



INTERNATIONAL UNIVERSITY OF SARAJEVO

AVICENNA: SEQUENCE SEGMENTS SIMILARITY BASED PROTEIN SECONDARY STRUCTURE PREDICTION METHOD

A THESIS SUBMITTED TO
THE FACULTY OF ENGINEERING AND NATURAL SCIENCES
OF
INTERNATIONAL UNIVERSITY OF SARAJEVO

BY

FARUK BERAT AKÇEŞME

STUDENT ID: o1210459

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
GENETICS AND BIOENGINEERING

JUNE 2016

Approval of the Faculty of Engineering and Natural Sciences

Prof Dr. Fuat GÜRCAN

Dean

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

Assoc. Prof. Dr. Mohamed Ragab Abdel GAWWAD

Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. Mehmet CAN

Mentor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

Prof. Dr. Kasim BAJROVIĆ

University of Sarajevo, Institute for Genetic Engineering and Biotechnology and Faculty of Science

Prof. Dr. Mehmet CAN

International University of Sarajevo, Faculty of Engineering and Natural Sciences

Assoc. Prof. Dr. Sabina SEMIZ

International University of Sarajevo, Faculty of Engineering and Natural Sciences

Assoc. Prof. Dr. Mohamed Ragab Abdel GAWWAD

International University of Sarajevo, Faculty of Engineering and Natural Sciences

Assoc. Prof. Dr. Ayla ARSLAN

International University of Sarajevo, Faculty of Engineering and Natural Sciences

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Faruk Berat AKÇEŞME



INTERNATIONAL UNIVERSITY OF SARAJEVO

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF
UNPUBLISHED WORK

Copyright © 2016 by International University of Sarajevo

AVICENNA: A SEQUENCE SEGMENTS SIMILARITY BASED PROTEIN
SECONDARY STRUCTURE PREDICTION METHOD

No part of this unpublished work may be reproduced, stored in a retrieval system, or transmitted, in any form or by any mean, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder and IUS Library.

Abstract

AVICENNA: A SEQUENCE SEGMENTS SIMILARITY BASED PROTEIN SECONDARY STRUCTURE PREDICTION METHOD

Faruk Berat Akçeşme

June, 2016

Understanding the tertiary structure of proteins can shed light in protein function. Therefore, protein structure determination has become one of the hot topics in structural biology and computational biology. Large-scale sequencing techniques generate a vast amount of protein sequences but number of structurally analyzed proteins still remains much smaller due to difficulties and cost of the experimental processes such as X-ray crystallography and NMR spectroscopy.

Prediction of protein secondary structure based on its amino acid sequence is a common technique used to predict protein's three dimensional structure. In this thesis, the problem of secondary structure prediction is analyzed and several aspects of secondary structure prediction are investigated. An efficient new algorithm "Avicenna" for predicting secondary structure is presented and it's superiority to the current methods is shown.

It is well known that even for two distant proteins, it is possible to find similar segments in amino acid sequences, and the corresponding secondary structure segments are mostly much more similar. Completeness of secondary structures in the current Protein Data Bank (PDB) library is examined with it's around 110.000 proteins whose three dimensional structures are solved for the prediction of the secondary structures of

proteins. To address this issue, we employ a comprehensive set of 80,592 non redundant proteins of length larger than 30 residues derived from PDB databank. We mostly find similar segments to a query segment of the amino acid sequence with levels of similarity ranging from full-match to 60% mismatches. Since corresponding secondary structure segments in average have much more higher matches, by the use of consensus of these secondary structure segments, it is possible to predict secondary structure of the query protein with an accuracy of 85% in average. Our results showed that the protein secondary structure prediction problem can in principle be solved based on the current PDB library by developing efficient similarity capturing and information aggregation techniques that can successfully predict secondary structures of unknown proteins.



To my daughter Elif...

Acknowledgements

I would like to express my sincere gratitude to my advisor Prof. Mehmet Can for the continuous support during my PhD study, for his encouragement, patience and motivation. I thank to him for the systematic guidance and his unconditional support. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would like to thank the Thesis Following Committee members Assoc. Prof. Dr. M. Ragab Abdel Gawwad and Assoc. Prof. Dr. Ayla Arslan and Defense Committee Chair, Prof. Dr. Kasim Bajrović and the member Assoc. Prof. Sabina Semiz. I would like to thank them for their valuable suggestions and comments to make this work a complete.

Finally, I wish to express my profound gratitude and sincere thanks to my wife. Most importantly, none of this would have been possible without her patience. I would like to thank my parents for supporting me morally throughout the writing of process of this thesis.

Lastly, it should be mentioned that all mistakes existing in this work, entirely belong to the author.

Table of Contents

Abstract	V
Acknowledgements	VIII
List of Figures	XIII
List of Tables	XV
List of Abbreviations.....	XVI
 CHAPTER I INTRODUCTION.....	 1
1.1. Proteins and Some Fundamental Concepts	2
1.1.1. Proteins.....	3
1.1.2. Why Study Proteins?.....	3
1.1.3. A Short Historical Perspective of Proteins.....	4
1.1.4. Amino Acids	5
1.1.5. Physicochemical Properties of Amino Acids.....	8
1.2. Protein Structure.....	10
1.2.1. Protein Structural Levels.....	11
1.2.1.1. Protein Primary Structure.....	13
1.2.1.2. Protein Secondary Structure.....	13
1.2.1.2.1. Properties of Alpha-Helices	16
1.2.1.2.2. Properties of Beta Sheets.....	18
1.2.1.2.3. Turns.....	19
1.2.1.3. Protein Tertiary Structure.....	20
1.3. Central Dogma	20
1.4. Protein Folding.....	21
1.5. Anfinsen`s Theory.....	24
1.6. Protein Similarity	26
1.7. Protein Sequence Similarity and Protein Secondary Structure Prediction..	27

1.8.	Experimental Determination of Protein Structures	29
1.8.1.	Determination of Protein Primary Structure (Protein sequencing)	29
1.8.2.	Protein Secondary Structure Determination.....	29
1.9.	Protein Secondary Structure Prediction Methods	31
1.9.1.	Chou-Fasman Method	33
1.9.2.	GOR Method.....	34
1.9.3.	Lim Method.....	35
1.9.4.	Machine Learning	36
1.9.4.1.	Predicting the Secondary Structure of Proteins Using Artificial Neural Networks	36
1.9.4.2.	Support Vector Machine Applications.....	38
1.9.5.	Using PSI-BLAST for Protein Secondary Structure Prediction	38
1.10.	Available Protein Secondary Structure Predictors	43
1.11.	Assessment of the Prediction Methods	45
1.12.	Protein Sequence Similarity Analyses and Sequence Alignment Methods	46
1.12.1.	Global and Local Alignment.....	47
1.12.2.	Pairwise Alignment	47
1.12.3.	Dot Matrix Method.....	48
1.12.4.	Dynamic Programming	48
1.12.4.1.	Needleman-Wunch Algorithm	49
1.12.4.2.	Smith–Waterman Algorithm	50
1.12.5.	Word Methods.....	50
1.12.6.	Multiple Sequence Alignment.....	52
1.13.	Using Sequence Similarity for the Prediction of Secondary Structure of Proteins.....	53
1.14.	Rationale and Objectives.....	55
1.14.1.	Significance of the Study	56
1.14.2.	Research Objectives	56
1.14.3.	Research Hypothesis	57

CHAPTER II THE METHOD	58
2.1. Motivation of the Method	59
2.1.1. Protein Database and Protein Sequence Similarity	60
2.1.2. Testing the Completeness of Protein Database for Using in Protein Secondary Structure Prediction	62
2.1.2.1. Super-Chain Identification	63
2.1.2.2. Secondary Structure Prediction of the Query Protein Using Super Chains	64
2.2. Protein Secondary structure Prediction Based on Similar Segments of Proteins in PDB	66
2.2.1. Deviation in Secondary Structure Lagging Considerably Behind the alterations in the Primary Structure	69
2.3. Introduction to the MAX-MATCH	72
2.3.1. Sequence Similarity Measures for Finding Similar Proteins and Protein... Segments	72
2.3.1.1. Sequence Similarity Measures	72
2.3.1.2. An Efficient Hamming Distance Based Measure for Finding Similar Proteins that Contain Similar Sequence Segments	76
2.3.2. A new Protein Secondary Structure Prediction Method: Avicenna	79
2.3.2.1. Reduction of the Search Space	83
2.3.2.2. Similar Segment Identification	83
2.3.2.3. Aggregation of the Information Obtained at Several Mismatch Levels.	89
2.3.2.4. Prediction of the Average Mismatches of the Secondary Structure Segments	91
CHAPTER III RESULT AND DISCUSSION	95
3.1. Structures are More Conserved Than Sequences	96
3.1.1. Full Secondary Structure Matches vs. Low Amino Acid Matches	99

3.2.	The Performance of Avicenna Protein Secondary Structure Prediction Method	104
3.2.1.	Reduction Strategies for the Search-Space	105
3.2.2.	Filtering Strategies.	106
3.2.2.1.	The performance of Avicenna Protein Secondary Structure Prediction Method on Filtered Search Space.....	107
4.	CONCLUSION	112
5.	REFERENCES	114



List of Figures

Figure 1.1.4-1 A skeletal model of a generalized amino acid. The amino group (blue), a carboxyl group (red) and an R group (green) are attached to a central carbon.....	6
Figure 1.2.1-1 Representations of the hierarchical organizations of protein structures: the secondary structure, super-secondary structure, motif, alpha-helix, beta-strand, alpha-hairpin, beta-hairpin and four-helical bundle. (Blundell and Campillo, 2014).....	12
Figure 1.2.1.1-1 The representation of atomic positions in primary structure of proteins (Blundell and Campillo, 2014).....	13
Figure 1.2.1.2-1 Ramachandran diagram of the torsion angles that specify protein backbone conformation. The alpha-helix and beta-strand regions are labeled. Data as in Lovell et al. 2003 Proteins 50:437, showing about 100,000 data points for general amino-acid types (Lovell et al., 2003)	15
Figure 1.2.1.2.1-1 Representation of alpha helix, Gray: hydrogen, Purple: R group, Orange: oxygen, Blue: Hydrogen, Black: Carbon, Green: Peptide bond (primary structure of protein, 2014)	17
Figure 1.2.1.2.2-1 A: Parallel, B: antiparallel and C: mixed beta sheets. The dashed lines indicate main chain hydrogen bonds between the beta strands.....	19
Figure 1.8.2-1 The eight structure classes' reduction into the three classes of helix (H), sheet (E), and coil (C), a) DSSP, eight types of secondary structure, b) DSSP, three types of secondary structure	30
Figure 1.9.5-1 Steps of PSI-BLAST search.....	40
Figure 1.9.5-2 BLOSUM 62 matrix (Henikoff S, and Henikoff J, 1992).....	42
Figure 1.9.5-3 Strategy of using PSSMs in protein secondary structure prediction by machine learning	43
Figure 1.12.5-1 k-letter word list generation.....	52
Figure 2.1.1-1 Growth of unique folds per year as defined by SCOP v1.75 (Murzin et al., 1995)	60
Figure 2.1.2.1-1 From the secondary structure sequence of the super-chain (protein A), the subsequence of the same length as the query protein is extracted. This subsequence is taken as the predicted secondary structure of the query (protein B).	64
Figure 2.2.1-1 Plateau that is generated by several thousand proteins shows the relation between structure and sequence.....	70

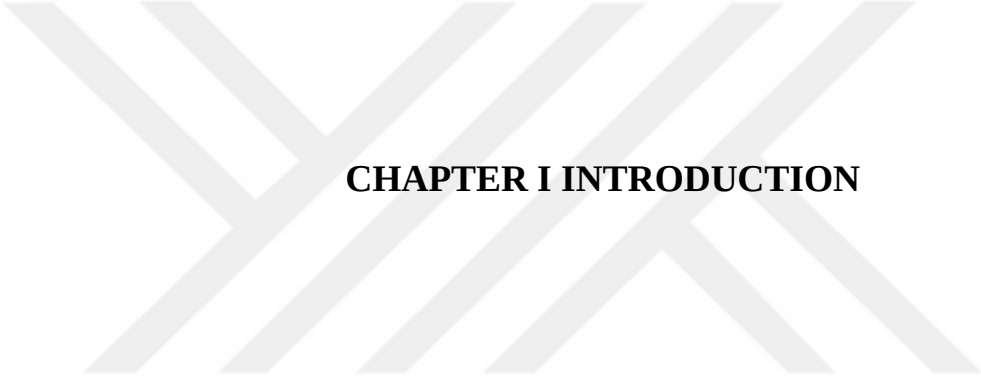
Figure 2.3.1.2-1 Short protein moves along to the long protein. The similarity score of MAX-MATCH is counted for each step. The maximum score is assigned as a similarity score.	78
Figure 2.3.2-1 Short protein moves along the long protein, the hamming distance between sequences are recorded in each step of movement. The previous score of hamming distance does have any effects to the following score since no gaps are introduced between residues.....	82
Figure 2.3.2.2-1 Segment length 17 and a mismatch level 9 is fixed. Then from the amino acid sequence of the query protein (PROTEIN A) a segment is taken from the prescribed length. Amino acid sequences of all proteins in the PDB database except the query protein (PROTEIN B).....	85
Figure 2.3.2.2-2 Each segment of the query sequence (Q) is taken as a motif. This motif and each of the segments of PDB proteins is subjected to the similarity analysis. Based on their MAX MATCH score they are placed to the proper mismatches (MM) baskets. This is repeated until the last segment of the query protein. In this case, the query protein has similar segments with proteins in PDB database where similar segments are 4, 5 and 9 mismatches.....	86
Figure 2.3.2.2-3 Empty column vectors whose length is the same as the length of the query are filled with the corresponding segments taken from the basket. This step is repeated for the each mismatch baskets. In the end, 9 mismatch baskets are emptied to the empty column.....	88
Figure 2.3.2.3-1 Each consensus mismatches column vectors are used to assign secondary structure to the query. Priority is given to the d-1 levels of mismatch according to their predicted secondary structure mismatches. The final decision vector is filled by the prediction with smallest average mismatches, and if there remain indecisive, these positions are filled by the level with the prediction with second smallest average mismatches, and so on.	90
Figure 3-1 Images 2SCP_B and 2HQ8_A are generated by Swiss-PdbViewer (Guex and Peitsch,1997).....	100
Figure 3-2 2 SCP_B and 3OX6_A are generated by Swiss-PdbViewer (Guex and Peitsch,1997)	102
Figure 3-3 2 SCP_B and 2K7C_A are generated by Swiss-PdbViewer (Guex and Peitsch,1997)	103
Figure 3-4 Protein Secondary Structure Predictors Performance Comparison, Q3 accuracies in percentage.....	105
Figure 3-5 Frequencies of prediction accuracies for the test set CB513.	108

List of Tables

<i>Table 1.1.5-1 Proteinogenic amino acids and their physicochemical properties*</i>	9
<i>Table 1.1.5-2 Essential, nonessential and conditional amino acids.</i>	9
<i>Table 1.10-1 Name of the famous secondary structure prediction methods and their descriptions and initial release data*</i>	44
<i>Table 1.11-1 Type of the most known protein secondary structure prediction methods</i>	46
<i>Table 2.1.1-1 In PDB the query implementation for removing similar sequences is based on pre-calculated clusters of protein chains*</i>	61
<i>Table 2.1.2.2-1 The percentage of random sample proteins of batch size 1000, to have a super-chain in the database, and their mean accuracies of predicted secondary structure.</i>	65
<i>Table 2.1.2.2-2 Summary of the accuracies of secondary structure prediction using the secondary structure sequences of host proteins</i>	66
<i>Table 2.3.2.4-1 Summary of the results obtained for protein 1Q2U of the 699 dataset.</i>	92
<i>Table 2.3.2.4-2 Summary of the results obtained for protein 1BROA of the 1189 dataset</i>	93
<i>Table 2.3.2.4-3 Summary of the results obtained for protein 1PKYA: 1 – 69 of the 1189 dataset.</i>	93
<i>Table 3-1 Probabilities in rows are conditional probabilities of mismatches of query secondary structure sequence segments of length 17, with and secondary structure sequence segments of same length cut from the protein</i>	98
<i>Table 3-2 Summary of the accuracies of secondary structure prediction on filtered dataset by the first filter using the secondary structure sequence segments of target proteins.</i>	107
<i>Table 3-3 Summary of the accuracies of secondary structure prediction on filtered dataset by the second filter using the secondary structure sequence segments of target proteins</i>	108
<i>Table 3-4 Summary of the accuracies of secondary structure prediction on filtered dataset by the third filter using the secondary structure sequence segments of target proteins</i>	109

List of Abbreviations

<i>3D</i>	<i>Three dimensional</i>
<i>ANNs</i>	<i>Artificial Neural Networks</i>
<i>BLAST</i>	<i>The Basic Local Alignment Search Tool</i>
<i>BLOSUM</i>	<i>Blocks Substitution Matrix</i>
<i>CASP</i>	<i>The Critical Assessment of protein Structure Prediction</i>
<i>CATH</i>	<i>Protein Structure Classification Database at UCL</i>
<i>D</i>	<i>Gate size</i>
<i>DNA</i>	<i>Deoxyribonucleic acid</i>
<i>DSSP</i>	<i>Hydrogen bond estimation algorithm</i>
<i>EVA</i>	<i>Evaluation of protein structure prediction servers</i>
<i>h</i>	<i>Basket label</i>
<i>HMMs</i>	<i>Hidden Markov Models</i>
<i>IUPAC</i>	<i>International Union of Pure and Applied Chemistry</i>
<i>k-NNs</i>	<i>k-Nearest Neighbors</i>
<i>m</i>	<i>Motif</i>
<i>MM</i>	<i>Mismatch</i>
<i>mRNA</i>	<i>Messenger ribonucleic acid</i>
<i>MSA</i>	<i>Multiple Sequence Alignment</i>
<i>NCBI</i>	<i>National Center for Biotechnology Information</i>
<i>PAM</i>	<i>Point accepted mutation</i>
<i>PDB</i>	<i>Protein Databank</i>
<i>prot</i>	<i>Non-redundant PDB dataset</i>
<i>PSI-BLAST</i>	<i>Position-Specific Iterative Basic Local Alignment Search Tool</i>
<i>PSSM</i>	<i>position-specific scoring matrix</i>
<i>q</i>	<i>Query</i>
<i>Q3</i>	<i>Three-state overall per residue accuracy</i>
<i>RNA</i>	<i>Ribonucleic acid</i>
<i>SCOP</i>	<i>Structural Classification of Proteins</i>
<i>SCR</i>	<i>structurally conserved region</i>
<i>SVMs</i>	<i>Support Vector Machines</i>
<i>t</i>	<i>Target</i>
<i>tRNA</i>	<i>Transfer ribonucleic acid</i>
<i>w</i>	<i>Window size</i>



CHAPTER I INTRODUCTION

1.1. Proteins and Some Fundamental Concepts

Molecular biology is defined as a field of biology that studies the structures, compositions, functions and interactions of cellular molecules (Lodish, 2008). Specifically, molecular biology is interested in the relationships between genes and their functional counterparts, proteins. These relations are described in several processes in the forms of central dogma which are valid for all forms of life. A biological process by which DNA is copied to RNA is called transcription. Transcribed RNA is used to produce proteins by another process called translation. These two important processes are the basic mechanisms of transferring biological information from cell to the divided cells which permit living organisms to exist and replicate. Transfer of biological information is possible with the help of specialized proteins inside the cells. A protein, as a complex molecule, consists of simpler components, amino acids, in the form of the chain which is commonly referred as the primary structure. Each amino acid and its position in the chain are determined by the genes located in the nucleus. The primary structure contributes to the determination of the protein's three-dimensional structure which allows a protein to perform its functions. A part of the relationship between linear amino acid sequences and three-dimensional structures of proteins is a result of energy minimization of the protein's total energy (Albert et al., 2013). This relation is subjected to the one of the most fundamental questions of molecular biology that cannot be simply answered with the current knowledge of biology or with the involvement of mathematical, chemical and physical sciences. Understanding the relationships between protein structural levels require interdisciplinary studies that employ computational techniques, biological knowledge, physics and chemistry. Computational biology; a new

branch of biology uses computational techniques and bioinformatics; creates tools to solve biology problems.

1.1.1. Proteins

Proteins are large, complex molecules that play many critical roles in living organisms. Proteins do most of the works in cells. Catalyzing metabolic reactions, responding to stimuli, transporting molecules from one location to another, binding, and switching are some of the important functions of the proteins. Moreover, proteins also serve as major structural elements of living systems. Particular recognition of different molecules is vital to protein functional capacity. Specific binding is represented by shape complementarities and polar interactions (Berg et al., 2002).

All chemical reactions in the living systems are needed to be catalyzed. Enzymes, as protein molecules mostly undertake this task. They have remarkable efficiency and can accelerate reactions as much as 17 orders of magnitude over simple buffer catalysis (Petsko and Ringe, 2004). Protein structures and their conformations are flexible to ligand binding and changes in pH. This flexibility enables proteins to play the role of a switcher to control cellular processes (Petsko and Ringe, 2004).

1.1.2. Why Study Proteins?

Proteins underpin every aspect of cellular and metabolic activities. Their structures and functions have an impact on a human endeavor such as medicine, nutrition, agriculture, and energy.

Treatment of various diseases, developing new disease detection methods and drug design highly depend on the level of structural and functional knowledge of proteins. To solve protein deficiency problems in organisms, to increase protein quality of the nutrition and to design complementary foods are other endeavors which will benefit from the improvement in protein science.

Moreover, proteins are important energy sources for most of the organisms. In the 21st century, proteins are successfully converted into biofuel and they started to be considered promising energy source for industrial applications.

1.1.3. A Short Historical Perspective of Proteins

Although there is a vast accumulation of knowledge of proteins, surprisingly the term protein has been introduced only 180 years ago (Hartley, 1951). Gerhardus Johannes Mulder work's on the composition of animal substances in 1839 is considered as the first study related to the proteins (Hartley, 1951). He showed the presence of carbon, hydrogen and oxygen in fibrin (Whitford, 2013). Progress in biochemistry enabled scientists to analyze the composition of structures of molecules. In 1873, Heinrich Hlasiwety and Josef Habermann stated the first time that proteins are made up of amino acids (Hunter, 2000). After 30 years, another protein scientist in protein history, Franz Hofmeister, proposed that polypeptides are amino acids linked by peptide bonds for the first time (Rosenfeld, 2012). In some sources in the literature, he is mentioned as the first scientist who crystallized a protein (Baldwin, 1996).

Other studies, including Emil Fisher on enzymes (Kunz, 2003), and Pauling's description of the geometry of the peptide bonds deserved to be mentioned as some of the landmarks in the early period of protein science (Mason, 1997).

Studies on proteins accelerated by the discoveries in chemistry and physics due to many developments in those disciplines were rapidly applied to the protein structural and functional analysis. In 1958, the first atomic structure of protein, myoglobin was determined by the use of X-ray by Max Perutz and John Kendrew (Mondal et al., 2015) and this was followed by explorations of the structures of larger proteins such as hemoglobin and lysozyme. Following these developments of protein structures, knowledge about protein molecules is extended. Remarkable advance in modern biochemistry and molecular biology provided a large amount of theoretical information to understand protein structure and function. Large scale study of protein's functions and structures started to called proteomics which refers to the studies that are considering entire set of proteins (Chandrasekhar et al., 2014).

1.1.4. Amino Acids

Amino acids are organic compounds containing amine (-NH_2) and carboxylic acid (-COOH) functional groups, usually along with a side-chain specific to each of them. Amino acids play central roles as building blocks of proteins and being intermediates in metabolisms ("Amino acid," 2015). Since proteins consist of amino acids, their chemical properties determine the biological activities of the proteins. Carbon, hydrogen, oxygen and nitrogen are the key elements of amino acids. R groups of amino

acids at the side chains make amino acids unique, since the other functional groups are the same for all amino acids. In an amino acid structure, the carbon atom attached to the carboxyl group is called alpha carbon. Some amino acids contain an amino group. In this type of amino acids, the amino group is bonded directly to the alpha carbon. This type of amino acids called alpha amino acids. A skeletal model of a generalized amino acid is shown in Figure 1.1.4-1.

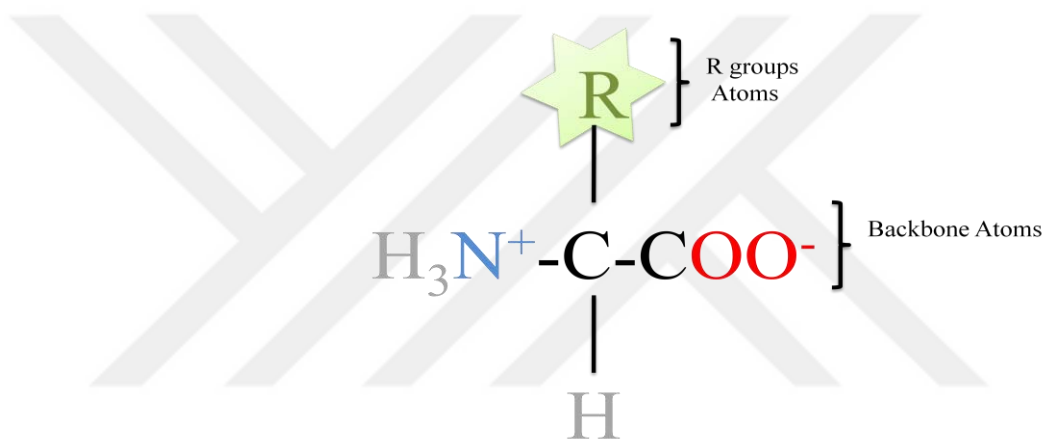


Figure 1.1.4-1 A skeletal model of a generalized amino acid. The amino group (blue), a carboxyl group (red) and an R group (green) are attached to a central carbon.

There are many ways of classifying amino acids , some of them are as follows: based on the core structural and functional groups' locations as alpha- (α -), beta- (β -), gamma- (γ -) or delta- (δ -) amino acids; or related to the physicochemical properties like polarity, pH level, and side-chain group types (aliphatic, acyclic, aromatic, containing hydroxyl or sulfur, etc.) (Wagner and Musso, 1983).

Amino acids are the second abundant constituent of human cells after the water molecules (Lathman, 1997). In addition to their role in forming proteins, amino acids

play many other critical roles in biological processes such as neurotransmitter transport and biosynthesis of some molecules. There are twenty-two amino acids incorporated into the polypeptide chain. These amino acids are called proteinogenic amino acids.

1.1.4.1. Proteinogenic Amino Acids

Proteinogenic amino acids represent a very small fraction of all known amino acids as precursors to proteins. They are also known as protein-building amino acids which are attached to each other to form polypeptides chain of proteins (O'Connor, 2010). There are 22 proteinogenic amino acids where twenty of them are encoded directly by triplet codons (Ochre, 2014). Three of them can be incorporated into the protein structures by exceptional translational machinery. Proteinogenic amino acids are further classified into the essential, non-essential, and a conditional group depending on their place of synthesis.

