54.7%.

Further, using HMM, noticeable improvement of secondary structure prediction accuracy was done by Bystroff, Thorsson and Baker [102]. Their secondary structure prediction accuracy (Q3 score) was 74.3% using homologous sequence information. They proposed a novel HMM, HMMSTR based on I-sites library of sequence-structure motifs. I-sites are short sequence-structure motifs that show strong correlation with local three dimensional structural elements of proteins. They obtained the I-sites library through exhaustive clustering of sequence segments of a non-redundant database of known structure. Their model applied highly branched topology discovered from the clustering process and captured recurrent local features of protein sequences and structures that are not confined within a particular protein family. In their HMM every I-sites motif was presented as a chain of Markov states each of which contained information about the sequence and structure attributes of a single position in the motif. Merging of these I-site motifs based on sequence and structural similarity created a network of states, in other words the it formed Markov process with hidden states, i.e., HMM. Each state is associated with four probability distributions such as probability of observing a particular amino acid, probability of being in a particular secondary structure, backbone angle region and structural context descriptor. Structural context descriptor describes the context of the residue; for example, it distinguishes beta strand in the middle of a sheet from one that at the end of a sheet. They developed three different models by clustering the I-sites motifs and using observed adjacencies in the database. First model was trained with sequence and secondary structure data. Second model was trained with sequence and

structural context data. Third model was produced based on hierarchical pairwise alignments and trained with sequence and backbone angel data. Collective name of these three models is HMMSTR. They used some heuristic criteria to delete or add hidden states and came up with a fairly complex model where protein 3D structure was modeled through the succession of I-site motifs [103].

Martin, Gibrat and Rodolphe [103] introduced a new type of HMM without prior knowledge. They chose the model from a collections of models based on the Q3 achieved in prediction, the *Bayesian Information Criterion* (BIC) value of the model and the statistical distance between models. Their model for secondary structure prediction referred to as *Optimal Secondary Structure prediction Hidden Markov Model* (OSS-HMM). Their final model has 36 hidden states, distributed as: 15 of them model α-helices, 12 of them model coil and 9 of them model β-strands. Organization of protein structures into secondary structure segments was reflected by the connection between hidden states and emission probabilities of their model. They used two main strategies for developing models: first, start building models from smallest size and gradually increase the size, i.e., the number of hidden states; second, start from a large model and gradually reduce size based on Q3 achieved in prediction, the *Bayesian Information Criterion* (BIC) value of the model and the statistical distance between models. In the first strategy they applied genetic algorithm (GA) for DNA sequence analysis [104, 105]. They applied four types of mutation such as addition of one hidden state, deletion of one hidden state, addition of one transition and deletion of one transition. They also applied cross-over, which involves exchanging several states between two HMMs. In order to automatically select an HMM topology, they applied a systematic approach in which when a new state is introduced all transitions between hidden states were initially allowed. Then the system was allowed to evolve. A big problem of applying GA to HMM topology is the

over fitting of the model towards learning data [104, 105]. To monitor and avoid this loop-hole, they used an independent set of structures which is never used in the cross-validation procedure. The model's Q3 score was 68.8% for single sequence and 75.5% for multiple sequence alignment. With a view to decreasing the complexity of manual generation of HMM.

Another approach of HMM capitalizing on GA was proposed by Won *et. el*. [101] to predict protein secondary structures. In their GA procedure, they developed models of HMMs consist of biologically meaningful building blocks. That is why they named their model as Block-HMM. Each block was labeled corresponding to one of the three secondary structures. Mutation and crossover were applied to these building blocks. They crossed over blocks not arbitrary number of states, which ultimately translated into exchange of different number of states. Mutation was an intra-block phenomenon. They applied another form of mutation called type-mutation which changes the secondary structure label of mutated block and consequently randomly generates new transition probabilities for that block. Baum-Welch algorithm was used after every step of GA to update the model parameters. Baum-Welch is a standard algorithm for HMM parameter estimation. It is an iterative or recursive algorithm that updates a given model closer to the optimal one by increasing a proxy of the log-likelihood after each iteration. However, the algorithm does not guarantee finding the optimal model. The finally accepted HMM captures several structural and sequence related properties of protein. It also calculates the probabilities associated with the prediction. Prediction was done by deducing the values of the hidden states a particular amino acid sequence belongs to and examining the secondary labels of the blocks that states are in. In order to enhance the accuracy of prediction further they used a 3-layer perceptron consisting of 3 input nodes, 3 hidden nodes and 3 output nodes. A very good aspect of their model is that it is capable of randomly generating sequences which matches natural situation. They

generated 1662 (equal to the number of training sequences) random sequences from the evolved HMM. In the generated sequences the overall secondary structure contents were 35.5% helices, 23.5% β-strands, and 44.5% coils whereas the training sequences had 35.3% helices, 22.8% β-strands, and 41.9% coils. They compared performance of their best HMM topology trained on all the 1662 training sequences with other leading predictors such as PSIPRED [106] under both single and multiple sequence using data common to the training of both models as well as data uncommon to both models. For single sequence prediction their accuracy was 68.6% and 69% with uncommon and common data respectively while the figures for PSIPRED were 67.3% and 67.6% respectively. On the other hand, for multiple sequence, their accuracy was 74.5% and 75% with uncommon and common data respectively while the figures for PSIPRED were 78.9% and 79.5% respectively. Therefore, we may loosely conclude that their prediction outperformed PSIPRED for single sequence but underperformed PSIPRED for multiple sequence, however no statistical significance of the differences in the performances were reported.

2.4.2.2 Artificial Neural Network:

Another well adopted approach of solving critical problem is *the divide and conquer rule*. This approach requires that we decompose a complex system into relatively simpler small elements, solve them, and then integrate the small solutions effectively to deduce the ultimate solution of the main problem. Networks are widely accepted tools to do so. However, networks may be of great varieties, all of them share some common attributes: a set of nodes and links or connections between those nodes. These nodes may be treated as small parts of the complex problem and the connections define how the solution of these small problems should be integrated. Nodes may be deemed as computational units which take certain input, process that input and yield some output. Artificial neural network (ANN) or simply neural network (NN) is a modern computational model

that stemmed from the scientific endeavor to mimic the brain which is a network of millions of unit cells, known as neuron. In brain, neurons receive signals from the synapses located on the dendrites or membrane of the neuron. When the signals are stronger than certain threshold, the receptor neuron is activated which then generates a signal through axon. This signal go to another neuron. This way the brain controls the functions of the body. The concept of real neuron is emulated in ANN. ANN can be consisted of layers, where each layer has single or multiple nodes and the nodes are connected through some weighed path. Here we may envision, nodes as neurons, the inputs to these nodes as synapses and weights of the connections as strength of signal. In each node the input information is processed by an appropriate mathematical function, known as activation function, and an output is generated, which may be the input for the other nodes when there are multiple layers of nodes in the NN.

In Figure 9 a single node of a NN is shown. Here we see that a single node is fed with multiple inputs and their associated weights. The activation function f_{sig} will process these inputs and yield an output. This output may be fed to another node with certain weights.

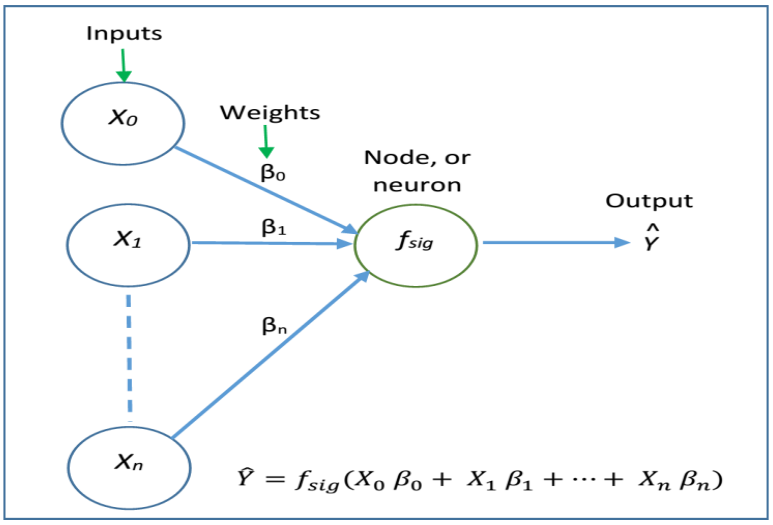

Figure 9: A single neuron or node of an ANN. f_{sig} is the activation function which determines the output. Here x_i are the features used, where n = number of feature and i = 1, 2, 3, …n. X_0 is known as bias term.

An ANN is usually a combination multiple nodes like this one shown in Figure 9. Numerous variants of NNs have been developed since the development of first NN by McCulloch and Pitts [107]. The differences between different variants of NNs may be in the activation functions used, topology of the network, algorithm employed for training, etc. In depth discussion of different types of NNs may be found in Haykin [108]. For SSP the most popular and widely used model is the feed-forward NN [ref?]. Feed-forward NN is made up of layers such as input layer, output layer and zero or more hidden layer(s) in between. Each layer consists of single or multiple nodes as shown in Figure 9. A multi-layer feed-forward NN is shown in Figure 10.

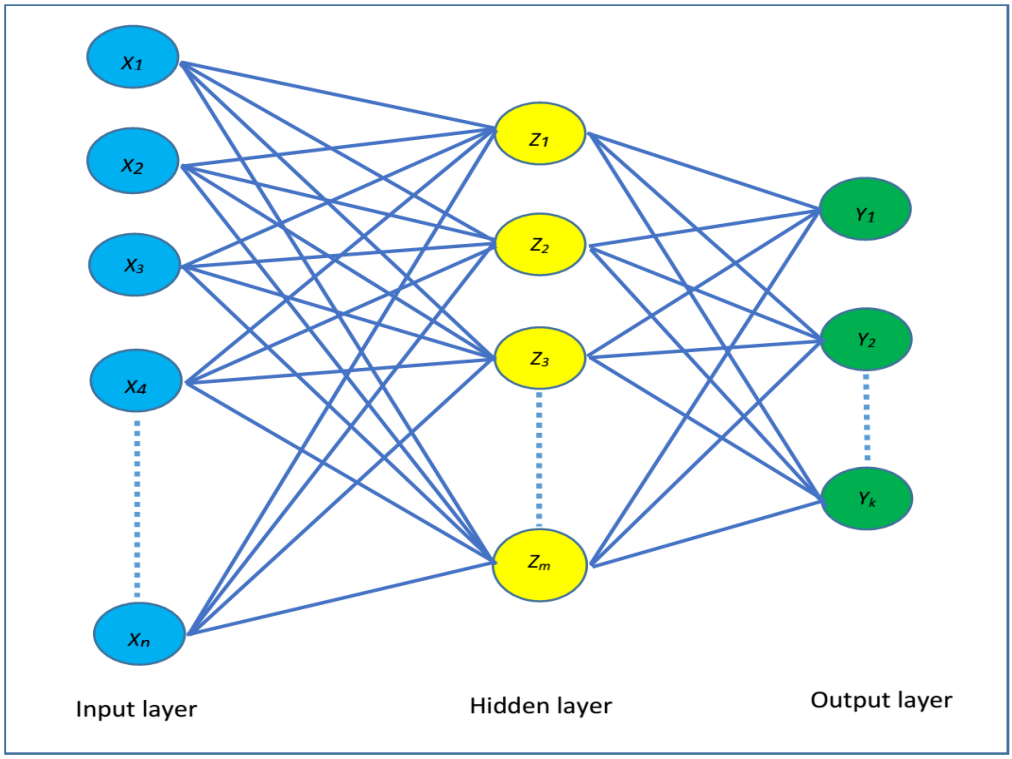

Figure 10: Schematic diagram of a single hidden layer feed-forward neural network with k different output. The units in the middle of the networks are known as hidden nodes. Each node has a derived feature Zm computed form the input of the preceding layer nodes. Here maximum value of m is the number of hidden nodes in a particular hidden layer.

ANN for SSP was first employed in 1988 by Qian and Sejnowski [109] with a view to leveraging on the information form the database of known protein sequences to predict SS. They developed a three layer (input, hidden and output) feed-forward multilayer perceptron ANN for SSP. Their network had 40 hidden nodes and three nodes in the output layer. They fed their ANN information with a window of size 13 residue where the target residue for predicting SS is the one in the center of the window. They chose a representative dataset of 106 sequences with limited identity. They also took special care while choosing sequences to ensure balanced combination of helix, sheet and coil in their data set. Their overall accuracy on a test set non-homologous to the training set was 64.3%.

