EVALUATION OF PROTEIN SECONDARY STRUCTURE PREDICTION ALGORITHMS ON A NEW ADVANCED BENCHMARK DATASET

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Molecular Biology and Genetics

by
Canan HAS

December 2011
İZMİR

We approve the thesis of **Canan Has**

Assoc. Prof. Dr. Jens ALLMER
Supervisor

Assoc. Prof. Dr. Bilge KARAÇALI
Committee Member

Assoc. Prof. Dr. Ahmet KOÇ
Committee Member

20 December 2011

Assoc. Prof. Dr. Ahmet KOÇ
Head of the Department of Molecular Biologly and Genetics

Prof. Dr. R. Tuğrul SENGER
Dean of the Graduate School of Engineering and Sciences

ACKNOWLEDGEMENTS

Words are not enough to explain my feelings now, while writing. I decided to be a bioinformatician without any idea how it could be. Life is quite interesting. Years passed over, and I tried to do my best to reach the goal. Sometimes I felt disappointed and decided to give up. However, I couldn't. Some people in my life and myself convinced me to continue and reminded me how much I want to do this job and how much I love to do that.

I have to express my deepest regards to my adviser Dr. Jens Allmer. His wisdom, knowledge, passion, curiosity on science and his personality inspired me a lot. I am so grateful that he gave me a chance to study with him.

I would like to thank Dr. Bilge Karaçalı, Dr. Ahmet Koç for being my committee members. I would like to thank Dr. Tuğkan Tuğlular, Dr. Bünyamin Akgül, Dr. Avni Kuru, Dr. Nazlı Arda for their contribution on my education.

I would like to all my friends, especially Aslı Kartal, Funda Akıncıoğlu, Özge Yoluk, Hande Dirim, Adem Kocaman for their unforgettable support in my life. I should also state my thanks to all IYTE-Bioinformatics Lab members and my friends in IYTE-MBG and IYTE-Computer Engineering departments.

To Hasan Has: Her zaman bana örnek olduğun, eğitim ve öğretimin önemini bana öğrettiğin için sana çok teşekkür ederim. Sen benim dedem olduğun için çok şanslıyım.

To Ercan Peker: Bütün ailemin desteğini, bana inancını hayatım boyunca derinden hissettim. Senin desteğin, sözlerin ve varlığın ise hep hayatımın tam ortasında bana hep yol gösterici oldu. Bu yüzden bu alanda çalışmanın, başarılara imza atmanın mutluluğu bana her adımımda elinden gelen yardımı yapan sevgili dayımla özel olarak paylaşılmalıydı.

To my beloved family: Hard to explain how much all of you mean for me. I am who I am with your support and love. "We" succeed not me. I would like to thank: My father, Erdal: Thanks that you are my father. You taught me to think others as much as myself, and be honest at first to myself. My mother, Azize: I couldn't get over any kind of problem without your light on my life. My brother, Hakan: I haven't met any body else like you; such a person who is so special, talented and mature.

ABSTRACT

EVALUATION OF PROTEIN SECONDARY STRUCTURE PREDICTION ALGORITHMS ON A NEW ADVANCED BENCHMARK DATASET

Starting from 1970s, researchers have been studying secondary structure prediction. However the accuracy of state-of art methods reach to approximately 80-85%. One of the reasons for that is related with the limitations in respect to datasets used for training or testing the algorithm. A number of databases with *n* number of experimentally determined proteins, which also contain the knowledge of functionality, biochemical properties and location annotation of proteins, will directly show us how the algorithms work on certain groups of proteins. This also ensures opportunity to users to determine the quality of algorithms on those datasets and to decide on which algorithm can be used for which type of proteins.

In this thesis, the objective is set through the development of a new and advanced protein benchmark database which contains functional and biochemical information of experimentally defined 64872 proteins in S2C database derived by ProteinDataBank (PDB). With this database, the seven available predictors are evaluated in respect to their performances on different datasets in terms of functionality and subcellular localization of proteins in the benchmark database. According to the results obtained on proposed benchmark datasets in compare to results on one of existing dataset, RS126, it was shown that grouping proteins into functions in their subcellular localizations have a great impact on deciding the accuracies of existing algorithms.

ÖZET

PROTEİN İKİNCİL YAPI TAHMİNİ ALGORİTMALARININ YENİ VE İLERİ KIYASLAMALI VERİTABANINDA DEĞERLENDİRİLMELERİ

1970'lerden bu yana araştırmacılar protein ikincil yapı tahmini çalışmaktadırlar. Fakat günümüzdeki metotların kesinliği yaklaşık olarak %80-85' e ulaşmaktadır. Bunun sebeplerinden biri algoritmanın eğitimi ve testinde kullanılan veri setlerinden kaynaklı kısıtlamalardır. N sayıdaki deneysel olarak tanımlanmış proteini ve aynı zamanda proteinlerin fonksiyonlarına, biyokimyasal özelliklerine ve lokasyon anotasyonlarına dair bilgileri içeren veri setleri bize algoritmalarının belli protein grupları üzerinde nasıl çalıştıklarını direk gösterecektir. Bu ayrıca kullanıcılara veri setlerinde algoritmalarının kalitelerini belirleme ve hangi algoritmanın hangi tip protein grubu üzerinde kullanılabileceğine karar verme konusunda olanak tanıyacaktır.

Bu tez çalışmasında, çalışma amacı deneysel olarak tanımlanmış ProteinVeriBankası (PDB) 'den elde edilmiş S2C veritabanında yer alan 64872 proteinin fonksiyonel ve biyokimyasal bilgilerini içeren yeni ve ileri protein veritabanı geliştirilmesi olarak belirlenmiştir. Bu veritabanı ile 7 ulaşılabilir tahmin algoritmasının performansları proteinlerin fonksiyonları ve lokalizasyonları bakımından farklı standart veri setleri üzerinde değerlendirilmiştir. Önerilen standart verisetlerinden elde edilen sonuçların var olan RS126 verisetinde elde edilmiş sonuçlarıyla kıyasına göre, proteinleri hücresel lokalizasyonlarındaki fonksiyonlarına göre gruplamanın var olan algoritmaların objektif değerlendirilmesinde büyük etkisi olduğu gösterilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES x

LIST OF TABLES xii

LIST OF ABBREVIATIONS xiii

CHAPTER 1. INTRODUCTION 1

1.1. Introduction to Structural Proteomics 1

1.2. Overview on Experimental Techniques 3

1.3. Computational Studies on Protein Tertiary Structure Prediction 5

1.4. Overview of The Protein Secondary Structure 7

1.5. Overview of Existing Algorithms 9

1.6. Overview on Assessment of Existing Algorithms (CASP) and Benchmark Datasets 11

1.6.1. RS126 12

1.6.2. CB396 12

1.6.3. PDB-REPRDB 13

1.6.4. UniqueProt 13

1.6.5. Critical Assessment of Methods of Protein Structure Prediction 14

1.7. Objective of The Study 14

CHAPTER 2. MATERIALS AND METHODS 18

2.1. Organization of the Materials and Methods 18

2.2. Database Construction 18

2.2.1. Protein Sources and Biochemical Properties 19

2.2.2. Protein Annotation 20

2.2.3. Protein Structure Sources 22

2.2.3.1. STRIDE 22

2.2.3.2. Secondary Structure Prediction Algorithms in Used..... 22
2.2.4. Protein Ontology 24
2.2.5. Protein Homology and Needleman-Wunsch Algorithm 25

CHAPTER 3. RESULTS AND DISCUSSION 27
3.1. Data Acquistion from PDB 27
3.2. Data Acquisition from UniProtKB and GeneOntology 28
3.3. Global Sequence Alignment of Proteins 30
3.4. Construction of Benchmark Datasets and Evaluation of The Algorithms 30