There are nine essential proteinogenic amino acids which cannot be synthesized in the human body. Nonessential amino acids can be synthesized by the mammalian organisms. Their presence does not depend on the nutrition uptake. Conditional amino acids are not considered as essential amino acids except in times of illness and stress. Both proteinogenic and non-proteinogenic amino acids have a very crucial role in many biological metabolisms.

In some biological processes, amino acids standard form present as the main actor in performing a specific function, whereas their non-standard forms may be seen in some other cases. For instance, the standard amino acid glycine is used to synthesize

porphyrin which is used in red blood cells as a pigment. The hydroxyproline synthesized from proline is the major component of the connective tissue collagen. Glutamic acid's role in the human brain as main excitatory neurotransmitters can be mentioned as one of the other examples of the important role of glutamine in the non-standard form (Sapolsky, 2005).

1.1.5. Physicochemical Properties of Amino Acids

Amino acids can be grouped based on their physicochemical properties as well. Most important physicochemical properties of amino acids are charge, hydrophilicity, hydrophobicity, and size (Creighton, 1993). Some amino acids including glycine (Gly), alanine (Ala), valine (Val), leucine (Leu), isoleucine (Ile), proline (Pro), phenylalanine (Phe), methionine (Met), and tryptophan (Trp) are hydrophobic amino acids which tend to be repelled from water. Proteins tend to have their hydrophobic residues buried in the middle of its structure. There are also some amino acids with R groups which have hydrophilic propensities. They can expose to the aqueous solvent. Amino acids can be positively charged. Proteins containing positively charged amino acids are enabled to bind to the negatively charged molecules. Lysine and Arginine are positively charged amino acids. Cystine is one of the unique amino acids with its special properties to form covalent disulfide bonds. Another amino acid, Proline is unique with forming a cycle to the polypeptide backbone. Some of the physicochemical properties of amino acids are shown in Table 1.1.5-1.

Table 1.1.5-1 Proteinogenic amino acids and their physicochemical properties*

Amino acid	Side Chain Polarity	Side Chain Charge (pH 7.4)	Hydropathy Index	Molecular Weight
Alanine	Nonpolar	neutral	1.8	89
Arginine	Basic polar	positive	-4.5	174
Asparagine	Polar	neutral	-3.5	132
Aspartic acid	Acidic polar	negative	-3.5	133
Cysteine	Nonpolar	neutral	2.5	121
Glutamic acid	Acidic polar	negative	-3.5	147
Glutamine	Polar	neutral	-3.5	146
Glycine	Nonpolar	neutral	-0.4	75
Histidine	Basic polar	positive (10%) neutral (90%)	-3.2	155
Isoleucine	Nonpolar	neutral	4.5	131
Leucine	Nonpolar	neutral	3.8	131
Lysine	Basic polar	positive	-3.9	146
Methionine	Nonpolar	neutral	1.9	149
Phenylalanine	Nonpolar	neutral	2.8	165
Proline	Nonpolar	neutral	-1.6	115
Serine	Polar	neutral	-0.8	105
Threonine	Polar	neutral	-0.7	119
Tryptophan	Nonpolar	neutral	-0.9	204
Tyrosine	Polar	neutral	-1.3	181
Valine	nonpolar	neutral	4.2	117

*Side Chain Polarity (Cooper and Hausman, 2004), Side Chain Charge (Cooper and Hausman, 2004), Hydropathy Index (Freifelder, 1983), Molecular Weight (Lehinger et al., 2000).

Table 1.1.5-2 Essential, nonessential and conditional amino acids.

Essential	Nonessential	Conditional
Histidine	Alanine	Arginine
Isoleucine	Aspartic acid	Cysteine
Leucine	Glutamic acid	Glutamine
Lysine	Selenocysteine	Glycine
Methionine		Proline
Phenylalanine		Serine
Threonine		Tyrosine
Tryptophan		Asparagine
Valine		

1.2. Protein Structure

Proteins were not recognized as a class of biological molecule until the eighteenth century. After Antoine Fourcroy observation which is the ability of molecules to coagulate or flocculate under treatments with acid or heat, proteins are referred as a distinct class of biological molecules (Osborne, 1911). Discovery of albumin from egg whites, blood serum albumin, and wheat gluten contribute to introducing this class as a distant biological class.

The historical belief, “flesh makes flesh” led scientist to think proteins as the most important nutrient for maintaining the structure of the body. Identification of glutamic acid by Karl Heinrich Ritthausen extends the knowledge on proteins in that direction. Later on, in early 19th century, proteins are described as polypeptides in the Franz Hofmeister's and Hermann Emil Fisher's works (Wieland and Bodanszky, 2012). Enzymes are started to be called as proteins in living organisms with the James B. Sumner's work that describes enzyme, urease (Carey, 2006).

Generally, a polypeptide chain that contains less than 40 residues is referred as a peptide. Larger polypeptide chains are called proteins (Brocchier and Karlin, 2005). Proteins fold into one or more specific conformations in order to perform their specific biological functions. Protein Folding is a consequence of intermolecular forces including pure ionic interactions, dipole interactions, hydrogen interactions, van der Waals forces and hydrophobic interactions. Three-dimensional structures of proteins need to be known in order to identify molecular functions of proteins. Understanding and analyzing the structures of molecules and their functions give a birth to a new area in biological

sciences called structural biology. The structures of molecules, including proteins, are analyzed by some techniques including X-ray crystallography, NMR spectroscopy, and dual polarization interferometry. Various techniques and latest technology enable structural biologists to obtain many molecular structures of proteins and protein databases get larger and larger day by day in last decades.

The structure of a protein is generally decomposed into a primary structure, secondary structure, tertiary structure, and quaternary structure. The terms primary, secondary, tertiary and quaternary first used by Kaj Ulrik Linderström-Lang in 1951 (Linderstrøm-Lang, 1952). The main elements of protein structures are provided by secondary structural elements.

Since biological functions of proteins highly depend on its three-dimensional structures, analysis or prediction their structures became a very hot task.

1.2.1. Protein Structural Levels

Proteins fold in three dimensions and their structure is organized hierarchically. The basic structure of the protein is called primary structure when its polypeptide chain is in linear form. The highest structure is referred as a quaternary structure where more than one polypeptide chains are present in a complex protein.

There is also a supersecondary structure which is combinations of alpha helixes and beta sheets connected with loops. This complex type of secondary structure form patterns which are called as motifs of the proteins. Sometimes the term “motif” is used to define

a supersecondary structure. There are several structural motifs in proteins which often include loops of variable length and unspecified structures. Some of the motifs names are as follows: beta-hairpin, beta-barrel, Leucine-rich repeat, Greek Key, Omega Loop, Helix-Loop-Helix and Zinc finger (Chiang, 2007).

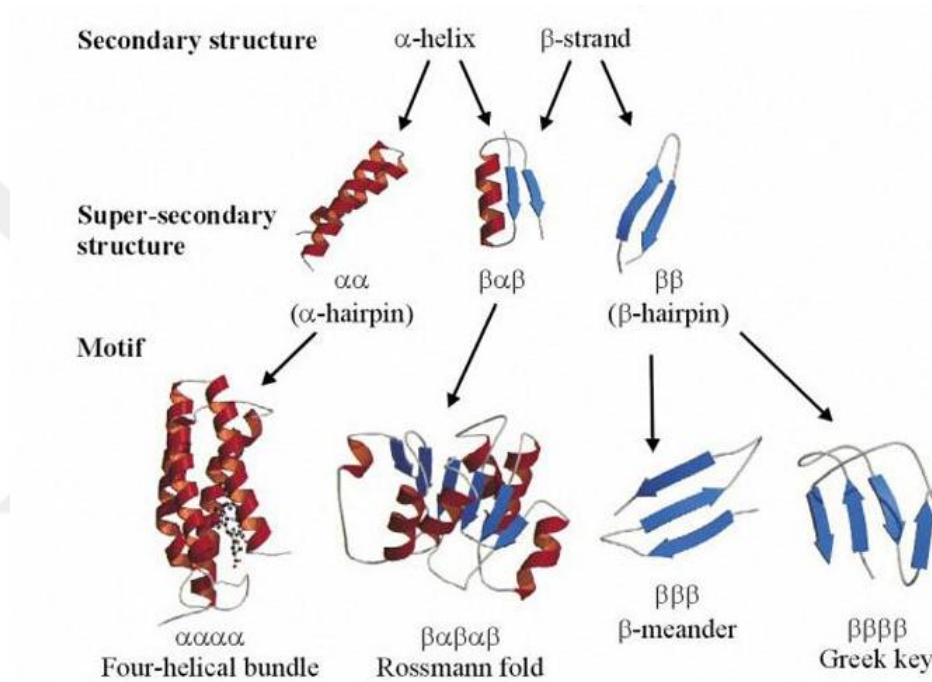


Figure 1.2.1-1 Representations of the hierarchical organizations of protein structures: the secondary structure, super-secondary structure, motif, alpha-helix, beta-strand, alpha-hairpin, beta-hairpin and four-helical bundle. (Blundell and Campillo, 2014)

Another important term for the structure of proteins is a domain. Structural domains are used to describe conserved part of given protein sequence which can function and exist independently of the rest of the protein (Linderstrøm-Lang, 1952) A protein may have more than one domains and a domain may appear in a variety of different proteins. Folded domains can serve as modules for building up large assemblies such as muscle fibers or cell wall components. Domains may exist in the different sizes. 90% of the

domains have less than 200 residues. (Islam, 1995) and the average of domain size is approximately 100 residues (Wheelan et al., 2000).

1.2.1.1. Protein Primary Structure

The primary structure of a protein is the linear sequence of its amino acid units. As its name refers, amino acids contain both a basic amino group and an acidic carboxyl group; this feature of amino acids allows the individual amino acids to join together into a long chain. Since this long chain consists of amino acids which are attached to each other with peptide bonds, this chain referred as polypeptide chain. Polypeptide chain has free groups on their both ends. These free groups give names for the ends as carboxyl terminus and amino terminus.

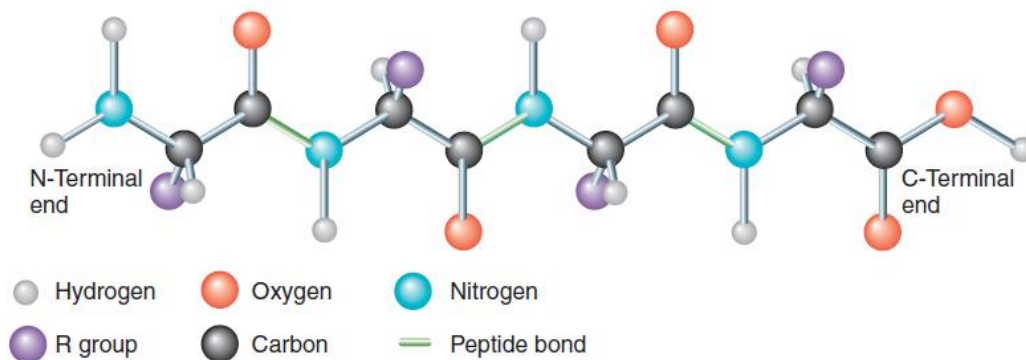


Figure 1.2.1.1-1 The representation of atomic positions in primary structure of proteins (Blundell and Campillo, 2014)

1.2.1.2. Protein Secondary Structure

In biochemistry and structural biology, the secondary structure is the general three-dimensional form of local segments of biopolymers. The secondary structure of proteins is stretches of proteins that have distinct atomic positions or distinct characteristic local

structural conformations, depending on positions of hydrogen bonds. Two possible stable conformations are defined as an alpha helix and a beta strand/sheet. These conformations are formed by hydrogen bonds that make conformations stable between the residue chains. Some parts of proteins are less stable and they have less regular folds. Less stable parts generally named as a coil or a loop and they have no regular shape. The secondary structure of a protein may also define as a consecutive fragment of regular and irregular structures. Mostly protein's secondary structures consist of all these three conformations but in many cases, proteins can have different propensity to an alpha helix or beta sheets. Distribution of conformations varies according to the type of proteins.

The secondary structure of proteins illustrates the structure of small regions of the peptide rather than describing specific atomic positions in three-dimensional space. There is a standard method, DSSP algorithm, for assigning secondary structure to the amino acids of a protein whose atomic-resolution coordinates of proteins are given.

One very useful form of visualizing energetically allowed region for backbone dihedral angles ψ (psi) against ϕ (phi) of amino acid residues in protein structure using a Ramachandran plot. It is developed by G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan in 1963 (Ramachandran et al., 1963). Ramachandran plot can be generated with a single protein sequence to represent its secondary structure. It may also use a database of many structures to represent propensities of amino acids for being in certain secondary structure conformation. Many studies have taken advantage of PDB database which contains thousands of proteins to produce more detailed and accurate ϕ, ψ plots

(Morris et al., 1992; Kleywegt and Jones, 1996; Hoof, 1997; Hovmöller, 2002; Lovell, 2003; Anderson et al., 2005; Ting et al., 2010)

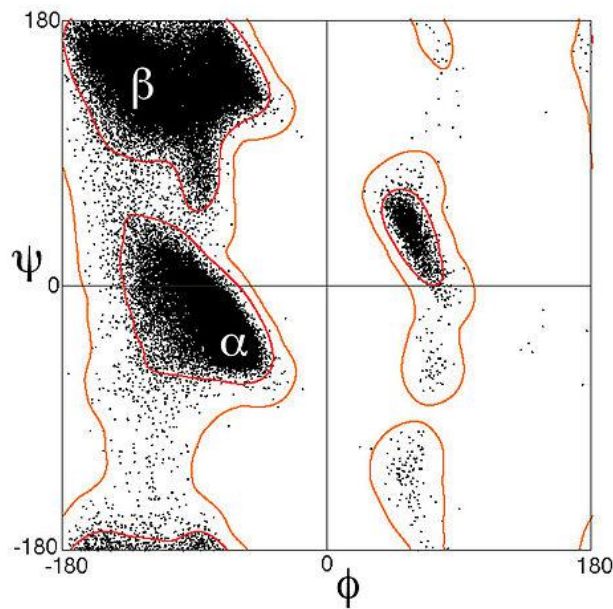


Figure 1.2.1.2-1 Ramachandran diagram of the torsion angles that specify protein backbone conformation. The alpha-helix and beta-strand regions are labeled. Data as in Lovell et al. 2003 Proteins 50:437, showing about 100,000 data points for general amino-acid types (Lovell et al., 2003)

Unlike nucleic acids, where secondary structure is defined by the hydrogen bonding between the nitrogenous bases, the secondary structure of proteins is defined by patterns of hydrogen bonds between backbone amides and carboxyl groups. However, the other structural features of secondary structure play the role in forming irregular hydrogen bonding. For instance, residue in protein helices adopt backbone dihedral plot in a specific region of the Ramachandran plot; along these lines, a portion of deposits with such dihedral edges is regularly called a "helix", paying little heed to whether it has the right hydrogen bonds.

1.2.1.2.1. Properties of Alpha-Helices

Alpha helices are compact, versatile and cylindrical structures. They are stabilized by hydrogen bonds between C=O and N-H groups in polypeptide chains. In the amino acid polypeptide chain, each residue (n) accepts hydrogen bonds from the amide nitrogen. In an alpha helix, the N-H group of an amino acid forms a hydrogen bond with the C=O group of the amino acid four residues earlier. The first and the last four residues in the helix do not form hydrogen bonds. The first residue in alpha helices is bonded to the following residue by N-H group. In the end of the alpha helix, the C=O group is positioned between the last two residues. The rest of residues between these two edges that are bonded by amide groups are hydrogen bonded.

In alpha helix structure, each turn is offset from the ones above and below, such that the N-H and C=O groups are brought into line with each other, this conformation resulted in a cylindrical shape. In the outside of the cylindrical structure of alpha helices there are protruding side chains that determine the interaction of the alpha helices. There is almost no free space within the helix, since α -helix is tightly packed. Depending on the side chains of the residues in alpha helices, helix may have interactions with the other folded protein chains that allow to protein accomplish its function.

The alpha helix has approximately -60 phi values and -50 psi values between its residues. Depending on the direction of the composition of the alpha helices, they are named as clockwise (right-handed) or counterclockwise (left-handed) spiral staircase. The alpha helices have no practical limit on their length. Hundreds of angstroms long alpha helices have been observed in some proteins. The length of the alpha helices and

protein stability are correlated. Typical helix contains about ten amino acids but the length can range from four to over forty residues (Shiga, 2005). All amino acids may exist in the alpha helices but the frequency of amino acids being in the alpha helix differs.



Figure 1.2.1.2.1-1 Representation of alpha helix, Gray: hydrogen, Purple: R group, Orange: oxygen, Blue: Hydrogen, Black: Carbon, Green: Peptide bond (primary structure of protein, 2014)

There are many statistical analyses on the amino acid propensity of being in secondary structure conformations. Some studies have shown that proline and glycine have fewer tendencies to be in the alpha helices. These studies indicated that due to the rigid side chain of the proline, alpha helices do not include proline with high frequency (Petsko and Ringe, 2004). In some cases, proline exists in the first helix turn where its geometry does not affect the hydrogen bonds. The rare existence of glycine in the alpha helices

causes problems since glycine has a single hydrogen atom which is insufficient to protect hydrogen bonds inside the helix (Petsko and Ringe, 2004).

1.2.1.2.2. Properties of Beta Sheets

Beta sheet is a form of the regular secondary structure of proteins which consists of beta strands. Beta strands are connected to each other with hydrogen bonds. Parallel or antiparallel beta strands are connected laterally and they generate a plated sheet. Unlike the alpha helix, beta sheet contains hydrogen bonds between backbone groups from residues that are distant from each other in the linear sequence. It is formed by hydrogen bonds between protein strands, rather than within a strand.

Beside regular strand, some strands can lie up both parallel and antiparallel. If it is the case, a sheet called mixed sheet.

The parallel sheet is generally small and buried, whereas antiparallel sheets frequently exposed to aqueous solutions. Since antiparallel beta sheets contain hydrogen bonds, they are considered as more stable structures. In the following figure 1.2.1.2.2-1 parallel, antiparallel and mixed sheets are illustrated.

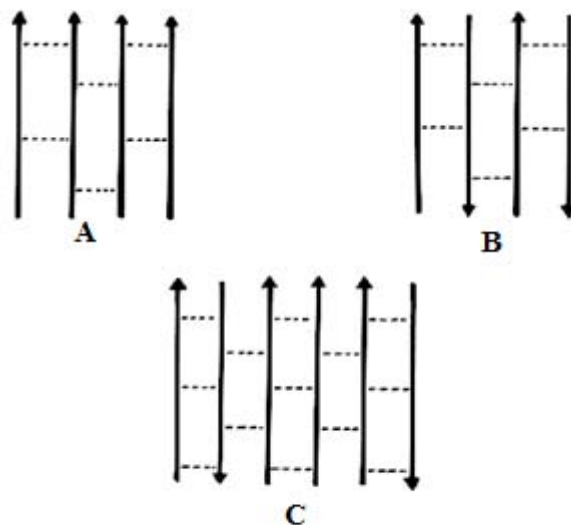


Figure 1.2.1.2.2-1 A: Parallel, B: antiparallel and C: mixed beta sheets. The dashed lines indicate main chain hydrogen bonds between the beta strands.

Successive side chains repeatedly point straight up, then straight down, in a fully extended β -strand. In many cases, beta sheets are not perfectly extended. Due to the chirality of their component amino acids, they may exhibit a twist which is associated with alternating fluctuation. The amino acids side chains that form the beta sheets point outward from the folds; this enables the structure to alternate its faces.

In beta sheets, the polypeptide chain is extended; it is though those amino acids which have a branch at the beta carbon can be accommodated more easily in a beta sheet.

1.2.1.2.3. Turns

Turns are not regular conformations but considered as a secondary structure element in proteins where the polypeptide chain reverses its overall direction. Turn occurs where $C\alpha$ atoms of two residues separated by few, one to five, peptide bonds and while the residues do not form regular secondary structure elements (Rose, 1985).

1.2.1.3. Protein Tertiary Structure

The tertiary structure of proteins has single polypeptide chain that is called backbone with single or multiple protein secondary structures. The tertiary structure may consist of a single domain or multi-domain structures, where a domain is defined as a part of the tertiary structure that can function and exist independently. The geometric shape of the tertiary structure is determined by interactions and bonds between its side chains. Atomic coordinates of proteins are used to determine either entire protein tertiary structure or a protein domain structure (IUPAC, 1997).

Proteins are composed of amino acids using information encoded in corresponding genes. Each protein has its own unique amino acid sequence. Amino acids in the sequences are encoded by a set of three-nucleotides which are called codons. Codons are the combinations of 4 nucleotides which designate amino acids. For instance three nucleotides, adenine-uracil and guanine are the code for one of the amino acids; methionine is the starting codon for all amino acid sequences. Some amino acids are specified by more than one codon, since nucleic acid sequence contains four nucleotides and the total number of the possible codons is 64.

1.3. Central Dogma

Central dogma is an explanation of the flow of genetic information from genome to a functional molecules that is, proteins. It was first stated by Francis Crick who is also a co-discoverer of the structure of the DNA molecule together with William Cochran and Vladimir Vand in 1952 (Cochran, 1952). The genetic information carried in genes in DNA first transcribed into mRNA by RNA polymerase. With the help of post-

transcriptional modification mechanism, mRNA converted into the mature mRNA. This important biological process called transcription. Mature mRNA is used as a template for protein synthesis by the complex molecular machine called ribosome. Ribosome links amino acids together in an order specified by messenger RNA (mRNA) molecule. The process that synthesizes protein from an mRNA template called translation. In this stage, mRNA is loaded onto the ribosome and is read three nucleotides at a time by matching each codon to its base pairing anticodon. Transfer RNA molecule is responsible for carrying the amino acid to its corresponding codon. Transfer RNA is charged by the enzyme called aminoacyl tRNA synthetase in order to match correct amino acid with the corresponding triple codon. Proteins are always synthesized from N terminus to C terminus. The amino acid chain started to be released from the ribosome and grow until the last triple codon is translated into the corresponding amino acids. This growing polypeptide chain is often referred as a nascent chain (Alberts, 2002).

Protein synthesis mechanism, as the last step of the central dogma, is described in four stages; initiation, elongation, translocation, and termination. Polypeptide chain folding starts for acquiring its native three-dimensional structure during to the translocation and ends up after termination stage. This is a millisecond phenomenon whose details are still remains unclear for the time being.

1.4. Protein Folding

Proteins are synthesized by a complex process called translation. A particular sequence is dictated by messenger RNA and synthesized by a large protein-nucleic acid complex

called a ribosome. In the meantime, polypeptide chain folds into three-dimensional structures to perform its function. Protein folding occurs just after translation. It is though that polypeptide chain of proteins has very crucial properties that limit protein to fold into a specific three-dimensional structure. The peptide bonds that tides two amino acids together behaves like double bounds since it keeps two amino acids in a fixed plane, no rotation is possible about the peptide bond itself. The only possible rotation for adopting different conformation is a rotation of the fixed planes of peptide bonds. As has already been described in the previous sections, there are two bonds namely $C\alpha$ -amino nitrogen bond and $C\alpha$ -carbonyl carbon bond in the amino acid structure. The values of two torsion angles, Φ angles, and Ψ which describe rotations of the polypeptide backbone are discrete and bounded since the backbone or side chain atoms would come close to each other where there is no balance in the charge of atoms. This would generate unstable structures.

Even values of Φ and Ψ angles are discrete and bounded; theoretically, a chain may fold into many conformations. For instance, if the values of Φ and Ψ angles are limited to only eight combinations, n -residue-long peptide would potentially have 8^n conformations. If we consider ten residue long polypeptides, about 8.6 million possible conformations would be formed. It is observed in the nature that only one or a few very closely related conformation is functional. This most stable functional conformation state is called native state of protein. In the native state, proteins are in the lowest free energy conformations.

Protein folding is assisted by other proteins called chaperones and chaperonins. They are found in all organisms from bacteria to humans. In many different organisms their mechanisms are almost identical (Lodish, 2000). Chaperones are generally located in every cellular compartment and organelles (Mathews, 2012). They bind to the proteins in which they will assist its folding. Their main function is to prevent aggregation of unfolded proteins since unfolded or partially folded protein has the intent to aggregate into water in the form of insoluble masses. All newly synthesized proteins have aggregation potential thereby chaperones are essential molecules for correct protein folding.

Chaperonins are the other important class of proteins for forming correct conformation. They are big cylindrical macromolecules and consist of two rings of oligomers. With these two rings, chaperonins can have tight peptide binding state and a relaxed peptide releasing state.

Aggregation of proteins can cause many health problems including prion-related illnesses such as Creutzfeldt-Jakob disease, bovine spongiform encephalopathy (mad cow disease) and amyloid-related illnesses such as Alzheimer's disease (Hammarström, 2003). Aggregated proteins can also cause intracytoplasmic aggregation diseases such as Huntington's and Parkinson's disease (Selkoe, 2003).

Any health problem caused by aggregation or misfolding of proteins called proteinopathies in medicine (Walker, 2000). Although, these diseases are extensively studied, some of their mechanisms still remain unclear. Analyzes of the secondary

structure of health related proteins will contribute to the understanding of their mechanism.

1.5. Anfinsen`s Theory

Anfinsen`s Experiment is a study on the relationship between a sequence of proteins and their three-dimensional shapes. Many different attempts including chemical, physical and statistical studies have been performed in order to explore this relation. Today, it is still one of the core subjects in protein structure science. The well-known works of Christian Anfinsen on the enzyme ribonuclease analyzed the mentioned relation in the 1950s and he came up with a theory that the native structure is determined by the protein's amino acid sequence, at least for small globular proteins (Anfinsen, 1972).

Ribonuclease is a single polypeptide chain consists of 124 amino acid residues. This chain is cross-linked by 4 disulfide bridges. The experiment`s motivation was to understand what conditions are required to restore the three-dimensional structure of proteins that are destroyed into a polypeptide chain (Berg, 2002).

He watched over the enzymatic activity of ribonuclease during his experiment for understanding how protein regains its structural properties.

Anfinsen effectively disrupted the non-covalent bonds between the residues with the help of urea and cleaves the disulfide bonds between cysteine residues by a reagent such as β -mercaptaethonol. He succeeded to obtain a polypeptide chain of ribonuclease protein which is in the form of random coils.

Anfinsen's observation was very critical for explaining the effect of primary structure on folding of proteins. Denatured ribonuclease slowly regained its enzymatic activity when it became oxidized by air, refolding was observed into a catalytically active form. Physicochemical properties of the chain are observed identical in refolded protein with those of native enzyme. Since 1950, this study still keeps its fascination to explain what information is needed for forming the active structure of proteins. Sequence depended on conformation is significant to reveal a connection between conformation and function of proteins.