Rost and Sander established new standard of SSP method introducing PHD method in 1993 [110]. They applied a set of feed forward neural networks trained by back-propagation algorithm [111] with non-redundant data set of 130 protein chains. Most important aspect of their method was that they used multiple sequence alignment (MSA) as evolutionary information instead of single sequences. They employed three level neural networks. First level predicts structure from sequence. In this level helices may be found with length less than 3 residue, which is too short as helices should be at least 3 residue long [22]. Second level refines the first level prediction further, for example, by converting the predicted helices that are too short into loops or by extending the helices with more adjacent residues to make the length 3. They introduced a reliability index (RI), which may have a normalized value between 0 - 9. If any residue within helices with length less than three has $RI \geq 4$, they added additional residue(s) to make the length 3, otherwise converted the helices to loop. Finally the third level, also named as jury level, averages the output from previous levels and finally decides the SS. Accuracy of PHD method was reported as 70.8%. If

larger data set is used along with PSI-BLAST [112], accuracy of PHD method may increase to 75% [113].

Another notably successful NN based SSP method is PSIPRED [106]. Instead of applying MSA, they used intermediate PSI-BLAST profiles as a direct input to their SSP model. This multi-stage prediction method consists of three stages: generating sequence profile, predicting initial SS, and finally filtering the predicted structure. PSIPRED achieved an overall accuracy between 76.5 to 78.3%. PSIPRED and PHD shared similar network topology. The improvement in PSIPRED over PHD may be attributed to the better alignment fed to the NN because of the filtering strategy applied by PSIPRED to exclude unrelated proteins and also in part to the increase in the size of database [114]. This is important to note that, although PSI-BALST is very sensitive to biases in the sequence data banks, because of its iterative nature and it may erroneously include repetitive sequences with low complexity that have biologically insignificant similarity into the intermediate profiles resulting in completely random sequences being matched with high confidence [106].

An important high accuracy recent work on SSP is SPINE-X by Faraggi *et al.* [23]. They employed a multi-step NN algorithm by combining SSP with prediction of real value residue solvent accessibility (RSA) and backbone torsion angles in an iterative manner. Their process started with generation of PSSM by running PSI-BLAST. They also collected seven physical parameters (PP) of the amino acids residues which includes hydrophobicity, polarizability, volume, iso-electric points, and so on for each residue. There are total 6 steps in their prediction process. In the first, fourth and last steps they predicted secondary structure, in the second step they predicted RSA and in the third and fifth steps, they predicted backbone torsion angles. They attempted to boost up the accuracy of any subsequent prediction step utilizing previous steps' predicted information as input features. They tested their accuracy on multiple set of data sets and

their overall accuracy ranges from 81.3 to 82% when DSSP assignment is used. However they also reported an accuracy of 83.8% if modified version of consensus based assignment method, SKSP [115], is used. In their paper, they also reported a detail comparative analysis of performance with respect to another high accuracy SSP method, PSIPRED [106]. They claimed that SPINE-X consistently makes 6% more accurate prediction in helical residues without over prediction while PSIPRED makes 3-5% more accurate prediction in coil residues, however PSIPRED over predicts coils by 7%. In SPINE-X paper, it is stated that the superior prediction result of their model may be attributed to the better prediction of real value torsion angles and multiple step training and prediction of SS.

Although NN based predictors reported the best accuracy for SSP problem, the maximum overall accuracy achieved is still around 80%.On the other hand theoretical limit of SSP is 88% [116]. Another important issue is that the accuracies of predicting beta structure of the best performing methods such as SPINE-X or PSIPRED are around 75%. Therefore, we say that there is scope for further improvement of the overall secondary structure prediction accuracy by improving only beta structure prediction accuracy.

2.4.2.3 Support Vector Machine:

Support Vector Machine (SVM) is a discriminative classifier that analyzes data and separates them into different classes by generating a separating hyper plane. It is a supervised machine learning approach first invented by Vapnik [117] in at Bell AT&T laboratories. To describe SVM, we may start with a simple binary classification problem. Let's say we have an input space $X \subseteq \mathbb{R}^n$, where $n \in \mathbb{N}$, an ouput space $Y = \{+1, -1\}$, and a training set T where $T = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\ldots (x_n, y_n)\} \in (X \times Y)$. Now the job of SVM is to derive a function that maps each element of *X* in to *Y*. In more formal language of classification problem,

we may state this problem as to find a decision rule that classifies each $x \in X$ into any one of the classes +1 or -1. Now we may build up a linear binary classification function f:$X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ in such a way that if $f(x) \geq 0$, x belongs to class +1 or else, *x* belongs to class -1. In linear situation we may write:

$$f(x) = (w.x) + b \tag{3}$$

where, $w \in \mathbb{R}^n$.

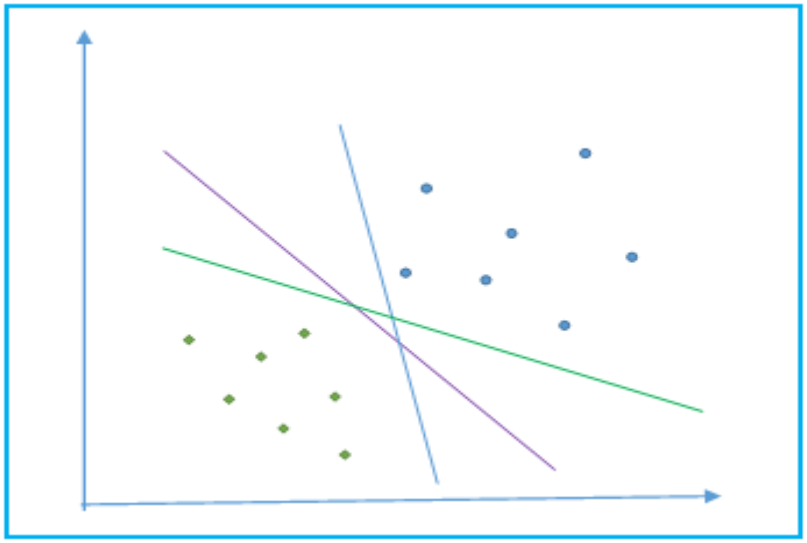

Figure 11: A simplified two class classification problem is shown here. The circular and diamond shaped data points belong to two different classes. The classes may be separated by many different decision boundaries as shown by the solid lines.

This function is known as discriminant function as it discriminates the class to which any $x \in X$ belongs. From geometric point of view of this problem, f represents all possible hyper planes that are capable of correctly classifying the input data. Modeling of this hyper plane depends on the value of *w* and *b* parameters, which are learned by the SVM algorithm from the training data set. A suitable separation is attained by the hyper plane that has the largest distance to the nearest training data points of any class. In general the higher the margin the lower the generalization error of the classifier. These nearest data points are generally known as *Support Vectors* and are very

important as they are the only ones that determines the final solution of the problem. SVM algorithm defines the optimal separation hyper-plane as the one that maximize the area between these *Support Vectors* of the two classes.

We may envision the problem through the simplified picture of Figure 12. Here we see two different types of points, one is circular and the one is diamond shape representing two different classes. Now the data points may be separated into two different classes through many different linear hyper plane indicated by solid lines. However SVM chooses the hyper plane which maximizes the boundary between these two classes as we see in Figure 12. This way SVM ensures the maximum distance between two classes.

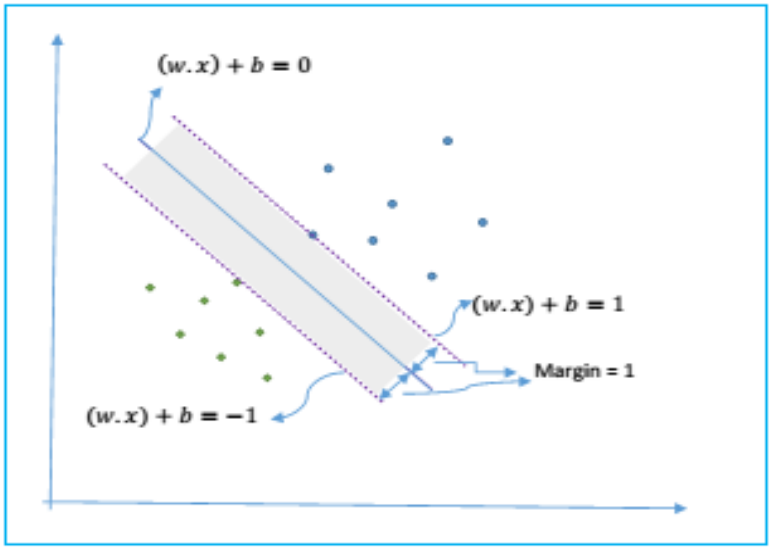

Figure 12: Support vector classifier. The solid line represents the decision boundary while the dashed lines are the boundaries of maximal margin area, shown as shaded area. Data points on the dashed lines are the support vectors. The 1 or -1 values on the right hand side of the equations of the margin boundaries represent the scaled distance from the decision line of the nearest points that belong to +1 and -1 class respectively.

However, not in every cases, all data points of different classes may be separated. In that case SVM still tends to maximize the margin allowing some points to be misclassified and assigning a

penalty for that. In other words, if there exists no hyper plane that is capable to classify the given all data points into two classes, a soft margin binary classifier may be used to create a hyper plane that partitions the data with maximum accuracy. When data is linear, a separating hyper plane may be used to divide the data. However it is more often the case than not that the data is far away from linearity and the datasets are inseparable. To solve such problem, we can make the procedure more flexible by enlarging the feature space applying basis expansions such as polynomials or splines [118]. In general linear boundaries in the enlarged space yield better training-class separation, and translate to nonlinear boundaries in the original space. The support vector machine classifier, empowered by kernel function, is a more flourished form of this technique where the dimension of the enlarged space is allowed to expand greatly. Once the data is expanded in higher dimensional space properly, often the classification by generating a maximal margin or optimal linear separating hyper plane becomes a trivial problem [119]. There are different kinds of kernel functions used in SVM, such as linear kernel, polynomial kernel and radial basis function (RBF) kernel or Gaussian kernel. The benefit of using RBF kernel is that it can automatically expand the feature space into as high as infinite dimension.

Although SVM is regarded as one of the most robust classifier in machine learning by many [120], it has not been widely used in SSP problem. Earliest work in this regard was done by Hua and Sun [19]. They used an RBF kernel based SVM classifier to predict three different secondary structures of protein. They developed six different binary classifier for this purpose, (H /~ H, E/~ E, C/~ C, H/~ E, E/~ C and C/ ~ H). They developed another classifier to combine the output of these six classifiers using voting mechanism to come up with the final prediction. They used only evolutionary information or PSSM and amino acid residue information as feature (total of 21 feature) in their prediction. They also used sliding window of 21 size. Their overall accuracy

was 73.5%. Later Guo and his colleagues bring about little changes to the predictor of Hua and Sun by using another layer of SVM that further rectifies the prediction using minimum sequence length criteria for each type of SS [121]. For example, a helix contains at least 4 consecutive residue pattern, and a sheet contains at least 3 consecutive residue pattern. Doing this further rectification they achieved an overall accuracy of 75.2%.

Wang *et al.* [4] proposed another SVM based SS predictor using RBF kernel and a sliding window of 15. They used propensity of any amino acid to be in H, E or C structure at a particular position in a protein and the hydrophobicity of the amino acids as features. They reported an overall accuracy of 78.44%.

Another notable SVM based SSP effort was from Ward and his colleagues [122]. They also trained binary SVMs to discriminate between two structural classes. The binary classifiers were then combined in several ways to predict multi-class secondary structure. They reported average three-state prediction accuracy of 77.07%. They also used PSSM or evolutionary information only as their feature. Their window size was 15 and they used 2 degree polynomial kernel instead of RBF kernel for feature space expansion. They also reported that their method was not very efficient in predicting beta strands.

3. Methodology

3.1 Prediction Method

We have discussed in chapter 2 that many different machine learning algorithms such as ANN, SVM, HMM, etc. have so far been applied to tackle the SSP problem. We have used binary SVMs in our investigation coupled with genetic algorithm. Details of the method is discussed in the following sections.

3.1.1 Classification Algorithm

We have trained three binary SVMs, E versus non-E (i.e., E/~E), C/~C and H/~H.. A description of SVM is given in literature review part. These three SVMs provide us the probability of each residue belongs to beta, coil and helix structure respectively. Combining these three predictors results into final three class prediction is a crucial challenge. We combine optimally these three binary predictors using a genetic algorithm (GA) to form final three class prediction. GA finds separate real value paramter for each class as an additive factor for each class probability given by three binary SVMs. For example if the probabilities that a particular residue belongs to E, C or H class are p_1, p_2 and p_3 respectively, GA founds three real values v_2, v_2 and v_3 and the revised class probabilities become (p_1+v_1), (p_2+v_2) and (p_3+v_3) for E, C or H class respectively. Finally the class, for which this revised probability is highest, is accpeted as the predicted class for that particular residue. We refer to our combined SVM predictor as cSVM.