CHAPTER 4. CONCLUSION 38

REFERENCES 39

LIST OF FIGURES

Figure		Page
Figure 1.1.	The relationship between protein structure and function is summarized. Small hexagons show the subfields of protein structure studies. Arrows between hexagons show the applications in respect to function studies by utilizing structure studies.	3
Figure 1.2.	The number searchable PDB entries solved by X-ray are shown in the graph. 66847 structures out of 70861 are determined by X-ray crystallography up to 2011 (red bar). It is also seen that the technical improvement in X-ray crystallography is mostly increased from 1976 to 2011, even though in 2011 there are less X-ray solved proteins in compare to 2010 (blue bar)	4
Figure 1.3.	In the MHC Class I Protein (PDB ID: 1a1n) (a) alpha helices (purple and pink), beta sheets (yellow) and turns (gray) are shown. (b) The Ramachandran plot shows the geometrical features of the protein by indicating the atomic angles Psi and Phi. Alpha helix region (Psi= -47, Phi = -57), beta sheet region (Psi = 113, Phi = -119) and turns (distinct region) lie on the plot.	8
Figure 1.4.	Torsion angles of a peptide are shown. Angle between N and alpha-C is called Phi, C and alpha C is called Psi, alpha-C and alpha C is called omega. Furthermore, in the figure, angle between alpha-C and side chain is marked as X.	8
Figure 1.5.	The left image is cytosolic human hexokinase and the right image is mitochondrial hexokinase. The sequence similarity of these proteins is 73.39%. However, they have different structures. This can be interpreted as proteins adopted different cellular conditions for their functions.	15
Figure 1.6.	Secondary structure elements and surrounded water molecules of aspartate proteases 4cms (gray) and 5er2 (blue) are shown. Even though they do not have similar sequences, they have highly similar tertiary structures and catalyze the same reaction in the cytosol	16

Figure 2.1. The entire computational process on database construction is shown. The rectangles denote data processing done by Java Programming. The lozenges show the processes data. All processed data is stored in mysql database. 19
Figure 2.2. The entire structure of database is shown..... 20
Figure 3.1. The data extracted from PDB was verified by Electron microscopy, Fiber Diffraction, FTIR, Neutron Diffraction, Powder diffraction, Solid and Solution states NMR and X-Ray crystallography. The most of the data was verified via X-Ray with 71821 entries, then 7775 solution-NMR and 1263 electron microscopy..... 27
Figure 3.2. Sub cellular localizations of proteins are shown. 29
Figure 3.3. Resolution (x-axis) and pI (y-axis) values of proteins in RS126 dataset are shown. Proteins that were verified by NMR are shown as -1 in x-axis. The rest of the proteins that were by X-Ray crystallography are shown on the right hand side with their resolution values. The distribution in the pI is related to the different subcellular localizations and functionalities of proteins..... 30
Figure 3.4. Resolution (x-axis) and pI (y-axis) values of proteins in Eukaryotic Glycoprotein dataset are shown. Proteins that were verified by NMR are shown as -1 in x-axis. The rest of the proteins that were by X-Ray crystallography are shown on the right hand side with their resolution values. 34
Figure 3.5. The cross similarities of proteins in eukaryotic glycoprotein group are shown as heatmap. Green color shows the maximum identity. Yellow tones show the similarity degree up to 80%. The red colors show the lowest similarities that are close to 50%. 34
Figure 3.6. Resolution (x-axis) and pI (y-axis) values of proteins in Eukaryotic Cytolic Translation protein dataset are shown. Proteins that were verified by NMR are shown as -1 in x-axis. The rest of the proteins that were by X-Ray crystallography are shown on the right hand side with their resolution values. 36

LIST OF TABLES

Table		Page
Table 2.1.	List of algorithms used in this study. Generation name, name of the algorithm, implementation, and sources are given.	23
Table 3.1.	Structural classes and percentage of the proteins	31
Table 3.2.	Q3 accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in Simple and CB1999 decomposition methods on RS126	32
Table 3.3.	SOV accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in Simple and CB1999 decomposition methods on RS126	33
Table 3.4.	Q3 accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in simple and CB1999 decomposition methods on Eukaryotic Membrane Glycoproteins	35
Table 3.5.	SOV accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in simple and CB1999 decomposition methods on Eukaryotic Membrane Glycoproteins	35
Table 3.6.	Q3 accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in simple and CB1999 decomposition methods on Eukaryotic Cytosolic Translation Proteins	37
Table 3.7.	SOV accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in simple and CB1999 decomposition methods on Eukaryotic Cytosolic Translation proteins	37

LIST OF ABBREVIATIONS

2D structure : Protein secondary structure

3D structure : Three dimensional protein structure (Tertiary structure)

CASP : Critical Assessment of Techniques for Protein Structure Prediction

DSSP : Dictionary of Secondary Structure of Proteins

HMM : Hidden Markov Model

HSSP : Homology-derived Secondary Structure of Proteins

NMR : Nuclear Magnetic Resonans

NN : Neural Network

PDB : Protein Data Bank

SVM : Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1. Introduction to Structural Proteomics

The term “proteome” was defined by Marc R. Wilkins in 1994 to explain the complete protein set synthesized by a genome (Liu and Hsu, 2005). With description of proteome, a new term “proteomics” was defined to mean qualitative or quantitative state changes of the proteome of a cell under different conditions. The overall stages of proteomics cover identification and quantification of proteins as well as characterization of their structure, functions and interactions by applying experimental techniques (Phizicky et al. 2003) .

Structural proteomics is one of the aspects of proteomics which deals with determination of spatial configuration of proteins, and thus structure-function relationship. The functionality of protein depends on its proper 3D structure, hence it depends on folding. Properly folded proteins play various roles in many biological processes including activities on molecular level such as catalysis or regulation, complex processes based on protein interactions. Known function and structure also provide a foundation to structure-based drug design (Floudas et al. 2006), applications on protein engineering in medicine, biotechnology via controlling signaling pathways (Petrey et al. 2005).

Inferring the functionality from the sequence or the structure has been studied for many years either by computationally or experimentally. As summarized by Whisstock and coworkers, Watson and coworkers, and Lee and coworkers (Lee et al. 2007; Watson et al. 2005; Whisstock et al. 2003) proteins which have similar sequences may have similar structures and function. On the other hand, similar structures are found with different sequences (Chothia and Lesk, 1986) .

Sequence-based methods depend on the observation that similar sequences may have similar functions. However, as pointed out by many researchers previously (Bork and Koonin, 1998; Karp, 1998) proteins of the same family can have several functions either by diverging to a related function or by gaining a complete new function.

Therefore, particularly in the absence of experimental data, sequence-homology based studies on function determination can be misleading. A good example can be given on eye lens proteins in duck which have high similarity with lactate dehydrogenase and enolase in the other tissues of the same organism (Wistow and Piatigorsky, 1987) . In addition to that, non-homologous proteins may possess similar functionalities according to the convergent evolution. Well-known enzymes chymotrypsin and subtilisin share the same catalytic pattern even though they display entirely different folding patterns i.e. tertiary structure. Since there is no significant similarity that can be detected between these two sequences, sequence-based methods can fail to assign function. A study on MJ0882 is demonstrated that MJ0882 is a methyltransferase even though there is no sequence homology with any of structurally known other methyltransferases (Huang et al. 2002). After structure determination of the protein via experimental methods, this functionality was investigated by biochemical assays. Thus, for such cases structure based method applications seem more promising than using sequence-based methods themselves (Jung and Lee, 2004).