The native structure of proteins, the functional three-dimensional structure, is the folded stage of proteins that possesses a minimum of free energy (Creighton, 1990). The state of lowest energy of the whole system called Gibbs Free energy which is determined by the totality of interatomic interactions of the residues in protein's polypeptide chain (Anfinsen, 1972). The validity of this observations and definitions are highly related to the environmental conditions of proteins called normal physiological milieu which includes constant temperature, solvent, pH, and ionic strength, the presence of metal ions or prosthetic groups.

Almost five decades past over Anfinsen's experiment. The number of solved proteins is enormous. Does today's knowledge of proteins support the theory? The facts revealed by this thesis work give a partial answer to this question

1.6. Protein Similarity

Although, the number of different species live on the Earth is not known, today, 2.1 million species have been classified in a systematic way. Most of the invertebrates remain as unclassified species (Archibald, 2009).

Living organisms are classified in a hierarchy by a Swedish naturalist Linnaeus in the 18th century. In the history, biologists had many attempts to classify organisms into different categories mostly by judging degrees of apparent similarity and difference that they can observe. “Greater the degree of physical similarity, the closer the biological relationship” principle was used in the early classification approaches. A static view of this assumption was questioned by the middle of the 19th century by a small number of radical naturalists, most notably Charles Darwin. His explanation for diversity which is largely a result of branching evolution has a big impact on classification strategies. Later on, the comparison of DNA sequences of genes and entire genome offer an interesting insight into the possible evolutionary relatedness which could be used to classify a living organism. The evolutionary relationship started to be used in phylogenetics trees which are first presented in the book *Elementary Geology*, by Edward Hitchcock (Hitchcock, 1860).

In the 21st century, scientists give more efforts to introduce similarity between DNA and protein sequences since similar biological functions are seen in organisms which have similar sequences. Today, phylogenetic trees are still in use. DNA and protein sequences are attached to its branches rather than different species.

Sequence similarity is a measure of an empirical relationship between sequences (Heringa, 2008). Although it is a well-accepted term and used for a long time, it still remains as an intuitive concept. Formulating a good similarity measure is a vital task to drive biological information derived based on similarity search.

Protein sequence similarity may be considered as a measurement to analyze the homology as well. Homolog sequence can be used to deduce a function and structure to the unknown proteins. In similarity analyses of proteins it is wished to (i) measure their similarity, (ii) determine the residue-residue correspondences, (iii) observe patterns of conservation and variability and (iv) infer evolutionary relationships (Kertész-Farkas, 2008).

1.7. Protein Sequence Similarity and Protein Secondary Structure Prediction

Protein sequence alignment is a task of identifying the similarity between two sequences. Defining similarity between protein sequences is also found to be very important for predicting protein structures. The similarity is a consequence of functional, structural and/or evolutionary relationship between two sequences (Pearson, 2013). Almost all protein pairs have similar segments in their sequences. They are called conserved sequences when similarity is seen across multiple species. Similar segments are generally identified in somehow related proteins; but they might also be seen among not closely related sequences. In evolutionary biology, a similarity in distantly related species is defined as convergent evolution which describes the independent evolution of a similar feature (Pagel, 1999). There are a few reports of convergent evolution on the

molecular basis since it is much more complicated to recognize it in sequence level. In chapter III, it is shown that how the similarity in amino acid sequences is more pronounced in secondary structures. Chothia and Lesk traced this phenomenon in tertiary structure. They called the collection of such regions the 'common core' (Chothia and Lesk, 1986). The effects of the alterations in amino acid sequences on the secondary and tertiary structures of even diverse proteins can be revealed by the comparative analysis of the structures of related proteins. For many years effect of residue alterations in amino acid sequences on the higher level of structures investigated with several sequence alignment algorithms which assume that sequences have a common origin and divergences have been widened by time due to the mutations. This assumption leads traditional alignment methods to consider sequence alignment as a variation of the string edit distance problem (Cao, 2010). Gaps are inserted between the residues so that identical or similar characters are aligned to edit the distance which is corresponding to mutations as insertion or deletion in a biological understanding. The justification of the introduction a gap into the sequence which triggers a penalty score that reducing the total score remains to be questionable. To avoid this drawback for the identification of structurally related positions in a collection of amino acid sequences, an efficient similarity measure “MAX-MATCH” for indentifying similar proteins which have common similar segments is presented in chapter II. Benefits of this similarity measure on the newly introduced secondary structure prediction method “Avicenna” accuracy are described in chapter II.

1.8. Experimental Determination of Protein Structures

Determination of protein structures determination by experimental methods is a challenging task due to the varying degrees of proteins stability. Structural determination of proteins is more difficult in the higher structural levels of proteins since there are more complex biochemical reactions and in the higher structural level of proteins.

1.8.1. Determination of Protein Primary Structure (Protein sequencing)

There are several techniques to determine the amino acid sequence of a protein. Most popular direct methods are mass spectrometry and Edman Degradation reaction. If the DNA sequence or mRNA sequence encoding specific proteins are known, it would be possible to generate an amino acid sequence from that.

1.8.2. Protein Secondary Structure Determination

The rough secondary structure content of a biopolymer can be estimated by the interaction between matter and electromagnetic radiation (Pelton and McLean, 2000). The most known method for assigning secondary structure to proteins is hydrogen bond estimation algorithm. This method is known as the dictionary of protein secondary structure (DSSP). It begins by identifying the intra-backbone hydrogen bonds of the protein using a purely electrostatic approach (Kabsch and Sander, 1983).

In DSSP, eight secondary structure types appear in a secondary sequence; H(α -helix), G(310-helix), I(π -helix), E(β -strand), B(isolated β -bridge), T(turn), S(bend), and -

(rest). The eight structure classes are usually reduced to three classes of the helix (H), sheet (E), and coil (C) (Rangwala and Karypis, 2011).

a	E	E	E	E	C	C	C	C	T	T	T	T	G	G	G	G	H	H	H	H	H	S	S	S	S	T	T	C	C	C
b	S	S	S	S	C	C	C	C	C	C	C	C	H	H	H	H	H	H	H	H	H	S	S	S	S	C	C	C	C	C

Figure 1.8.2-1 The eight structure classes' reduction into the three classes of helix (H), sheet (E), and coil (C), a) DSSP, eight types of secondary structure, b) DSSP, three types of secondary structure

The major experimental methods used in three-dimensional protein structure determinations are X-ray Crystallography and Nuclear Magnetic Resonance (NMR). X-ray is a more widely used method, around 90% of the protein structures available in the Protein Data Bank (PDB) have been determined by X-ray crystallography (Berman, 2000).

Besides X-ray and NMR, another method for structural analyses of proteins is an electron microscope. Each of these three methods has its advantages and limitations.

Crystallography begins with a crystallizing a protein. Proteins can be prompted to form crystal when they become supersaturated. In this form, protein molecules held together by non-covalent interaction and these interactions allow protein molecules to be packed in a repeating array (Rhodes, 2006).

After crystallization, molecules are placed in a monochromatic (single wavelength) X-ray beams. Molecules repeatedly exposed to the X-ray beam, while changing its rotation, each exposure provides an image which represents the diffraction of the beam (Rhodes, 2006). Once data have been obtained, the rest of the method is performed by

sophisticated computer programs that analyze the generated data. Most of the proteins can be crystallized but some cannot be crystallized for X-ray crystallography. A general problem in protein crystallization is to find optimum crystallization conditions. These conditions include a combination of a right pH, ionic strength, temperature, and protein concentration, the presence of various salts, ligands or additives (McPherson, 2004). This could be count as one of the biggest limitation of the method. This limitation can be overcome with using electron microscope which does not require the protein to be crystallized. Although electron microscope seems useful for protein structural analyses, resolution is generally limited in comparison to the resolution obtained by crystallography. Nuclear Magnetic Resonance gives high resolution but requires much larger amounts of material which is not possible to provide in all cases.

Although, many methods are available to determine a protein structure, they are still very costly and time-consuming. For most of the studies, protein structure analyses remain as one of the biggest challenges.

1.9. Protein Secondary Structure Prediction Methods

Secondary structure prediction of proteins is an effort to deduce a secondary structure for proteins from their amino acid sequences, using insight obtained from already known structures. Since Anfinsen's theory, it is believed that conformation of proteins is determined predominantly by their amino acid sequence. Many efforts have been devoted predicting the secondary structure of a protein from its amino acid sequence up to date.

Since Pauling (Pauling, 1951) suggested that proteins are formed by certain local conformational patterns like helices and strands, many methods have been proposed to predict the secondary structure of proteins with a satisfactory accuracy (Ptitsyn, 1969; Pain and Robson, 1970; Robson and Pain, 1971). Different attempts can be put into the following 5 groups: (1) statistical information (Nagano, 1973; Chou & Fasman, 1974; Nagano and Hasegawa, 1975; Garnier et al., 1978; Schulz and Schirmer, 1979; Levin et al., 1986; Gibrat et al., 1987; Biou et al., 1988; Kanehisa, 1988; Levin and Garnier, 1988; Fasman, 1989; Garrett et al., 1991; Muggleton et al., 1992); (2) physicochemical properties (Lim, 1974; Ptitsyn and Finkelstein, 1983); (3) Sequence patterns (Cohen et al., 1983, 1986; Taylor and Thornton, 1983; Rooman et al., 1989, 1991; Sternberg and King, 1990; Rooman and Wodak, 1991; Presnell et al., 1992) (4) multi-layered (or neural) networks (Bohr et al., 1988, 1990; Qian and Sejnowski, 1988; Holley and Karplus, 1989; Bossa and Pascarella, 1990; Kneller et al., 1990; Hirst and Sternberg, 1992; Maclin and Shavlik, 1993; Stolorz et al., 1992; Zhang et al., 1992); and (5) evolutionary conservation (Maxfield and Scheraga, 1979; Zvelebil et al., 1987; Frampton et al., 1989; Benner and Gerloff, 1990; Barton et al., 1991; Niermann and Kirschner, 1991; Ouzounis and Melvin, 1991; Musacchio et al., 1992; Russell et al., 1992; Gibson et al., 1993).

Although there are some attempts before the 1970s, (Guzzo, 1965; Prothero, 1966; Schiffer and Edmundson, 1967; Kotelchuck and Scheraga 1969) no prediction method was introduced which includes three main elements of the secondary structure. During 1970s more accurate prediction methods with statistical assessments based on probability parameters derived from known solved structures was proposed. Most

remarkable one was the Chou and Fasman method. They announced that they are able to predict the secondary structure of proteins with up to 80% accuracy (Chou and Fasman, 1974). However, several years later, Kabsch and Sander showed that their result was overestimated due to the usage of the small size of data (Kabsch and Sander, 1983). So now, statistical methods used to predict a secondary structure are at the most about 60-65% accurate (Mount, 2004) which is not enough considered to be of satisfactory accuracy.

1.9.1. Chou-Fasman Method

The very first secondary structure prediction algorithm which relies predominantly on probability parameters was developed by Peter Y. Chou and Gerald D. Fasman in 1970s (Chou and Fasman, 1974, 1978). A set of known protein structures which comes by X-ray crystallography analysis was used to calculate the relative frequencies of each amino acid in alpha helices, beta sheets and turns (coils).

The Chou-Fasman algorithm consists of several simple steps. The first step is assigning an appropriate set of parameters to all residues in the amino acid sequence. The following step is scanning to the entire amino acid sequence and identifying regions where 4 out of 6 contiguous residues have a high probability of being in an alpha helix or three of five amino acids gave a high probability of being in a β -strand. The probability of residues being in a secondary structure units in both directions determines the type of structural units until the prediction value drops below a specific value. The region in the query sequence which contains overlapping alpha helices and beta sheets are taken to be helices in the case that the average probability of being in alpha is greater

than the beta sheet probability for that region. It is a beta sheet if the average of the probability of being in the beta strand is greater than the probability of being in an alpha helix. This approach allowed Chou-Fasman method to predict the secondary structure of a set of proteins roughly with 50-60% accuracy.

1.9.2. GOR Method

GOR method was developed by three scientists; Garnier J, Osguthorpe DJ, and Robson B. in 1978 (Garnier et al., 1978). Similar to Chou Fasman, GOR method is based on probability parameters derived from empirical studies of known protein tertiary structures solved by X-ray crystallography. GOR method is more rigorous than Chou-Fasman-like propensities with its calculation of the probability of a given amino acids. Unlike Chou Fasman method, GOR method considers the given secondary structure element in its propensity calculations. This method is using a principle of information theory and calculates probability values for specific amino acids based on the neighboring amino acids up to eight residues ahead. This approach is more sensitive and accurate than Chou and Fasman method (Garnier et al., 1978). It is suggested that weak contribution of all neighbors can end with strong effect to the performance of prediction. GOR method was more accurate for alpha helices than a sheet and its success is approximately 65%, till the study alpha helices are frequently mispredicted as loops or disorganized regions (Lewis, 1970). In the other hand, beta sheets were frequently predicted as loops or disorganized regions. GOR method has been updated many times. The most remarkable version of the update of the method, GOR V, is based on the current database of 513 domains recommended by Cuff and Barton (Cuff, 1999). The

average Q3 score of GOR V is 71% which is 5% less accurate than the neural-network based methods of PhD and PSIPRED (Sen et al., 2005).

1.9.3. Lim Method

Secondary structure prediction problem was also thought to be solved by stereochemical rule-based approaches (Mona, 2001); one of the techniques of that type was developed by Lim (Lim, 1974). Compared to statistical methods, Lim method considers a longer range of interactions between residues.

If the protein sequence is $r_1, r_2, \dots, r_N$, then for the i th residue, the following pairs and triples are considered particularly important for helical regions: (r_i, r_{i+1}) , (r_i, r_{i+3}) , (r_i, r_{i+4}) , (r_i, r_{i+1}, r_{i+4}) , (r_i, r_{i+3}, r_{i+4}) . Residues three and four apart are considered, as they lie on the same face of an α -helix. Similarly, the pair (r^j, r^{j+2}) contains residues on the same face of the beta strand.

Pairs and triplets of particular amino acids are then deemed as compatible or incompatible with helices and strands based on their physicochemical properties such as size, hydrophobicity, charge and its ability to form hydrogen bonds. For instance, if the query sequence has a residue with a hydrophobic feature in every three to four residues, it is considered to be an alpha helix, since it is known that one side of the helix is hydrophobic.

1.9.4. Machine Learning

Machine learning is a field of computer science that evolved from the computational learning theory and pattern recognition (“Michalski,” 2013). Since 1959, it is widely used in many different disciplines. Neural networks, as one of the most important application of machine learning, use learning algorithms that are inspired by the understanding of how the brain learns. Today it is used in solving many biological problems and understanding biological processes including predicting the structure and function of complex biological molecules. The tremendous increase in biological data accelerates the usage of machine learning tools in biology.

1.9.4.1. Predicting the Secondary Structure of Proteins Using Artificial Neural Networks

In many biological problems, collected biological data is enough big to be used as a training data to map specific input signals to the desired output. Artificial Neural Network (ANN) is used in protein secondary structure prediction as well. In the particular secondary structure prediction problem, generally sliding windows with 13-17 residue sequences are considered as input (Chatterjee, 2011).

The general approach of using a neural network in secondary structure prediction is modifying the central amino acid of each input by a weighting factor. Each of the possible secondary structures is represented as an output. Information is summed and sent to a hidden layer. In the hidden layer, signal is transformed into a number close to either 1 or zero. Output close to zero indicates that the corresponding unit is not

predicted whereas output close to 1, secondary structure unit is indicated prediction (Mount, 2001).

In the training process of a neural network, values of the weights are modified by a set of sequences with known secondary structures. The program is optimized with weight values where most of the residues are correctly predicted.

The first remarkable usage of a neural network in secondary structure prediction was introduced by Rost and Sander ((Rost and Sander, 1993). They name their algorithms as PHD, which is trained on a test set of profiles of multiple sequences (Rost and Sander, 1993). They claimed 72.2% overall accuracy.

Recently, accuracy obtained by using neural network was shown to be increased to about 75% with using larger databases and PSI-BLAST (Jones, 1999).

In his study, David Jones used a two-stage neural network to predict protein secondary structure. He used the benefits of sensitive sequence comparison based on the new gapped version of BLAST published by Altschul in 1997 (Altschul, 1977). Indeed, the PSI-BLAST algorithm uses a standard Smith- Waterman search which has the ability to detect distant homologous sequences. When PSI-BLAST is running for finding similarities to the query, it generates an intermediate PSI-BLAST profile. Rather than extracting the sequences, Jones used this intermediate profile as a direct input. Jones algorithm is known by its online software tool, PSIPRED (McGuffin, 2000). The Q3 success of PSIPRED was evaluated as 73.4 %. It is claimed in its latest version that; PSIPRED 3.2 achieves an average Q3 score of 81.6% (PSIPRED, 2016).

1.9.4.2.Support Vector Machine Applications

Support Vector Machines (SVM) is supervised learning systems with associated learning algorithms that use a hyper space of linear or non-linear functions in a high dimensional feature space (Vapnik and Cortes, 1995). Learning algorithms are used to classify the data and analyze a regression since its first introduction by Vapnik and co-workers. It outperformed most other systems in a wide variety of applications including some biological applications such as protein function prediction, protease functional site recognition, transcription initiation site prediction and gene expression data classification (Yang, 2004).

SVMs have been shown to be superior to the early PHD prediction method which is using neural networks. Hua and Sun quoted a Q3 score of 73% with their secondary structure prediction method based on SVMs (Hua and Sun, 2001). Their SVMs use a quadratic kernel function to map the data to a feature space. This space represents pair interactions between features and it increases particularly the accuracy of helix prediction. The probability estimates are also used to combine binary classifiers in SVMs systems. As it is used in neural networks, PSI-BLAST profiles are also used in SVMs in training process with the latest sequence database, and the overall accuracy of Q3 was reached to 75% (Guo, 2004).

1.9.5. Using PSI-BLAST for Protein Secondary Structure Prediction

Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) derives a profile from multiple sequence alignment of sequences (Altschul 1997). It uses a protein-protein BLAST algorithm to collect similar sequences with a given score

threshold and generate matrix called position-specific scoring matrix (PSSM). This matrix is used to further search the database. When new similar sequences are found, matrix is updated for subsequent iterations. The PSSM is used as an input to train machine learning and prediction performed based on that.

Steps of PSI-BLAST

This program is used to find distant relatives of a protein. First, a list of all closely related proteins is created. These proteins are combined into a general "profile" sequence, which represents significant features present in these sequences. A query against the protein database is then run using this profile, the other more distant proteins are found and a larger group of proteins are used to construct another profile. Described process is repeated until no similar proteins are found.

By including related proteins in the search, PSI-BLAST is much more sensitive in finding a distant relationship than a standard protein-protein BLAST.

The length of PSSM equals the length of the query sequence; the query sequence is taken as a template for the alignment whereas identical sequences are discarded. In most of the studies, there are no identical sequences to the query sequence since the dataset that are used to generate PSSMs are non-redundant (Jones, 1999). PSI-BLAST continues until no new proteins with defined E-value are found. Since the conservation is usually indicative of the formation of repetitive motifs such as the secondary structures, this information was found useful in structure prediction of proteins.

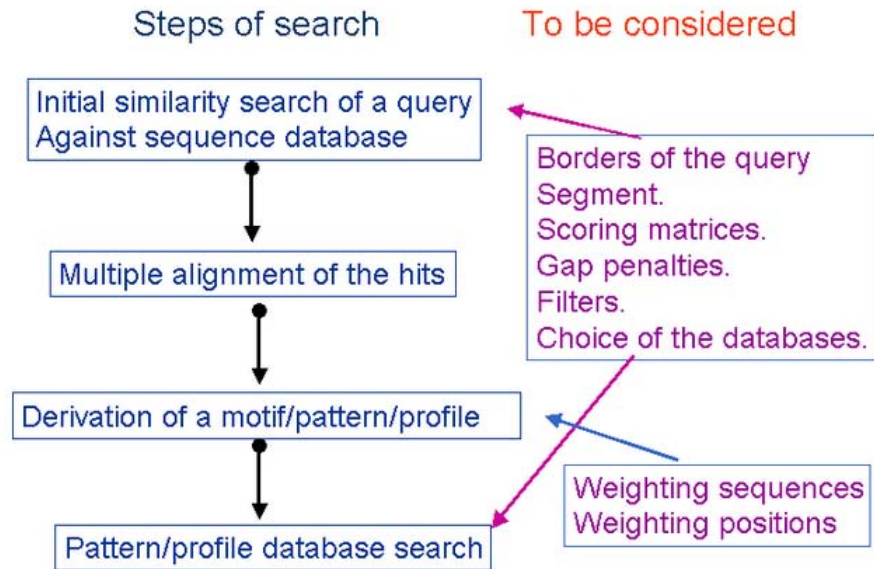


Figure 1.9.5-1 Steps of PSI-BLAST search

Position Specific Scoring Matrix (PSSM)

The position weight matrix was introduced by American geneticist Gary Stormo and colleagues in 1982 (Stormo et al., 1982). PSSM has found a good alternative to the consensus sequence. Consensus sequences had previously been used to represent patterns in biological sequences but had difficulties in the prediction of new occurrences of these patterns.

Advantages of using PSSM can be listed as follows:

- Weights sequence according to observed diversity specific to the family of interest
- Minimal Assumption
- Easy to compute
- The PSSM indicates whether a given residue in the query sequence is conserved

PSI-BLAST uses BLOcks SUBstitution Matrix (BLOSUM Matrix)

The BLOSUM (BLOcks SUBstitution Matrix) matrix is a substitution matrix used for sequence alignment of proteins. BLOSUM matrices are used to score alignments between sequences which have no globular similarity in their entire sequence, in other words, it scores divergent protein sequences. Similarity found on local alignments. BLOSUM matrices were first introduced in a paper by Henikoff (Henikoff et al., 1992). They scanned the BLOCKS database for conserved regions of protein families, they did not allow any gaps as it is in the PAM (Point accepted matrix) and then counted the relative frequencies of amino acids and their substitution probabilities. Then, they calculated a log-odds score for each of the 210 possible substitution pairs of the 20 standard amino acids. All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins like the PAM Matrices.

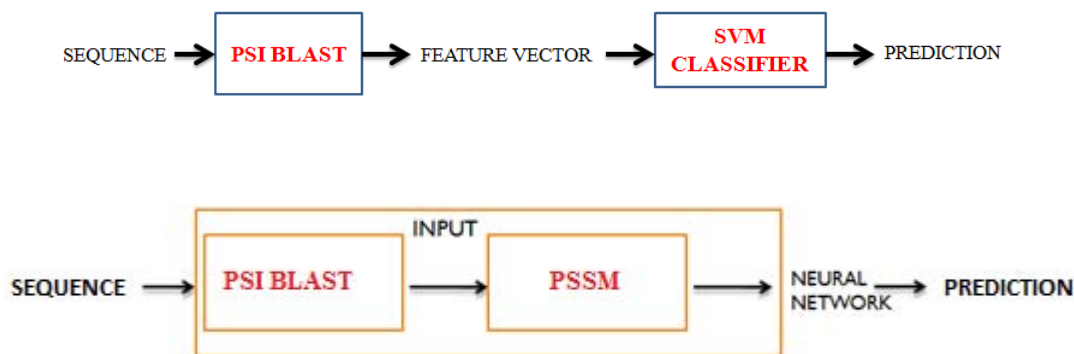


Figure 1.9.5-3 Strategy of using PSSMs in protein secondary structure prediction by machine learning

1.10. Available Protein Secondary Structure Predictors

There are many secondary structure predictors in the web which are frequently used by the biologists for many different purposes. Many of them are receiving query from their users by e-mails whereas few of them are providing results online. Some of the predictors are downloadable and they work standalone. In the following table, some famous predictors name, initial release data are given and their methods are shortly described.

Table 1.10-1 Name of the famous secondary structure prediction methods and their descriptions and initial release data*

Name	Method description	Type	Initial release
SPIDER2	Method developed based on iterative Deep Neural Network	Webserver/downloadable	2015
RaptorX-SS8	This predictor uses conditional neural fields from PSI-BLAST profiles	Webserver/downloadable	2011
NetSurfP	Profile-based neural network	Webserver/downloadable	2009
GOR(GORV)	Information theory	Webserver	2002
Jpred	Neural network assignment	Webserver	1998
PREDATOR	Knowledge-based database comparison	Webserver	1997
PredictProtein	Profile-based neural network	Webserver	1992
PSIPRED	two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST	Webserver	1999
SymPred	an improved dictionary based approach which captures local sequence similarities in a group of proteins	Webserver	2004
YASSPP	Cascaded SVM-based predictor using PSI-BLAST profiles	Webserver	2006
PSSpred	Multiple backpropagations neural network predictors from PSI-BLAST profiles	Webserver/downloadable program	2012
Frag1D	Prediction of both secondary structure and Shape Strings using profile based fragment matching	Webserver/downloadable program	2010

*SPIDER2 (Heffernan, 2015), RaptorX-SS8 (Wang et al., 2011), NetSurfP (Petersen et al., 2009), GOR (GORV) (Kloczkowski, 2002), Jpred (Drozdetskiy, 2015)

1.11. Assessment of the Prediction Methods

Comparisons of the performance of various prediction techniques had been successfully done by some platforms. One of the well known platforms is the Evaluation of Protein Structure Prediction Servers (EVA) but it is not active since 2002. Another is CASP (Critical Assessment of Techniques for Protein Structure Prediction). Assessments are performing every 2 years since 1994 as a fully blinded testing of structure prediction methods (Moult, 1999). Protein sequences for which the structure is about to be solved by X-ray are distributed to registered members of the community, who submit models before there is any release of the experimental data. Collected models are then evaluated by the automated methods. There has been little or no progress in secondary structure prediction over the past decade; that is why secondary structure prediction performance assessment was dropped from CASP after CASP5. CASP and EVA web servers are different in terms of their addressing questions. While CASP is addressing question “how well can experts predict protein structures with the help of machines?” the EVA is addressing question “how well can automatic servers predict protein structures?” (Koh, 2003). As shown in Table 3 performances of some famous methods for secondary structure prediction are compared and their year of publications and generation are shown (Rost, 2000).

Table 1.11-1 Type of the most known protein secondary structure prediction methods.

Type of the method	Method Name	Year	Generation of the method (Rost, 2000)	Claimed Q3 Score
Statistical	Chou and Fasman	1974	First	77
	Garnier	1978	First	57
Neural Network	Qian and Sejnowski	1988	Second	64.3
	PHD (Rost and Sander)	1994	Third	72.2
	PSIPRED (Jones)	1999	Third	76.3
	Sspro (Pollastri et al.)	2002	Third	76
Nearest Neighbor	NNSSP (Sal and Solovyev)	1995	Third	72.2
	Predator (Frish and Argos)	1997	Third	75
Hidden Markow Model	SAM-T99 (Karplus et al.)	1998	Third	73.9
	Bystroff (Bystroff et al.)	2000	Third	74.3

1.12. Protein Sequence Similarity Analyses and Sequence Alignment Methods

Most of the sequence similarity search methods are developed based on alignment strategy between query and target proteins. Database and query sequence match are subjected to statistical assessment which infers homology. Since there are an enormous number of sequences of various lengths, sequence alignment requires computational approaches. Dynamic Programming, including heuristic algorithms or probabilistic methods, is widely used as a computational application for finding sequence similarity.