3.1.2 Meta Predictor

In addition to the cSVM predictor, we also developed a meta predictor. Our meta predictor combines the 3 class prediction of our cSVM with the prediction from SPINE X. We combined these two predictors in the manner shwon in Figure 13.

```
1. Generate secondary structure probabilities and classes by cSVM.
2. IF cSVM's output class is E THEN
        3a.ACCEPT E as the output of MetaSSPred
   ELSE
        3b. ACCEPT SPINE X's output as the output of MetaSSpred
   ENDIF.
```

Figure 13: Algorithm for combining cSVM and SPINE X.

We refer to this meta secondary structure predictor as MetaSSPred which is our final predictor.

3.1.3 Genetetic Algorithm for Combining Binary SVMs

GA is an evolutionary learning based heuristic algorithm which maintains a population of individuals for each iteration. Each individual in the population, comonly known as chromosome, represents a potential solution to the problem to be solved. Each chromosome in the population is assessed on the basis of some fitness function. In our case, the fitness function was overall accuracy of three class prediction (Q_3). We call the value of the fitness function for a particular chromosome as the the fitness of that chromosome. Each chromosome in a population is assigned a probability to be survived in the next generation based on this fitness. The higher the fitness, the more likely the chromosome will survive in the next generation or participate in mutation or crossover. Crossover and mutation mechanism will be discussed soon in this section. We call this fitness based probability as survival probability. In the first iteration the population was generated

randomly where each chromosome is a 36 bit long binary number, which essentially contains the three additive factors, each 12 bit long. In the next iteration, a new generation of population is created from the previous generation after some evolutionary transformation. The evolutionary transformation contains three mehanisms – elite preservation, crossover and mutation. Certain percentage of chromosomes are preserved and included in the next generation based on higher fitness. These preserved chromosomes are elites. This way top solutions from previous generation are always preserved. In crossover technique, pair of chromosoms are selected based on survival probabilities of the chromosomes. Then a site for crossover is randomly selected. Finally, each crossover produces 2 new chromosomes. The crossover mechanism is demonstrated in Figure 14.

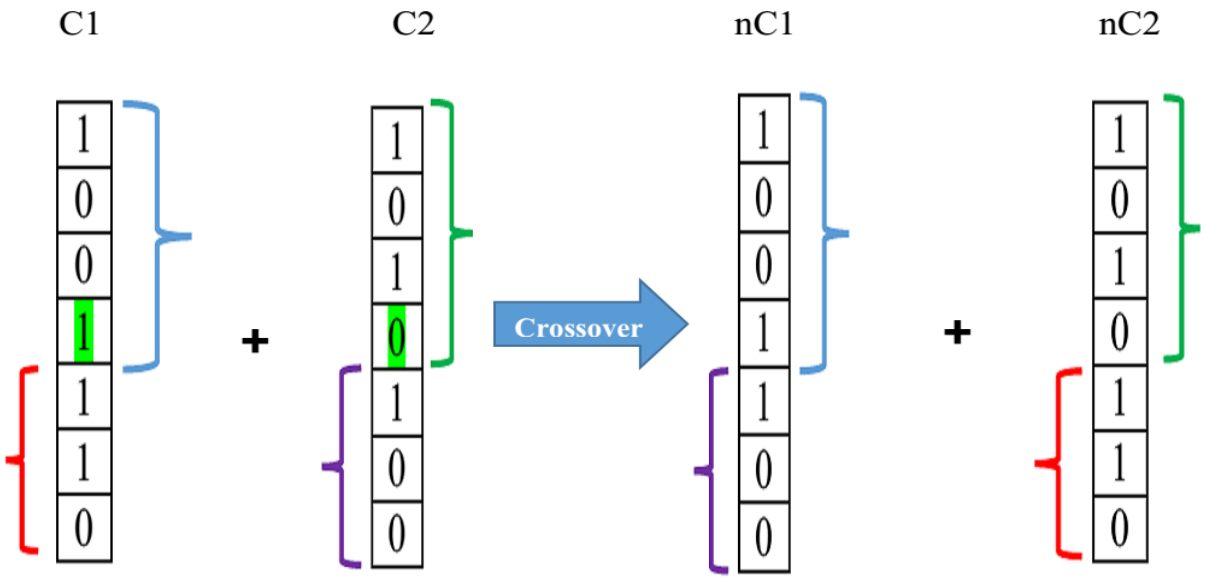

Figure 14: This figure demonstrate a cross over operation. C1, and C2 are survival probability based selections as crossover candidates. nC1, and nC2 are two new chromosomes created after crossover. The green highlighted bit in C1 and C2 indicate the randomly selected crossover site. Similar color curly braces show the origin of the part in new chromosome.

Mutation selects a mutation candidate based on survival probability from the previous generation and then randomly selects a mutation site and flips the bit value. This way a new generation of chromosomes is created and again the fitness for each chromosome is calculated until some targeted fitness is achieved or the predefined number of iteration ends or the improvement of fitness becomes stagnant. The pseudo code for genetic algorithm is shown in Figure 15.

1. Randomly form the initial population
2. Compute the fitness to evaluate each chromosome
3. Select pairs to mate from best-ranked individuals and replenish the new generation by
    a. Preserving elites
    b. Applying crossover operator
    c. Applying mutation operator
4. Check for termination criteria, else go to step #2

Figure 15: Pseudo code for GA.

3.2 Data Collection

A very important component of protemoic research, specially of protein structure prediction methods, is the data set used. Quality of the research highly depends on the data purity, resolution, degree of similarity between test and training data sets, etc. Here steps towards obtaining the training and test data sets are described.

3.2.1 Training Data Set Preparation

We collected protein sequences from PDB [123] and culling server [124] with the following specifications:

- sequence length ≥ 50
- x-ray resolution ≥2.5 Å
- sequence identity ≤ 30%
- method- refinement R factor: 0 to 0.25

We got 6521 protein sequences after this search in PDB. Usually direct culling from PDB doesn't satisfy all such criterion mentioned above. Therefore, we also ran BLASTclust [125] to ensure that the identity cut-off criterion is met. BLASTClust is a software to cluster protein or nucleotide sequences. The program starts with pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already present in the cluster. Before running BLASTclust, sequences with same ID however different in case were manually removed to avoid error. User may specify the degree of seqeunce identity cut-off to develop different clusters. We have used 25% identity cut-off. One seqeunce from each cluster was finally kept aside for further processing. After all these filtering, we had 2150 sequences left. Our SS assignment method was DSSP.

Table 1*:* A summary of the secondary structure composition of T552 test dataset

Secondary Structure	Residue Count	Percentage
Beta	27,229	18.2%
Coil	76,959	51.6%
Helix	44,905	30.2%
Total	149,093	100.0%

Therefore, we then ran DSSP [22] to collect secondary structure assignment for each residue using the collected dataset. We discarded the seqeunces for which DSSP failed to fully assign secondary structure. We also discarded seqeunces where we found mismatch between the length of sequence as given by DSSP assignment and the that of PDB fasta format data. We refined this data set further to discard the protein sequences that contain un-known amino acids labelled as "X" and the sequences which contain amino acids of unknown coordinates. After all these refinements, we obtained 554 sequences (T554). Finally, we discarded 2 more sequences for which one of our used features, torsion angles fluctuation cannot be predicted. Therefore, our final training data set consists of 552 sequences with no more than 25% identity among themselves. From now on we will call these dataset as T552. A summary of the secondary structure composition of T552 is given in Table 1.

3.2.2 Test Data Set Preparation

We have two different test datasets. First, we have collected CB513 dataset [126] for testing purpose. Then we ran BLASTclust at 25% identity cut-off on T552 and CB513 to ensure that this dataset is independent of our training dataset. Here we extracted 475 seqeunces (CB475) from CB513 at 25% identity cut-off with respect to T552. After further refinement based on failure to generate angle fluctuations or ASA for some sequences, we had 471 seqeunces as our first test set (CB471). A summary of the secondary structure composition of CB471 is given in Table 2.

Table 2*:* A summary of the secondary structure composition of CB471 test dataset.

Secondary Structure	Residue Count	Percentage
Beta	17,037	22.8%
Coil	31,908	42.7%
Helix	25,843	34.5%
Total	74,788	100.0%

We collected another comparatively new independent dataset with 25% identity cut-off criteria, prepared in 2014. This dataset consists of 295 seqeunces. From now on we will call these dataset as N295. N295 dataset was used to further confirm the robustness of our predictors. A summary of the secondary structure composition of N295 is given in Table 3.

Table 3*:* A summary of the secondary structure composition of N295 test dataset.

Secondary Structure	Residue Count	Percentage
Beta	16,052	26.2%
Coil	25,199	41.2%
Helix	19,913	32.6%
Total	61,164	100.0%

3.3 Features

A crucial factor for classification problem is feature set. We collected a comprehensive and independent set of residue level features which may sufficinetly capture sequence information, evolutionary information as well as structural information of the amino acids in the protein seqeunces. A brief discussion on each category of feature is given below:

Amino acid: this is simple the the information about the particular amino acid on certain position of a protein seqeunce. Twenty different amino acids are marked by distinct integers 1 through 20.

Physiochemical properties: Each amino acid has 7 unique properties, combinedly named as physiochemical properties. They are steric parameter, polarizability, hydrophobicity, isoelectric point, helix probability and sheet probability [127]. These parameters influence possible structure of an amino acid residue in a protein sequence. For example, a hydrophobic residue is more likely to be inside the core of a globular protein.

Position specific scoring matrix: Position-sspecific scoring matrix (PSSM) is a kind of scoring matrix used in protein BLAST [128] searches in which amino acid substitution scores are given separately for each position in a protein multiple sequence alignment. This score captures similarities between protein query sequences and all sequences in one or more protein databases. Therefore, PSSM represents valuable evolutionary information at each position of the protein seqeunce. PSSM was generated by running PSIBLAST [112].

Monogram and bigram: These matrices were proposed by Sharma and his colleagues [129]. Monogram and bigram are derived from PSSM score to infer structural information from sequence level evolutionary information.

Disorder probability: It gives the probability of amino acid residue being disordered, i.e., having no well defined three dimensional structure. Disorder probabilities (DPs) were calculated from DisPredict [130].

Accessible surfae area: Accessible surface area (ASA) is the surface area of a biomolecule that is accessible to the solvent in which the molecule is dissolved. Conformational dynamics of proteins which is crucial for their diverse functionalities, is strongly correlated with the ASA of each of the residue of a protein [131, 132]. ASA is directly related to the protein-protein interactions [133, 134] and is also an important factor in beta pair formation [135]. Therefore, we used a very recently developed high accuracy ASA predictor, REGAd3p [136], to predict ASA in our work. REGAd3p uses regularized exact regression with 3rd degree polynomial kernel and also applied GA to optimize the weights computed by regularized regression.

Torsion angle (φ, ψ) fluctuations: Torsion angle fluctuations (AF) represent the flexibility of protein backbone as derived from the ensembles of NMR structure. These are two very important

features as only two torsion angles- φ, ψ are sufficient for a nearly complete description of the backbone of a protein structure [137].

Terminal indocator: First five and last five residues in any sequence are considered as terminal residues. For the first or last one we assigned -1 and 1 respectively as terminal information and then gradually increased or increased the value by 0.2 as we moved forward from the starting terminal or move back-ward from the end terminal. For example, the second and the penultimate residue gets -0.8 and 0.8 respectively as terminal information value. All other intermediate residues have 0 terminal value. A complete list of features is shown in Table 4.

Table 4*:* A list of all features used in this research.

Category	Feature count
Amino acid (AA)	1
Physiochemical properties (PP)	7
Position specific scoring matrix (PSSM)	20
Monogram (MG)	1
Bigram (BG)	20
Disorder probability (DP)	1
Accessible surfae area (ASA)	1
Torsion angles (φ, ψ) fluctuation (AF)	2
Terminal indocator (TI)	1

We used these features in a variety of combinations in our search to come up with an optimal feature set. Some features were kept in all models, whereas some were excluded in some models to gauge the impact of the excluded features in our prediction. Different sets of features we used in our model search are listed in the Table 5.

Table 5*:* Description of the feature sets used.

Name of the feature set	# features	Details of the features (feature count)
f_{29}	29	AA(1), PP (7), PSSM(20), and TI(1)
f_{31}	31	AA(1), PP (7), PSSM(20), DP(1), ASA(1) and TI(1)
f_{33}	33	AA(1), PP (7), PSSM(20), DP(1), ASA(1), AF(2) and TI(1)
f_{51}	51	AA(1), PP (7), PSSM(20), MG(1), BG(20), DP(1) and TI(1)

3.4 Performance Evaluation

To compare and evaluate the performance of each predictors we used 4 performance criteria: accuracy, precision, recall and overprediction rate. To measure these matrices we need to know the definition of true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

Table 6*: Evaluation criteria.*

Measure	**Formula**	**Evaluation Focus**
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall effectiveness of a classifier
Precision	$\frac{TP}{TP + FP}$	Agreement on class of the data labels with the positive labels given by the classifier
Recall (Sensitivity)	$\frac{TP}{TP + FN}$	Effectiveness of a classifier in identifying positive labels
Over predicton	$\frac{\#Predicted\ class}{\#Actual\ class}$	Measures whether higher accuracy for particular class is due to over prediction or not
Overall Precision*	$\frac{\sum_{c=1}^{n} Precision_c}{n}$	Average agreement on class of the data labels with the positive *n* labels given by the classifier
Overall Recall*	$\frac{\sum_{c=1}^{n} Recall_c}{n}$	Avergae effectiveness of a classifier in identifying positive *n* labels

* This definition is usually known as macro measure.