Current structure-based methods on function determination aim to find global similar structures by determining all folding classes or any structural similarities, particularly functional sites (Watson et al. 2005). Furthermore, structure based methods utilize information on sequence similarity to detect patterns. The study done by Han and Baker (Han et al. 1996) showed the correlation between structure and function by searching the sequence patterns. They investigated the recurring sequence patterns and their correlation with structure by mapping similar sequences into different proteins which adopt similar 3D structure and functionality

Ultimately, inferring the function by using structure based approaches with the support of evolutionary information of sequences may lead us to come closer to the goal. Since the structure influences protein functionality, learning about the structure will help us to decipher functionality of proteins in many aspects summarized as in Figure 1.1.

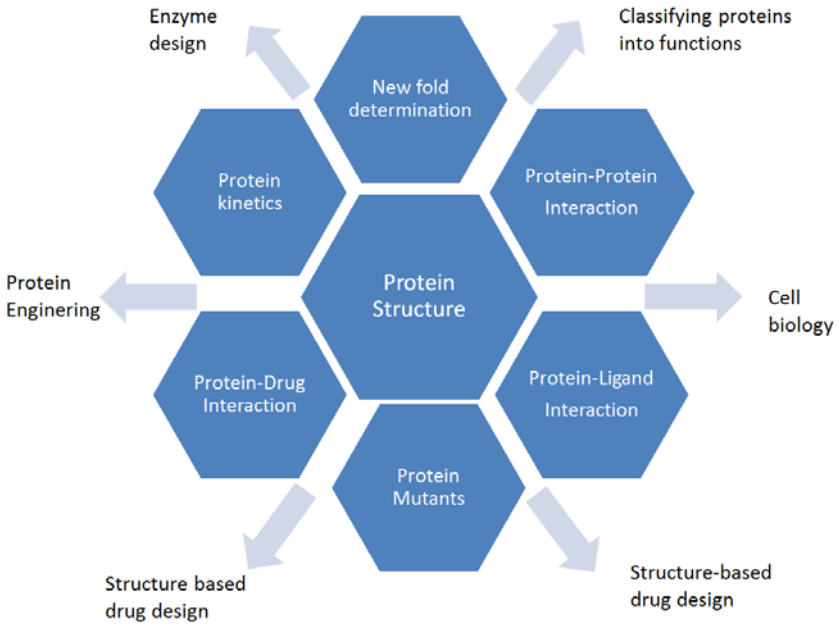

Figure 1.1. The relationship between protein structure and function is summarized. Small hexagons show the subfields of protein structure studies. Arrows between hexagons show the applications in respect to function studies by utilizing structure studies.

To determine the structure, firstly, experimental techniques should be examined. In the next section, major experimental techniques will be introduced.

1.2. Overview on Experimental Techniques

As mentioned in previous section, structural proteomics aims to determine the spatial configuration of proteins. According to the records in PDB, 70861 protein structures have been determined. X-ray crystallography, NMR, and electron microscopy are the experimental techniques to achieve this goal.

X-ray crystallography is one of the most powerful techniques because of providing atomic coordinates of each amino acid of a protein. Most of the experimentally known structures in PDB up to now are determined by X-ray crystallography as shown in Figure 1.2.

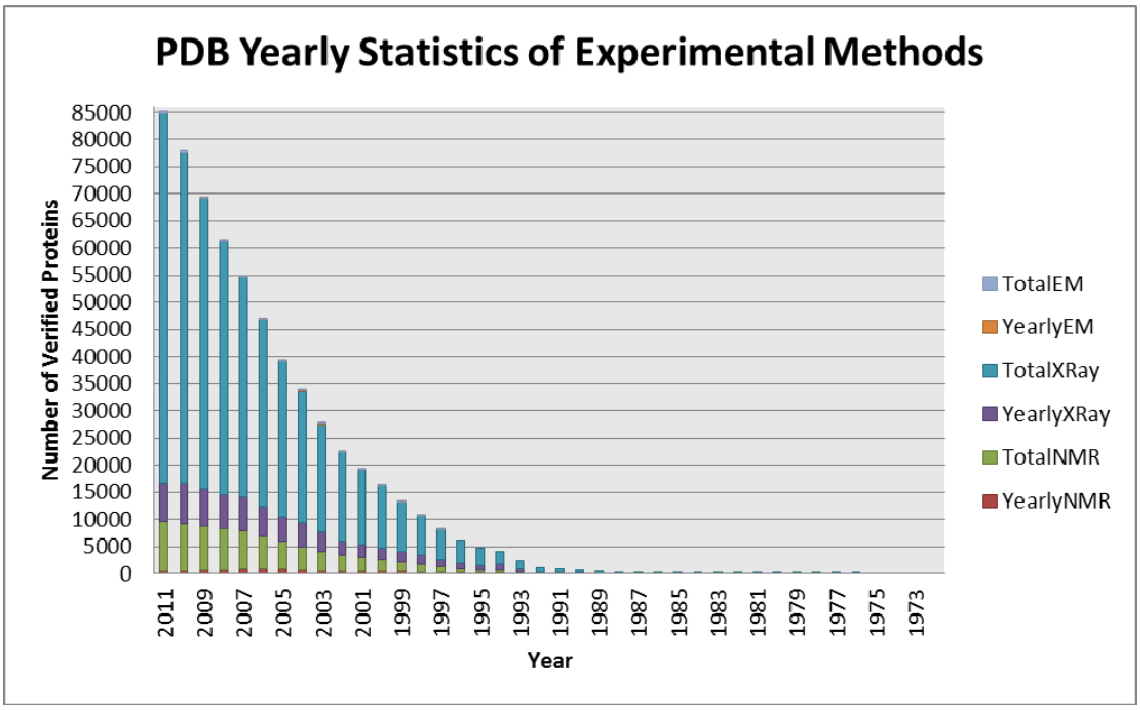

Figure 1.2. The number searchable PDB entries solved by X-ray are shown in the graph. 66847 structures out of 70861 are determined by X-ray crystallography up to 2011 (red bar). It is also seen that the technical improvement in X-ray crystallography is mostly increased from 1976 to 2011, even though in 2011 there are less X-ray solved proteins in compare to 2010 (blue bar)

Basically, the technique depends on obtaining crystals of proteins under certain pH, temperature, salt and cofactor conditions. The collected crystals are further analyzed by image scanning. In order to reduce error rate and increase the efficiency for low-quality protein sample, crystallography has been automated (Norin and Sundström, 2002). However, the type of the proteins, the crystallization conditions needed to form crystals limit the success. Membrane proteins, multi domain proteins and cofactor bounded proteins are hard to be crystallized. Also, proteins may lose their functionality because of dissimilarity between physiological conditions and solution conditions (Liu and Hsu, 2005).

Another technique to determine tertiary structure is NMR which assigns resonance values to each amino acid residue. This technique takes the advantage of using aqueous solutions to dissolve proteins which is similar to physiological conditions. Thus, they keep their natural structure. However, the first challenge with NMR is the time required to measure chemical shifts, to collect the data, and to analyze spectrum data (Liu and Hsu, 2005). Also, conventional NMR technique cannot be

applied to proteins that their molecular weights are more than approximately 25-30 kDa. To overcome this problem TROY-NMR has been developed (Fernandez and Wider, 2003).

In PDB, some of the entries were determined by electron microscopy. This technique enables us to determine the structure of bulge proteins and heterogeneous proteins especially membrane proteins with high resolution up to 3.5 A ° (Ubarretxena-Belandia and Stokes, 2010). Nevertheless, the problems with sample preparation and image processing of acquired data limit the success of the technique.

Although these techniques have recent improvements, the gap between known protein sequences and known structures is constantly increasing due to the rapidly increasing number of known sequences and technical problems mentioned above. In order to predict the structures of proteins that might not be verified by experimental methods in a short period of time, computational studies have been employed.