Specificity, sensitivity, and speed are the three important parameters to consider when designing a similarity search method. Speed is particularly important when the sequence of interest is subjected to a large scale database.

1.12.1. Global and Local Alignment

Alignment approaches fall into two categories: Global alignments and local alignments.

The final goal of calculating global alignment is to form a global optimization between sequences that are forced to align each other in their entire length.

The assigning similarity to the sequences which are widely divergent is not subjected to the global alignment but to local alignment. Although it is difficult to calculate local similarity due to additional challenge of identifying the regions of similarity, it is generally preferable.

1.12.2. Pairwise Alignment

The aim of the alignment is to find the best-matching piecewise local or global alignments of two query sequences. It can only be used between two sequences at a time. It is especially useful for the methods for those which search a database for sequences with high similarity to a query. There are different approaches to produce pairwise alignment; dot matrix methods, dynamic programming and word methods (Mount, 2004).

1.12.3. Dot Matrix Method

It is a preliminary method to compare two sequences for detecting possible alignment of characters between the sequences. It is particularly useful for finding direct or inverted repeats in DNA and protein sequence, however; it is a very slow method for analyzing large scale data (Mount, 2004). The method is qualitatively and conceptually simple. For aligning two sequences with a dot matrix, the two sequences are written along the top row and leftmost column then dots are placed at any points where the characters match. The diagonal line appears in the matrix where two close related sequences are subjected to analyze. Parallel diagonal lines represent repetitive sequences in the matrix. Although this method is still in use, there are many limitations of the methods such as lack of clarity, non-intuitiveness, difficulty extracting match summary statistics and match positions on the two sequences (Gibbs, 1997).

1.12.4. Dynamic Programming

Dynamic programming can be applied to produce global alignments via the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970), and local alignments via the Smith-Waterman algorithm (Smith and Waterman, 1981) which are two most known algorithms for sequence alignment. In most of the usage, protein alignments use a substitution matrix to assign scores to amino-acid matches or mismatches. Gaps are introduced between some of the residues in order to reach the optimum number of matches between the sequences. Gaps assumed as mutations (insertion or deletions) should be penalized in order to not lose the sense of biological similarity.

1.12.4.1. Needleman-Wunch Algorithm

The Needleman–Wunsch algorithm is an efficient alignment algorithm used to align proteins or nucleotide sequences. It is also known as the optimal matching algorithm and the global alignment. The algorithm essentially divides a large problem (e.g. the full sequence) into a series of smaller problems. It is one of the most used algorithms especially when the quality of the global alignment is of the utmost importance.

The dynamic programming matrix is defined with three different steps. The first step is initialization step where the first row and first column of the matrix can be initially filled with 0. If the gap score is assumed, the gap score can be added to the previous cell of the row or column. The second stage of the method is filling the matrix cells from the upper left-hand corner of the matrix. In order to assign a highest possible score for each cell of the matrix, it is required to know the neighboring scores (diagonal, left and right) of the current position. According to the defined scoring matrix, each cell needs to be filled with the match or mismatch score to the diagonal value. Similarly, the gap score to the other neighboring values needs to be added. In the end, each cell will have three different values, from that take the maximum among them and fill the i th and j th position with the score obtained. The last step of the method is called trace back where the alignment is generated with following the highest scores from the bottom right corner to the upper left corner. It is important to consider that there may be two or more alignments possible between the two sample sequences.

1.12.4.2. Smith–Waterman Algorithm

This is a sequence alignment algorithm proposed by Temple F. Smith and Michael S. Waterman in 1981, for determining similar regions between two strings of protein or nucleotide sequences. It is able to guarantee to find the optimal local alignment with its desirable properties. The negative scoring matrix cells are set to zero in the algorithm which renders local alignment visible. It is the one of the main differences to the Needleman-Wunsch Algorithm. Highest score is chosen to start backtracking whereas alignment finishes into the cell with score zero. Improved versions of the algorithm are available (Gotoh, 1982; Altschul et al., 1990; Altschul and Erickson., 1986).

1.12.5. Word Methods

Word method is a heuristic method that does not guaranteed an optimal alignment score. It is also called a k-tuple method since this method uses series of short subsequences in a query.

Word methods are very efficient in terms of speed. It is found very useful in large-scale database search where it is understood that a large proportion of the candidate sequences will have essentially no significant match with the query sequences. Since it is much faster than dynamic programming, most of the algorithms designed for the database search use words method. The most famous implementation of the method in large-scale database search is FASTA and BLAST (Lipman and Pearson, 1985).

FASTA use local sequence alignment to find matches or similar database sequences. Algorithm begins with an observation on the pattern of word hits, word to word matches

of a given length, and marks potential matches before performing an optimized search algorithm like Smith-Waterman. The parameter k-tupe is used to optimize the size taken for a word which controls the sensitivity and speed of the algorithm. Optimizing a k-tupe is critical since increasing it decrease number of background hits that are found. From the word hits that are returned, the program looks for segments that contain a cluster of nearby hits. These segments are investigated for possible matches.

The BLAST algorithm is today known with its variety of sub-algorithms which provides a number of tools optimized for particular types of queries. BLASTN, BLASTP, PSI-BLAST can be considered as most popular programs in BLAST family.

If you are ready to sacrifice a little bit from accuracy, with its higher speed BLAST is a good alternative to FASTA. Similarly to FASTA, BLAST also uses advantages of word methods to search in large scale databases but evaluates only the most significant word matches, rather than every word match as does FASTA. K-length is optimized in most of the BLAST implementation and set up as a default. Implementation of BLAST can be found in many web portals today.

BLAST begins with removing sequence repeats in the query sequence since it is thought that these regions might give high scores that confuse the program to find the actual significant sequences. As a next step, the k-letter word list is generated in the query sequence as it is shown in the following diagram.

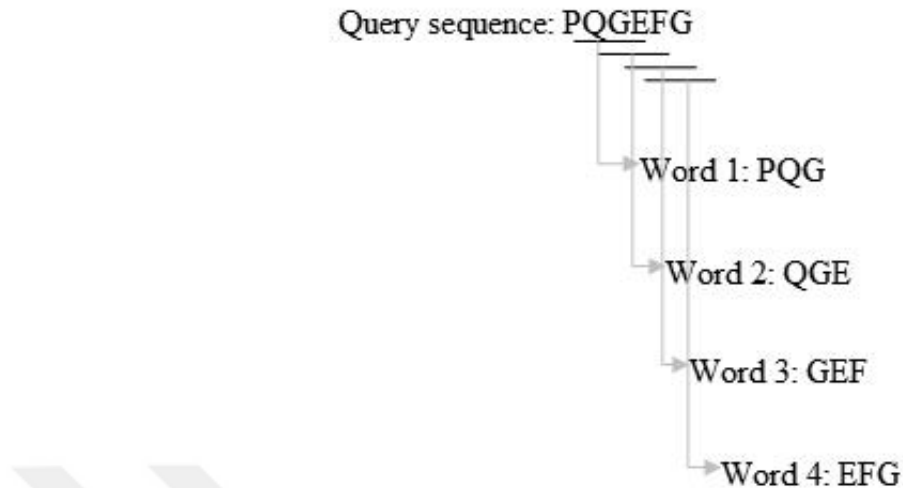


Figure 1.12.5-1 k-letter word list generation.

In the following steps, possible matching words are listed. It is one of the main different steps between BLAST and FASTA. As it is mentioned, FASTA cares about all of the common words in the dataset whereas BLAST cares about only the high-scoring words. BLAST is organizing the remaining high-scoring words into an efficient search tree and repeats to match words and to organize the remaining high score words for each k -letter word in the query sequence. After completing all word of the query, database sequences are searched for exact matches with the remaining high-scoring words. The exact matches are extended to the both direction to obtain high-scoring segment pairs. All pairs that are giving enough high score are listed and their significance is evaluated.

1.12.6. Multiple Sequence Alignment

When more than two sequences are subjected to the alignments, it is called multiple sequence alignment (MSA). In general, the input set of query sequence is assumed to

have similarities which are descended from common ancestors. It is generally much more difficult to align more than two sequences. Most MSA methods use heuristic methods. They are not designed to obtain global optimal matches since identifying the optimal alignment between more than a few sequences of moderate length is prohibitively computationally time consuming.

Multiple sequence alignment is also found very efficient to be used in some secondary structure prediction methods. The similarity between two sequences does not provide information to generate a profile which is thought that it carries similar biological characters. These profiles are used in PSI-BLAST and it has potential benefits to the secondary structure prediction has shown in subsection 1.9.5. Multiple sequence alignment is also used to obtain a consensus of secondary structure sequences of structurally known proteins during the prediction processes (Do, 2008).

1.13. Using Sequence Similarity for the Prediction of Secondary Structure of Proteins

The most important achievement in the development of the secondary structure prediction algorithms is to provide some information about the fold of the target protein. There are some attempts to predict the fold directly from the predicted secondary structure patterns (Rost, 1995; Di Francesco et al., 1997). Recently secondary structure prediction became an important tool in augmentation of the fold recognition methods (Bowie et al., 1991; Ouzounis et al., 1993; Wilmanns and Eisenberg, 1993; Chou and Zhang, 1994; Yi and Lander, 1994; Alexandrov et al., 1996; Fischer and Eisenberg, 1996; Jaroszewski, et al., 1998).

Although protein folding problem is on a long way to the solution, there are protein structure prediction methods, such as comparative modeling and threading that work well, but unfortunately only for some special cases. These methods are based on the paradigm that proteins with similar sequences fold into similar structures and usually have similar functions (Sali et al., 1990). Traditional methods of sequence analysis define the similarity on the level of amino acids treated as letters from an abstract alphabet (Waterman, 1995, Lin et al., 2010). Threading methods extend the notion of protein sequence similarity by including the compatibility between interaction preferences of amino acids in different protein structures expressed in propensity matrices. Underlying rules of protein folding is widely accepted as a secret of nature, and researchers benefit from the reality that, as long as they have similar sequences, they should fold to similar structures.

The usefulness of these analogy based methods are arguable since in nineties; even the most optimistic estimate was claiming that only 10% of all existing protein topologies are currently known (Chothia, 1992). Today this percentage is not improved significantly. But the low percentage of known structures is not so severe for local structure prediction any more. Almost all regular structures are known and thus, there is no limit on the accuracy of analogy based secondary structure prediction algorithms.

Analogy based approaches are referred to as the nearest-neighbor algorithm (Levin et al., 1986; Nishikawa and Ooi, 1986; Salzberg and Cost, 1992; Zhang et al., 1992; Geourjon and Deleage, 1994; Yi and Lander, 1994; Salamov and Solovyev, 1995). The

core of all these procedures is the definition of similarity, just as in the fold recognition methods.

In the following chapter, a new protein secondary structure prediction method which benefits from similar segments are presented and each step of the method is described in details.

1.14. Rationale and Objectives

Secondary structure prediction, as one of the hottest topics in bioinformatics, still remains as an unsolved problem. Various methods have been proposed to solve the problem of predicting proteins secondary structure from the amino acid sequences including k-Nearest Neighbors (k-NNs), Hidden Markov Models (HMMs), Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs). In this thesis, an efficient algorithm that depends on segments level similarity for predicting secondary structure is presented and shown to be superior to the current methods. The main motivations of this thesis are based on following two observations:

1. In spite of the tremendous increase of the number of characterized proteins, there is an apparent saturation in the occurrence of the new folds while the number of similar proteins whose structures are solved raised.
2. Variation in secondary structures of proteins lagging considerably behind the deviation into amino acid sequence.

These two observations raise an important question: is the current structure library already complete enough in a principle to solve the protein secondary structure prediction problem?

This thesis is a partial answer to this question.

1.14.1. Significance of the Study

Relying on the above mentioned two observations this thesis proposes a method to improve the rate of reliability of the protein secondary structure prediction. Meanwhile it expands the knowledge about the relationship between protein sequences and protein structures. It creates a new set of database that provides more reliable and objective prediction performance for secondary structure prediction algorithms. High accuracy of predicting secondary structure of proteins by the new technique proposed in this thesis will be a significant step towards the successful application on medicine and biotechnology.

1.14.2. Research Objectives

The main objective of this thesis is to create a tool to predict secondary structures of Proteins with high accuracy. This goal achieved through a proper understanding of the relations between primary and secondary structures of proteins via computational techniques with biological interpretation. This understanding shed some light on the question how local secondary structure evolves along by the time depending on the modifications in primary sequence segment base search of similarities helps to identify

embedded local structural codes of proteins based solely on sequence information. The new tool creates an opportunity to examine the existing methods and their performance in protein secondary structure prediction. It also makes the assessment of difficulties which limits the prediction accuracy possible.

1.14.3. Research Hypothesis

“In this thesis it is claimed that protein secondary structure prediction problem can in principle be solved based on the current PDB library by developing efficient sequence segment similarity detection and information aggregation techniques that can recover secondary structures of unknown proteins from their amino acid sequences”.



CHAPTER II THE METHOD

2.1. Motivation of the Method

Due to the fact that the structure of a protein plays a central role in its function, structure prediction is a subject of great interest in biology. Since the first determination of a protein sequence by Sanger in 1952 (Sanger, 1952), protein primary sequences and their relations with other structural levels have been studied extensively. On the other hand, exploding number of known protein sequences increased the necessity for accurate prediction algorithms, since the determination of a protein structure by experimental methods is timely and costly. Since 1970s, secondary structure prediction of proteins has been performed by various methods by using statistical information, physicochemical properties, sequence patterns, multi-layered (or neural) networks and evolutionary conservation (Arjunan et al., 2012).

In 1992, only about one-seventh of new protein sequences had sequence similarities to known structures (Bork et al., 1992). Although detection of significant similarities to proteins of known structure considered as one of the most reliable methods for prediction (Rost and Sander, 1993) the necessity of finding similar proteins to the new protein sequence was a drawback of the method due to a low number of similar sequences present in the database. For that reason, it was preferable to develop methods for predicting the secondary structure of proteins in low similarity database.

Today there are millions of sequences in a non-redundant (NR) database of protein sequences in NCBI (Rekapalli, 2012), including the complete genomes of more than 4000 different species. The sequences are different in length that range from 6 to almost 37,000 amino acid residues (Berman and Helen, 2000).

2.1.1. Protein Database and Protein Sequence Similarity

Today, one of the most known protein database, PDB (Brookhaven Protein Data Bank, 2016) is holding more than 110.000 proteins whose secondary structures are analyzed by X-ray or NMR. Just only in 2015, about 10.000 new proteins were added to the database (PDB statistics, 2016). The growth rate is much faster than before, but still far from reaching the number of newly sequenced proteins. Unlike the number of new protein sequences, the percentage of new folds in new entries to PDB, the secondary structure of which has been never seen in the current PDB library decreased and no new unique fold has been reported since 2008. As it is shown in the following table the number of the newly seen secondary structure was 27% in the year 1995, %5 in 2001 (Zhang and Skolnick, 2005). Since 2008, no new unique fold have been reported (“PDB statistics,” 2016)

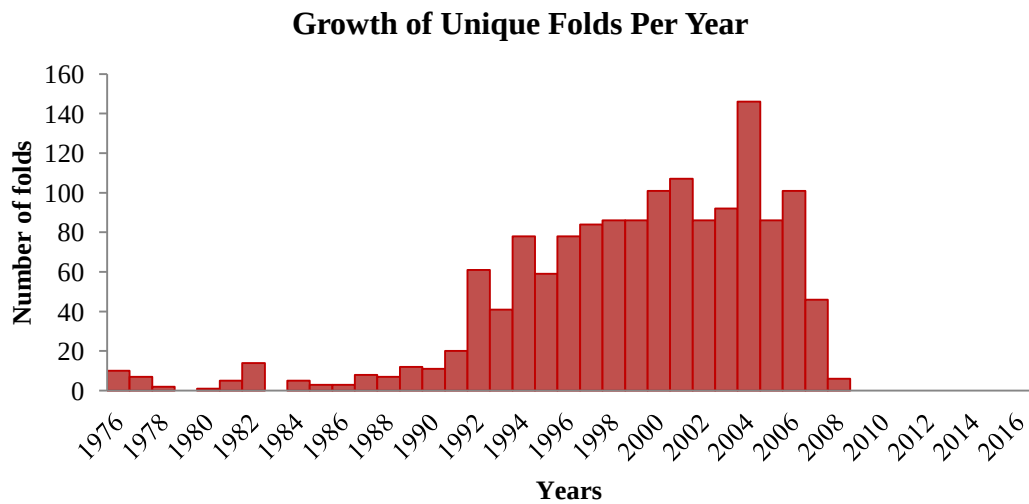


Figure 2.1.1-1 Growth of unique folds per year as defined by SCOP v1.75 (Murzin et al., 1995)

As mentioned in one of the papers published by Rost (Rost and Sander, 1993) which is cited by more than 3000 papers, in 1992 only about one-seventh of the new sequences

had significant sequence similarities to a known structure. Today, finding similar proteins to a newly sequenced protein is much easier due to the huge number of structurally solved proteins. In the table 2.1.1-1, the number of non-redundant sequences as determined by blastclust (Kuzniar et al., 2010) at several levels of sequence identity is shown.

Table 2.1.1-1 In PDB the query implementation for removing similar sequences is based on pre-calculated clusters of protein chains.*

Sequence Identity in percentage	100	95	90	70	50	40
Number of clusters	65521	45904	43745	38683	33238	29430

**All protein chains of at least 20 amino acids are clustered by blastclust at 100%, 95%, 90%, 70%, 50%, 40%, and 30% sequence identities*

In line with the growth of database, many attempts have been made to classify sequential and structural motifs into the protein families which share a common evolutionary origin. These families reflected by related functions and similarities in protein sequence or structure. Most prominent of these classifications are PROSITE domains (Falquet et al., 2002), SCOP Structural classification of Proteins (Cuff et al., 2009), and CATH Class Architecture Topology Homology (Fox et al., 2015).

The tremendous increase in the number of proteins, the progress of the classification efforts, the apparent saturation of new folds and increase in the number of similar proteins whose structures are solved raises an important question: is the current structure

library already complete enough in principle to solve the protein secondary structure prediction problem?

This thesis work gives a partial answer to this question. To assess the level of the completeness, the strategy will be directed to the comprehensive, large-scale PDB set. Since, in PDB, sequence similarity is defined on a chain basis, many structures in the PDB contain multiple protein chains. In this study duplications, and short proteins of length less than 30 residues are eliminated, and identical proteins are removed. The outcome is a database of 80,592 proteins.

2.1.2. Testing the Completeness of Protein Database for Using in Protein Secondary Structure Prediction

By employing a variety of structure comparison tools (Taylor et al., 2004; Shindyalov and Bourne, 1998; Holm and Sander, 1996; Kihara and Skolnick, 2003) completeness of the PDB issue has been partially addressed by many authors (Murzin et al., 1995; Orengo et al., 1997; Yang and Honig, 2000; Harrison et al., 2002; Kihara and Skolnick, 2003; Presnell et al., 1992; Rychlewski and Godzik, 1997; Frishman and Argos, 1997; Lin et al., 2010). The completeness of the protein structures in the current Protein Data Bank (PDB) library for use in secondary structure prediction of unknown sequences is examined in this section. To deal with this issue, ten 1000 protein chains batches are randomly chosen from PDB.

In PDB database, some large amino acid chains contain other smaller chains as a subsequence. These large chains are called super chains.

The protein secondary structure prediction procedure relying on larger PDB database consists of two steps: Identification of a query proteins super-chains and extraction of the secondary structure sequence of the query protein from the secondary structure sequence of the super-chain.

2.1.2.1. Super-Chain Identification

Super-chain that contains the query protein as a sub-chain are identified from the solved protein structures in the PDB by searching super-chain that contains the query protein as a subsequence from the sample set which includes the query protein in it. This is done by finding the common subsequence of the two proteins. If the length of the common subsequence is the same as the length of the query protein, if the host protein is at least five residues longer than the query, this protein is assigned as a super-chain of the query protein. During super-chain identification process, the address of the starting amino acid in super-chain is also recorded.

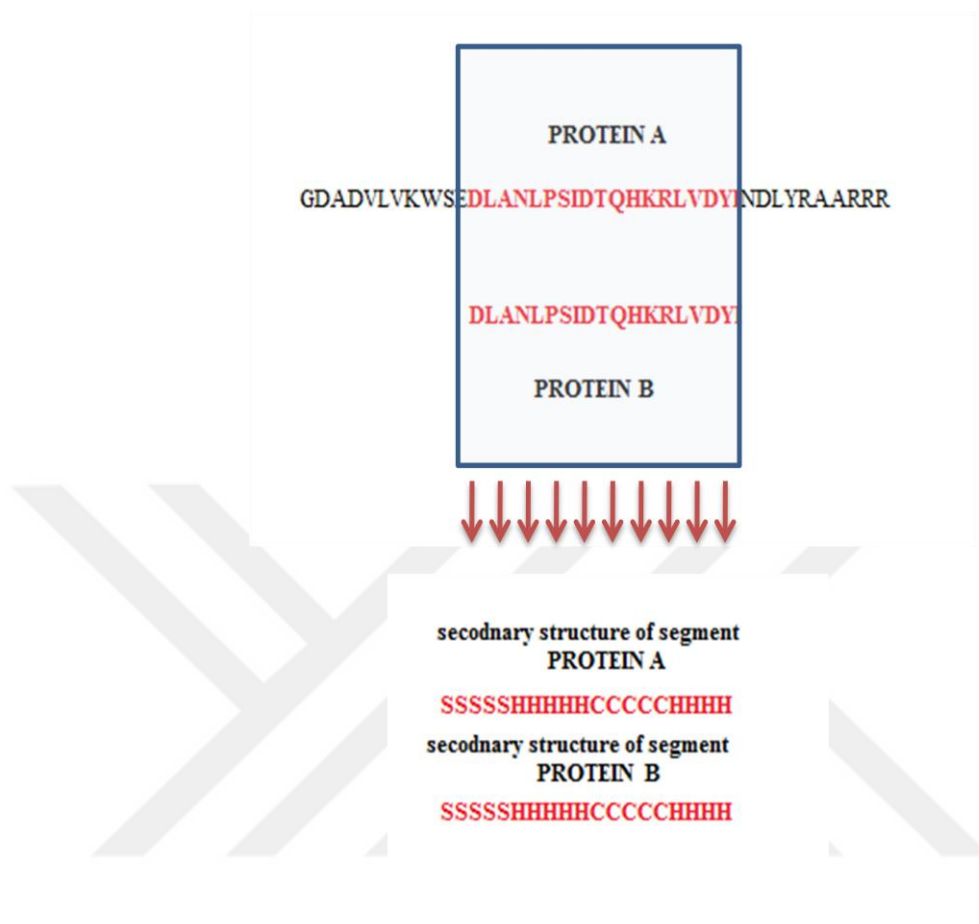


Figure 2.1.2.1-1 From the secondary structure sequence of the super-chain (protein A), the subsequence of the same length as the query protein is extracted. This subsequence is taken as the predicted secondary structure of the query (protein B).

2.1.2.2.Secondary Structure Prediction of the Query Protein Using Super Chains

Address of the starting amino acid of the query protein, in the corresponding sequence of the super-chain is detected from the records of the previous step. From the secondary structure sequence of the host protein, the subsequence of the same length as the query is extracted starting from the recorded address as shown in Figure 2.1.2.1-1. This secondary structure segment is taken as the predicted secondary structure of the query. If more than two super-chains do exist for the query, their consensus is the predicted as the

secondary structure of the query. The percentage of sample proteins that have a super-chain in the database and their mean accuracies are given in Table 2.1.2.2-2.

Table 2.1.2.2-1 The percentage of random sample proteins of batch size 1000, to have a super-chain in the database, and their mean accuracies of predicted secondary structure.

Sample	% Super-chains	% Accuracy
1	18.7	67.32
2	17.9	89.80
3	15.7	86.89
4	16.3	70.12
5	17.4	90.59
6	23.1	90.91
7	17.9	89.80
8	15.7	86.89
9	19.7	75.46
10	18.2	71.33
Average	18.06	81.91

The technique is repeated for well-known datasets such that CB513 (Cuff and Barton, 1999), FC699 (Kurgan et al., 2018), 640 (Chen et al., 2008), 25PDB (Kurgan and Homaeian, 2006), SCOP (SCOP, 2016), and 1189 (Wang and Yuan, 2000). It is seen that sequences of around 18% of protein chains in the batch are present in other chains of PDB dataset. The average prediction accuracy of this method is found to be 90%. Therefore, an unknown protein has a chance of 20% to have a super-chain in Protein Data Bank (PDB), and if this is the case, there is a possibility that its secondary structure is predicted with around 80% accuracy. The percentage of sample proteins that have a host in the database and their mean accuracies are given in the Table 2.1.2.2.-2.

Table 2.1.2.2-2 Summary of the accuracies of secondary structure prediction using the secondary structure sequences of host proteins.

Data Set	# of Proteins	% Super-chain	% Accuracy
CB513	513	55.17	94.50
FC699	858	71.91	79.19
640	640	58.13	82.72
25PDB	1670	29.76	79.19
SCOP	10294	11.42	90.91
1189	1092	51.10	81.70
Average		23.25	88.05

It is seen that around 20% of proteins, far secondary structures can be predicted with a mean of 82%. This accuracy around 6% higher for specially designed sets of proteins.

80% of proteins in PDB do not have host super-chain in PDB. Although prediction accuracies high enough, for the secondary structure of a query protein there is only 20% chance to be predicted in this way. In the coming sections, secondary structure prediction method by the use of smaller conserved segments is presented.

2.2. Protein Secondary structure Prediction Based on Similar Segments of Proteins in PDB

The completeness of PDB for the secondary structure prediction using super-chains is questionable. This leads us to search completeness in smaller scale, segment based. In the literature there are some efforts that try to predict secondary structure of a query using segment similarity.

Although protein folding problem is on a long way to a solution, there are protein structure prediction methods, such as comparative modeling and threading that work well, but unfortunately only for some special cases. These methods are based on the

paradigm that proteins with similar sequences fold into similar structures and usually have similar functions (Sali et al., 1990). Traditional methods of sequence analysis define the similarity on the level of amino acids treated as letters from an abstract alphabet (Waterman, 1995; Lin et al., 2010). Threading methods extend the notion of protein sequence similarity by including the compatibility between interaction preferences of amino acids in different protein structures expressed in propensity matrices. Underlying rules of protein folding is widely accepted as a secret of nature, and researchers benefit from the reality that, as long as they have similar sequences, they should fold to similar structures (Presnell et al., 1992; Rychlewski and Godzik, 1997; Frishman and Argos, 1997; Lin et al., 2010).

In many studies, protein secondary structure had shown that it is a most beneficial feature to predict the tertiary structure of proteins (Fischer et al., 2001; Gong and Rose 2005; Meiler J, Baker, 2003; Rost, 2001). Even it has been proposed that secondary structure alone may be sufficient for accurate prediction of a protein's tertiary structure (Przytycka et al., 1999).