TP is the number instances that are labelled as positive and are actually positive. FP is the number instances that are labelled as positive and are actually negative. TN is the number instances that are labelled as negative and are actually negative. FN is the number instances that are labelled as negative and are actually positive. Calculations of these measures along with their evaluation focus are presented in Table 6.

3.5 In Search of an Appropriate Model

Although SVM may be directly used for three class classification, we choose to use three binary SVMs so that we may attain a balanced accuracy in all three classes. Another reason for using binar SVMs is that SVM was built for binary classification problem and performance may degrade if used for multi class classification problem. We carried out experiment to compare the performance of prediction using f_{29} and f_{51} as feature sets. Again we trained the model with 90% of the 554 training data and tested on hold out 10% dataset. Then we also tested on CB513 dataset. The reults are shown in Table 7. Here we see that f_{51} feature predicts slighly better than f_{29} in all three classes for both test data set of CB513 and 10% hold out from 554, with one exception only: f_{29}feature based C/~C model performed slightly better on CB513.

Table 7*:* Comparison of performance of models trained with f_{29} and f_{51} feature sets.

Model	Feature Set	Accuracy on CB513	Accuracy on 10% Hold out of T554
E/~E	f_{29}	81.84%	82.99%
E/~E	f_{51}	83.10%	83.44%
C/~C	f_{29}	73.04%	63.41%
C/~C	f_{51}	72.57%	63.83%
H/~H	f_{29}	80.91%	74.82%
H/~H	f_{51}	82.51%	75.07%

To further justify the superior performance of f_{51} feature over f_{29}, we went for 10 fold cross validation (FCV) for both feature set based models along with optimization using RBF kernel. The result of this investigation is shown in Table 8. Here we see that, though initially on a small hold out test set, f_{51} feature was predicting better than f_{29} feature, on an average using f_{51} has no advantage over using f_{29}. Only in case of E/~E, f_{51} based model performed slightly better than f_{29} based model. This eventually establishes that bigram or monogram are not very useful features for SSP as including them does not improve accuracy, rather in some cases accuracy decreased. Moreover, using f_{51} feature is costlier as well. Therefore, we decided not to use f_{51} feafture set, particularly bigram and monograms, in any further investigation.

Table 8*:* Comparison of the performance of using f_{29} and f_{51} features sets. CB475 dataset was extracted from CB513 dataset to ensure that the test set is no more than 25% similar to training set to ensure more robust comparison.

Model	Feature Set	10 FCV Accuracy on 554 Dataset	Accuracy on CB475 Test Dataset
E/~E	f_{29}	82.64%	81.96%
E/~E	f_{51}	82.58%	82.46%
C/~C	f_{29}	63.54%	74.06%
C/~C	f_{51}	61.81%	73.03%
H/~H	f_{29}	73.81%	78.66%
H/~H	f_{51}	73.66%	78.12%

We also investigated the efficacy of f_{31} feature set. The outcomes are compared with f_{29} based models' results. Result of this investigation is presented in Table 9. In this comparision we see that f_{31} underperfroms in predicting E or C, but significantly overperforms in predicting H class, when compared with those of f_{29} feature based predictions. Using only 2 more features is not that costly as well. Therefore, we decided not to discard f_{31} at this point. It is to be noted that f_{31} contains all the features used in f_{29}. In addition f_{31} contains two more features – DP and ASA.

Table 9: Comparison of the performance of using f_{29} and f_{31} features sets. CB475 dataset was extracted from CB513 dataset to ensure that the test set is no more than 25% similar to training set to ensure more robust comparison.

Model	Feature set	Accuracy on CB475 Test Dataset
E/~E	f_{29}	81.96%
E/~E	f_{31}	81.49%
C/~C	f_{29}	74.06%
C/~C	f_{31}	72.92%
H/~H	f_{29}	78.66%
H/~H	f_{31}	83.76%

Finally, we also investigated the efficacy of using f_{33} feature sets by comparing the outcomes of the models trained by f_{33} with those of f_{29} and f_{31} based models.

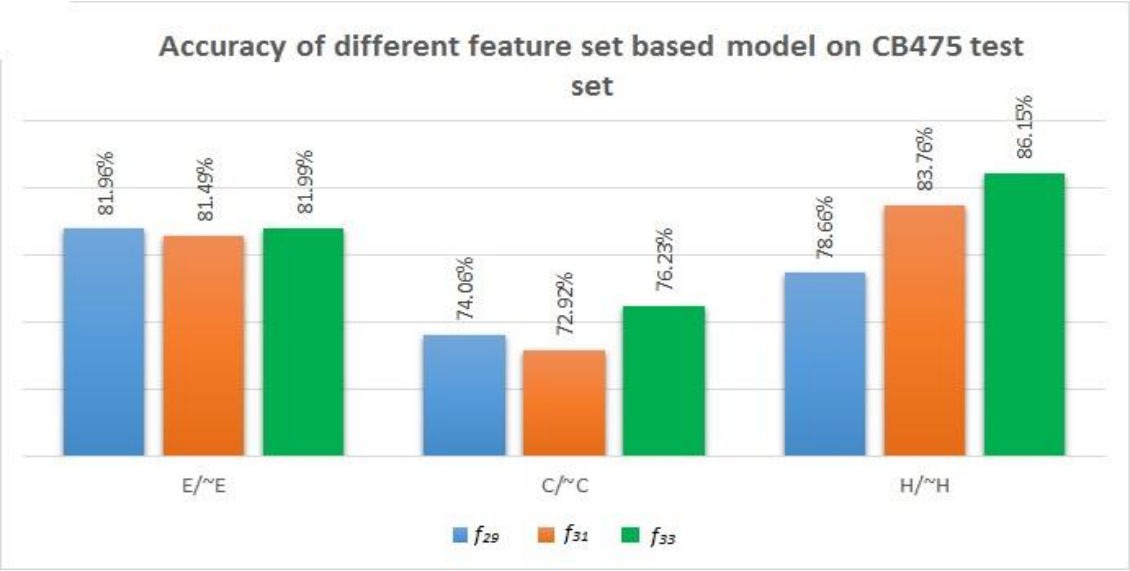

Figure 16: Binary class accuracies on CB475 dataset for different feature set based models.

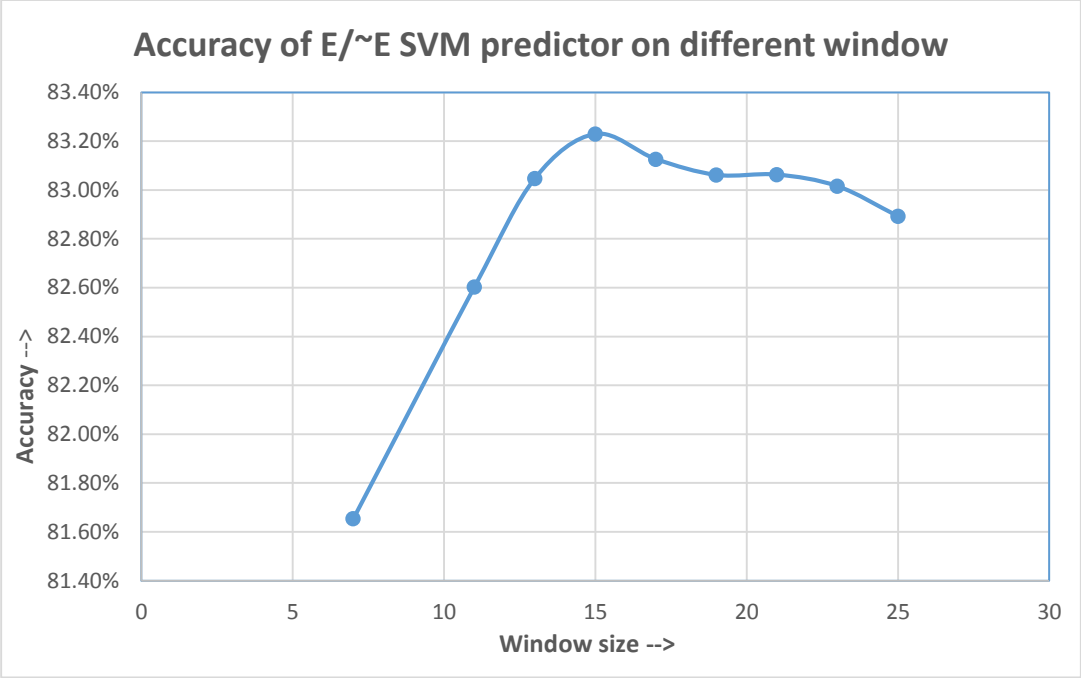

Figure 17*:* Accuracy of E/~E SVM predictor at different window size.

We also tried with various window size to obtain an optimal window for our predictor. For this we only used E/~E SVM as we priritized the enhancement of accuracy of E class prediction over other classes. Figure 17. shows the performance of using different window size. Based on this analysis we choose a wndow size of 15 for our final model.

4. Results and Discussion

Our final test datasets were CB471, extracted from CB513, and N295, collected by ourselves. In order to assess the perfromance of each predictors, we calculated the overall accuracy (Q_3) of each predictors as well as accuracy for beta, coil and helix class (Q_E, Q_C and Q_H respectively) along with precision and recall. We also calculated the over prediction rate for each class to investigate whether any higher measure is due to overprediction. In the following sub sections comparative perfromance of cSVM, SPINE X and MetaSSPred is discussed in light of accuracy, precision, recall and over prediction measures for two different test datasets.

4.1 Performance on CB471 Test Dataset

Accuracies on CB471 dataset are presented in Table 10. To facilitate comparison, in Figure 18, accuracies as well as overprediction rates of each predictors for each class is presented in bar chart. In Figure 18 we see that Q_E of MetaSSPred is significantly higher than those of other predictors. More specifically, Q_E of MetaSSPred is 20.9% improvement over than that of SPINE X. Further, MetaSSPred does not heavily under or over predicts beta compared to other two methods. Therefore, MetaSSPred is certainly a better predictor for the beta class. Poor performanceLower Q_E of SPINE X here may be attributed to the very high under prediction rate (24.6%).

Table 10*:* Accuracy of secondary structure prediction on CB471 test dataset.

	Accuracy				Standard Deviation of Class Wise Accuracies
Model	**Q_E**	**Q_C**	**Q_H**	**Q_3**	
cSVM	63.7%	80.6%	75.7%	75.1%	8.7%
SPINE X	59.3%	81.4%	81.9%	76.5%	12.9%
MetaSSPred	71.7%	76.0%	80.1%	76.4%	4.2%

In case of Q_C, SPINE X gives the highest score (81.4%), cSVM is just after SPINE X (with a Q_C of 80.6%) and MetaSSPred's Q_C is 76.0% only. If we look at the over prediction rates of this class, we find that both cSVM and SPINE X highly over predict coil class (13.4% and 14.5% respectively). On the other hand, MetSSPred's over prediction rate is very low, only 1.9%. This is a strong reason why the Q_Cs of cSVM and SPINE X are higher than that of MetaSSPred.