1.3. Computational Studies on Protein Tertiary Structure Prediction

Challenges in experimental techniques to determine tertiary structure of proteins have directed researchers to computational prediction techniques. To bridge the gap between the number of known sequences and structures, *ab initio* prediction, homology modeling, and fold recognition are the state of the art for computational approaches on 3D structure prediction even though they are not extremely successful (Baker et al. 2003; Fiser, 2004; Rost and O Donoghue, 1997).

The first technique *ab initio* is able to predict tertiary structure of a protein without prior knowledge about structure. *ab initio* technique is based on direct tertiary structure prediction from the amino acid sequence by using physico-chemical properties. It has been proposed that to form a native functional protein, the structure should satisfy minimum free energy (Van Gunsteren, 1993). Therefore, mean-force potentials (Van Gunsteren, 1993) and physical potentials (Brünger and Nilges, 1993) such as bonds, angles, Van der Waals, and electrostatic non-bonded measurements must be calculated. However, due to the lack of highly accurate experimental results in respect to these measurements, inferring basic parameters to calculate minimum free energy is complicated. These methods use a scoring function that discriminates the possible conformer model among proposed conformations. However, one of the

bottlenecks of the technique is related to the size of search space. There are few methods proposed to reduce search space by discrete representation. Despite the advantage of having small search space, sampling can end up with the loss of accurate conformers. Even though ab initio can be applicable to any protein and have a significant improvement (Xu et al. 2009), it still remains problematic on its basis. Besides the inaccuracy of the obtained parameter values, required computation power is enormously high (Allen et al. 2001).

When the homology can be detected, comparative modeling can be applicable and efficient. Comparative modeling or, in other terms, homology modeling is claimed to be the most powerful method according to the critical assessment analysis (Mariani et al. 2011). The method is simply based on using templates derived from previously determined structures with similar sequences to target. Finding the suitable template for comparative modeling is directly related to finding a known structure with at least 30% (Kopp and Schwede, 2004) sequence similarity to target protein due to the last studies. Furthermore, the low resolution X-ray or NMR data can cause errors on template selection. Even though the problems on template selection or resolution quality of the template are solved, building 3D model can fail since initial problems of alignment are hard to correct. To overcome problems of comparative modeling, template-free search methods, in other terms, fold recognition methods are developed which computes the conformation probabilities of each fold (Xu et al. 2009).

The last approach is fold recognition. Determination of structural elements by certain localization of each amino acid's side chain and the arrangement of these structural elements presents folding pattern which is the part of protein architecture. To discover these folding patterns, the sequence is threaded on the sequence profile of a given fold by assessing the fitness of fold features (solvent accessibility, secondary structure, and environment) and input sequence (Baker et al. 2003). Thus, without searching a sequence complement or conformer computations, proteins that have same folding patterns can be detected. However, learning the folding pattern of a protein does not refer to determination of the structure. By considering that there are many distinct mechanisms such as recruitment of chaperones, post translation modification, natural disorder of the protein, and environmental conditions which directly affect protein folding fold recognition technique may not guarantee to find the most correct folding form (Watson et al. 2005).

Despite the improvements in 3D structure prediction, predicting the structural class of proteins, deciding on the architecture and topology, and predicting the structures of unknown folding patterns encompass the success of these algorithms. Therefore, this has created a research bottleneck to overcome the limitations by simplifying the problem into secondary dimension (Zhang, 2008).

As a result, secondary level (2D) has been popular over last decades. In the rest of the Chapter 1, protein structure prediction algorithms and datasets used to train and test these algorithms will be discussed in general after giving a brief introduction on protein secondary structure.

1.4. Overview of The Protein Secondary Structure

Protein architecture is composed of the chain of the amino acids. The side chain of an amino acid contributes to the general biochemical property and functionality of the protein as well as its structure. At the beginning of 50s, Pauling and Corey started protein structure studies that can be considered the start of understanding protein biology. They examined (Pauling and Corey, 1951a) the chain forms assembled into different conformations in globular proteins and synthetic polypeptides (Pauling and Corey, 1951a). They observed that the inter-hydrogen bonding between amino acid side chains yield two common stable structural elements: α-helix and β-sheet. Besides that, the irregular parts between helices and sheets are referred as coil or turn structure. The same research group established that in order to form a structural element, there must be distinct geometrical features. In Figure 1.3 (a), MHC Class I Protein monomer is given as an example of the helix, beta sheet and turn element types with color codes.

In the studies of Pauling and Corey (Pauling and Corey, 1951b, 1951c), the geometrical features of each structural element were examined. It was found out that in order to form helix or sheet, turn and bridge structures must be formed. The hydrogen bond formation between the CO of residue i and the NH residue of i+n refers “turn” where n may vary from 3 to 5. Repeating turns form alpha helix (Figure 1.3 pink parts) types including 310 helices, π helices. Also, turns can stay as single or bend. On the other hand, the term “bridge” points the hydrogen bond formation between non-adjacent amino acid residues. Repeating bridges constitutes beta sheets (Figure 1.3 yellow parts)

which later form anti parallel or parallel sheets, extended strands or isolated beta bridges.

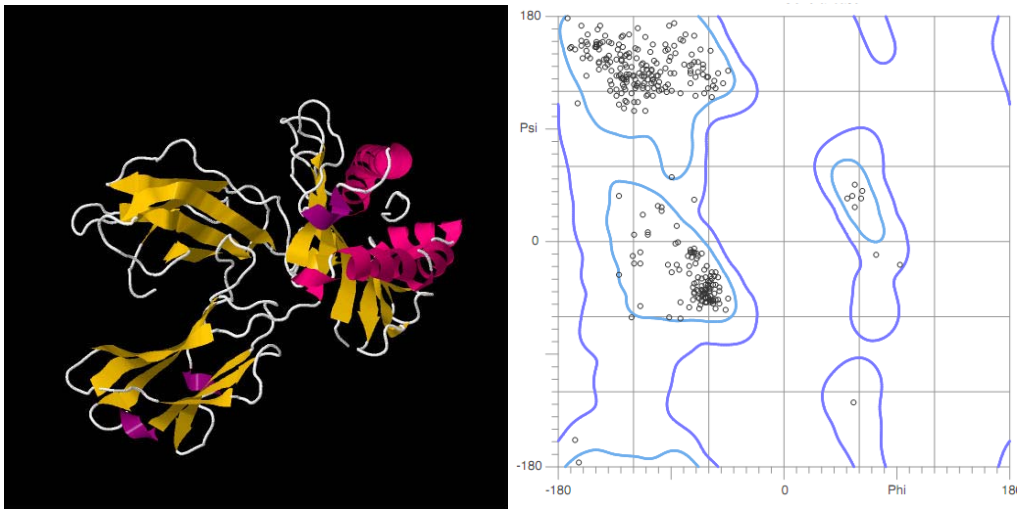

Figure 1.3. In the MHC Class I Protein (PDB ID: 1a1n) (a) alpha helices (purple and pink), beta sheets (yellow) and turns (gray) are shown. (b) The Ramachandran plot shows the geometrical features of the protein by indicating the atomic angles Psi and Phi. Alpha helix region (Psi= -47, Phi = -57), beta sheet region (Psi = 113, Phi = -119) and turns (distinct region) lie on the plot.

The decision on when an alpha helix or beta sheet forms is directly related to several parameters such as solvent accessibility, backbone psi-phi angles as shown in Figure 1.4, Cα positions, and hydrogen bond patterns. In Figure 1.3(b), psi-phi angles of each amino acid in MHC Class I protein are shown in the Ramachandran plot. The middle left region shows the amino acids which are found the alpha helices and turns that bound each alpha helix. The upper left shows the amino acids that are seen the beta sheet formations. The rest of the plot denotes the single turns.