As discussed in the previous chapter, many different methods are available to predict the secondary structure of proteins. New generation methods can be classified into two categories: template-based methods and machine learning methods (Bondugula and Xu, 2007). In the template-based methods, protein sequences whose structure is known are used as templates for a query protein. Prediction performed by finding alignments between query and template sequences. The nearest-neighbor method is counted in this category which uses a database of proteins with known structures to find similar protein

to predict secondary structure proteins. This method simply finds nearest neighbors sequences or sequence segments in the databases to the query protein with counting matched residues between query and target protein sequences. Another category, machine learning methods, generates learning models to classify sequence profiles into different patterns. In the previous chapter, all machine learning methods are discussed extensively. For a decade, there are no many attempts to develop template based methods for protein secondary structure prediction since it was shown that there were no significantly similar proteins to the query proteins (discussed in the previous section). This limitation put template based method aside for a long time to us it for predicting the secondary structure of proteins.

Today, although PDB is much larger, it is still not yet easy to find similar protein to the query. The very promising fact is that the low percentage of known structures is not so severe for local structure prediction. Almost all regular structures are known and thus, there is no limit on the accuracy of analogy based secondary structure prediction algorithms.

Although protein secondary structure prediction problem is addressed decades before tertiary structure prediction, it is interesting that, except some pioneering works (Rychlewski and Godzik 1997; Lin et al., 2010), until recently, no attempt have been done to use the same technique in protein secondary structure prediction as in tertiary structure. To date, there has been no systematic demonstration that this is possible; the exploration of this issue provides one of the motivations for this study.

Our investigation in PDB revealed that, completeness of PDB in prediction of secondary structure of proteins whose structure is unknown may be seen in smaller scale level of segments. It is widely accepted that high similarity in primary sequence segments would have high similarity in secondary sequence segments. Alongside the increased possibility of finding full matching primary sequence segments to the query segments, even primary sequence segments with several mismatches have a chance to correspond secondary structure sequence segments with much lower mismatches. This observation rises a hope of completeness of PDB in smaller scale, in the level of size 7-45 residues.

2.2.1. Deviation in Secondary Structure Lagging Considerably Behind the alterations in the Primary Structure

Although it is accepted in general that mutations, insertions, and deletions in the amino acid sequences produce alterations in secondary structures (Greer, 1981; Lesk and Chothia, 1982; Read et al., 1984), comparative analyses on the structures of related proteins showed that effects of alterations in the amino acid sequences are not equally reflected to the secondary and tertiary structures (Perutz, et al., 1965).

Mutations in DNA that do not significantly alter the phenotype are called silent mutations. This type of mutations is generally seen in non-coding regions, they occur within exons as well. Although there are studies which claim that silent mutations can have an effect on subsequent protein structure and activity (Komar, 2007), there are very few quantitative analyses which show how often silent mutations alter the timing and rate of protein folding which can lead to functional impairments (Zhang, 2012).

According to our observations, the lag of alterations in secondary structure cannot be associated with silent mutations. It is observed that some dissimilar protein sequence segments have identical or very similar secondary structure sequence segments. These dissimilarities are too much to be explained by silent mutations only. The observation that is secondary structure changes lagging considerably behind to the primary structure was announced first time as a structural conversion during protein evolution by Salemma and his colleagues in 1977 (Salemma et al., 1977).

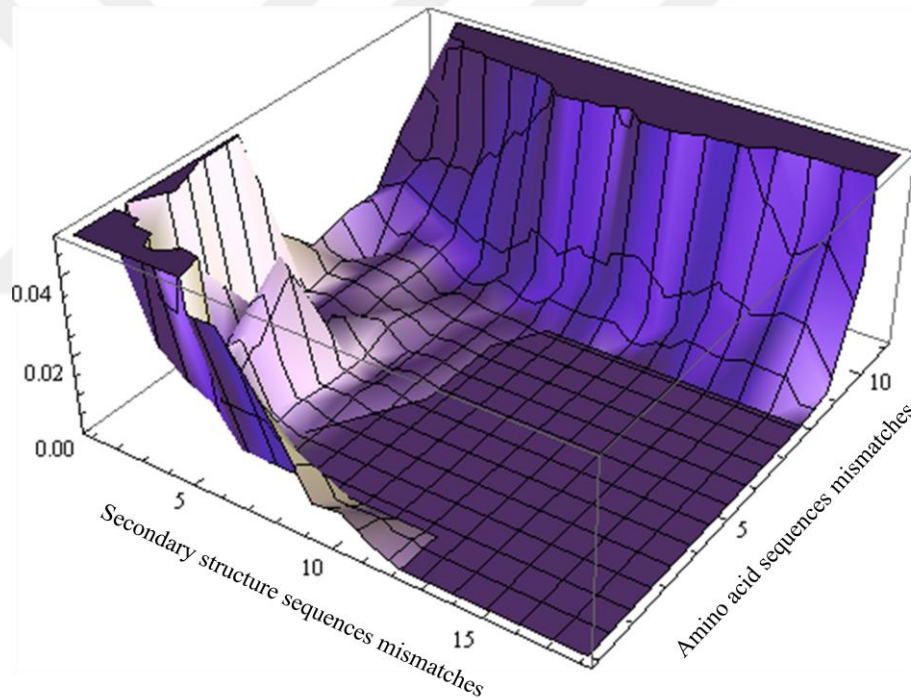


Figure 2.2.1-1 Plateau that is generated by several thousand proteins shows the relation between structure and sequence.

The plateau in Figure 4 3 summarizes our observation on the lag of secondary structure sequence alterations. Coordinate axis in the front represents the sequence mismatches between two segments of equal lengths of 17 residues. Coordinate axis on the left represents mismatches of the corresponding to the two secondary structure sequence

segments that corresponds the two primary sequence segments. The height is the number of segment pairs with mentioned properties.

Front wall tells that segments that have fully matched amino acid sequences mostly have matched secondary structures. Mismatches of secondary structure sequences are not more than 12 for a segment size 17 along this wall where primary sequence segments are in full matches. The left wall tells that segments that have amino acid sequences with mismatches from 1 to 10 may still have fully matched secondary structures. The importance of the left wall relies on the fact that segments that have amino acid sequences with mismatches from 2 to 8 hardly have secondary structures with more than five mismatches. The front corner represents the number of segment pairs whose primary as well as secondary sequences are in the full match. The back wall shows that if amino acid sequences of two segments of length 17 have 10 mismatches, then secondary structures of these segments may have from 0 to 16 mismatches.

The above front corner and left wall facts of the plateau, suggest that there are a large number of segment pairs whose primary as well as secondary sequences are in the strong match positions. In this study, this idea is used for the secondary structure prediction of the unknown proteins. For each segment of the length of a window size of a query protein, segments of the same size whose amino acid chains have 0-10 mismatches with the query segment are searched in the database. The secondary structure sequence of the query segment is assigned as the consensus of those secondary structure sequences of segments with several match levels. For several datasets, window

sizes from 7 to 45 are tried. No significant differences have been seen in accuracies. The window size 17 is adopted for most of our research.

2.3. Introduction to the MAX-MATCH

Avicenna Protein Secondary Structure Prediction Method base on similar sequence segments is introduced in two parts. In the first part sequence similarity measure for finding homologous proteins are introduced. Than MAX-MATCH similarity measure is described and its search space reduction ability is presented. In the second part, Avicenna, protein secondary structure prediction method is described in the following three steps: (1) Similar Segment Identification. (2) The consensus of segments matches to the each query windows (3) Aggregating information obtained from several mismatch levels.

2.3.1. Sequence Similarity Measures for Finding Similar Proteins and Protein Segments

Many similarity measures and alignment strategies are proposed for finding similar proteins. In this subsection some of them are introduced.

2.3.1.1. Sequence Similarity Measures

a) Hamming Distance

Given two l-mers v and w , the Hamming distance between them, $dH(v,w)$, is defined as the number of positions that differ in the two strings.

For example, Hamming distance between the two sequences

AKSVVEGPLPLFFSRAVIDTNED

MDRVSLRPEEIEELREAFRENF

Is $23-5=18$, since there are five matches **VPRND** in 23 residues.

b) Longest Common Sequence

This similarity finds the longest sequence of contiguous or disjoint elements common to the strings. This similarity strategy is found in the computer algebra package MATHEMATICA (Wolfram Mathematica, 2011) as a building function and the following example is generated by that building function.

AKSVVEGPLPLFFRAVDTNEDNNI

MDRSLRPEEIEELREAFREFD

Longest common sequence between sequences is: “**SLPLRAED**”

The similarity value of these two proteins can be designated as 8 for this measure.

c) Longest Common Subsequence

It is a dynamic programming approach which is designed to find the longest subsequence that occurs in both sequences. Consider the two strings:

AKSVVEGPLPLFRAVDTNEDNNI

MDRSLRPEEIEELREAFREFD

The longest contiguous subsequence is FR. Hence, the similarity value of these two proteins is 2 for this measure. This similarity is close to Smith-Waterman local alignment algorithm with its motivation to find local similarity.

d) Needleman-Wunsch Similarity

The Needleman–Wunsch algorithm is an efficient alignment algorithm used to align protein or nucleotide sequences. The algorithm essentially divides the full sequence into a series of smaller problems. Algorithm solves to reconstruct a larger problem by using the solutions of smaller problems. The Needleman-Wunsch similarity measure is also known as the optimal matching algorithm and the global alignment. It is one of the most used algorithms especially when the quality of the global alignment is of the utmost importance. In the following example, Needleman-Wunsch similarity was calculated by using one of the most popular tools of NCBI Needleman-Wunsch Global Align Protein Sequences (Needleman and Wunch, 1990). NCBI uses BLOSUM62 as a substitution matrix and sequence identity was found 2 while positive score is calculated as 21%. In the example, 3 gaps are used in order to reach the optimum alignment.

```

AKSVVEGPLPLFFRAVDTNEDNNI
  +  P  +           E  +
MDRSLR-PEEIEELREAFREFD

```

e) Smith-Waterman Algorithm

This is a sequence alignment algorithm proposed by Temple F. Smith and Michael S. Waterman in 1981, for determining similar regions between two strings of proteins or

nucleotide sequences. It is able to guarantee to find the optimal local alignment with its desirable properties. The negative scoring matrix cells are set to zero in the algorithm which renders local alignment visible. It is the one of the main difference to the Needleman-Wunsch Algorithm. Highest score is chosen to start backtracking whereas alignment finishes into the cell with score zero. Improved versions of the algorithm are available (Gotoh et al., 1982; Erickson et al., 1986)

In the following example, two sequences are aligned with smith Waterman algorithm. In silico simulation of the molecular biology experiments web page (NCBI, BLAST, 2016) are used to perform the following example where 4 matches are found between following two sequences.

```

-----AKSVVEGPL--PLFFRAVDTNEDNNI
      |   |       ||   |
MDRSLRPEE-IEE-LREA--FREFD-----

```

f) Damerau Levenshtein Distance, Edit Distance

In 1966, Vladimir Levenshtein introduced the notion of the edit distance between two strings as the minimum number of editing operations needed to transform one string into another, where the edit operations are the insertion of a symbol, deletion of a symbol, and substitution of one symbol for another (Levenshtein, 1966). For example, the sequence

AKSVVEGPLPLFFRAVDTNEDNNI

can be transformed into

MDRSLRPEEIEELREAFREFD

with 21 editing operations. This implies that these edit distance between the two strings is at most 21. Unlike Hamming distance, edit distance allows one to compare strings of different lengths. In the computer algebra package MATHEMATICA (Wolfram Mathematica, 2012), there is a built-in function for the similarity of strings named Damerau Levenshtein distance which gives the edit distance of the two strings (Kelemen et al., 2008).

All of the above similarity measures are good for certain algorithms. If the job is to find similar proteins which have common similar segments, none of them is efficient as the MAX-MATCH similarity measure described below.

2.3.1.2. An Efficient Hamming Distance Based Measure for Finding Similar Proteins that Contain Similar Sequence Segments

Hamming distance is a good tool to define a similarity measure that reveals of existence of identical segments but it works on two strings of equal lengths. To define a similarity measure for two strings of unequal lengths using hamming distance, we allowed short protein to move along the long protein. The hamming distances H_i is computed at each step i . Then hamming distance based similarity S_i is computed at step i ;

$$S_i = \frac{N_s - H_i}{N_s} \quad (1)$$

where N_s is the length of the shorter sequence of the protein pair.

Then take the biggest of S_i as the MAX-MATCH similarity of the two sequences.

$$\text{MMSim} = \text{Max}_i(S_i) \quad (2)$$

In the following diagram (Diagram 4-1), it is demonstrated that how short protein moves along the long protein. In each step, similarity score is calculated and the highest score is assigned as the MAX-MATCH similarity score of the sequences which are subjected to the similarity analyses.

Two proteins chosen for this example, the maximum match related similarity is found to be 0.36. Therefore it is concluded that MMSim of these two proteins is 0.36.

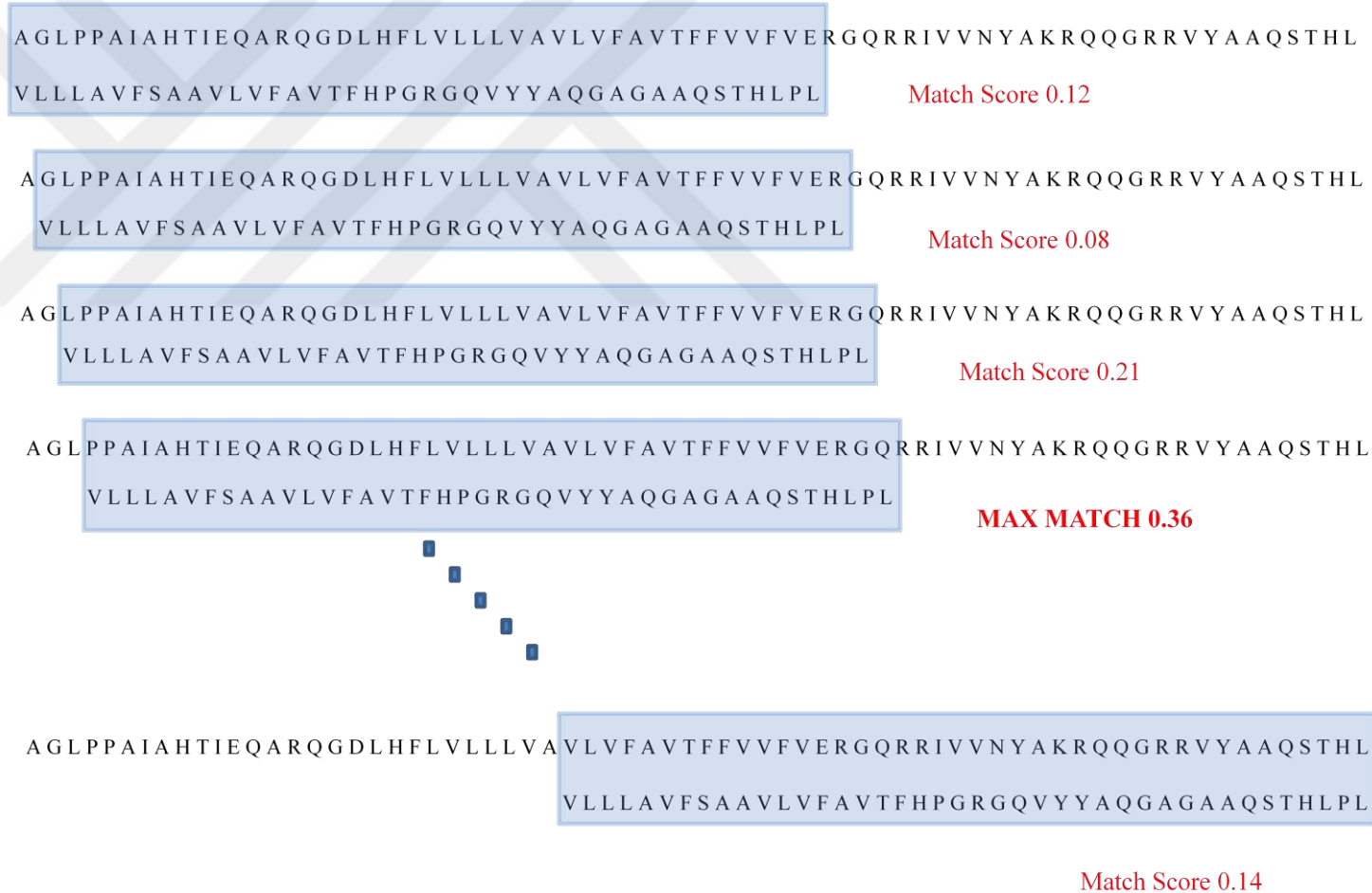


Figure 2.3.1.2-1 Short protein moves along to the long protein. The similarity score of MAX-MATCH is counted for each step. The maximum score is assigned as a similarity score.

2.3.2. A new Protein Secondary Structure Prediction Method: Avicenna

In a previous section (section 2.1.2.2.) it has been shown that around 20% of proteins in a non-redundant subset of 80,529 protein chains from PDB dataset have larger protein chains that contain them as subsequences, and using the secondary structure sequences of these larger protein chains, it is possible to predict the structures of these smaller proteins with an accuracy of around 80% (Table 2.1.2.2.-1).

We postulated that for the remaining 80% of proteins that do not have larger proteins to include them as subsequences, shorter segments with high similarity may be a strong reason for a higher similarity of corresponding secondary structure segments.

It is observed in the large database PDB that, there are a large number of segment pairs whose primary sequences have several mismatches where their secondary sequences are in the full match.

For purposes of demonstration, we have chosen the protein 2SCP_B (Vijay-Kumar and Cook, 1992) from the CB513 database and protein 2K7C_A (Mukherjee et al., 2007) from PDB database. In this example, 2K7C_A is moving along to the 2SCP_B. hamming distance between two sequences is found enough big to count these sequences as similar sequences. Similar segments of proteins are highlighted in red color.

2SCP_B

SDLWVQKMKTYFNRIDFDKDGAITRMDFESMAERFAKESEMKAEHAKVLM
DSLTGVWDNFLTAVAGGKGIDETTFINSMKEMVKNPEAKSVVEGP LPLFFR
AVDTNEDNNIS RDEYGIFFGMLGLDKTMAPASFDAIDTNNDGLLSLEEFVIA
GSDDFFMNDGDSTNKVFWGPLV

2K7C_A

ADMIGVKELRDAFREFDTNNGDGEISTSELREAMRKLLGHQVGHRDIEEIIRD
VDLNGDGRVDFEEFVRMMSR

These two proteins have similar segments of lengths 17 with eight mismatches.

LPLFFRAVDTNEDNNIS

LRDAFREFDTNNGDEIS

However corresponding secondary structure sequences

2SCP_B

CHHHHHHHHHHHHHHHHHCCCCCEECHHHHHHHHHHHHHHHHHCCCCCHHHHHH
HHHHHHHHHHHHCHHHCCCCCEEHHHHHHHHHHHHHHCCCHHHCHHHHHCH
HHHHHHHHHHHHCCCCCEECHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHCCCCCCE
EHHHHHHHHHHHHHHHHCCCCCHHHHHCCCCC

2K7C_A

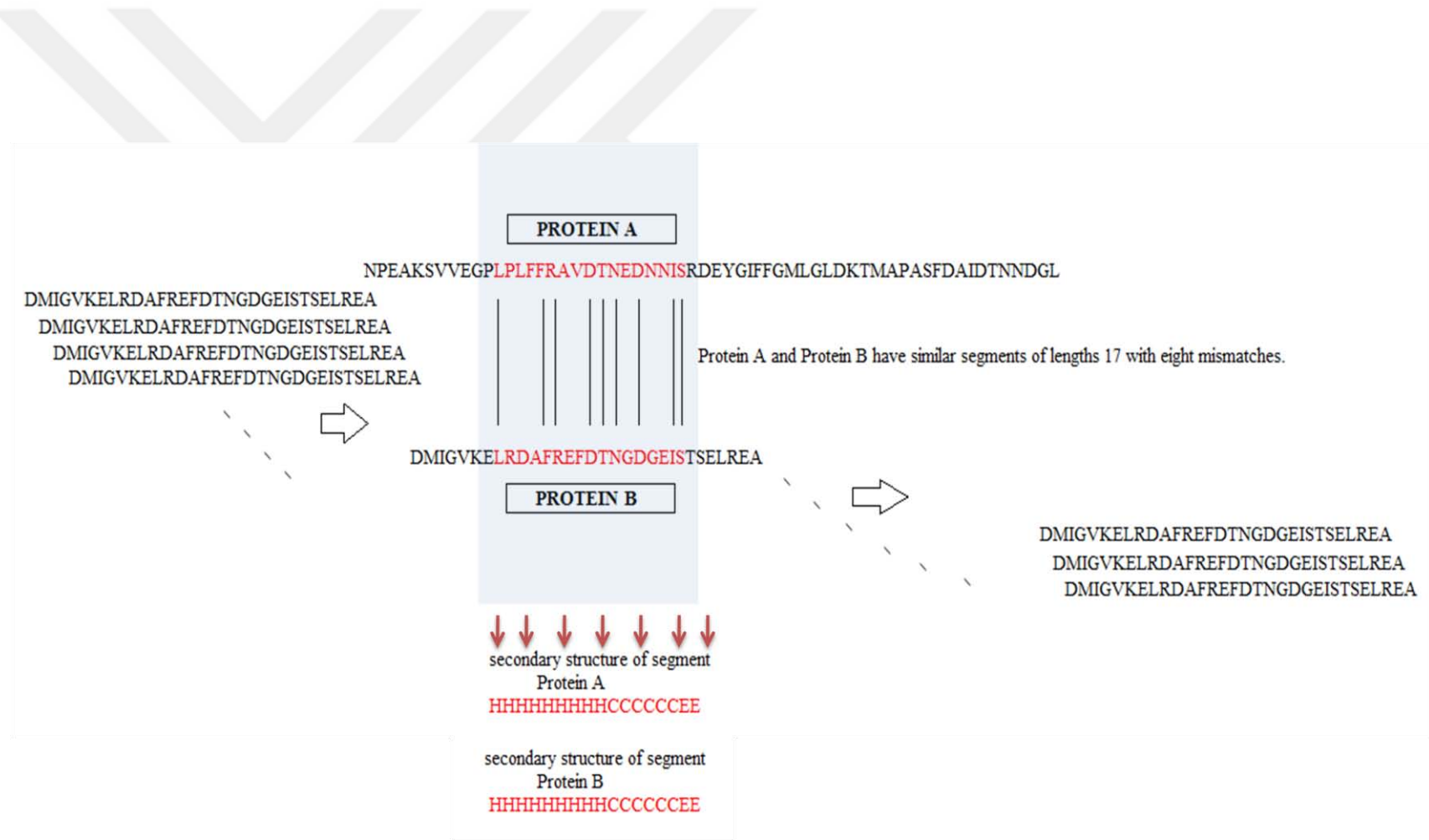
CCCCCHHHHHHHHHHHHHCCCCCEECHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHC
CCCCCEEHHHHHHHHHH

Have similar segments of lengths 17 which match fully.

HHHHHHHHHHCCCCCEE

HHHHHHHHHHCCCCCEE

This observation is used for the secondary structure prediction of the unknown proteins. For each segment of the length of a window size of a query protein, segments of the same size whose amino acid chains matches the query segment in match levels 17-7 are searched in the database. The addresses of these target segments are recorded in II registry. The corresponding secondary structure segments are called from the recorded addresses in eleven registries. The consensus of these secondary structure segments is assigned as a vote for the predicted secondary structure segment corresponding to the query segment for each registry. When this is done for all segments of size 17 of the query protein, overlapping votes are aggregated to obtain a prediction for the secondary structure sequence of the whole protein for each of eleven registries. Afterwards, these eleven predictions are aggregated.



Protein A and Protein B have identical secondary structure sequence in the their corresponding similar amino acid segments

Figure 2.3.2-1 Short protein moves along the long protein, the hamming distance between sequences are recorded in each step of movement. The previous score of hamming distance does have any effects to the following score since no gaps are introduced between residues.

In the following diagram, it is demonstrated that how primary sequence segment of length 17 of the query moves along the target protein in PDB. For the query protein A and protein B, similar sequence segments of length 17 with eight mismatches and their corresponding secondary structure sequences which have full-match are shown as an example.

In practice the PDB database is too large to finish this similar segment search in reasonable time. Therefore the first stage of algorithm is the reduction of the search space using the most appropriate similarity measure MAX-MATXH for the purpose.

2.3.2.1. Reduction of the Search Space

MAX-MATCH similarities are measured between a query and each of sequences in PDB dataset. Score of MAX-MATCH for each sequence are recorded and the most similar 1000 proteins are chosen as a sub-sequence. Only these proteins are used for finding similar segments to the segments of query proteins.

In the sequel, the main steps of the second stage of Avicenna are given as a flow diagram.

2.3.2.2. Similar Segment Identification

The process for similar segment identification and secondary structure prediction can be summarized as follows:

- I. *Collect addresses of similar segments of amino acid sequences*
 - a) Set a window size w , and a gate size d
 - b) Prepare d empty baskets; for each level of mismatches from zero to $d-1$.
 - c) Choose one of the proteins of the test data set CB513 as a *query*.

- d) Take the segment of the first w amino acids of the primary sequence of the query protein as *motif*.
- e) Go to the first protein *prot* of the database and take the segment of the first w amino acids of the primary sequence of this *prot* as a *target*.
- f) Compute the hamming distance of *motif* and *target*. Let it be $h = 0, 1, \dots, w$.
- g) If the hamming distance is less than d , put the address

$$\{\{m\}, \{s, t\}\}$$

Into the basket labeled as h . In this address, m is the position of the first amino acid of the *motif* in the primary sequence of the *query*, s is the label number of the *prot* in the dataset, and t is the position of the first amino acid of the *target* in the primary sequence of the *prot*.

- h) Shift one residue down in the *prot* and repeat *e-g*.
- i) Do the same for all proteins in the reduced search space of 80,592 proteins.
- j) Shift one residue down on the query and do the same for all residues of the query.