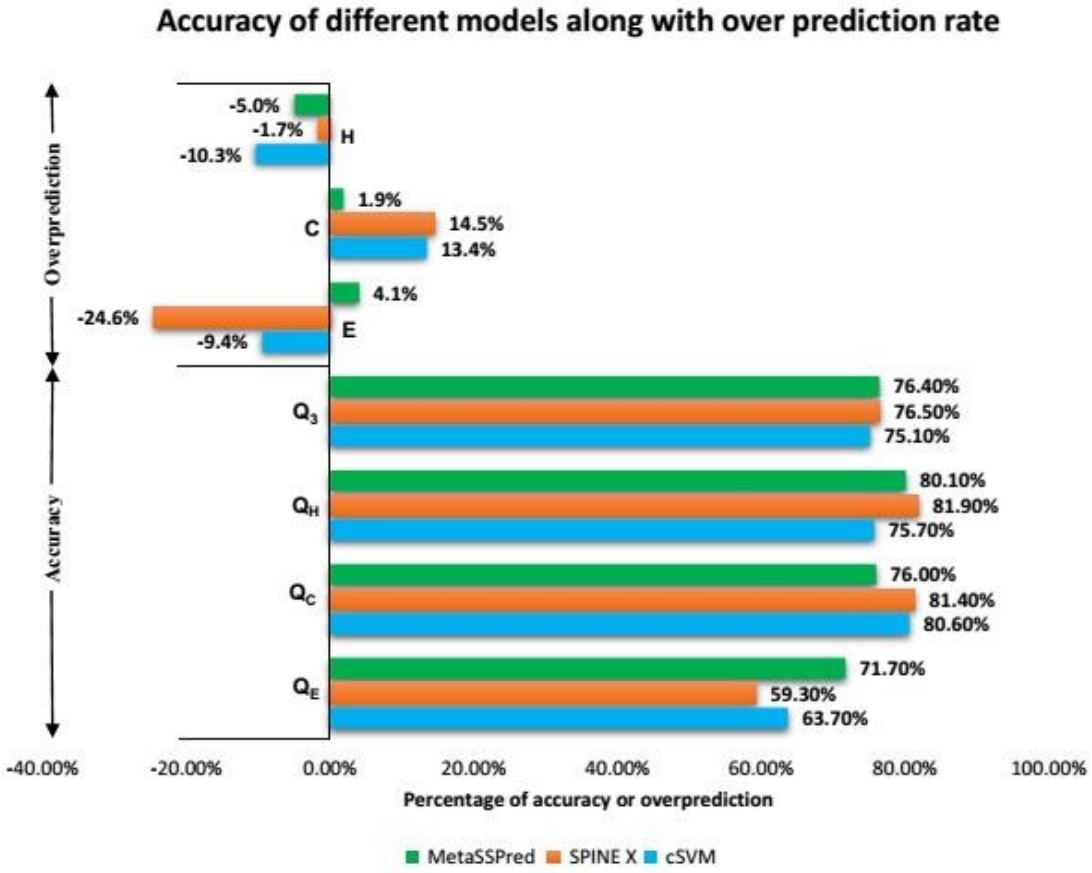

Figure 18*:* Comparison of accuracy along with over prediction rate on CB471 dataset.

SPINE X comes up with the highest Q_H (81.90%) and MetaSSPred closely follows SPINE X with 80.10%. cSVM is the worst performer in this case. The reason for low Q_H of cSVM may be attributed to the fact that it under prediction predicts of helix by 10.3%. In Q_3 measure,

MetaSSPred and SPINE X are almost equal with a 0.1% gap in favor of SPINE X. cSVM is also not far behind. Overall, comparatively MetaSSPred appears as a very balanced predictor yielding good accuracies for all three classes separately as well as in Q_3.However, it seems that it is very difficult to find the best model based on this accuracy measure only as no single predictor has highest score in all four measures here. Therefore, we will now focus on precision and recall. Precision and recall measures for CB471 datset are presented in Table 11.

Table 11: Precision and recall of secondary structure prediction on CB471 test dataset.

Model	Measure	Beta (E)	Coil (C)	Helix (H)	Overall
cSVM	Precision	70.3%	71.0%	84.5%	75.3%
SPINE X	Precision	78.7%	71.1%	83.3%	77.7%
MetaSSPred	Precision	68.9%	74.6%	84.3%	75.9%
cSVM	Recall	63.7%	80.6%	75.7%	73.3%
SPINE X	Recall	59.3%	81.4%	81.9%	74.2%
MetaSSPred	Recall	71.7%	76.0%	80.1%	75.9%

In Figure 19 precision and recall values are plotted grouping by class for all predictors to facilitate visual perception. Here in Figure 18, we see that SPINE X has the highest precision while theand lowest recall value for beta class is the lowest. Gap between the precision and recall score of SPINE X for beta class is 7%. This suggests that overall SPINE X gives lower false positives, however it gives a very high false negatives in case of beta prediction. Therefore, for any application, where the cost of failure to detect beta residue is high, SPINE X may not be suitable. MetaSSPred provides relatively high balanced precision and recall value for beta prediction with a gap of only 2.8% between recall and precision. The proposed MetaSSPred provides the highest recall value for beta class among the three predictors discussed. Therefore, applications where detecting betas are very important, MetaSSPred would perform well. cSVM has a recall value 7.4% higher than

that of SPINE X, however, on the other hand, cSVM achieves 10.7% lower precision compared to SPINE X. Therefore, if cost of failure to identify beta is higher, we should prefer cSVM to SPINE X.

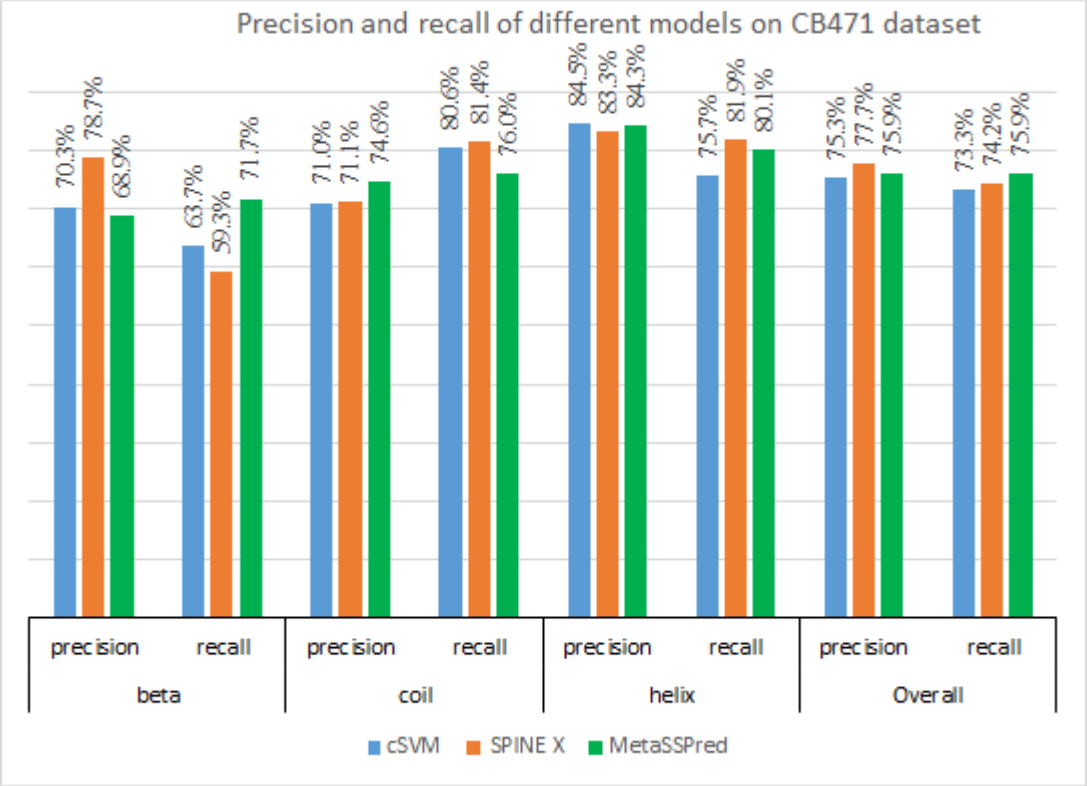

Figure 19: Precision and recall on CB471 dataset obtained for different predictors.

In coil prediction, MetaSSPred provides balanced precision and recall score with a gap of 1.4% only, and its precision is the highest among all predictors. SPINE X and cSVM provides 7.1% and 6.0% higher recall score than that of MetaSSPred. However, MetaSSPred provided 4.9% and 5.1% improvement in precision score than those of SPINE X and cSVM respectively. The reason behind this phenomenon is that both cSVM and SPINE X highly over predict coil class which are 13.4% and 14.5% higher prediction rate respectively.

In helix prediction, precisions of each class are very close. Although cSVM provides the highest precision for helix, it gives the lowest recall. Therefore, if failure to detect helix is very costly, we should prefer other predictors to cSVM. SPINE X and MetaSSPred are close competitors for helix prediction accuracies. Both of them have good and balanced precision and recall score in this case.

Overall, both precision and recall for MetaSSPred are equal and at the higher end. On the other hand Precision and recall for SPINE X widely varies due to high over or underprediction rates across classes. cSVM gives lowest precision and recall.

4.2 Performance on N295 Test Dataset

We will start our analysis of the performances of our predictors on N295 dataset with accuracy measure. Table 12 shows the accuracy of different predictors.

Table 12*:* Accuracy of secondary structure prediction on N295 test dataset.

	Accuracy				Standard Deviation of Class Wise Accuracies
Methods	**Q_E**	**Q_C**	**Q_H**	**Q_3**	
cSVM	65.7%	82.7%	74.3%	75.5%	8.5%
SPINE X	62.5%	82.2%	80.6%	76.5%	10.9%
MetaSSPred	74.4%	77.5%	79.0%	77.2%	2.3%

To have a more comprehensive view, these accuracies and overprediction rates are shown in Figure 20. Both SPINE X and cSVM give very poorcomparatively lower Q_E score and both of them highly under predict beta (22.5% and 15.0% under prediction respectively). On the other hand MetaSSPred gives the highest Q_E almost without over or under prediction of beta class. Here improvement in Q_E achieved by MetaSSPred over SPINE X is 19.0%. Therefore, MetaSSPred is clearly the best predictor for beta class for this N295 dataset.

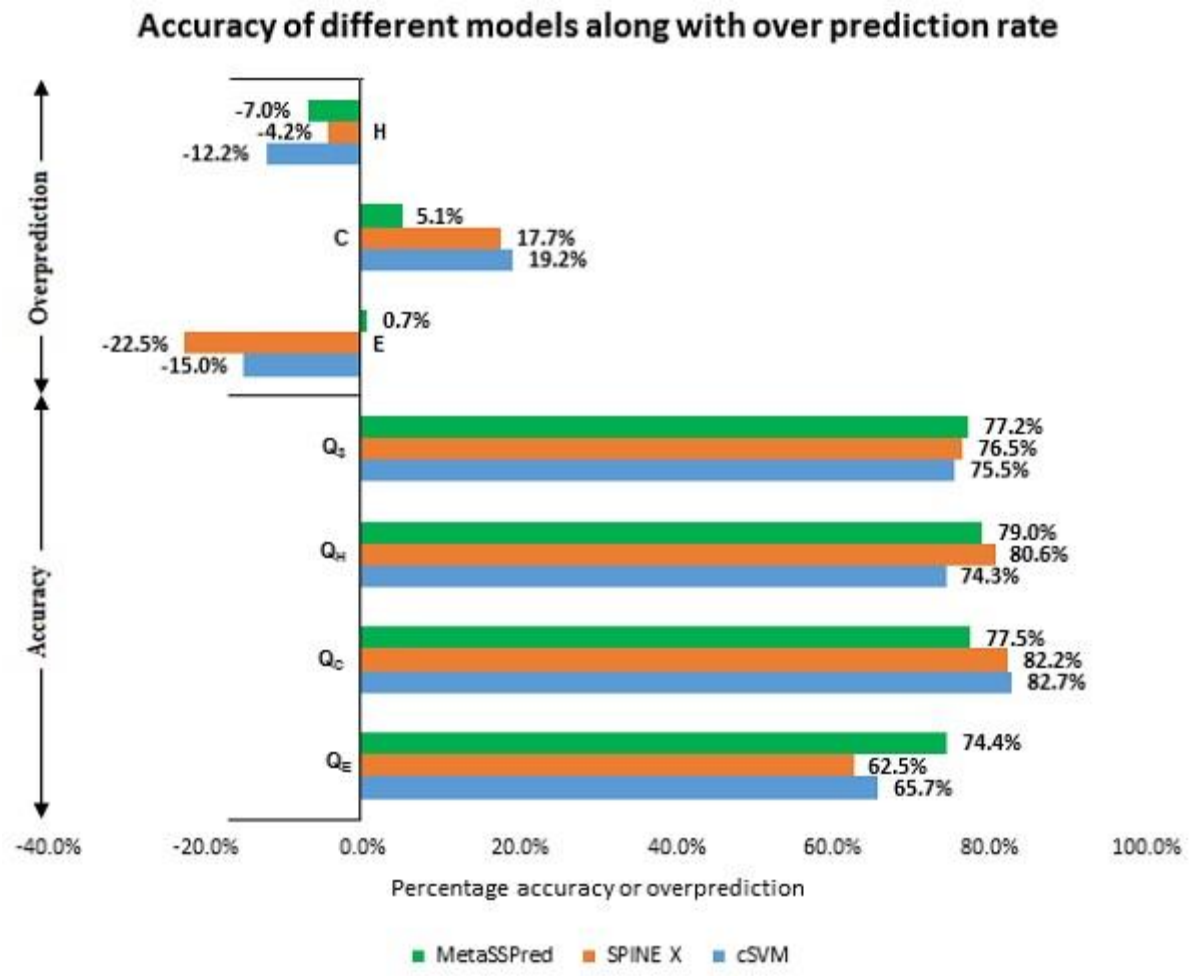

Figure 20: Comparison of accuracy along with over prediction rate on N295 dataset.