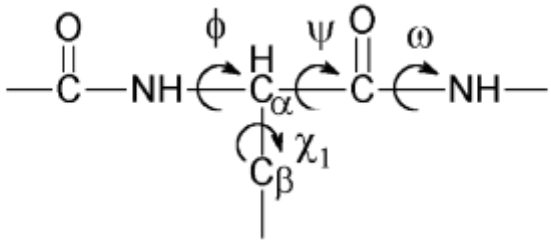

Figure 1.4. Torsion angles of a peptide are shown. Angle between N and alpha-C is called Phi, C and alpha C is called Psi, alpha-C and alpha C is called omega. Furthermore, in the figure, angle between alpha-C and side chain is marked as X.

In order to assign a secondary structure to X-ray solved protein, structure assignment algorithms to atomic coordinates of psi, phi, omega, and X angles have been developed. Currently in PDB, Dictionary of Secondary Structure of Proteins -DSSP (Kabsch and Sander, 1983) and STRIDE (Heinig and Frishman, 2004) are used to define experimentally solved protein structures. The results obtained by assignment algorithms are verified with the number of helical and strand segments measured by circular dichroism (CD) or Fourier Transform Infrared Spectroscopy (FTIR) spectra. Here, the details of the techniques are not given. However, both these techniques give an idea on the approximate percentile of each structural element in spectra without given positions of them. The studies show the correlation between structural characteristics presented by CD or IR and structure assignment methods (Sreerama et al. 1999). Therefore, these assignments are used to train or test the algorithms as experimental examples.

1.5. Overview of Existing Algorithms

The general idea behind all secondary structure prediction algorithms is to investigate the preferences of segments of amino acids for certain structural element and present the structure of the protein. Do all existing algorithms achieve this goal? The answer is: “No”. Despite the fact that structure prediction problem is simplified into secondary structure, structure prediction problem still remains challenging.

The existing algorithms are grouped into three generation. (Pirovano and Heringa, 2011) The first generation algorithms Chou- Fasman and GOR (Garnier et. al 1978) were started to be proposed after studies of Györgyi and coworkers on the likelihood of proline existence in alpha helices of keratin, myosin, epidermin and fibrinogen (Szent-Gyorgyi and Cohen, 1957). Basically, these earlier methods aimed to correlate amino acids and structural elements by calculating the frequency of existence of each amino acid in a certain structural element. While Chou and Fasman calculated the residue compositions for each structural element and assigned the prediction for the structural element with the highest score, GOR tried to predict the structure of each 17 residue length of amino acid chains. However, due to the lack of sufficient number of experimentally verified structures and consideration the impact of long-range

interaction on structure formation, the accuracies of these algorithms were limited to 60-65% (Kabsch and Sander, 1983b) This led the researchers to develop second generation algorithms.

Unlike the first generation algorithms, second generation algorithms incorporate the structural state of central amino acid in 11-21 adjacent residues with machine learning algorithms such as neural networks networks (Bohr et al. 1988; Qian and Sejnowski, 1988) , statistical information (Gibrat et. al 1987) and nearest neighbor algorithms (Kabat and Wu, 1973). The performances of first and second generation of algorithms are shown in Figure 1.4. Rost and Sander compared the performances (Rost and Sander, 2000) of the first and second generation algorithms on K&S 62 dataset. (Kabsch and Sander, 1983b) The result shows that there is approximately 10% difference between tested first generation algorithms Chou-Fasman, GOR, and Lim and second generation algorithms ALB and Scheider. The result does not reflect the reality since the used dataset contains only 62 proteins. According to the observation of accuracy results of first and second generation algorithms, it has been realized that the window length 11-21 is not sufficient to detect long range interactions and beta sheet predictions are not local as helix formations. Therefore, to capture the whole picture and comprehends the long- range interactions and non-local formations (Cuff and Barton, 1999), evolutionary information has been incorporated into the third generation predictors.

The third generation predictors use sequence similarity information either by implementing multiple sequence alignment (Zheng and Kurgan, 2008) and local similarity information as profiles (Cole et al. 2008; Cuff and Barton, 2000). Machine learning techniques and artificial intelligence techniques such as support vector machines (SVM) (Hua and Sun, 2001; Nguyen and Rajapakse, 2005; Kim and Park, 2003; Shoyaib et al. 2007; Ward et al. 2003), neural networks (NNs) (Babaei et al. 2010; Kakumani et al. 2008; Qian et al. 1988), k- nearest neighbors (Madera et al. 2010; Qu et al. 2011), hidden markov models (HMM) (Kumar and Raghava, 2009; Malekpour et al. 2004) are improved in respect to their initial architectures. There is also another approach in this generation which is called meta predictor. Meta predictors (Cole et al. 2008; Palopoli et al. 2009) have been developed to merge the few successful algorithms on different aspects. Taking the advantage of artificial intelligence and evolutionary information, a remarkable success was obtained in compare to first two generation algorithms. However, the caps of beta sheets and alpha helices were more poorly

predicted than the inner structural elements (Rost and Sander, 1993). Furthermore, some long- range interactions were lost even though evolutionary information was used. The long range interaction formation is forced by the environmental conditions to become stabilized. All these pitfalls of development or evaluation of the algorithms are directly correlated with the limitations regarding the data.

1.6. Overview on Assessment of Existing Algorithms (CASP) and Benchmark Datasets

One of the main pitfalls of protein secondary structure prediction is data quality and features. During development of an algorithm, system is trained by the data to learn potential relations on being a certain structural element. On the other hand, to test the accuracy of the algorithm a new set which does not contain any data in training set must be generated. It is called test set. Previously proposed data to train and test the algorithms and also to evaluate the existing algorithms and newly developed ones are limited. Therefore, cross validation or jack-knife test emerged to generate training and test sets with n number of times repeating. These techniques could be explained briefly. Jack- knife splits the whole data into two groups; n-1 proteins as training, 1 protein for test. It repeats splitting for n times. In contrast, cross-validation divides the data equally into m groups. (m-1)n/m number of proteins are considered as training, and n/m proteins are named as test sets. This process repeats for m times until each protein is in the test group once. The competence of these techniques is not the part of the study. Thus, the weakness or strength of them will not be discussed.

Here, starting from the historical ones to recently cited datasets will be given to state an issue in respect to data. Notice that there are other datasets which were used either to investigate the knowledge by training or assess the performance by testing. However, considering the availability and the number of citations, the datasets below were worth mentioning. Furthermore, by exploring the general features of the datasets here, the reasons for establishing a new benchmark will be become clear. Finally, the performances of these datasets on algorithm prediction performances were compared with newly proposed benchmark dataset.

1.6.1. RS126

This dataset is one of the datasets that has been studied widely. 126 non-redundant protein chains make up the RS126 dataset (Rost and Sander, 1993). All proteins were verified by X- ray crystallography with 2.5 A°. While constructing the RS126 training and test sets, proteins that have lower than 25% pairwise similarity for chains with more than 80 amino acids were considered. It has to be mentioned here that this threshold aims to generalize the methods which incorporate sequence similarity into protein secondary structure prediction. To explain it more clearly, it is assumed that the prediction performance on highly homologous proteins would be high as in homology modeling in 3D structure prediction. Accordingly, the same good performance should be taken from non-homologous proteins. Another feature of the dataset is related to the creation of training and test sets. In the construction of RS126, the authors tried jack-knife test which was repeated 130 times until each protein has been used once. However, it failed since their algorithm was based on the neural network. Because they could not use only one test set with 20 proteins, they finally performed 7 fold cross validation. Therefore, they set 7 different test sets with 19 different proteins.

Although they tried to produce non-homologous protein dataset, there were 11 proteins which have less similarity but the same folding structure i.e. almost the same functionality. For instance, immunoglobin proteins 1fldh and 1mcpl have highly similar few blocks but low overall percentage similarity. This can be interpreted as to conserve the functionality; final tertiary structure was conserved more than sequence.