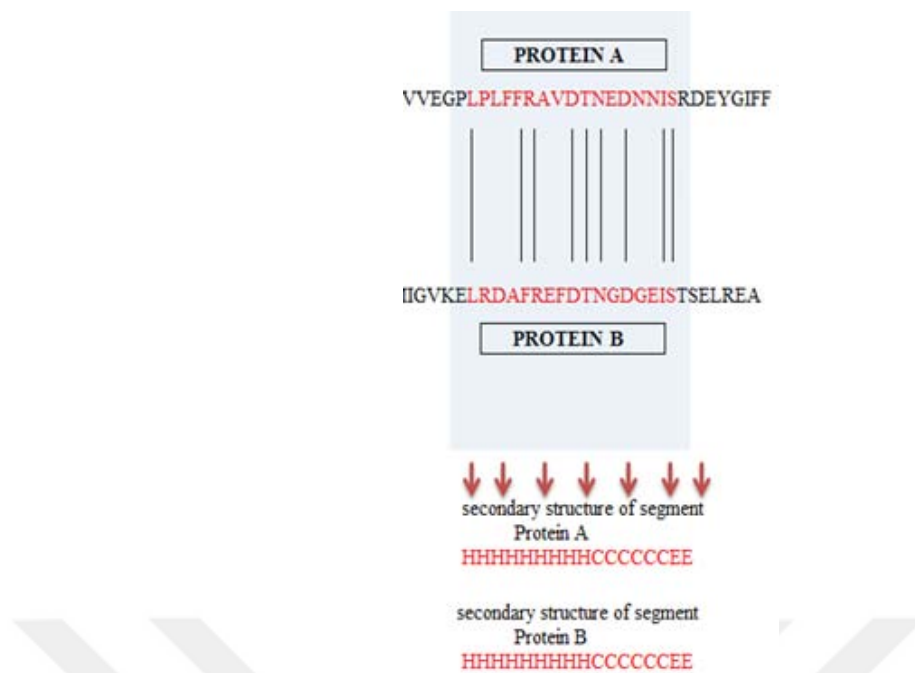


Figure 2.3.2.2-1 Segment length 17 and a mismatch level 9 is fixed. Then from the amino acid sequence of the query protein (PROTEIN A) a segment is taken from the prescribed length. Amino acid sequences of all proteins in the PDB database except the query protein (PROTEIN B)

The addresses of segments of the same sizes and have the set number of mismatches with the segment from the query are noted as it is seen in above figure 2.3.2.2-1 where the fixed length is 17, and mismatch level is 9 (minimum match level 7). Then addresses at the secondary structure sequences of the host proteins (protein B) are visited. At the address, the segment of the same length as the query protein segment is extracted. This segment is taken as the predicted secondary structure of the query segment. The consensus of these secondary structure segments is assigned as a vote for the predicted secondary structure segment corresponding to the query segment. When this is done for all segments of size 17 of the query protein, overlapping votes are aggregated to obtain a prediction for the secondary structure sequence of the whole protein.

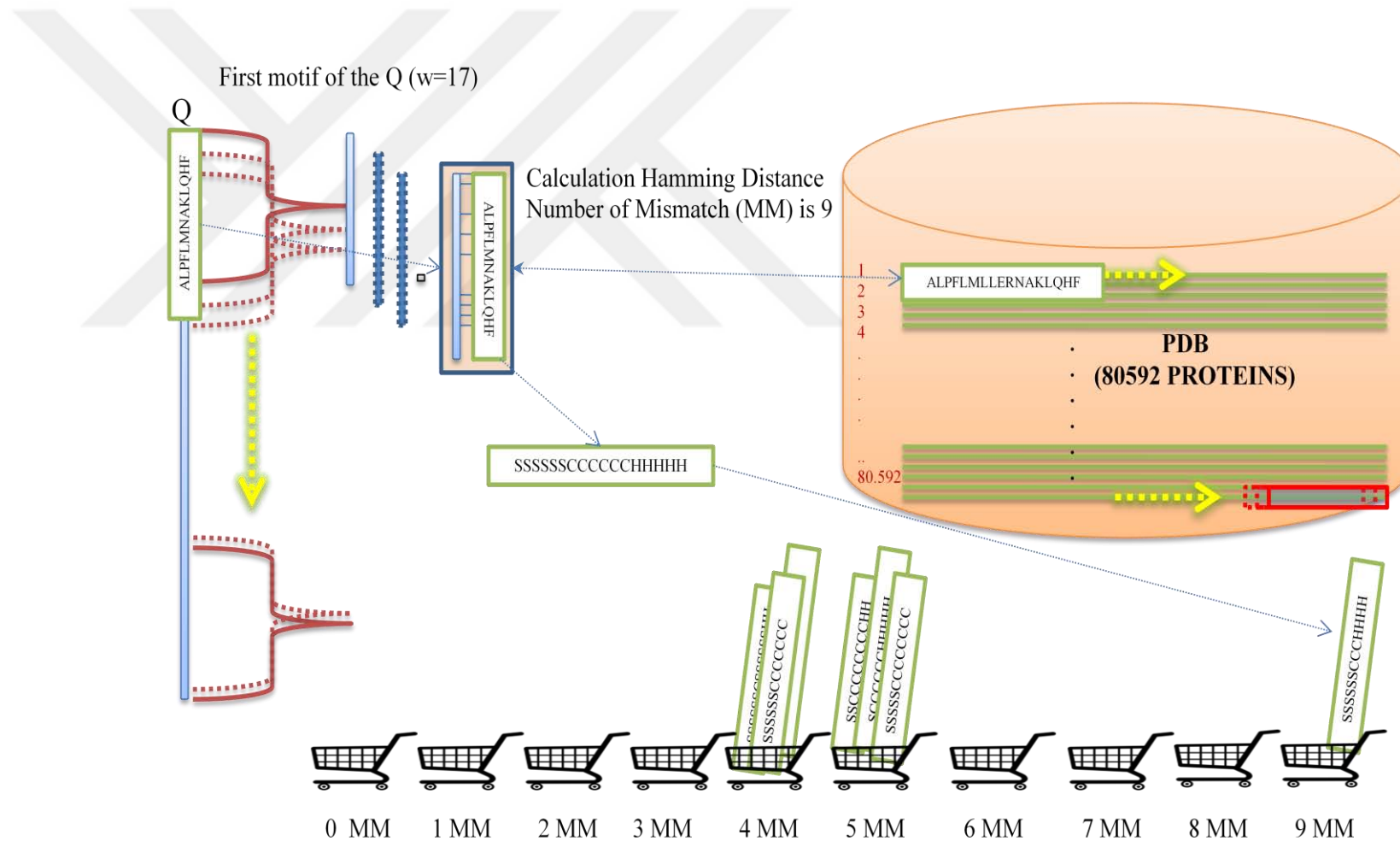


Figure 2.3.2.2-2 Each segment of the query sequence (Q) is taken as a motif. This motif and each of the segments of PDB proteins is subjected to the similarity analysis. Based on their MAX MATCH score they are placed to the proper mismatches (MM) baskets. This is repeated until the last segment of the query protein. In this case, the query protein has similar segments with proteins in PDB database where similar segments are 4, 5 and 9 mismatches.

Our next step is to perform a prediction of the secondary structure of the query. This prediction will be done for each level of mismatches independently, using the information accumulated in the relevant basket.

- II. Open one of the baskets. Bring secondary structure segments from collected addresses in the basket, and get the consensus.

From the basket of zero mismatches: Let the number of the addresses in this basket is nb , assume $nb > 0$.

- a) Prepare empty column vectors whose length is the same as the length of the query and the number is " nb ".
- b) Take one of the addresses from the basket with label " 0 "
$$\{\{m\}, \{s, t\}\}$$
- c) Go to the secondary structure sequence of *proteins* in the search space. Take the segment of length w that starts at the residue t .
- d) Paste this segment to an empty column vector starting at the row m downward; *do not change anything on this vector afterward.*
- e) Do the same for all addresses in the basket with label " 0 ".
- f) Find consensus vector of all nb column vectors at hand.

This consensus vector is the prediction of the secondary structure of the query obtained from segments that are in full-match with the segments of the query. Even the basket is not empty; there is a possibility that some rows of the consensus vector are empty, in this case, the secondary structures of these residues are said "indecisive". These gaps may be filled with decisions from other levels of mismatches.

Repeat the same for all other d mismatch levels.



Figure 2.3.2.2-3 Empty column vectors whose length is the same as the length of the query are filled with the corresponding segments taken from the basket. This step is repeated for the each mismatch baskets. In the end, 9 mismatch baskets are emptied to the empty column

2.3.2.3. Aggregation of the Information Obtained at Several Mismatch Levels.

In general, the decision obtained from the full-match level ($d=0$) may be due to the closely homologous relatives of the query. Since our aim is to underline the role of similar segments of the conserved regions, we will disregard the information of that level in one of the filtering strategies that are described in section 3.2.2. To aggregate the other levels, they are sorted according to their predicted average secondary structure mismatches.



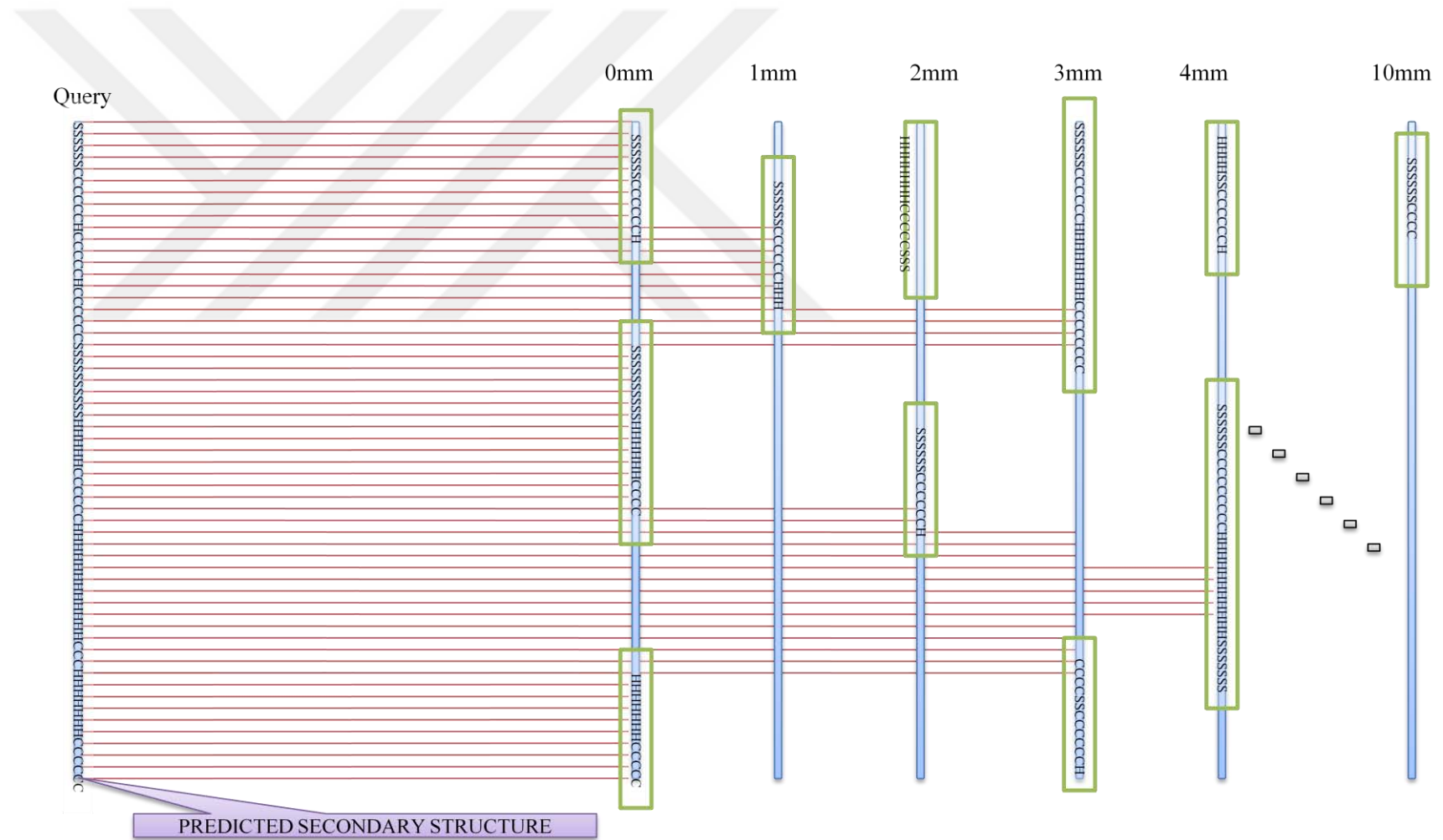


Figure 2.3.2.3-1 Each consensus mismatches column vectors are used to assign secondary structure to the query. Priority is given to the $d-1$ levels of mismatch according to their predicted secondary structure mismatches. The final decision vector is filled by the prediction with smallest average mismatches, and if there remain indecisive, these positions are filled by the level with the prediction with second smallest average mismatches, and so on.

2.3.2.4. Prediction of the Average Mismatches of the Secondary Structure Segments

From the information of addresses in baskets, it is also possible to estimate average mismatch of secondary structure segments at each mismatch level. Secondary structure segments from addresses are collected in a basket. These segments are then classified according to the first number in the address $\{m\}, \{s, t\}$. These are the secondary structure segments for the segment of length w of the query starting at the residue m . The consensus of all these segments is the most probable secondary structure estimation for the corresponding segment of the query. Take this consensus segment as the secondary structure estimation for the corresponding segment of the query, find the average mismatches of all other segments according to the consensus segment. Do the same for all valid residue positions m . The average of these mismatches is an estimate for the average mismatches of secondary structure segments collected. The accuracy of the secondary structure prediction is closely related to this average mismatch. Therefore we give priority to the predictions at the $d-1$ levels of mismatch according to their predicted secondary structure mismatches. The final decision vector is filled by the prediction with smallest average mismatches, and if there remain indecisive, these positions are filled by the level with the prediction with second smallest average mismatches, and so on. A segment length and a mismatch level are fixed. Then from the amino acid sequence of the query protein, a segment is taken from the prescribed length. Amino acid sequences of all proteins in the PDB database except the query protein and proteins that contain the query protein as subsequences are visited.

Some Examples

Table 2.3.2-4, 2.3.2-5, and 2.3.2-6 summarize this technique for the protein 1Q2U (Huai et al. 2003), 1BROA (Hecht et al., 1994), and 1PKYA: 1 – 69 (Mattevi et al., 1995). The first column is the number of mismatches; the second column gives the number of the proteins in the PDB dataset that have segments which make mismatches given in the first column. The third column is the total number of segments supplied by the proteins in the second column.

The fourth column gives the number of residues left undecided by the segments with given mismatches in the first column. The fifth column gives the average mismatches of collected secondary structure segments. The last column is the achieved accuracies of secondary structure prediction using the secondary structure sequence segments of target proteins. Protein 1Q2U has 287 residues in its amino acid chain. The following table summarizes its similar segments in a different level of mismatches.

Table 2.3.2.4-1 Summary of the results obtained for protein 1Q2U of the 699 dataset.

Number of mismatches	Number of Host proteins	Segments	indecisive	Average mismatches	Accuracy
0	2	13	194	1.20	77.24
1	2	103	101	0.40	81.18
2	2	93	38	0.81	84.34
3	2	66	63	0.90	80.36
4	2	51	79	0.25	79.81
5	2	70	68	0.08	77.63
6	2	46	111	0.00	75.00
7	2	36	111	1.26	75.00
8	71	227	40	0.80	62.35
9	438	1297	0	0.64	50.17
10	2408	13723	0	0.98	59.58

Protein 1BROA has 277 residues in its amino acid chain. Following table summarizes its similar segments in different level of mismatches.

Table 2.3.2.4-2Summary of the results obtained for protein 1BROA of the 1189 dataset.

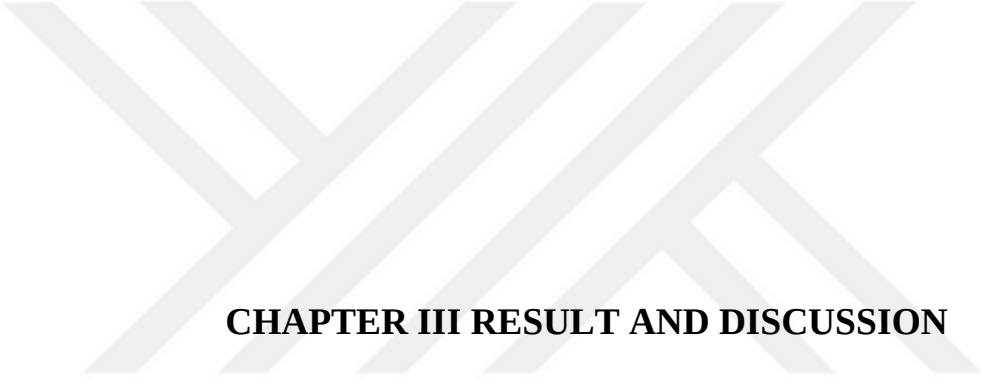
Number of mismatches	Number of Host proteins	Segments	indecisive	Average mismatches	Accuracy
0	5	1135	0	2.68	86.28
1	8	187	103	2.00	90.23
2	3	46	140	1.38	88.32
3	4	65	100	1.65	84.18
4	8	114	80	2.00	84.26
5	10	124	40	1.83	84.39
6	11	191	30	1.85	83.40
7	11	203	16	1.59	84.67
8	13	253	26	1.53	88.05
9	39	331	13	2.93	86.36
10	174	1067	2	7.96	68.70

Protein1PKY HAS 470 residues in its amino acid chain and 69 of them are taken as an entry in the 1189 database. The following table summarizes its similar segments in a different level of mismatches.

Table 2.3.2.4-3Summary of the results obtained for protein 1PKYA: 1 – 69 of the 1189 dataset.

Number of mismatches	Number of Host proteins	Segments	indecisive	Average mismatches	Accuracy
0	0	0	-	-	-
1	0	0	-	-	-
2	2	3	34	0.47	85.71
3	3	10	25	0.17	88.64
4	3	16	24	0.47	84.44
5	32	94	24	0.26	88.89
6	42	219	0	0.36	94.20
7	46	210	0	0.31	92.75
8	49	469	1	0.32	94.12
9	100	559	0	0.73	86.96
10	347	1412	0	0.76	63.77

Comparison of first and last columns gives an idea of primary and secondary structures. As revealed by the plateau in Figure 2.2.1-1, corresponding to the high mismatch levels in the primary sequence of column 1, is not reflected the fifth column which is the average secondary structure mismatches of collected segments. Because of the high match levels of secondary structure segments, the high accuracies in column six are achieved. Aggregation of the decisions at every level are collected. Each step of the method is described and steps are illustrated with the diagrams in the following sections.



CHAPTER III RESULT AND DISCUSSION

3.1. Structures are More Conserved Than Sequences

Time dependent changes of amino acid sequence of proteins that become fixed in populations are mainly due to alterations of single amino acid residues as insertions or deletions. Since most secondary and tertiary structures of proteins are mainly determined by their amino acid sequences (Anfinsen, 1973), secondary and higher-order structure are expected to alter along these changes (Chan and Dill, 1990). The extent of higher-order structural change due to alterations in the amino acid sequences depend also on the type and location of these alterations (Illergard et al., 2009; Haspel, 2003). While some single changes completely disrupt higher order structure, the others that conserve the physicochemical properties of the protein may slightly affect the structure (Matthews, 1995). It is observed that similar proteins have more conserved structure than their sequences. This fact can even be observed in highly divergent protein families within their conserved regions (Huang, et al., 2012). On the other hand, the development of sequencing technology and three dimensional structure determination tools led the accumulation of examples of similar proteins with distinct structures (Grishin, 2001).

Studies which claim that the local structural determinants are frequently encoded in short segments of protein sequence are promising to rely on segments based similarity search for assigning a structure to the proteins (Menke, 2009). However, it is not easy task to identify embedded local structural codes based solely on sequence information (Yang and Wang, 2003; Chothial and Lesk, 1986). This thesis spread some light to the question how local secondary structure evolves along with the time dependent primary sequence changes and how the relation between amino acid sequence and secondary

structure develops. The lag in secondary structure alterations behind the changes in primary sequence is the main inspiration of the introduced method in this thesis. We test one of the claims “secondary and tertiary structures are more conserved than amino acid sequence” which is often used in discussions about proteins. This claim is supported by our observations of conditional probability landscape in the table below (Table3-1).

In table below, probabilities in rows are conditional probabilities of mismatches of a query secondary substring of length 17, with and secondary substring of same length cut from the proteins from the database given the mismatches of corresponding secondary structure substrings in the first column. For example second column of the table gives the probabilities of secondary substring full matches given primary amino acid substring mismatches in the first column. The highest probability of 37% of the secondary structure full matches appears at the row of 3 mismatches of the primary substring.

0-1 Probabilities in rows are conditional probabilities of mismatches of query secondary structure sequence segments of length 17, with and secondary structure sequence segments of same length cut from the protein

mm	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	0.19	0.21	0.23	0.11	0.14	0.08	0.02	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
1	0.25	0.26	0.21	0.1	0.11	0.05	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.26	0.26	0.14	0.14	0.11	0.06	0.05	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
3	0.37	0.25	0.11	0.11	0.1	0.03	0.02	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
4	0.2	0.2	0.13	0.19	0.18	0.04	0.04	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
5	0.17	0.21	0.21	0.21	0.19	0.07	0.03	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
6	0.15	0.13	0.23	0.23	0.18	0.19	0.02	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
7	0.15	0.14	0.15	0.15	0.09	0.06	0.12	0.02	0.	0.	0.01	0.01	0.	0.	0.	0.	0.	0.
8	0.16	0.12	0.12	0.08	0.09	0.09	0.04	0.05	0.04	0.03	0.04	0.03	0.02	0.03	0.02	0.01	0.01	0.03
9	0.05	0.05	0.04	0.04	0.06	0.06	0.06	0.06	0.07	0.06	0.07	0.07	0.08	0.07	0.05	0.04	0.03	0.04
10	0.01	0.02	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.08	0.09	0.08	0.08	0.08	0.07	0.06	0.04	0.03

3.1.1. Full Secondary Structure Matches vs. Low Amino Acid Matches

In the sequel some examples of full secondary structure matches of short segments corresponding to short primary structure segments with high mismatches are given. For sampled proteins primary and secondary structures are presented and the similar segments are highlighted with red color. 3D images are generated by using Swiss-PdbViewer (Guex and Peitsch, 1997).

a) The substring of length 17 of the protein 2SCP_B starting at the residue 7,

KMKTYFNRIDFDKDGAI

has six mismatches with the substring

KAITCFNTLDFNKNGQI

of protein 2HQ8_A starting at the residue 139.

PDB code: 2SCP_B

SDLWVQ**KMKTYFNRIDFDKDGA**ITRMDFESMAERFKESEMKAEHAKVLMDSL
TGVWDNFLTAVAGGKGIDETTFINSMKEMVKNPEAKSVVEGPLPLFFRAVDTN
EDNNISRDEYGIFFGMLGLDKTMAPASFDAIDTNN
DGLLSLEEFVIAGSDFFMNDGDSTNKVFWGPLV

PDB code: 2HQ8_A

MPEITESERAYHLRKMKTRMQRVDVTGDGFISREDYELIAVRIAKIAKLSAEKA
EETRQEFLRVADQLGLAPGVRISEEA AVNATDSLLKMKGEEKAMAVIQSLIM
YDCIDTDKDG YVSLPEFKAFLQAVGPDLTDD**KAITCFNTLDFNKN**GQISRDEFL
VTVNDFLFGLEETALANAFYGDLDV

while the corresponding subsequences of the secondary structures have a full match:

HHHHHHHHHCCCCCEE

HHHHHHHHHCCCCCEE

PDB code: 2SCP_B

CHHHHHHHHHHHHHHHHHHHCCCCCEECHHHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHH
HHHHHHHHHHCHHHCCCCCEEHHHHHHHHHHHHHHHHHHCCCCCHHHCHHHHHCCCHHHHH
HHCCCCCEECHHHHHHHHHHHHHHHHHHHCCCCCCCCCHHHHHHHHHHHCCCCCEEHHHHHH
HHHHHHHHHHCCCCCHHHHHCCCCC

PDBcode: 2HQ8_A

CCCCCHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCEEHHHHHHHHHHHHHHHHHHHHCCC
HHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCEEHHHHHHHHHHHHHHHHHHHHCCCCCHHHHH
CHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHCCC
CCCEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCC
EEE



Figure 0-1 2SCP_B and 2HQ8_A are generated by Swiss-PdbViewer (Guex and Peitsch,1997)

b) The substring of length 17 of the protein 2SCP_B starting from the residue 96

LPLFFRAVDTNEDNNIS

has eight mismatches with the substring

LRDAFREFD TN GDGEIS

of protein 3OX6_A starting from the residue 96

PDB code: 2SCP_B

SDLWVQKMKTYFNRIDFDKDGAITRMDFESMAERFAKESEMKAEHAKVLMDS
LTGVWDNFLTAVAGGKGIDETTFINSMKEMVKNPEAKSVVEGPLPLFFRAVD
T
NEDNNISRDEYGIFFGMLGLDKTMAPASFDAIDTNNDGLLSLEEFVIAGSDFFMN
DGDSTNKVFWGPLY

PDB code: 3OX6_A

MDRSLRPEEIEELREAFREFDKDKDGYINCRDLGNCMRTMGYPTEMEILISQ
QINMNLGGHVDFFDFVELMGPKLLAETADMIGVKELRDAFREFD
TNGDGEIST
SELREAMRALLGHQVGHRDIEEIIRDVDLNGDGRVDFEEFVRMMSR

while the corresponding subsequences of the secondary structures are in full-match:

HHHHHHHHCCCCCEECH

HHHHHHHHCCCCCEECH

PDB code: 2SCP_B

CHHHHHHHHHHHHHHHHHCCCCCEECHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHH
HHHHHHHHHHCHHHCCCCCEEHHHHHHHHHHHHHHHHCCCCCHHHCHHHHHCCCC
HHHHHHHHHHCCCCCEECHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHCCCCCEEHHHHHH
HHHHHHHHHHCCCCCHHHHHCCCCC

PDB code: 3OX6_A

CCCCCHHHHHHHHHHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHHHHHCCCCCHHHHH
HHHHHHHHCCCCCCCCCHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHCCCC
CCCCCEECHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHCCCCCCCCCEEHHHHHHHH
CC

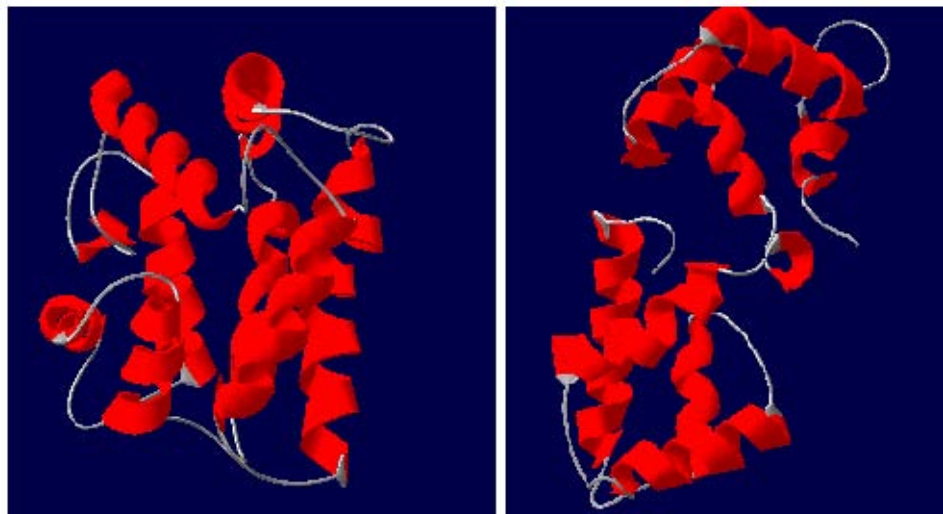


Figure 0-2 2 SCP_B and 3OX6_A are generated by Swiss-PdbViewer (Guex and Peitsch,1997)

C) The substring of length 17 of the protein 2SCP_B starting from the residue 99

FFRAVDTNEDNNISRDE

has eight mismatches with the substring

AFREFDTNGDGEISTSE

of protein 2K7C_A from the residue 12.