In coil prediction, cSVM and SPINE X give almost similar accuracy with a gap of 0.5% in favor of cSVM. On the other hand, Q_C of MetaSSPred is the lowest among all. Q_C of SPINE X is 6.1% higher than that of MetaSSPred. If we look at the over prediction rates, we see that both cSVM and SPINE X highly over predicts coil, by 17.7% and 19.2% respectively. MetaSSPred also over predicts coil, however by only 5.1%. Therefore, the higher Q_C of cSVM and SPINE X than that of MetaSSPred may be because of over predictions by the first two methods mainly.

Highest Q_H is obtained from SPINE X, and MetaSSPred is closely following SPINE X by a gap of 1.6%. Q_H score of cSVM is significantly lower than those of two others. If we look at the

over prediction rate, we see that cSVM under predicts helix by 12.2%. This may be a strong reason why cSVM gives so poor Q_H. Other two methods, SPINE X and MetaSSPred also under predict helix. Interestingly, slightly lower Q_H of MetaSSPred results from its around 3% of higher under prediction rate compared to that of SPINE X. Overall, MetaSSPred provides the highest Q_3 score for N295 dataset. To further gauge the performnaces of our predictors on N295 dataset, now we will focus on the precision and recall scores. Precision and recall scores of all three predictors on N295 dataset are presented in Table 13. Class-wise grouped precision and recall values for all predictors are also shown in a bar chart in Figure 21 to ease the comparison.

Table 13*: Precision and recall* of secondary structure prediction on N295 test dataset.

Methods	Measure	Beta (E)	Coil (C)	Helix (H)	Overall
cSVM	Precision	77.3%	69.4%	84.6%	77.1%
SPINE X	Precision	80.7%	69.8%	84.1%	78.2%
MetaSSPred	Precision	73.9%	73.7%	84.9%	77.5%
cSVM	Recall	65.7%	82.7%	74.3%	74.2%
SPINE X	Recall	62.5%	82.2%	80.6%	75.1%
MetaSSPred	Recall	74.4%	77.5%	79.0%	77.0%

We see in Figure 21 that SPINE X gives the highest precision and lowest recall in beta prediction. The reason behind such imbalance prediction is that SPINE X highly under predicts (by 22.5%) beta residues. Second highest precision is given by cSVM and again it also has a recall score comparatively much lower than that of MetaSSPred. Main reason is again under prediction. cSVM under predicts beta residues by 15%. On the other hand, MetaSSPred gives a very balanced precision and recall and it has the highest recall score for beta prediction. Over prediction rate of MetaSSPred is only 0.7% for beta residues. Therefore, false positive rate of MetaSSPred is also very low.

In case of coil prediction, precision scores of SPINE X and cSVM are comparatively lower than that of MetaSSPred. On the other hand, recall score of MetaSSPred is lower than that of both cSVM and SPINE X in coil prediction. The reason is again that cSVM and SPINE X over predicts coils by 19.2% and 17.7% respectively. MetaSSPred also over predicts coil, however by only 5.1%. Therefore, false positive rates of SPINE X and cSVM are higher than that of MetaSSPred. MetaSSPred gives the highest precision for helix prediction. Other predictors are also very close. Highest recall is given by SPINE X and MetaSSPred closely follows it. If we look at the over prediction rate, we see that all three methods under predict helix, however under prediction rate is lowest for SPINE X. Considering precision, recall and over prediction rate, it seems that SPINE X is the best method for helix prediction and then comes MetaSSPred.

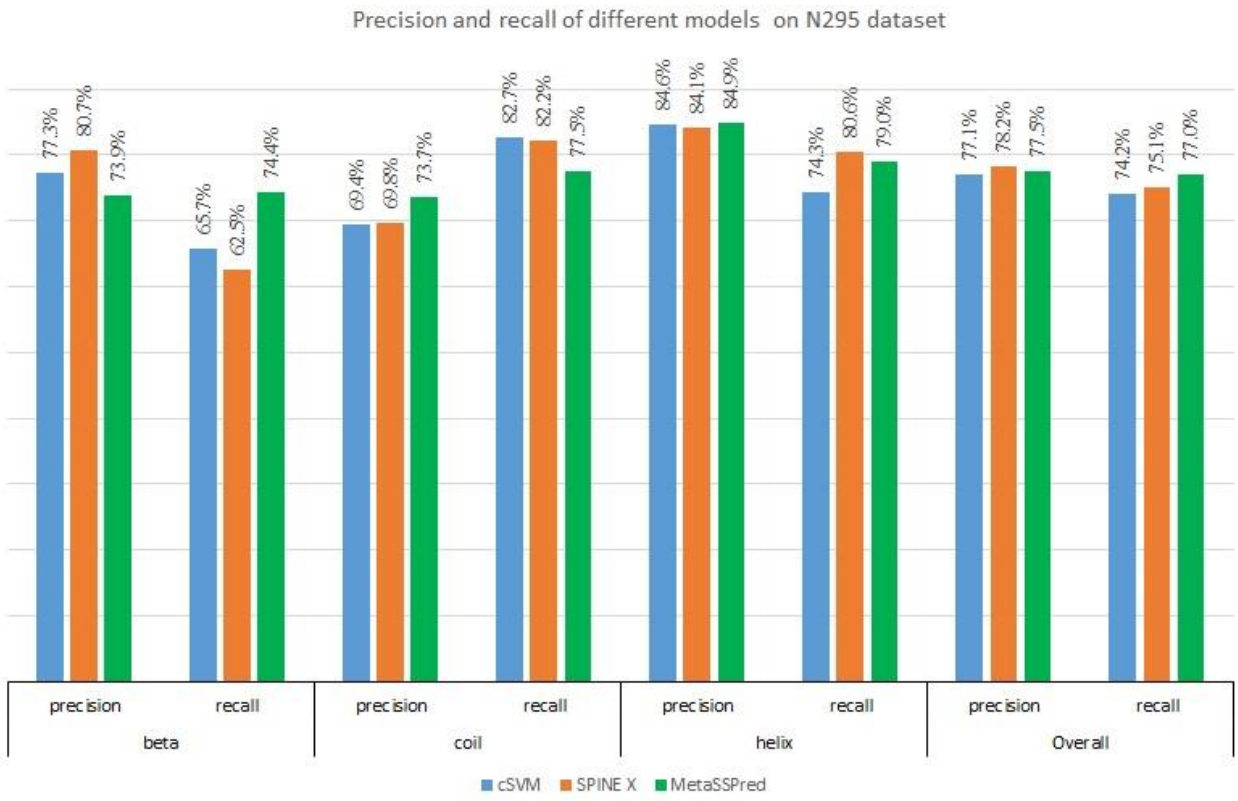

Figure 21: Precision and recall on N295 dataset obtained for different predictors.

Overall precision scores of all three methods are very close. Highest precision score is given by SPINE X, and highest recall score is given by MetaSSPred. However the gap between precision and recall scores of SPINE X is comparatively wide (3.1%).

On the other hand, precision and recall scores of MetaSSPred are close (0.5% apart), which indicates that MetaSSPred is comparatively more balanced predictor with low rate of over or under prediction.

4.3 Overall Ranking of the Predictors

In this section, we will try to summarize the performance of each predictor by ranking them with respect to Q_3, Q_E, Q_C, Q_H, overall and class wise precision recall and absolute over/under prediction rate prediction rate for each test dataset. Higher absolute over/under prediction rate

Table 14: Rank of all predictors across different performance measure on CB471 test data set.

	Rank (higher point better)		
Measure	**cSVM**	**SPINE X**	**MetaSSPred**
Q_3	1	3	2
Q_E	2	1	3
Q_C	2	3	1
Q_H	1	3	2
Precision (E)	2	3	1
Precision (C)	1	2	3
Precision (H)	3	1	2
Overall precision	1	3	2
Recall (E)	2	1	3
Recall (C)	2	3	1
Recall (H)	1	3	2
Overall recall	1	2	3
Absolute over prediction (E)	2	1	3
Absolute over prediction (C)	2	1	3
Absolute over prediction (H)	1	3	2
Total	**24**	**33**	**33**

yields lower ranking, while higher scores for all other measures result in higher ranking. Absolute value of over prediction is taken assuming that both over and under prediction are equally bad. Best scorer in any measure gets 3 point, second best 2 and the third gets 1. The ranks of testing CB471 and N295 are presented in Table 14 and Table 15 respectively.

We see in Table 14 that SPINE X and MetaSSPred performed equally on CB471 test dataset. One the other hand performance ranking of cSVM is comparatively much lower. However in many parameters, cSVM is better than SPINE X. For example, recall of E or precision of H is better in cSVM compared to those of SPINE X. Therefore, our test result on CB471 justifies the development of meta predictor.

Table 15: Rank of all predictors across different performance measure on N295 test data set.

	Rank (higher point better)		
Measure	**cSVM**	**SPINE X**	**MetaSSPred**
Q_3	1	2	3
Q_E	2	1	3
Q_C	3	2	1
Q_H	1	3	2
Precision(E)	2	3	1
Precision(C)	1	2	3
Precision(H)	2	1	3
Overall precision	1	3	2
Recall(E)	2	1	3
Recall(C)	3	2	1
Recall(H)	1	3	2
Overall recall	1	2	3
Absolute over prediction(E)	2	1	3
Absolute over prediction (C)	1	2	3
Absolute over prediction (H)	1	3	2
Total	**24**	**31**	**35**

We see in Table 15 that performance of MetaSSPred is the best on N295 test dataset, whereas SPINE X ranked second. Therefore, overall we may conclude that MetaSSPred is the best predictor among the three discussed here across different dataset. In other words, MetaSSPred is a more generalized predictor with balanced accuracy across all secondary structure class.

5. Conclusion

In this section, the outcomes of our investigation is briefly summarized and some future directions for further improvement are also suggested.

5.1 Summary of Outcomes

Comparing the performance of different classifier in multiclass classification is a complex task as assigning cost of missclassification in multiclass classification is not a straight forward. Overall accuracy of multiclass classification alone is not a good measure to decide on the performance of such a classifier. Becasue overall accuracy may be higher even when the classifier fails to correctly classify any member of a minority class in an imbalanced dataset. Therefore, we also have calculated the precision and recall measures for each class as well as for overall classfication. We have tested three different models here on two different datasets independent of our training dataset. Our basic model was a combined version of the three binary class SVMs, which were optimally combined into a multiclass classifier (cSVM) using GA. We compared the performance measures of our classifier with those of SPINE X, which is the state-of-the-art secondary structure predictor in terms of reported accuracy. We have found that, although SPINE X claimed Q_3 score higher than 80% in their own prepared dataset, in none of the test datasets those we used prepared, SPINE X did achieve such 80% Q_3 score . Q_3 score of SPINE X was 76.5% on both CB471 and N295 datasets in our investigation. For our cSVM, we obtained 75.5% and 74.2% Q_3 score based on CB471 and N295 datasets respectively. However, we observed that Q_E scores of SPINE X were comparatively lower for both CB471 and N295 datasets and which were found to 59.3% and 62.5% respectively to be exact. On the other hand, our cSVM provided better Q_E scores than SPINE X

on both datasets which and the scores were were 63.7% and 65.7% respectively. We also have observed that SPINE X highly under predicted helix by 24.6% and 22.5% for dataset CB471 and dataset N295 respectively. Q_C score of cSVM and SPINE X for both datasets were close. For example, Q_C of cSVM were 80.6% and 82.7% and those of SPINE X were 81.4% and 82.2% for dataset CB471 and dataset N295 respectively. On CB471, SPINE X gave higher accuracy in coil than cSVM. However, the difference was by only 0.8%. On the other hand, cSVM gave higher coil accuracy on N295 dataset and the gap was again just only 0.5% only. In helix prediction, SPINE X performed better than cSVM on both the test datasets. Gaps in Q_H of SPINE X and cSVM were 6.2% and 1% based on dataset CB471 and N295 respectively.

In a nutshell, SPINE X was better for helix prediction. On the other hand cSVM was better for beta prediction. In coil prediction, both are alomost equally accurate. Therefore, we found an opportunity to combine cSVM and SPINE X to achieve better accuracy in all three classes and developed our meta predictor, MetaSSPred, combining the result of cSVM and SPINE X. The outcome is found to be very promising.