1.6.2. CB396

This dataset was proposed by Cuff and Barton (Cuff and Barton, 1999). The remarkable feature of this non-redundant dataset is that it does not consider percentage similarity. The percentage similarity is changeable according to the length and the composition of the sequences (Brenner et al. 1998). This means that even if the percentage similarity seems low, this might not reflect the reality. To overcome this problem, they first aligned sequences by Needleman-Wunsch algorithm (see Chapter 2). Then, sequences are randomized, and re aligned again for more than 100 times. Finally, SD or in other terms Z score is computed by using calculated mean and standard

deviation of the randomized alignment for the first alignment V. SD score removes the bias on the alignment and helps to yield a non-redundant dataset. If the score is lower than 5, this means that there is low similarity. Therefore, the total numbers of sequences were reduced. Even though these proteins have less sequence similarity, it is understood that they share the same folding pattern because of having the same functionality. There is some filtering that was performed after reducing the size by alignment. For instance, multi domain segments were removed, the resolution was limited to 2.5 A°. Also, RS126 proteins were excluded from that dataset. The reason is that if the algorithm trained by RS126 would be tested with CB396, there would not be a bias. The proteins which are lack of DSSP or other information regarding annotation were also removed. Consequently, they produced CB396 non-redundant dataset.

1.6.3. PDB-REPRDB

Unlike previous datasets, PDB-REPDF is an algorithm to generate datasets from proteins in PDB. This algorithm considers nine different criteria. They can be listed such as resolution, number of chains, R-factor, the ratio of non-standard amino acids, ratio of residues with only C-α coordinates, number of residues with only backbone coordinates, number of amino acids, which experimental technique would be requested by user etc (Noguchi and Akiyama, 2003). The main property of this algorithm is selecting membrane proteins.

1.6.4. UniqueProt

UniqueProt (Mika and Rost, 2003) is an algorithm which generates unbiased dataset for the user. It uses the information from HSSP (homology derived structures of proteins) database to discriminate the proteins with homologous and non-homologous structures. It would be necessary to mention about HSSP. HSSP (Schneider and Sander, 1996) aligns the 3D structures of the proteins in PDB by performing modified Smith-Waterman local alignment. Therefore, it detects the similarities within the proteins. UniqueProt also incorporates the subcellular localization and functionality information by calculating the HSSP curve from BLAST output. However, this property has not

been clarified in the research paper well. The threshold to determine similar and non-similar groups is decided due to the HSSP curve.

1.6.5. Critical Assessment of Methods of Protein Structure Prediction

How well can we predict secondary structure of a protein when we develop or upgrade an algorithm? How can we assess the performance of predictors and evaluate the outcomes for the same, new sequences at the same time? How can we state and scale tasks for predictors? How can we adjust the solvability of tasks for the predictors? These questions address experts to organize a meeting which is called Critical Assessment of Methods of Protein Structure Prediction (CASP) (Moult, 2005). The progress of CASP can be summarized as: (1) Experimetalists submit sequences whose structures are nearly verified by experimental methods to CASP organization. (2) Sequences are distributed to predictor developers. (3) The experimental structures are announced and the evaluation results are declined in the meeting Asilomar.

The help of CASP is to remove the tendency on using already known sequences to predict their structures. However, yearly defined targets in CASP are limited number. Therefore, it is hard to evaluate predictors' performances in a statistically significant manner. Another issue is there is no option to evaluate algorithms for distinct tasks, for instance, proteins whose lengths are smaller than X or proteins which have several domains or proteins which have different or same functions. Different tasks with high numbers of proteins ensure to observe the changes on performances of predictors.

1.7. Objective of The Study

In this study, the main hypothesis is stated on "environmental conditions directly affect the functionality and therefore structure". Until now, most of the algorithms incorporate the evolutionary information to increase the accuracy. Due to the CASP or EVA results, an improvement has been detected with the utilization of sequence homology. However, there are other important examples showing low sequence similarity but almost similar structures and functions. Few different examples would be good to be given. The first example can be given for hexokinases. In the Figure 1.5,

cytosolic (left) and mitochondrial (right) hexokinases are shown. The sequence identity of these proteins is considerably moderate at 74% according to BLAST. The protein on the left is cytosolic hexokinase and the protein on right is mitochondrial hexokinase. As seen, even though they are both hexokinase, their functions are specialized for certain biochemical reactions which are specific to their sub cellular locations. Locating in either cytosol or mitochondria directly affects the tertiary and secondary structures of these two proteins pointing that the cell is compartmented to various units to carry different functions that are firmly related to the tertiary structures of proteins.

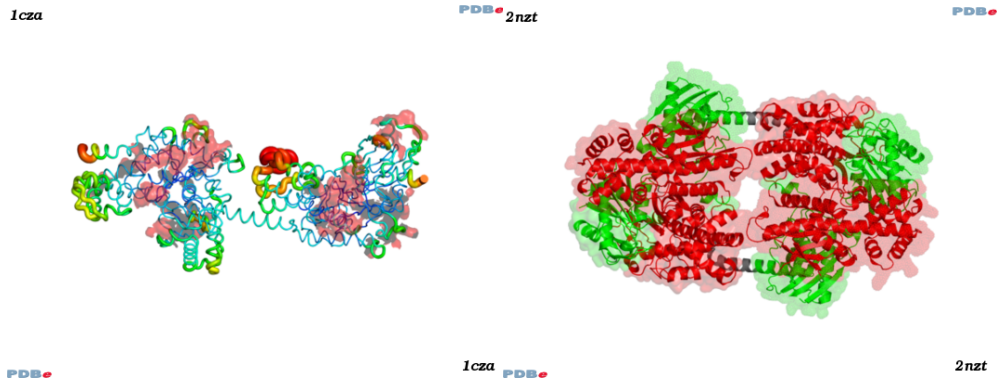

Figure 1.5. The left image is cytosolic human hexokinase and the right image is mitochondrial hexokinase. The sequence similarity of these proteins is 73.39%. However, they have different structures. This can be interpreted as proteins adopted different cellular conditions for their functions.

Another example is of aspartat proteases. 4cms and 5er2 have the same functionality and same cellular localization cytosol. In Figure 1.6, the structural alignment of these proteins was shown by using Astex visualization tool. The gray helices and sheets belong to 4cms, the blue ones belong to 5er2. The overlapping helices and sheets are shown both gray and blue. The E-value of the BLAST is 5e-17 points a moderate sequence similarity. Even though a significantly high sequence similarity is not observed, these proteins have similar structures because of having same functionality in the same subcellular location.

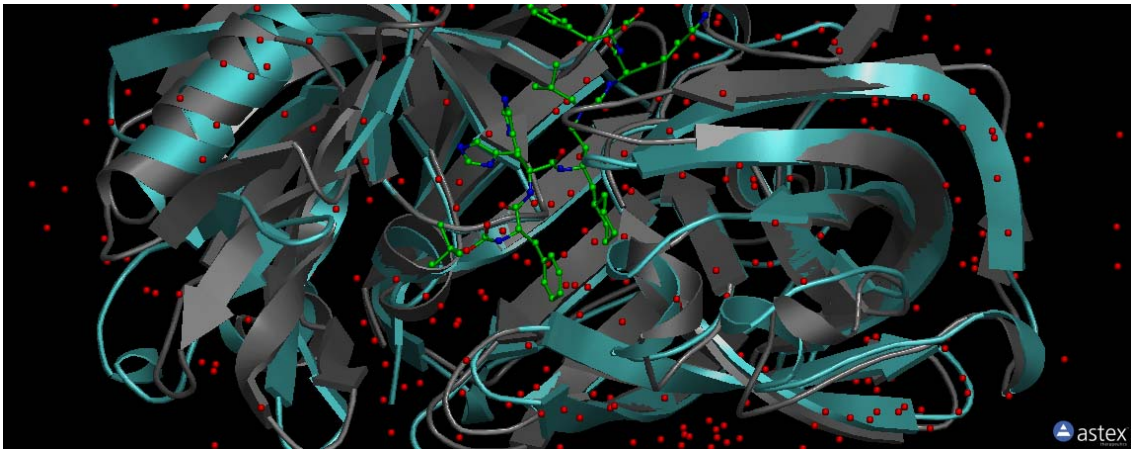

Figure 1.6. Secondary structure elements and surrounded water molecules of aspartate proteases 4cms (gray) and 5er2 (blue) are shown. Even though they do not have similar sequences, they have highly similar tertiary structures and catalyze the same reaction in the cytosol

During the development of PHD algorithm (Rost and Sander, 1993), some bad predictions were obtained. One of them was the prediction of SH3 protein. The predictor assigns the sheet formation to a part of chain 4. However, that part forms helix in the solution that it is functional. It can be concluded that environment has a great impact on the native functional structure.