PDB code: 2SCP_B

SDLWVQKMKT YFNRI DFDKDGAI TRMD FESMAERFAKESEMKA EHAKVL MDS
LTGVWDNFLTAVAGGKGIDETTFINS MKEMVKNPEAKSVVEGPLPL**FFRAVDT**
NEDNNISRDEYGIFFGMLGLDKTMAPASFDAIDTNNDGLLSLEEFVIAGSDFFMN
DGDSTNKVFWGPLV

PDB code: 2K7C_A

ADMIGVKELRD**AFREFDTNGDGEISTSE**LRAMRKLLGHQVGHRDIEEIIRDVDL
NGDGRVDFEEFVRMMSR

structures are in full match:

HHHHHCCCCCEECHHH

HHHHHCCCCCEECHHH

PDB code: 2SCP_B

CHHHHHHHHHHHHHHHHHCCCCCEECHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHH
HHHHHHHHHHCHHHCCCCCEEHHHHHHHHHHHHHHHHCCCCCHHHCHHHHCCHHHHHH
HHHCCCCCEECHHHHHHHHHHHCCCCCHHHHHHHHHHHCCCCCEEHHHHHHHH
HHHHHHHHHHCCCCCHHHHHCCCCC

PDB code: 2K7C_A

CCCCCHHHHHHHHHHHHHCCCCCEECHHHHHHHHHHHCCCCCHHHHHHHHHHH
HCCCCCEEHHHHHHHHHHHC

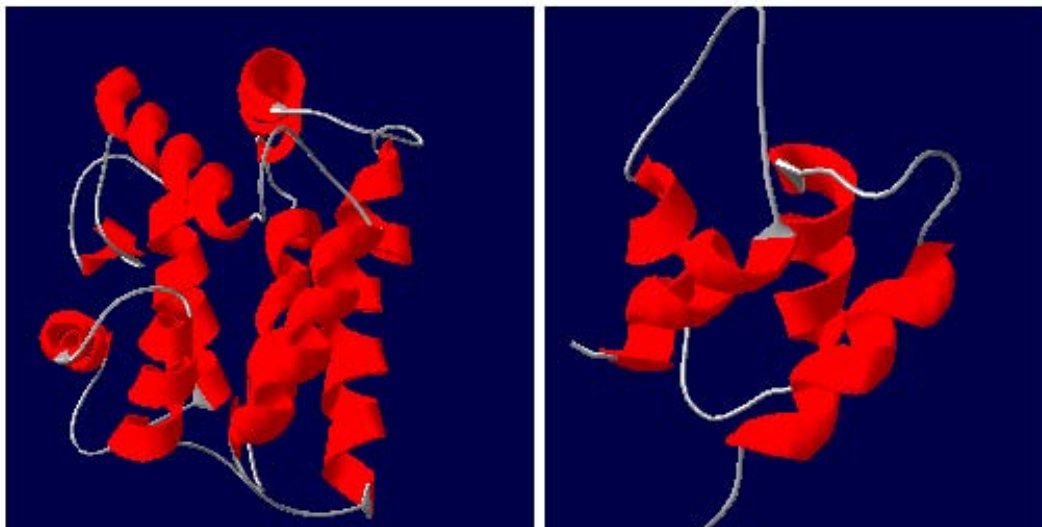


Figure 0-3 2SCP_B and 2K7C_A are generated by Swiss-PdbViewer (Guex and Peitsch,1997)

3.2. The Performance of Avicenna Protein Secondary Structure Prediction Method.

One of the main reasons which make our study unique among the prediction algorithms is that, rather than using multiple sequence alignment for detecting conserved cores we use 17 long substrings to find similar non-consecutive matches in two sequences. We also observe that similar substrings found by our algorithm are classified in the same superfamily by SCOP. The superfamilies are designed to find proteins which have structural evidence to support a common evolutionary ancestor but may not have detectable sequence homology.

Avicenna method performance is tested with more than 10,000 proteins and achieved 83% three-state overall per residue accuracy (Q3). This is the highest Q3 accuracy recorded up to date in the literature.

For testing Avicenna performance non-redundant dataset CB513 (Cuff and Barton, 1999), FC699 (Kurgan et al., 2008), 640 (Chen et al., 2008), 25PDB (Kurgan and Homaeian, 2006), SCOP (SCOP, 2016), and 1189 (Wang and Yuan, 2000) are used to have as much protein diversity as possible. In the figure 3-4, Avicenna accuracy is compared with the other famous predictors.

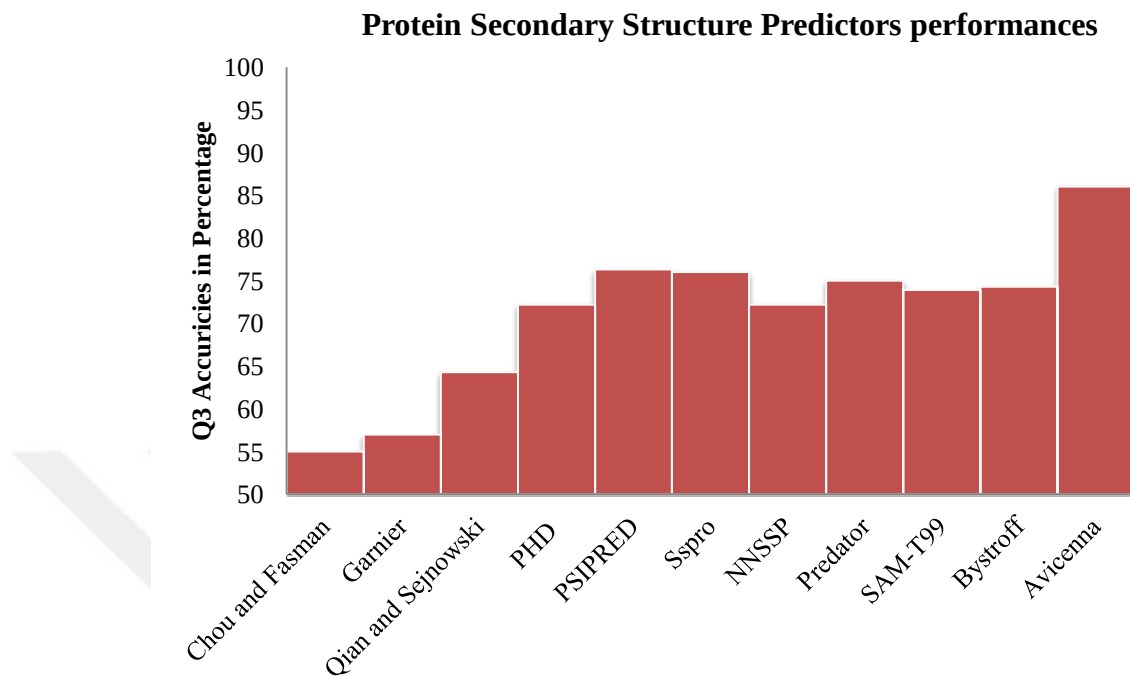


Figure 0-4 Protein Secondary Structure Predictors Performance Comparison, Q3 accuracies in percentage.

3.2.1. Reduction Strategies for the Search-Space

As described in the method part, Avicenna starts with finding similar proteins from PDB database by using MAX-MATCH similarity measure. Most similar 1000 proteins are chosen as a sub-dataset and only these proteins are used for finding similar segments to the segments of query proteins. If any residue remains indecisive, these are the remedies: 1) enlarge the search space to 5000 or 1000 most similar proteins. 2) Increase the mismatch threshold. The first needs more computational time whereas the second brings segments with low similar segments that lower the prediction accuracy.

In general, for predicting proteins secondary structures with Avicenna, the window size is chosen as $w=17$ and lowest allowable primary match is kept at 8. Standard performance, the three-state overall per residue accuracy (Q3) was used to assess prediction success.

3.2.2. Filtering Strategies.

Although identical chains are removed from the database for avoiding redundancy during the MAX-MATCH search, 3 different filtering strategies are applied to the search space. It is though that described filtering methods enable us to underline the essence of short conserved domains in the prediction of secondary structure.

Filter I

Each query protein subjected to the similarity analysis with “Longest common subsequence” (described in Chapter II, Section 2.3.1.1). Proteins sequences that are 95% similar in their residues, and at least 95% similar in their lengths with the query, are removed from the search space. It is though that some of the entries in dataset might be still identical even though some of their residues are different.

Filter II

Avicenna performance is also tested without using identical segments (full matches) during the voting step for assigning secondary structure to the query proteins. All the segments that are identical to any query amino acid segments are removed from the collected set of segments.

Filter III

The sequence which contains any other entire query sequence as a subsequence (it is named as a super-chain in chapter 4) is excluded from the search space, in other words, the query sequence that is present in any other sequence as an embedded sequence is detected and removed from the search space.

3.2.2.1. The performance of Avicenna Protein Secondary Structure Prediction Method on Filtered Search Space.

Each of the filters described in the previous section is applied to the PDB dataset and performance of Avicenna is evaluated. Avicenna's performance tested with Filter 1, the number of proteins in databases and their Q3 average accuracies are given in following Table 3-2.

Table 0-2 Summary of the accuracies of secondary structure prediction on filtered dataset by the first filter using the secondary structure sequence segments of target proteins.

Dataset	Number of proteins	% Accuracy (Q3)
CB513	513	96.33
640	640	84.45
25PDB	1670	83.40
FC699	858	85.35
1189	1092	85.35
RANDOM	500	94.75
AVERAGE		86.58

Avicenna method's performance is also tested with the Filter 2 and 78% three-state overall per residue accuracy (Q3) average is achieved without using the identical 17 long segments.

Table 0-3 Summary of the accuracies of secondary structure prediction on filtered dataset by the second filter using the secondary structure sequence segments of target proteins.

Dataset	Number of proteins	% Accuracy (Q3)
CB513	513	87.2
640	640	77.23
25PDB	1670	74.32
1189	1092	77.78
RANDOM	500	80.72
AVERAGE		77.83

For CB513 test dataset, only 16% of the prediction accuracies remain under 70 % accuracy while more than half of the proteins predicted with more than 90% accuracy.

In the following figure 3-5 frequencies of prediction accuracies are given for CB513.

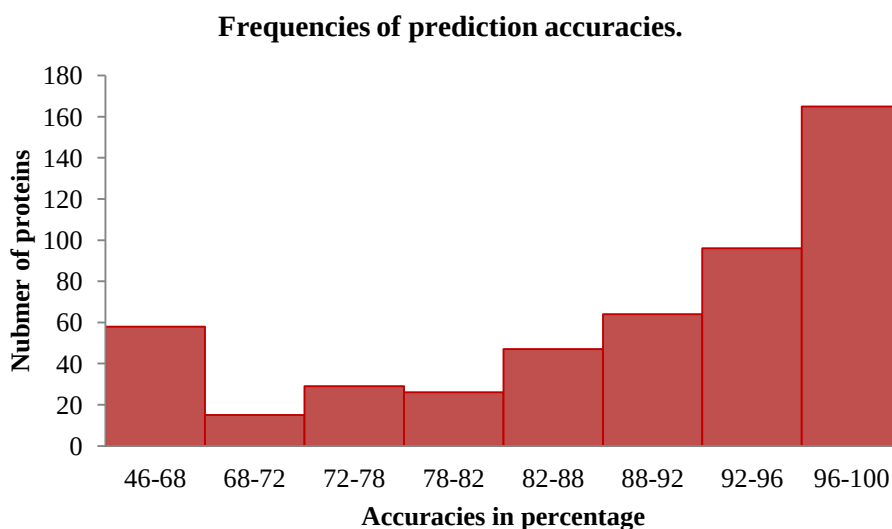


Figure 0-5 Frequencies of prediction accuracies for the test set CB513.

Filter 3 was used as well to testing the performance of Avicenna.

The number of proteins in databases and their Q3 average accuracies are given in the Table 3-4.

Table 0-4 Summary of the accuracies of secondary structure prediction on filtered dataset by the third filter using the secondary structure sequence segments of target proteins

Dataset	Number of proteins	% Accuracy (Q3)
CB513	513	89.06
640	640	78.84
25PDB	1670	79.19
1189	1092	86.77
RANDOM	500	86.19
AVERAGE		84.01

In the history of secondary structure prediction of proteins, since the first secondary structure prediction method was formulated, the accuracy of the prediction was improved gradually from ~62% (Chou and Fasman, 1974), (Garnier et al., 1978), to ~70% (Rost and Sander, 1993) and to ~76% (Jones, 1999). By using position specific scoring matrices and training neural networks David Jones succeeded to design one of the most famous methods but accuracy 76 % still remained far from to be satisfactory. He also claimed that they found that the accuracy of PSI-BLAST alignments (when compared to alignments based on structure comparison) is significantly higher than any other method they have tried for automatic multiple sequence alignment at that time.

Although there are many different attempts in the literature, no method can exceed to accuracy of PSIPRED (Jones D., 1999). Błażewicz and his colleagues introduced new machine learning methods such as LAD, LEM2 and MODLEM and obtained 60-70% accuracy (Błażewicz et al., 2007). Other group of scientist used two-stage feed forward neural network for prediction of protein secondary structures and get 73.6 overall accuracy for the dataset RS126 (Alirezaee et al., 2012). One another remarkable accuracy was obtained by Baldi and his colleagues (Baldi et al., 2010). They proposed

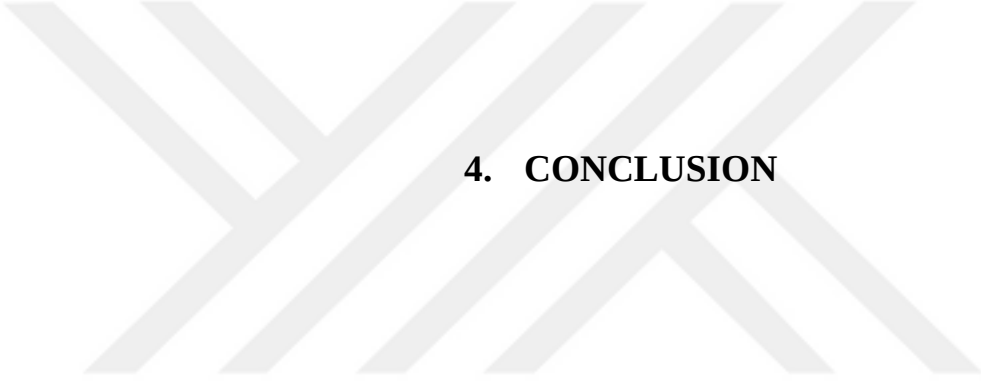
new family of adaptive graphical model architectures for learning non-causal sequence translations and record 71.8 on a set of 35 proteins from the 1998 edition of CASP.

There are also other studies which focused on specific secondary structure elements prediction and their accuracies are also remain low (Chamorro, et al., 2010; Cheng et al., 2007, Zheng and Kurgan, 2008). A new technique is recently proposed to improve the accuracy of predicted secondary structures. One of them published in The Nature Scientific Report in 2015 by Heffernan and his colleges (Wang et al., 2016), which achieves up to 82% overall accuracy with an iterative technique that starts with already predicted secondary structures of proteins, and improves it by iteration over backbone torsion angles.

Apparently there is a limit of the accuracies of the predicted secondary structures, and this limit cannot be bigger than accuracies of experimental structure determination using X-ray crystallography or NMR. It is reported that there is 5% error margin for these techniques (Colloc'h et al., 1993). Another source of limit for accuracy is due to the nature of the problem. Although the progress in the techniques for the prediction of the secondary structures of proteins from their amino acid sequences are remarkable, predicting secondary structures of proteins based on amino acid sequences with an acceptable accuracy is unrealistic due to the poor correlation among primary/secondary sequence identities. In spite of the mentioned limitations, this study is unique with its claim that current PDB database is big enough and diverse to develop a segments similarity-based approach to predict the secondary structures of proteins from their

amino acid sequences with an acceptable accuracy if efficient algorithms to detect similar segments are developed.





4. CONCLUSION

Prediction of protein secondary structure based on its amino acid sequence is a common technique used to predict protein's three dimensional structure. In many studies, protein secondary structure had shown that it is a most beneficial feature to predict the tertiary structure of proteins (Fischer et al., 2001; Gong and Rose 2005; Meiler J, Baker, 2003; Rost, 2001). Even it has been proposed that secondary structure alone may be sufficient for accurate prediction of a protein's tertiary structure (Przytycka et al., 1999). In this thesis, the problem of secondary structure prediction is analyzed and several aspects of secondary structure prediction are investigated. An efficient new algorithm "Avicenna" for predicting secondary structure is presented and it's superiority to the current methods is shown. The algorithm developed in this study is proved to be able to identify similar segments of the proteins in the search space to the segments of the query protein. Assembling corresponding secondary sequence segments of the search space proteins in an appropriate way, secondary structure of the query protein is predicted with 83% accuracies in average. Obviously the use of the MAX-MATCH similarity for finding similar segments, to reduce the search space, majority voting in the assembly of corresponding secondary sequence segments of the search space, creating decisions at the several mismatch levels and aggregating of these decisions to create the final predicted secondary structure sequence, played important roles in the achievement of this high overall average accuracy. It should be underlined that the main structural advantage benefited is the lag in the extent of the secondary structural alterations, compared to the changes in the amino acid sequences. New similarity measures to reduce search space, new segment similarity definitions are among two topics of further studies.

5. REFERENCES

"Amino Acid". Cambridge Dictionaries Online. Cambridge University Press. 2015. Retrieved 2015-07-03.

Adamczak, R. (2009). Dimensionality reduction of pssm matrix and its influence on secondary structure and relative solvent accessibility predictions. *World Academy Of Science, Engineering And Technology*, 58, 657-664.

Akcesme, B., & Akcesme, F. B. (2014). Regression Analysis to Predict the Secondary Structure of Proteins. *SouthEast Europe Journal of Soft Computing*, 3(2), 8.

Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M. & Walter, P. (2013). *Essential cell biology*. Garland Science.

Alexandrov, N. N., Nussinov, R., & Zimmer, R. M. (1996). Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In *Pac. Symp. Biocomput* (Vol. 96, pp. 53-72).

Alirezaee, M., Dehzangi, A., & Mansoori, E. (2012). Predicting the secondary structure of proteins by cascading neural networks. *International Journal of Artificial Intelligence & Applications*, 3(6), 39.

Almassy, R. J., & Dickerson, R. E. (1978). Pseudomonas cytochrome c551 at 2.0 Å resolution: enlargement of the cytochrome c family. *Proceedings of the National Academy of Sciences*, 75(6), 2674-2678.

Alterman, M. A., & Hunziker, P. (Eds.). (2012). *Amino acid analysis: methods and protocols*. Humana Press.

Altschul, S. F., & Erickson, B. W. (1986). Optimal sequence alignment using affine gap costs. *Bulletin of mathematical biology*, 48(5-6), 603-616.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25, 3389-3402.

Anfinsen C. (1973) Principles that govern the folding of protein chains. *Science*;181:223–230.

Archibald, J. D. (2009). Edward Hitchcock's pre-Darwinian (1840) "Tree of life". *Journal of the History of Biology*, 42(3), 561-592.

- Arjunan, S. N. V., Deris, S., & Illias, R. M. (2012). Prediction of protein secondary structure. *Jurnal Teknologi*, 35(1), 81-90.
- Baldi, P., Brunak, S., Frasconi, P., Pollastri, G., & Soda, G. (2000). Bidirectional dynamics for protein secondary structure prediction. In *Sequence Learning* (pp. 80-104). Springer Berlin Heidelberg.
- Baldwin, R. L. (1996). How Hofmeister ion interactions affect protein stability. *Biophysical journal*, 71(4), 2056.
- Barillot, E., Calzone, L., Hupe, P., Vert, J. P., & Zinovyev, A. (2012). *Computational systems biology of cancer*. CRC Press.
- Barton, A. F. (1991). *CRC handbook of solubility parameters and other cohesion parameters*. CRC press.
- Benner, S. A., & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Advances in enzyme regulation*, 31, 121-181.
- Berg, J. M., Tymoczko, J. L., & Stryer, L. (2002). *Biochemistry* 5th Edition.
- Berman, H. M., Henrick, K., Nakamura, H., & Markley, J. L. (2009). The worldwide protein data bank. *Structural Bioinformatics*, 44, 293.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1), 235-242.
- Berman, H., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12), 980-980.
- Biou, V., Gibrat, J. F., Levin, J. M., Robson, B., & Garnier, J. (1988). Secondary structure prediction: combination of three different methods. *Protein Engineering*, 2(3), 185-191.
- Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M., Fredholm, H., Lautrup, B., & Petersen, S. B. (1990). A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS letters*, 261(1), 43-46.
- Bondugula R, Xu D: MUPRED: A tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. *Proteins*. 2007, 66 (3): 664-670.

Bork, P., Ouzonis, C., Sander, C., Scharf, M., Schneider, R. & Sonnhammer, E. (1992). What's in a genome? *Nature* (London) , 358, 287.

Bossa, F., & Pascarella, S. (1990). PRONET: a microcomputer program for predicting the secondary structure of proteins with a neural network. *CABIOS*, 5, 319-320.

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164 –170.

Brocchieri, L., & Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic acids research*, 33(10), 3390-3400.

Cao, M. D., Dix, T. I., & Allison, L. (2010). A genome alignment algorithm based on compression. *BMC bioinformatics*, 11(1), 1.

Carey, C. W. (2006). *American scientists*. Infobase Publishing.

Chao, E. D. (2002). Critical Analysis of Secondary Structure Prediction Algorithms. *Computational Molecular Biology* Professor Doug Brutlag, Stanford University June, 6.

Chamorro, A. E. M., Divina, F., Ruiz, J. S. A., & Cortés, G. A. (2010, September). Alpha helix prediction based on evolutionary computation. In *IAPR International Conference on Pattern Recognition in Bioinformatics* (pp. 358-367). Springer Berlin Heidelberg.

Chandrasekhar, K., Dileep A. , Ester Lebonah D., Pramoda K. (2014) A Short Review on Proteomics and its Applications. *International Letters of Natural Sciences* 17, 77-84

Chatterjee, P., Basu, S., Kundu, M., Nasipuri, M., & Plewczynski, D. (2011). PSP_MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines. *Journal of molecular modeling*, 17(9), 2191-2201.

Chen, K. E., Kurgan, L. A., & Ruan, J. (2008). Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *Journal of computational chemistry*, 29(10), 1596-1604.

Cheng, J., & Baldi, P. (2007). Improved residue contact prediction using support vector machines and a large feature set. *BMC bioinformatics*, 8(1), 1.

- Chiang, Y. S., Gelfand, T. I., Kister, A. E., & Gelfand, I. M. (2007). New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage. *Proteins: Structure, Function, and Bioinformatics*, 68(4), 915-921.
- Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, 357(6379), 543.
- Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4), 823.
- Chou, K. C., & Zhang, C. T. (1994). Predicting protein folding types by distance functions that make allowances for amino acid interactions. *Journal of Biological Chemistry*, 269(35), 22014-22020.
- Chou, K. C., & Zhang, C. T. (1994). Predicting protein folding types by distance functions that make allowances for amino acid interactions. *Journal of Biological Chemistry*, 269(35), 22014-22020.
- Chou, P. Y., & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry*, 13(2), 222-245.
- Chou, P. Y., & Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annual review of biochemistry*, 47(1), 251-276.
- Cochran, W.; Crick, F. H.; Vand, V. (1952). "The structure of synthetic polypeptides. I. The transform of atoms on a helix". *Acta Crystallographica* 5 (5): 581–6. doi:10.1107/S0365110X52001635
- Cohen, C., & Szent-Gyorgyi, A. G. (1957). Optical rotation and helical polypeptide chain configuration in α -proteins. *Journal of the American Chemical Society*, 79(1), 248-248.
- Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., & Fletterick, R. J. (1986). Turn prediction in proteins using a pattern-matching approach. *Biochemistry*, 25(1), 266-275.
- Cooper, G. M., & Hausman, R. E. (2004). *The cell: A molecular approach*. Washington, D.C.: ASM Press.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

- Creighton, T. E. (1990). Protein folding. *Biochemical journal*, 270(1), 1.
- Creighton, T. H. (1993). Chapter 1. Proteins: structures and molecular properties. San Francisco W.H Freeman
- Cronin Jr, J. J., & Taylor, S. A. (1994). SERVPERF versus SERVQUAL: reconciling performance-based and perceptions-minus-expectations measurement of service quality. *The Journal of Marketing*, 125-131.
- Cuff, A. L., Sillitoe, I., Lewis, T., Redfern, O. C., Garratt, R., Thornton, J., & Orengo, C. A. (2009). The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic acids research*, 37(suppl 1), D310-D314.
- Cuff, J. A., & Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 34(4), 508-519.
- Dai, Q., Li, Y., Liu, X., Yao, Y., Cao, Y., & He, P. (2013). Comparison study on statistical features of predicted secondary structures for protein structural class prediction: From content to position. *BMC bioinformatics*, 14(1), 1.
- Di Francesco, V., Garnier, J., & Munson, P. J. (1997). Protein topology recognition from secondary structure sequences: Application of the hidden Markov models to the alpha class proteins. *Journal of Molecular Biology*, 267(2), 446-463.
- Do, C. B., & Katoh, K. (2008). Protein multiple sequence alignment. *Functional Proteomics: Methods and Protocols*, 379-413.
- Fair, D. S., Marlar, R. A., & Levin, E. G. (1986). Human endothelial cells synthesize protein S. *Blood*, 67(4), 1168-1171.
- Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., & Bairoch, A. (2002). The PROSITE database, its status in 2002. *Nucleic acids research*, 30(1), 235-238.
- Finkelstein, A. V., & Ptitsyn, O. B. (1971). Statistical analysis of the correlation among amino acid residues in helical, β -structural and non-regular regions of globular proteins. *Journal of molecular biology*, 62(3), 613-624.
- Fischer, D., & Eisenberg, D. (1996). Protein fold recognition using sequence derived predictions. *Protein Science*, 5(5), 947-955.

Fischer, D., Elofsson, A., Rychlewski, L., Pazos, F., Valencia, A., Rost, B., ... & Dunbrack, R. L. (2001). CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 45(S5), 171-183.

Fox, N. K., Brenner, S. E., & Chandonia, J. M. (2014). SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic acids research*, 42(D1), D304-D309.