MetaSSPred significantly increases Q_E for both datasets. Q_E score of MetaSSPred on CB471 and N295 were 71.7% and 74.4% respectively. This is 20.9% and 19.0% improvement over the Q_E scores given by SPINE X on CB471 and N295 datasets respectively. Improvements of Q_E scores by MetaSSPred over those of cSVM were also significant- 12.6% and 13.3% on CB471 and N295 datasets respectively. However this improvement in Q_E brought some cost for coil prediction mainly. For example Q_C scores of MetaSSpred were 5.4% and 4.7% lower in absolute value than those of SPINE X on CB471 and N295 datasets respectively. Average drop of Q_H score in MetaSSPred over SPINE X was 2.1%. Overall accuracy of MetaSSPred, decreased by 0.1% in absolute value on CB471 dataset, however increased by 0.9% in absolute value on N295

dataset compared to those of SPINE X. MetaSSPred also decreased the volatility of accuracies across thre secondary structure classes. For example, standard deviations of accuracies across three classes were 12.9% and 10.9% on CB471 and N295 test sets respectively for SPINE X as shown in Table 10 and Table 12 respectively. . On the other hand for the same data sets standard deviations of three class accuracies were 4.2% and 2.3% respectively for MetaSSPred as seen in Table 10 and Table 12 respectively Precision and recall gap volatility also decreased in MetaSSPred. For example, standard deviations of the gaps between respective precision and recall scores across three secondary structure class are 10.0%, 15.0% and 3.7% for cSVM, SPINE X and MetaSSPred respectively on CB471 dataset. Same volatility reduction in the gaps between precision and recall was observed for N295 dataset. Standard deviations of such gaps for cSVM, SPINE X and MetaSSPred on N295 dataset are 14.0%, 15.3% and 4.9% respectively. Therefore, we may conclude that MetaSSPred is a more balanced secondary structure predictor compared to SPINE X.

5.2 Scope for Further Improvement

We have observed that though our MetaSSPred gives more balanced SSP accuracies across three secondary structure classes, overall accuracies of MetaSSPred on different datasets were not significantly different from those of SPINE X. We suggest following measures to improve further the over all accuracy of SSP without compromising the balance achieved here:

- Use of boosting while training the SVMs. Boosting is simply testing the training set on the model and finding the data points, where the model fails to predict correctly and then adding those data points repeatedly to the training set and retraining the model. This process continues untill the test error becomes constant or does not reduce further.

- Using consensus secondary structure instead of DSSP assignment. Since different assigning methods assign SS based on differenrt factors, overall error in assignment should be lower. Some notable SS assignment methods are KAKSI [138], STRIDE [139], P-SEA [140], etc.
- It is also a good idea to further investigate the efficacy of the feature sets bigram and monograms utilizing the power of boosting together.

References

1. Hunter, L., *Molecular biology for computer scientists. In L. Hunter, editor, Artificial Intelligence for Molecular Biology*. 1993: AAAIPress. 1-46.
2. Branden, C. and J. Tooze, *Introduction to protein structure*. 1999, New York: GarlandPublishing, Inc.
3. Pietzsch, J., *The importance of protein folding*. Nature Publishing Group.
4. Wang, L.-H., et al., *Predicting Protein Secondary Structure by a Support Vector Machine Based on a New Coding Scheme.* Genome Informatics 2004. **15**(2): p. 181-190.
5. Gu, J. and P.E. Bourne, *Structural Bioinformatics*. 2nd ed. 2009.
6. Strandberg, B., R.E. Dickerson, and M.G. Rossmann, *50 Years of Protein Structure Analysis.* Journal of Molecular Biology, 2009. **392**: p. 2-32.
7. PDB. *Protein Data Bank*. 2014 [cited 2014 March 24]; Available from: http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/nature_of_3d_structural_data.html.
8. Wüthrich, K., *Protein Structure Determination in Solution by NMR Spectroscopy.* The Journal of BIological Chemistry 1990. **265**(36): p. 22059-22062.
9. Chopra, G., C.M. Summa, and M. Levitt, *Solvent dramatically affects protein structure refinement.* PNAS, 2008. **105**(51): p. 20239 –20244.
10. Metfessel, B.A. and P.N. Saurugge. *Pattern recognition in the prediction of protein structural class*. in *Proceeding of the Twenty-Sixth Hawaii International Conference on*. 1993. Hawaii: IEEE Xplore.
11. PDB. *Protein Data Bank*. 2015 [cited 2015 10 January 2015]; Available from: http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/nature_of_3d_structural_data.html.
12. Guo, J., et al., *A Novel Method for Protein Secondary Structure Prediction Using Dual-Layer SVM and Profiles.* Proteins: Structure, Function and Bioinformatics, 2004. **54**: p. 738-743.
13. Baker, D. and A. Sali, *Protein Structure Prediction and Structural Genomics.* Science, 2001. **294**.
14. Duan, Y. and P.A. Kollman, *Computational protein folding: From lattice to all-atom.* IBM Systems Journal 2001. **40**(2): p. 297–309.
15. Sander, I.M., J.L. Chaney, and P.L. Clark, *Expanding Anfinsen's Principle: Contributions of Synonymous Codon Selection to Rational Protein Design.* Journal of American Chemical Society, 2014. **136**: p. 858−861.
16. Anfinsen, C.B., *Principles that Govern the Folding of Protein Chains.* Science, New Series, 1973. **181**(4096): p. 223-230.
17. Chiang, Y.-S., et al., *New Classification of Supersecondary Structures of Sandwich-like Proteins Uncovers Strict Patterns of Strand Assemblage.* Proteins, 2007. **68**: p. 915-921.
18. Ward, J.J., et al., *Secondary structure prediction with support vector machines.* Bioinformatics, 2003. **19**(13): p. 1650–1655.
19. Hua, S. and Z. Sun, *A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach.* Journal of Molecular Biology, 2001. **308**: p. 397-407.
20. Voet, D. and J.G. Voet, *Biochemistry*. 4th ed. 2010: John Wiley & Sons, Inc.

21. Pauling, L., R.B. Corey, and H.R. Branson, *The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain*, in *PNAS*. 1951. p. 205-2011.
22. Kabsch, W. and C. Sander, *Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.* Biopolymers, 1983. **22**: p. 2577-2637.
23. Faraggi, E., et al., *SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles.* Journal of Computational Chemistry, 2012. **33**(3): p. 259-267.
24. Edison, A.S., *Linus Pauling and the planar peptide bond.* Nature Structural Biology, 2001. **8**(3): p. 201-202.
25. Anfinsen, C.B., *Studies on the Principles that Govern the Folding of Protein Chains.* 1973. **181** p. 223-230.
26. Anfinsen, C.B., et al., *The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain.* PNAS, 1961. **47**(9): p. 1309-1314.
27. Guzzo, A.V., *The influence of amino-acid sequence on protein structure.* Biophysical Journal, 1965. **5**: p. 809-822.
28. Levinthal, C., *Are There Pathways For Protein Folding?* Journal of Chemical Physics 1968. **64**: p. 44-45.
29. Bystroff, C., et al., *Local sequence-structure correlations in proteins.* Protein Engineering 1996. **7**(4): p. 417-421.
30. Baker, D. and A. Sali, *Protein Structure Prediction and Structural Genomics.* Science, 2001. **294**(5540): p. 93-96.
31. Arakaki, A.K., Y. Zhang, and J. Skolnick, *Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment.* Bioinformatics, 2004. **20** (7): p. 1087–1096.
32. Wieman, H., et al., *Homology-Based Modelling of Targets for Rational Drug Design.* Mini-Reviews in Medicinal Chemistry, 2004. **4**: p. 793-804.
33. Berg, J.M., J.L. Tymoczko, and L. Stryer, *Biochemistry*. 5th ed. 2002, New York: W. H. Freeman & Company.
34. Bidargaddi, N.P., *Hybrid Computational Models for Protein Sequence Analysis and Secondary Structure Prediction*. 2006, Monash University, Australia.
35. Alberts, B., et al., *Molecular Biology of the Cell*. 2002, New York: Garland Science.
36. Bernal, J.D. and D. Crowfoot, *X-Ray Photographs of Crystalline Pepsin.* Nature, 1934. **133**(3369): p. 794-795.
37. Kendrew, J.C., et al., *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.* Nature, 1958. **181**: p. 662–666.
38. Stretton, A.O.W., *The First Sequence: Fred Sanger and Insulin.* Genetics 2002. **162**: p. 527–532.
39. PDB. *Protein Data Bank*. 2014 [cited 2014 March 12]; Available from: http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=explMethod-nmr&seqid=100.
40. Socci, N.D., W.S. Bialek, and J.N. Onuchic, *Properties and origins of protein secondary structure.* American Physical Society, 1994. **49**: p. 3440-3443.
41. Meiler, J. and D. Baker, *Coupled prediction of protein secondary and tertiary structure.* PNAS, 2003. **100**(21): p. 12105-12110.
42. DePristo, M.A., P.I.W.d. Bakker, and T.L. Blundell, *Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography.* Structure, 2004. **12**: p. 831–838.
43. Acharya, K.R. and M.D. Lloyd, *The advantages and limitations of protein crystal structures.* (Elsevier)TRENDS in Pharmacological Sciences, 2005. **26**: p. 10-14.
44. Berg, J.M., et al., eds. *Biochemistry*. 5th ed. 2002, W. H. Freeman and Company.

45. Zhanhua, C., et al., *Protein subunit interfaces: heterodimers versus homodimers.* Bioinformation, 2005. **1**(2): p. 28-39.
46. LI, R., et al., *The three-dimensional structure of NAD(P)H:quinone reductase, a flavoprotein involved in cancer chemoprotection and chemotherapy: Mechanism of the two-electron reduction.* PNAS, 1995. **92**(19): p. 8846-8850.
47. Cooper, G.M., *The Cell: A Molecular Approach*. 2nd ed. 2000, Sunderland (MA): Sinauer Associates.
48. Giardina, B., et al., *The Multiple Functions of Hemoglobin.* Critical Reviews in Biochemistry and Molecular Biology, 1995. **30**(3): p. 165-196.
49. Berry, E.A., et al., *STRUCTURE AND FUNCTION OFCYTOCHROME bc COMPLEXES.* Annual Review of Biochemistry 2000, 2000. **69**: p. 1005-1075.
50. Sami, A.J., *Structure-function relation of somatotropin with reference to molecular modeling.* Current Protein & Peptide Science, 2007. **8**(3): p. 283-292.
51. Gu, L.-H. and P.A. Coulombe, *Keratin function in skin epithelia: a broadening palette with surprising shades.* Current Opinion in Cell Biology, 2007. **19**(1): p. 13-23.
52. Bränd´en, C.-I. and T.A. Jones, *Between objectivity and subjectivity.* Nature, 1990. **343**: p. 687-689
53. Wlodawer, A., et al., *Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures.* FEBS Journal, 2007. **275**: p. 1-21.
54. Wider, G., *Structure Determination of Biological Macromolecules in Solution Using NMR spectroscopy.* BioTechniques, 2000. **29**: p. 1278–1294
55. Frankel, R.I., *Centennial of Rontgen's discovery of x-rays.* West J Med, 1996. **164**(6): p. 497-501.
56. Bragg, W.L., *The structure of crystals as indicated by their diffraction of X-rays.* Proceedings of the Royal Society of London, 1913. **A.89**: p. 248-277.
57. Lawson, D. *A Brief Introduction to Protein Crystallography*. 2014 [cited 2014 March 24]; Available from: http://www.jic.ac.uk/staff/david-lawson/xtallog/summary.htm.
58. Kleywegt, G.J., *Validation of protein crystal structures.* Acta Crystallographica Section, 2000. **D56**: p. 249-265.
59. Stoddard, B.L., G.K. Farber, and R.K. Strong, *The facts and fancy of microgravity protein crystallization.* Biotechnology and genetic engineering review, 1993. **11**.
60. Kahn, R., et al., *Macromolecular Crystallography with Synchrotron Radiation: Photographic Data Collection and Polarization Correction* Journal of Applied Crystallography, 1982. **15**: p. 330-337.
61. Frauenfelder, H., S.G. Sligar, and P.G. Wolynes, *The energy landscapes and motions of proteins.* Science, 1991. **254**(5038): p. 1598-603.
62. McCammon, J.A. and S.C. Harvey, *Dynamics of Proteins and Nucleic Acids*. 1988, Cambridge, England: Cambridge University Press.
63. Li, C. and M. Liu, *Protein dynamics in living cells studied by in-cell NMR spectroscopy.* FEBS Letters 2013. **587**(8): p. 1008–1011.
64. Ringe, D. and G.A. Petsko, *Study of protein dynamics by X-ray diffraction.* Methods in Enzymology, 1986. **131**: p. 389–433.
65. Rejto, P.A. and S.T. Freer, *Protein conformational substates from X-ray crystallography.* Progress in Biophysics and Molecular Biology, 1996. **66**(2): p. 167–196.
66. Burling, F.T., et al., *Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. .* Science, 1996. **271**(5245): p. 72–77.
67. Wilson, M.A. and A.T. Brunger, *The 1.0 AÊ Crystal Structure of Ca2‡-bound Calmodulin: an Analysis of Disorder and Implications for Functionally Relevant Plasticity.* Journal of Molecular Biology, 2000. **301**: p. 1237-1256.