Specifically, it has been assumed that two level grouping proteins into their subcellular locations, and then functions in that location might result in much precise secondary structure prediction under the light of such examples.

The main hypothesis is to construct an advanced benchmark database not only to assess the quality of the algorithms, but also to show the weaknesses of the algorithms on which types of proteins and to direct users for the selection of appropriate algorithms. Therefore, main objectives of this study can be listed as:

1. Extraction the sequence and STRIDE secondary structure assignment of available experimentally verified 64872 proteins up to 2010 from PDB based-S2C database.
2. Retrieving the annotation on accession numbers, ontology information and subcellular localization information of sequence and UniProt accession number unique-proteins from UniProt and GeneOntology.
3. Finding the global pair-wise similarity of proteins by using Needleman-Wunsch algorithm according to the suitable BLOSUM substitution

matrix which is selected due to the identity between sequences, and normalizing the alignment scores.

4. Grouping proteins according to taxas such as eukaryotic, bacterial, viral and archaeal. To demonstrate the evaluation of seven algorithms from each generation, two sample datasets are generated and compared with one of widely used benchmark dataset RS126.

CHAPTER 2

MATERIALS AND METHODS

2.1. Organization of the Materials and Methods

In the concept of this study, a new and advanced database construction is aimed. The main feature of the dataset is to group proteins due to the localization information supported with functionalities and biochemical properties of proteins in that cellular compartment. Therefore, the data of interest is collected from various sources. To evaluate the performances of eight publicly available algorithms in compare to experimentally verified structure for each protein, the predictions are gathered and parsed.

Mysql 5.5.8 open source database in WAMP 2.1 web development server was used as database platform and programs to retrieve and process the data were written in JAVA programming language.

In following section, the database construction is explained by providing information about data sources including protein PDB identifier, chain number, sequence and experimental structure assignment STRIDE, alternative names of the protein, ontology information, and sequence similarity of each protein pairs. Moreover, the seven algorithms used in this study were briefly explained. To assess the performances of these algorithms in each dataset is explained in the last part of the chapter.

2.2. Database Construction

The computational design schema of the database including data acquisition processes, filtering processes and data storage is shown in Figure 2.1. The data sources and process details are given in next sections.

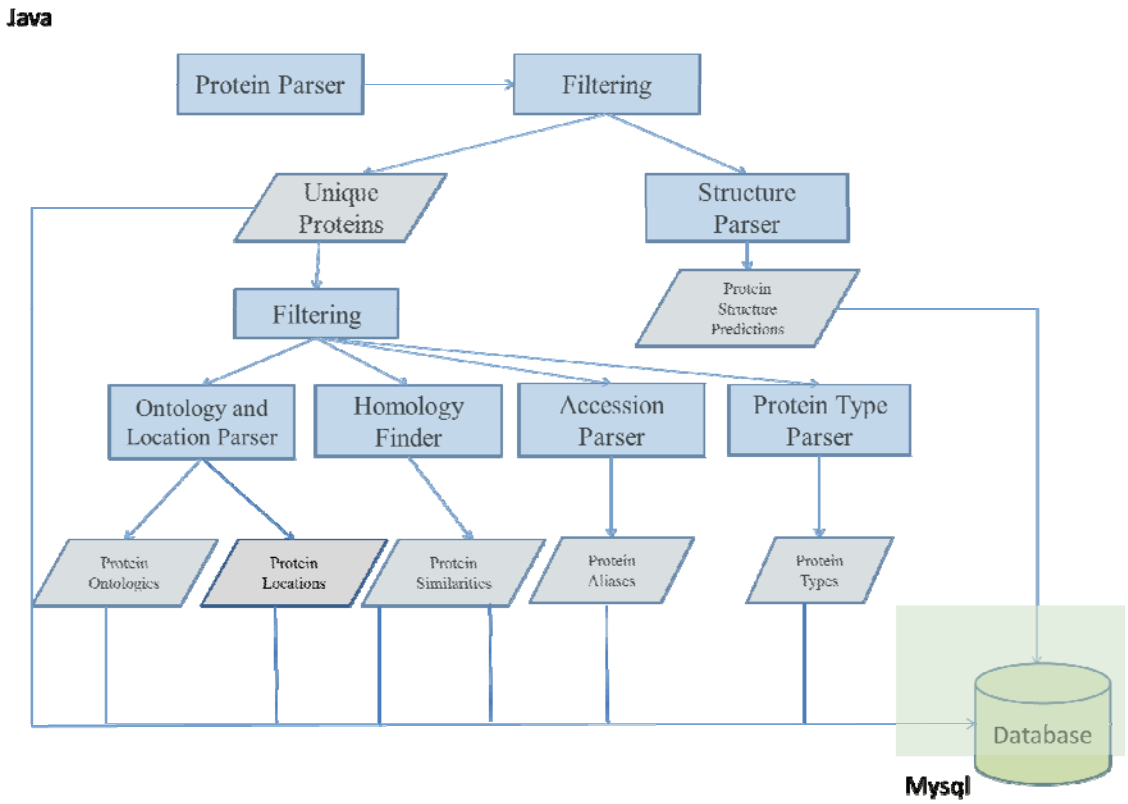

Figure 2.1. The entire computational process on database construction is shown. The rectangles denote data processing done by Java Programming. The lozenges show the processes data. All processed data is stored in mysql database.

2.2.1. Protein Sources and Biochemical Properties

PDB (http://www.pdb.org/pdb/home/home.do) was used as the central source for experimentally verified protein structures. Because of the complex structure of PDB format, were downloaded from S2C database which is available on http://dunbrack.fccc.edu/Guoli/s2c/ .

PDB identifier, protein sequences and STRIDE structures of each chain, literature references, additional information such as mutant amino acids or isolation environment of the protein, and its UniProt accession number(s) were presented in S2C file format. Since the parts of interest are PDB identifier, protein sequence, STRIDE structure, and UniProt accession numbers, these parts were extracted for each protein by using ProteinParser in Figure 2.1. Each chain of protein was considered as a protein. Notice that one chain was selected out of identical chains of a protein while parsing S2C data file of the protein in order to prevent redundancy. Thus, in the rest of the paper

term “protein” would be used as sequence unique chains of a protein. Besides that, isoelectric point of each protein was calculated to determine the biochemical properties in respect to behavior in the aqueous solutions. All of the information of a protein which has unique PDB ID, ChainID and UniProtID was inserted in the table “Proteins” shown in Figure 2.2.