Fox, N. K., Brenner, S. E., & Chandonia, J. M. (2015). The value of protein structure classification information—Surveying the scientific literature. *Proteins: Structure, Function, and Bioinformatics*, 83(11), 2025-2038.

Frampton, J., Leutz, A., Gibson, T. J., & Graf, T. (1989). DNA-binding domain ancestry. *Nature*, 342, 134.

Freifelder, D. (1983). *Physical Biochemistry* (2nd ed.). W. H. Freeman and Company

Frishman, D., & Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. *Proteins-Structure Function and Genetics*, 27(3), 329-335.

Fulcher, J. (2008). *Computational intelligence: a compendium* (Vol. 21). L. C. Jain (Ed.). Heidelberg: Springer.

Fürst, P., & Stehle, P. (2004). What are the essential elements needed for the determination of amino acid requirements in humans?. *The Journal of nutrition*, 134(6), 1558S-1565S.

Gandhi, K., Kumar, A., Gosewade, S., Kaushik, R., & Lal, D. (2013). X-Ray Crystallography and Its Applications in Dairy Science: A Review. *Research & Reviews: Journal of Dairy Science & Technology*, 2(1), 21-33.

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of molecular biology*, 120(1), 97-120.

Garrett, R. C., Thornton, J. M. & Taylor, W. R. (1991). An extension of secondary structure prediction towards the prediction of the tertiary structure. *PE BS Letters*, 280, 141-146

Gaur, D. (2016, February), Primary Structure Of Protein. Retrieved from <https://drgrpoinstitute.wordpress.com/tag/primary-structure-of-protein/>

Geourjon, C., & Deleage, G. (1994). SOPM: a self-optimized method for protein secondary structure prediction. *Protein Engineering*, 7(2), 157-164.

Gibbs, A. J., & McIntyre, G. A. (1970). The diagram, a method for comparing sequences. *European Journal of Biochemistry*, 16(1), 1-11.

Gibrat, J. F., Garnier, J., & Robson, B. (1987). Further developments of protein secondary structure prediction using information theory: new parameters and consideration of residue pairs. *Journal of molecular biology*, 198(3), 425-443.

Gibrat, J. F., Madej, T., & Bryant, S. H. (1996). Surprising similarities in structure comparison. *Current opinion in structural biology*, 6(3), 377-385.

Gong, H., & Rose, G. D. (2005). Does secondary structure determine tertiary structure in proteins?. *Proteins: Structure, Function, and Bioinformatics*, 61(2), 338-343.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, 162(3), 705-708.

Greer, J. (1981). Comparative model-building of the mammalian serine proteases. *Journal of molecular biology*, 153(4), 1027-1042.

Guo, J., Chen, H., Sun, Z., & Lin, Y. (2004). A novel method for protein secondary structure prediction using dual-layer SVM and profiles. *PROTEINS: Structure, Function, and Bioinformatics*, 54(4), 738-743.

Guzzo, A. V. (1965). The influence of amino acid sequence on protein structure. *Biophysical journal*, 5(6), 809.

Hammarström, P., Wiseman, R. L., Powers, E. T., & Kelly, J. W. (2003). Prevention of transthyretin amyloid disease by changing protein misfolding energetics. *Science*, 299(5607), 713-716.

Harrison, A., Pearl, F., Mott, R., Thornton, J. & Orengo, C. (2002) A fast method for reliably recognising the fold of a protein structure, *J. Mol. Biol.* 323, 909 –926.

Harrison, C., Ketchen, A. M., Redhead, N. J., O'Sullivan, M. J., & Melton, D. W. (2002). Replication failure, genome instability, and increased cancer susceptibility in mice with a point mutation in the DNA ligase I gene. *Cancer research*, 62(14), 4065-4074.

Hartley, H. (1951). Origin of the Word Protein'. *Nature*, 168, 244.

Hecht, H. J., Sobek, H., Haag, T., Pfeifer, O., & Van Pée, K. H. (1994). The metal-ion-free oxidoreductase from *Streptomyces aureofaciens* has an α/β hydrolase fold. *Nature Structural & Molecular Biology*, 1(8), 532-537.

Heffernan, R., Paliwal K. Lyons J., Dehzangi A., Sharma A., Wang j., Sattar A., Yang Y. and Zhou Y. (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Scientific Report

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22), 10915-10919.

Heringa, J. (Jul 2008) Sequence Similarity. In: eLS. John Wiley & Sons Ltd,

Hirst, J. D., & Sternberg, M. J. (1992). Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry*, 31(32), 7211-7218.

Hobohm, U., & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Science*, 3(3), 522-524.

Holley, L. H., & Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences*, 86(1), 152-156.

Holm, L., & Sander, C. (1995). Dali: a network tool for protein structure comparison. *Trends in biochemical sciences*, 20(11), 478-480.

Holm, L., & Sander, C. (1996). Mapping the protein universe. *Science*, 273(5275), 595-602.

Hooft, R. W., Sander, C., & Vriend, G. (1997). Objectively judging the quality of a protein structure from a Ramachandran plot. *Computer applications in the biosciences: CABIOS*, 13(4), 425-430.

Hovmöller, S., Zhou, T., & Ohlson, T. (2002). Conformations of amino acids in proteins. *Acta Crystallographica Section D: Biological Crystallography*, 58(5), 768-776.

Hua, S., & Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8), 721-728.

Huai, Q., Sun, Y., Wang, H., Chin, L. S., Li, L., Robinson, H., & Ke, H. (2003). Crystal structure of DJ-1/RS and implication on familial Parkinson's disease 1. *FEBS letters*, 549(1-3), 171-175.

Hudgins, RR; Jarrold MF (1999). "Helix Formation in Unsolvated Alanine-Based Peptides: Helical Monomers and Helical Dimers". *Journal of the American Chemical Society* 121 (14): 3494–3501.doi:10.1021/ja983996a

Hunter, G. K. (2000). *Vital forces: the discovery of the molecular basis of life*. Academic Press.

Islam, S. A., Luo, J., & Sternberg, M. J. (1995). Identification and analysis of domains in proteins. *Protein Engineering*, 8(6), 513-526.

IUPAC, (1997). *Compendium of Chemical Terminology*, 2nd ed. (the "Gold Book")

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2), 195-202.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577-2637.

Karypis, G. (2006). YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 64(3), 575-586.

Kazakov, V. A., & Migdal, A. A. (1988). Recent progress in the theory of noncritical strings. *Nuclear Physics B*, 311(1), 171-190.

Kertész-Farkas, A. (2008). *Protein Classification in a Machine Learning Framework* (Doctoral dissertation, University of Szeged).

Kihara, D. & Skolnick, J. (2003) The PDB is a covering set of small protein structures. *J. Mol. Biol.* 334, 793– 802.

King, R. D., & Sternberg, M. J. (1990). Machine learning approach for the prediction of protein secondary structure. *Journal of molecular biology*, 216(2), 441-457.

Kleywegt, G. J., & Jones, T. A. (1996). Phi/psi-chology: Ramachandran revisited. *Structure*, 4(12), 1395-1400.

Kloczkowski, A., Ting, K. L., Jernigan, R. L., & Garnier, J. (2002). Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins: Structure, Function, and Bioinformatics*, 49(2), 154-166.

- Kneller, D. G., Cohen, F. E., & Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *Journal of molecular biology*, 214(1), 171-182.
- Koehl, P., & Levitt, M. (1999). Structure-based conformational preferences of amino acids. *Proceedings of the National Academy of Sciences*, 96(22), 12524-12529.
- Koh, I. Y., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Eswar, N., ... & Rost, B. (2003). EVA: evaluation of protein structure prediction servers. *Nucleic Acids Research*, 31(13), 3311-3315.
- Komar, A. A. (2007). SNPs, silent but not invisible. *DNA*, 5(3).
- Kotelchuck, D., & Scheraga, H. A. (1969). The influence of short-range interactions on protein conformation, II. A model for predicting the α -helical regions of proteins. *Proceedings of the National Academy of Sciences*, 62(1), 14-21.
- Kunz, H. (2003). Emil Fischer—Unequalled Classicist, Master of Organic Chemistry Research, and Inspired Trailblazer of Biological Chemistry. *Angewandte Chemie International Edition*, 42(3), 252-252.
- Kurgan, L. A., & Homaeian, L. (2006). Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. *Pattern Recognition*, 39(12), 2323-2343.
- Kurgan, L., Cios, K., & Chen, K. (2008). SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC bioinformatics*, 9(1), 1.
- Kuzniar, A., Dhir, S., Nijveen, H., Pongor, S., & Leunissen, J. A. (2010). Multi-netclust: an efficient tool for finding connected clusters in multi-parametric networks. *Bioinformatics*, 26(19), 2482-2483.
- Latham, M. C. (1997). Human nutrition in the developing world (No. 29). Food & Agriculture Org..
- Lehninger, A. L., Nelson, D. L., & Cox, M. M. (2000). *Lehninger principles of biochemistry*. New York: Worth Publishers.
- Lesk, A. M., & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *Journal of molecular biology*, 136(3), 225-230.

Lesk, A. M., & Chothia, C. (1982). Evolution of proteins formed by β -sheets: II. The core of the immunoglobulin domains. *Journal of molecular biology*, 160(2), 325-342.

Levin, J. M., & Garnier, J. (1988). Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 955(3), 283-295.

Levin, J. M., Robson, B., & Garnier, J. (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS letters*, 205(2), 303-308.

Levitt, M. (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences*, 106(27), 11079-11084.

Liljas, L., Piškur, J., Lindblom, G., Nissen, P., & Kjeldgaard, M. (2009). *Textbook of structural biology* (pp. 377-380). Singapore, Singapore: World Scientific.

Lim, V. I. (1974). Algorithms for prediction of α -helical and β -structural regions in globular proteins. *Journal of Molecular Biology*, 88(4), 873-894.

Lin, H. N., Sung, T. Y., Ho, S. Y., & Hsu, W. L. (2010). Improving protein secondary structure prediction based on short subsequences with local structure similarity. *Bmc Genomics*, 11(4), 1.

Linderstrøm-Lang, K. (1952). *Proteins and enzymes* (Vol. 6). Stanford university press.

Lipman, D. J., & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, 227(4693), 1435-1441.

Lodish, H. (2008). *Molecular cell biology*. Macmillan.

Lodish, H., Berk, A., & Zipursky, S. L. (2000). *Molecular cell biology* 4th edition.

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I., Word, J. M., Prisant, M. G., ... & Richardson, D. C. (2003). Structure validation by $C\alpha$ geometry: ϕ , ψ and $C\beta$ deviation. *Proteins: Structure, Function, and Bioinformatics*, 50(3), 437-450.

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I., Word, J. M., Prisant, M. G., ... & Richardson, D. C. (2003). Structure validation by $C\alpha$ geometry: ϕ , ψ and $C\beta$ deviation. *Proteins: Structure, Function, and Bioinformatics*, 50(3), 437-450

Maclin, R., & Shavlik, J. W. (1993). Using knowledge-based neural networks to improve algorithms: Refining the Chou-Fasman algorithm for protein folding. In *Multistrategy Learning* (pp. 87-107). Springer US.

Mason, S. F. (1997). The science and humanism of Linus Pauling (1901–1994). *Chem. Soc. Rev.*, 26(1), 29-39.

Mathews, C. K., van Holde, K. E., & Ahern, K. G. (2000). *Biochemistry*. 2000. San Francisco: Benjamin Cummings.

Mattevi, A., Valentini, G., Rizzi, M., Speranza, M. L., Bolognesi, M., & Coda, A. (1995). Crystal structure of *Escherichia coli* pyruvate kinase type I: molecular basis of the allosteric transition. *Structure*, 3(7), 729-741.

Matthews B. (1995) Studies on protein stability with T4 lysozyme. *Adv Protein Chem*;46:249–278.

Maxfield, F. R., & Scheraga, H. A. (1979). Improvements in the prediction of protein backbone topography by reduction of statistical errors. *Biochemistry*, 18(4), 697-704.

McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, 16(4), 404-405.

McMillan, B. J., Schnute, B., Ohlenhard, N., Zimmerman, B., Miles, L., Beglova, N., ... & Blacklow, S. C. (2015). A Tail of Two Sites: A Bipartite Mechanism for Recognition of Notch Ligands by Mind Bomb E3 Ligases. *Molecular cell*, 57(5), 912-924.

McPherson, A. (2004). Introduction to protein crystallization. *Methods*, 34(3), 254-265.

Meiler J, Baker D: Coupled prediction of protein secondary and tertiary structure. *Proceedings of the National Academy of Sciences of the United States of America*. 2003, 100 (21): 12105-12110.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Mona S. (2001). *Predicting Protein Secondary and Supersecondary Structure*. (pp 29-7). CRC Press, LLC

Mondal, M. N. I., Al Mamun, M., & Rahman, M. Z. (2015) An Approach for Protein Secondary Structure Prediction Using Neural Network. 3(6),

Morris, A. L., MacArthur, M. W., Hutchinson, E. G., & Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics*, 12(4), 345-364.

Moult, J., Hubbard, T., Fidelis, K., & Pedersen, J. T. (1999). Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins: Struct*

Mount, D. W., & Mount, D. W. (2001). Bioinformatics: sequence and genome analysis (Vol. 2). New York: Cold spring harbor laboratory press.

Mount, W. D.(2004). Bioinformatics: sequence and genome analysis (II. Edition). New York:: Cold spring harbor laboratory press.

Muggleton, S., King, R. D., & Stenberg, M. J. (1992). Protein secondary structure prediction using logic-based machine learning. *Protein Engineering*,5(7), 647-657.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4), 536-540.

Musacchio, A., Gibson, T., Rice, P., Thompson, J., & Saraste, M. (1993). The PH domain: a common piece in the structural pathwork of signalling proteins. *Trends in biochemical sciences*, 18(9), 343-348.

Musacchio, A., Noble, M., Pauptit, R., Wierenga, R., & Saraste, M. (1992). Crystal structure of a Src-homology 3 (SH3) domain.

Nagano, K. (1973). Logical analysis of the mechanism of protein folding: I. Predictions of helices, loops and β -structures from primary structure. *Journal of molecular biology*, 75(2), 401-420.

Nagano, K., & Hasegawa, K. (1975). Logical analysis of the mechanism of protein folding: III. Prediction of the strong long-range interactions. *Journal of molecular biology*, 94(2), 257-281.

Nakai, K., Kidera, A., & Kanehisa, M. (1988). Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Engineering*,2(2), 93-100.

Nath, A., Chaube, R., & Subbiah, K. (2013). An insight into the molecular basis for convergent evolution in fish antifreeze Proteins. *Computers in biology and medicine*, 43(7), 817-821.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1), 31-88.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3), 443-453.

Niermann, T., & Kirschner, K. (1991). Improving the prediction of secondary structure of 'TIM-barrel' enzymes. *Protein engineering*, 4(3), 359-370.

Nishikawa, K., & Ooi, T. (1986). Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 871(1), 45-54.

Nishikawa, K., & Ooi, T. (1986). Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 871(1), 45-54.

O'Connor, C. M., Adams, J. U., & Fairman, J. (2010). *Essentials of cell biology*. NPG Education, Cambridge.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, 5(8), 1093-1109.

Osborne, T. B., Mendel, L. B., & Ferry, E. L. (1911). Feeding experiments with isolated food-substances (No. 156). Carnegie institution of Washington.

Ouzounis, c. a., and Melvin, w. t. (1991). Primary and secondary structural patterns in eukaryotic cytochrome P-450 families correspond to structures of the helix-rich domain of *Pseudomonas putida* cytochrome P-450cam. *European Journal of Biochemistry*, 198(2), 307-315.

Ouzounis, C., Sander, C., Scharf, M., & Schneider, R. (1993). Prediction of protein structure by evaluation of sequence-structure fitness: aligning sequences to contact profiles derived from three-dimensional structures. *Journal of molecular biology*, 232(3), 805-825.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756), 877-884.

Pain, R. H., & Robson, B. (1970). Analysis of the code relating sequence to secondary structure in proteins. *Nature*, 227, 62-63.

Pauling, L., Corey, R. B., & Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4), 205-211.

Pearson, W. R. (2013). An introduction to sequence similarity (“homology”) searching. *Current protocols in bioinformatics*, 3-1.

- Pelton, J. T., & McLean, L. R. (2000). Spectroscopic methods for analysis of protein secondary structure. *Analytical biochemistry*, 277(2), 167-176.
- Periti, P. F., Quagliarotti, G., & Liquori, A. M. (1967). Recognition of α -helical segments in proteins of known primary structure. *Journal of molecular biology*, 24(2), 313-322.
- Perutz, M. F., Kendrew, J. C., & Watson, H. C. (1965). Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *Journal of Molecular Biology*, 13(3), 669-678.
- Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M., & Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC structural biology*, 9(1), 1.
- Petsko, G. A., & Ringe, D. (2004). *Protein structure and function*. New Science Press.
- Pidwirny, M. (2006). "Species Diversity and Biodiversity". *Fundamentals of Physical Geography*, 2nd Edition
- Presnell, S. R., Cohen, B. I., & Cohen, F. E. (1992). A segment-based approach to protein secondary structure prediction. *Biochemistry*, 31(4), 983-993.
- Prevelige Jr, P., & Fasman, G. D. (1989). Chou-Fasman prediction of the secondary structure of proteins. In *Prediction of protein structure and the principles of protein conformation* (pp. 391-416). Springer US.
- Prothero, J. W. (1966). Correlation between the distribution of amino acids and alpha helices. *Biophysical journal*, 6(3), 367.
- Przytycka T, Aurora R, Rose GD: A protein taxonomy based on secondary structure. *Nature Structural Biology*. 1999, 6 (7): 672-682.
- Ptitsyn, O. B. (1969). Statistical analysis of the distribution of amino acid residues among helical and non-helical regions in globular proteins. *Journal of molecular biology*, 42(3), 501-510.
- Ptitsyn, O. B., & Finkelstein, A. V. (1983). Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*, 22(1), 15-25.
- Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of molecular biology*, 202(4), 865-884.

Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*, 7(1), 95-99.

Rangwala, H., & Karypis, G. (Eds.). (2011). *Introduction to Protein Structure Prediction: Methods and Algorithms* (Vol. 18). John Wiley & Sons.

Read, R. J., Brayer, G. D., Jurasek, L., & James, M. N. (1984). Critical evaluation of comparative model building of *Streptomyces griseus* trypsin. *Biochemistry*, 23(26), 6570-6575.

Reece, J. B., Meyers, N., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V. & Cooke, B. N. (2011). *Campbell Biology: Australian Version*, 9e.

Reeds, P. J. (2000). Dispensable and indispensable amino acids for humans. *The Journal of Nutrition*, 130(7), 1835S-1840S.

Rekapalli, B., Wuichet, K., Peterson, G. D., & Zhulin, I. B. (2012). Dynamics of domain coverage of the protein sequence universe. *BMC genomics*, 13(1), 634.

Rhodes, G. (2006) *Crystallography Made Crystal Clear*, Third Edition: A Guide for Users of Macromolecular Models, 3rd Ed., Academic Press

Robson, B., & Pain, R. H. (1971). Analysis of the code relating sequence to conformation in proteins: possible implications for the mechanism of formation of helical regions. *Journal of molecular biology*, 58(1), 237-257.

Rooman, M. J. & Wodak, S. (1991). Weak correlation between predictive power of individual sequence patterns and overall prediction accuracy in proteins. *Proteins: Struct. Funct. Genet.* 9, 69-78.

Rooman, M. J., Kocher, J. P. A., & Wodak, S. J. (1991). Prediction of protein backbone conformation based on seven structure assignments: influence of local interactions. *Journal of molecular biology*, 221(3), 961-979.

Rooman, M. J., Wodak, S. & Thornton, J. M. (1989). Amino acid sequence templates derived from recurrent turn motifs in proteins: critical evaluation of their predictive power. *Protein Eng.* 3, 23-27

Rose GD, Gierasch LM, Smith JA. (1985). "Turns in peptides and proteins". *Adv Protein Chem* 37: 1–109. doi:10.1016/S0065-3233(08)60063-7. PMID 2865874

Rosenfeld, L. (2012). *Origins of Clinical Chemistry: The Evolution of Protein Analysis*. Elsevier.

- Rost B: (2001). Review: Protein secondary structure prediction continues to rise. *Journal of Structural Biology*, 134 (2-3): 204-218.
- Rost, B., & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of molecular biology*, 232(2), 584-599.
- Rost, B., Sander, C., Casadio, R., & Fariselli, P. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Science*, 4(3), 521-533.
- Rost, B., Yachdav, G., & Liu, J. (2004). The predictprotein server. *Nucleic acids research*, 32(suppl 2), W321-W326.
- Russell, J. B., O'connor, J. D., Fox, D. G., Van Soest, P. J., & Sniffen, C. J. (1992). A net carbohydrate and protein system for evaluating cattle diets: I. Ruminant fermentation. *Journal of Animal Science*, 70(11), 3551-3561.
- Rychlewski, L., & Godzik, A. (1997). Secondary structure prediction using segment similarity. *Protein Engineering*, 10(10), 1143-1153.
- Salamov, A. A., & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of molecular biology*, 247(1), 11-15.
- Salemme, F. R., Miller, M. D., & Jordan, S. R. (1977). Structural convergence during protein evolution. *Proceedings of the National Academy of Sciences*, 74(7), 2820-2824.
- Sali, A., & Blundell, T. L. (1990). Definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *Journal of molecular biology*, 212(2), 403-428.
- Šali, A., Overington, J. P., Johnson, M. S., & Blundell, T. L. (1990). From comparisons of protein sequences and structures to protein modelling and design. *Trends in biochemical sciences*, 15(6), 235-240.
- Salzberg, S., & Cost, S. (1992). Predicting protein secondary structure with a nearest-neighbor algorithm. *Journal of molecular biology*, 227(2), 371-374.
- Sapolsky, R. M. (2005). *Biology and human behavior: The neurological origins of individuality*. Teaching Company.
- Schiffer, M., & Edmundson, A. B. (1967). Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophysical journal*, 7(2), 121.

- Schulz, G. E., & Schirmer, R. H. (1979). Patterns of folding and association of polypeptide chains. In *Principles of Protein Structure* (pp. 66-107). Springer New York.
- Selkoe, D. J. (2003). Folding proteins in fatal ways. *Nature*, 426(6968), 900-904.
- Sen, T. Z., Jernigan, R. L., Garnier, J., & Kloczkowski, A. (2005). GOR V server for protein secondary structure prediction. *Bioinformatics*, 21(11), 2787-2788.
- Shindyalov, I. N., & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein engineering*, 11(9), 739-747.
- Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., ... & Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in bioinformatics*, 3(3), 265-274.
- Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., ... & Lehtinen, S. (2014). CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research*, gku947.
- Simossis, V. A., & Heringa, J. (2004). Integrating protein secondary structure prediction and multiple sequence alignment. *Current Protein and Peptide Science*, 5(4), 249-266.
- Simossis, V. A., & Heringa, J. (2004). Optimally segmented consensus secondary structure prediction. *Bioinformatics*.
- Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America*, (4), 1029-1034.
- Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1), 195-197.
- Stolorz, P., Lapedes, A., & Xia, Y. (1992). Predicting protein secondary structure using neural net and statistical methods. *Journal of Molecular Biology*, 225(2), 363-377.
- Stormo, G. D., Schneider, T. D., Gold, L., & Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Research*, 10(9), 2997-3011.
- Blundell T. and Campillo N, (2016). Wellcome Images, Retrieved from <http://bigpictureeducation.com/proteins>
- Taylor, W. R., & Thornton, J. M. (1983). Prediction of super-secondary structure in proteins.

Taylor, W. R., Flores, T. P., & Orengo, C. A. (1994). Multiple protein structure alignment. *Protein Science*, 3(10), 1858-1870.

Uttamapinant, C., White, K. A., Baruah, H., Thompson, S., Fernández-Suárez, M., Puthenveetil, S., & Ting, A. Y. (2010). A fluorophore ligase for site-specific protein labeling inside living cells. *Proceedings of the National Academy of Sciences*, 107(24), 10914-10919.

Wagner, I., & Musso, H. (1983). New naturally occurring amino acids. *Angewandte Chemie International Edition in English*, 22(11), 816-828.

Walker, L. C., & LeVine, H. (2000). The cerebral proteopathies. *Molecular neurobiology*, 21(1-2), 83-95.

Wang Z, Zhao F, Peng J, Xu J. (2011) Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics*.

Wang, Z. X., & Yuan, Z. (2000). How good is prediction of protein structural class by the component-coupled method?. *Proteins: Structure, Function, and Bioinformatics*, 38(2), 165-175.

Waterman, M.S. (1995) *Introduction to Computational Biology: Maps, Sequences and Genomes (Interdisciplinary Statistics)*. Chapman and Hall, London.

Wheeler, S. J., Marchler-Bauer, A., & Bryant, S. H. (2000). Domain size distributions can predict domain boundaries. *Bioinformatics*, 16(7), 613-618.

Whitford, D. (2013). *Proteins: structure and function*. John Wiley & Sons.

Wieland, T., & Bodanszky, M. (2012). *The world of peptides: a brief history of peptide chemistry*. Springer Science & Business Media.

Wilmanns, M., & Eisenberg, D. (1993). Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold. *Proceedings of the National Academy of Sciences*, 90(4), 1379-1383.

Wolfram Research, Inc. (n.d.). WolframAlpha. Retrieved May 4, 2011, from <http://www.wolframalpha.com/input/?i=average+temperature+in+Panama+2001>

Xiang, Z. (2006). Advances in homology protein structure modeling. *Current protein & peptide science*, 7(3), 217.

- Yang, A. S., & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *Journal of molecular biology*, 301(3), 665-678.
- Yang, Z. R. (2004). Biological applications of support vector machines. *Briefings in bioinformatics*, 5(4), 328-338.
- Yi, T. M., & Lander, E. S. (1994). Recognition of related proteins by iterative template refinement (ITR). *Protein Science*, 3(8), 1315-1328.
- Zhang, X., Mesirov, J. P., & Waltz, D. L. (1992). Hybrid system for protein secondary structure prediction. *Journal of molecular biology*, 225(4), 1049-1063.
- Zhang, Y., & Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4), 1029-1034.
- Zhang, Z., Miteva, M. A., Wang, L., & Alexov, E. (2012). Analyzing effects of naturally occurring missense mutations. *Computational and mathematical methods in medicine*.
- Zhang, X., Mesirov, J. P. and Waltz, D. L. (1992) Hybrid system for protein secondary structure prediction., *J. Mol. Biol.*, 225, 1049–1063.
- Zheng, C., & Kurgan, L. (2008). Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. *BMC bioinformatics*, 9(1), 1.
- Zhou, T., Shu, N., & Hovmöller, S. (2010). A novel method for accurate one-dimensional protein structure prediction based on fragment matching. *Bioinformatics*, 26(4), 470-477.
- Zou, Z., & Zhang, J. (2015). No genome-wide protein sequence convergence for echolocation. *Molecular biology and evolution*.
- Zvelebil, M. J., Barton, G. J., Taylor, W. R., & Sternberg, M. J. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *Journal of molecular biology*, 195(4), 957-961.