68. Kuriyan, J., et al., *Effect of Anisotropy and Anharmonicity on Protein Crystallographic Refinement -An Evaluation by Molecular Dynamics* Journal of Molecular Biology, 1986. **190**: p. 227-254.
69. Kuriyan, D.J., et al., *Exploration of disorder in protein structures by X-ray restrained molecular dynamics.* Proteins: Structure, Function, and Bioinformatics, 1991. **10**: p. 340-358.
70. Swaminathan, G.J., et al., *Crystal Structures of Oligomeric Forms of the IP-10/CXCL10 Chemokine.* Structure, 2003. **11**: p. 521–532.
71. PDB. *X-Ray Crystallography*. [cited 2014 May 23]; Available from: http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/nature_of_3d_structural_data.html.
72. Garman, E., *Cool data: quantity AND quality.* Acta Crystallographica Section D Biological Crystallography, 1999. **D**(55): p. 1641-1653.
73. Yee, A.A., et al., *NMR and X-ray Crystallography, Complementary Tools in Structural Proteomics of Small Proteins.* Journal of American Chemical Society, 2005. **127**: p. 16512-16517.
74. Wüthrich, K., *NMR - This Other Method for Protein and Nucleic Acid Structure Determination* Acta Crystallographica Section D: Biological Crystallography, 1995. **51**: p. 249-270
75. Saunders, M., A. Wishnia, and J.G. Kirkwood, *The Nuclear Magnetic Resonance Spectrum of Ribonuclease* Journal of The American Chemical Society, 1957. **79**(12): p. 3289–3290.
76. Dyson, H.J. and P.E. Wright, *Insights into protein folding from NMR.* Annual Review of Physical Chemistry, 1996. **47**: p. 369-395.
77. Farrow, N.A., et al., *Characterization of the Backbone Dynamics of Folded and Denatured States of an SH3 Domain.* Bichemistry 1997. **36**(9): p. 2390-2402.
78. Kay, L.E., *Protein dynamics from NMR.* Nature Structural Biology, 1998(NMR supplement): p. 513 - 517.
79. Bax, A. and S. Grzesiek, *Methodological Advances in Protein NMR* Accounts of Chemical Research 1993. **26**(4): p. 131–138.
80. Sattler, M., J.r. Schleucher, and C. Griesinger, *Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients.* Progress in Nuclear Magnetic Resonance Spectroscopy, 1999. **34**: p. 93–158.
81. LawrenceA.Kelley, S.P. Gardner, and M.J. Sutcliffe, *An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies.* Protein Engineering, 1996. **9**(11): p. 1063-1065.
82. Kosol, S., et al., *Structural Characterization of Intrinsically Disordered Proteins by NMR Spectroscopy.* Molecules 2013. **18**: p. 10802-10828.
83. Sakakibara, D., et al., *Protein structure determination in living cells by in-cell NMR spectroscopy.* Nature, 2009. **458**: p. 102-106.
84. Fulton, A.B., *How Crowded Is the Cytoplasm?* Cell, 1982. **30**: p. 345-347.
85. McGuffee, S.R. and A.H. Elcock, *Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm.* PLoS Computational Biology, 2010. **6** (3): p. e1000694.
86. Zimmerman, S.B. and S.O. Trach, *Extimation of Macromolecule Concentrations and Excluded Volume Effects for the Cytoplasm of Escherichia coli.* Journal of Molecular Biology, 1991. **222**: p. 599-620
87. Chou, J.J. and R. Sounier, *Solution Nuclear Magnetic Resonance Spectroscopy.* Electron Crystallography of Soluble and Membrane Proteins Methods in Molecular Biology, 2013. **955**: p. 495-517.
88. Wang, S., et al., *Solid-state NMR spectroscopy structure determination of a lipid-embedded heptahelical membrane protein.* Nature Methods, 2013. **10**: p. 1007-1012.
89. *MEMBRANE PROTEINS OF KNOWN STRUCTURE DETERMINED BY NMR*. 2014 May 23]; Available from: http://www.drorlist.com/nmr/MPNMR.html.

90. Clore, G.M. and A.M. Gronenborn, *Determining the structures of large proteins and protein complexes by NMR.* Trends in Biotechnology, 1998. **16**(22-34).
91. Yu, H., *Extending the size limit of protein nuclear magnetic resonance.* PNAS, 1999. **96**: p. 332–334.
92. Dyson, H.J. and P.E. Wright, *Unfolded Proteins and Protein Folding Studied by NMR.* Chemical Reviews, 2004. **104**: p. 3607–3622.
93. Kay, L.E., *Advances in Magnetic Resonance. NMR studies of protein structure and dynamics.* Journal of Magnetic Resonance, 2005. **173**: p. 193-207.
94. Ghahramani, Z., *An Introduction to Hidden Markod Models and Bayesian Networks.* International Journal of Pattern Recognition and Artificial Intelligence 2001. **15**(1): p. 9-42.
95. Kundu, A., Y. He, and P. Bahl, *Recognition of handwritten word: First and second order hidden Markov model based approach.* Pattern Recognition, 1989. **22**(3): p. 283-297.
96. Rabiner, L.R. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition* in *Proceedings of the IEEE*. 1989.
97. Eddy, S.R., *What is a hidden Markov model?* Nature Biotechnology 2004. **22**(10): p. 1315-1316.
98. Durbin, R., et al., *Biological Sequence Analysis*. 1998, New York: Cambridge University Press.
99. Rodr´ıguez, L.J. and I.e. Torres, *Comparative Study of the Baum-Welch and Viterbi Training Algorithms Applied to Read and Spontaneous Speech Recognition*, in *First Iberian Conference, IbPRIA 2003*. 2003, Springer Berlin Heidelberg: Puerto de Andratx, Mallorca, Spain. p. 847-857.
100. Kiyoshi Asai, S. Hayamizu, and K.i. Handa, *Prediction of Protein Secondary Structure by the Hidden Markov Model.* Computer Applications in the Biosciences, 1993. **9**(2): p. 141-146.
101. Won, K.-J., et al., *An Evolutionary Method for Learning HMM Structure: Prediction of Protein Secondary Structure.* BMC Bioinformatics, 2007. **8**(357).
102. Bystroff, C., V. Thorsson, and D. Baker, *HMMSTR: a Hidden Markov Model for Local Sequence-Structure Correlations in Proteins.* Journal of Molecular Biology, 2000. **301**: p. 173-190.
103. Martin, J., J.-F. Gibrat, and F. Rodolphe, *Analysis of an optimal hidden Markov model for secondary structure prediction.* BMC Structural Biology, 2006. **6**(25).
104. Yada, T., et al., *DNA Sequence Analysis using Hidden Markov Model and Genetic Algorithm.* Genome Informatics, 1994. **5**(178-179).
105. Won, K.-J., A. Prügel-Bennett, and A. Krogh, *Training HMM Structure with Genetic Algorithm for Biological Sequence Analysis.* Bioinformatics, 2004. **20**(18): p. 3613-3619.
106. Jones, D.T., *Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices.* Jounal of Molecular Biology 1999. **292**: p. 195-202.
107. McCulloch, W.S. and W. Pitts, *A logical calculus of the ideas of immanent in nervous activity.* Bulletin ofMathematical Biophysics, 1943. **5**: p. 115-133.
108. Haykin, S., *Neural Netwroks - A comprehensive Foundation*. Second ed. 2005, India: Pearson Prentice Hall.
109. Qian, N. and T.J. Sejnowski, *Predicting the Secondary Structure of Globular Proteins Using Neural Network Models* Journal of Molecular Biology, 1988. **202**: p. 865-884.
110. Rost, B. and C. Sander, *Prediction of Protein Secondary Structure at Better than 70% Accuracy.* Journal of Molecular Biology, 1993. **232**: p. 584-599.
111. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning internal representations by error propagation*. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, ed. D.E. Rumelhart and J.L. McClelland. Vol. 1. 1986, Cambridge MA: MIT Press.
112. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Research, 1997. **25**(17): p. 3389–3402.
113. Przybylski, D. and B. Rost, *AlignmentsGrow,SecondaryStructurePrediction Improves.* PROTEINS: Structure, Function, and Genetics, 2002. **46**: p. 197-205.

114. Bruni, R., *Mathematical Approaches to Polymer Sequence Analysis and Related Problems*. 2011, New York: Springer.
115. Zhang, W., A. Keith Dunker, and Y. Zhou, *Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks.* Proteins 2008. **71**(61-67).
116. Aydin, Z., Y. Altunbasak, and H. Erdogan, *Bayesian Protein Secondary Structure Prediction With Near-Optimal Segmentations*, in *IEEE Transactions on Signal Processing*. 2007. p. 3512-3525.
117. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine Learning, 1995. **20**(3): p. 273-297.
118. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. 2nd ed. 2009: Springer.
119. Mezghani, D.B.A., S.Z. Boujelbene, and N. Ellouze, *Evaluation of SVM Kernels and Conventional Machine Learning Algorithms for Speaker Identification* International Journal of Hybrid Information Technology 2010. **3**(3): p. 23-34.
120. Yang, Z.R., *Biological applications of support vector machines.* BRIEFINGS IN BIOINFORMATICS, 2004. **5**(4): p. 328-338.
121. Guo, J., et al., *A Novel Method for Protein Secondary Structure Prediction Using Dual-Layer SVM and Profiles.* PROTEINS: Structure, Function, and Bioinformatics 2004. **54**: p. 738 –743.
122. Ward, J.J., et al., *Secondary structure prediction with support vector machines.* Bioinformatics, 2003. **19**(13): p. 1650-1655.
123. PDB. *Protein Data Bank*. 2014 [cited 2014 September 27]; Available from: http://www.rcsb.org/pdb/secondary.do?p=v2/secondary/search.jsp#AdvancedSearch.
124. Wang, G. and R.L.D. Jr, *PISCES: recent improvements to a PDB sequence culling server.* Nucleic Acids Res. , 2005. **33**: p. 94-98.
125. NCBI, *Using BLASTClust to Make Non-redundant Sequence Sets*, in *NCBI* 2014.
126. Cuff, J.A. and G.J. Barton, *Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction.* PROTEINS: Structure, Function, and Genetics, 1999. **34**: p. 508-519.
127. Meiler, J., et al., *Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks.* Journal of Molecular Modeling, 2001. **7**(9): p. 360-369.
128. NCBI. *BLAST*. 2015 [cited 2015; Available from: http://blast.ncbi.nlm.nih.gov/Blast.cgi.
129. Sharma, A., et al., *A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition.* Journal of Theoretical Biology, 2013. **320**: p. 41-46.
130. Iqbal, S. and M.T. Hoque, *DisPredict: A Predictor of Disordered Protein from Sequence using RBF Kernel.* Tech. Report TR-2014/1, 2014.
131. Marsh, J.A., *Buried and Accessible Surface Area Control Intrinsic Protein Flexibility.* Journal of Molecular Biology, 2013. **425**(17): p. 3250 - 3263.
132. Zhang, H., et al., *On the relation between residue flexibility and local solvent accessibility in proteins.* Proteins, 2009. **76**(3): p. 617 - 36.
133. Lee, B. and F. Richards, *The interpretation of protein structures: estimation of static accessibility.* J Mol Biol, 1971. **55**(3): p. 379-400.
134. Connoly, M., *Solvent accessibility surfaces of protein and nucleic acids.* Science, 1983. **221**: p. 709 - 713.
135. Cheng, J. and P. Baldi, *Three-stage prediction of proteinβ-sheets by neural networks, alignments and graph algorithms.* Bioinformatics, 2005. **21**(1): p. i75–i84.
136. Iqbal, S., A. Mishra, and M.T. Hoque, *Improved Prediction of Accessible Surface Area Results in Efficient Energy Function Application.* Tech. Report TR-2015/1, 2015.
137. Zhang, T., E. Faraggi, and Y. Zhou, *Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction.* Proteins, 2010. **78**(16): p. 3353–3362.

138. Martin, J., et al., *Protein secondary structure assignment revisited: a detailed analysis of different assignment methods.* BMC Structural Biology, 2005. **5**(17).
139. Heinig, M. and D. Frishman, *STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins.* Nucleic Acids Research, 2004. **1**(32): p. W500–W502.
140. G.Labesse, et al., *P-SEA:anewefficientassignmentof secondarystructurefromCOLtraceofproteins.* CABIOS, 1997. **13**(3): p. 291-295.

Vita

The author was born in Jessore, Bangladesh. He obtained his Bachelor's degree in 2008 from Bangladesh University of Engineering and Technology. He joined the University of New Orleans computer science graduate program in 2013 and worked as a research assistant under Dr. Md Tamjidul Hoque of the UNO computer science department, working on the protein secondary structure prediction project as part of his computer science thesis.