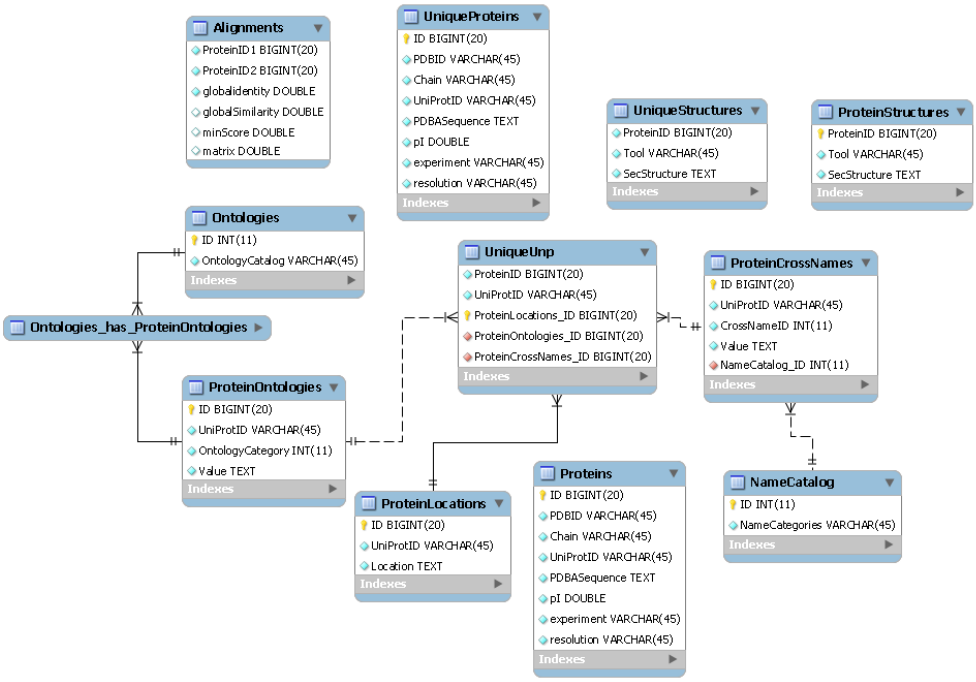

Figure 2.2. The entire structure of database is shown.

Afterwards, however, it was realized that the same protein sequence had been entered into PDB with different PDBIDs. Therefore, a filtering was performed by selecting a representative of a sequence which has the highest resolution if it is verified by X-ray. Filtering results were inserted into a new table called “UniqueProteins”.

2.2.2. Protein Annotation

After a protein is submitted to databases, a unique identifier is assigned by database to enable access to information via that identifier. That unique identifier is called accession number. Each database refers the cross references of corresponding protein.

In this study, it is aimed to keep all available accession numbers, recommended short and full names, alternative names and synonyms. The reason for that was to find missing protein localization information. By searching all names of a protein, we can collect the accessible literature sources to extract hidden localization information. Moreover, users can reach the data by querying alternative names of a protein instead just using PDBID.

The main source of accession numbers was Universal Protein Resource (UniProt). UniProt is accepted as one of the most comprehensive protein catalog with 18.184.507 protein entries in Release 2011 Oct, 19. Since UniProt accession numbers were obtained from PDB derived S2C database, alternative names and accession number were extracted from UniProtKB fetch by querying these UniProt accession numbers. Although several cross references exist in UniProtKB, in this study accession numbers of sequence databases were only taken. Sequence database section includes EMBL, GenBank and DDBJ accession numbers. Table "NameCatalog" was created, although each protein might have entries of different sequence databases.

Before querying UniProt accession numbers that are recorded in PDB, second filtering was performed. The reason for that was different proteins are represented by same UniProt accession number as family accession number. Within the family, members have 90% intra cluster sequence similarity. This filtering was designed to select one of the family members as a representative of the family. This selected entry is the protein which has the highest resolution value to increase the confidence of the experiment. Hence, for each protein, its sequence database entries were parsed as shown in Figure 2.1, and then the protein id, its UniProt accession number, sequence database accession numbers, and the corresponding identifier of that sequence database in the "NameCatalog" were inserted into "ProteinCrossNames" table as shown in Figure 2.2.

Other information which was extracted from UniProt was subcellular localization. It was stored in table "ProteinLocalizations" as Proteins shown in Figure 2.2. Some of the proteins have several localizations. This turns out be elaborative. Therefore, these proteins were eliminated.

2.2.3. Protein Structure Sources

As mentioned earlier, this study is based on the evaluation of secondary structure prediction algorithms. The algorithms, which were selected to be evaluated, were explained in this section. It must be underlined that the selection was performed due to the availability of algorithms. It was unfortunate to many of the algorithms that were introduced in the Chapter 1 are not accessible now.

2.2.3.1. STRIDE

The prediction models of the algorithms must be compared with experimentally defined structure. The experimental structures are obtained by a structure assignment algorithm such as DSSP, STRIDE so on. PDB has both STRIDE and DSSP assignments. However, in this study STRIDE was preferred because it consideres the hydrogen bonding patterns as well as backbone geometry.

STRIDE information for each protein was parsed from S2C data. The information was stored in "ProteinStructures" table. However, because of the experimental errors measurements are not accurate; STRIDE assignment did not work well. Dash symbol was assigned to unpredictable amino acids. Furthernore, some of the protein sequences were less than 20 amino acids. Since the smallest protein has 20 amino acids, the entries whose lengths' are less than 20 were eliminated. Also, proteins that have UNP aminoacids such as X were discarded. Consequently, STRIDE structures with its related protein identifiers were inserted into "UniqueStructures" table shown in Figure 2.2.

2.2.3.2. Secondary Structure Prediction Algorithms in Used

In this study, it would have been aimed to assess algorithms of different generations. On the down side, because of the lack of the algorithms and obstacles on server connections such as email requirement or operating system requirements, there

were only eight algorithms in hand. The features of algorithms, their generation, and developers were given in Table 2.1.

Table 2.1. List of algorithms used in this study. Generation name, name of the algorithm, implementation, and sources are given.

Generation	Name	Algorithm	Citation
1st	GOR	Position dependent structural element propensity calculations	(Garnier et al. 1978)
3rd	GORIV	Window based approach which utilizes all possible pair frequencies for each structural element	(Garnier et al. 1996)
	HNN	Neural network learning - based on regression models for pattern recognition.	(Guermeur and Gallinari, 1997)
	PHD	Two layer neural network architecture & multiple sequence alignment based	(Rost and Sander, 1993)
	PREDATOR	Database- derieved statistics for non-local interaction among structural elements	(Frishman and Argos, 1996)
	SIMPA96	Approach is based on short length of sequence homology using nearest neighbor	(Levin et al. 1986)
	SOPMA	Sequence similarity that is on level of structural classes and sequence level.	(Geourjon and Deleage, 1994)

The first algorithm used was GOR(Garnier et al. 1978). As namely mention in the Chapter 1, this algorithm is a first generation algorithm. It splits the protein into 17 residue- length windows. The algorithm calculates the structure of the central amino acid. The reason for that is the structure type of the central amino acidwhich influences the structure type of the adjacent amino acids. Afterwards, to increase the accuracy of the algorithm, hydrophobic triplets were searched. However, including that information did not affect the accuracy. GOR is also important to being the first algorithm to understand the importance of using evolutionary information.

The rest belongs to the third generation. The first algoritm is also upgraded version of GOR, GORIV (Garnier et al. 1996). In this version, all pair frequencies are computed with a 17-window. Also, database size is increased. HNN (Guermeur et al. 1998) is another algorithm which uses neural network learning technique. Sequence-to-structure and structure-to-structure networks incorporates the local statistical information. This information is calcuted by regression models. PHD (Rost and Sander, 1993) is another algorithm which is also based on neural network. It takes multiple sequence alignment output as an input to the sequence-to-structure and structure-to

structure network. On the other hand, the algorithm PREDATOR (Frishman and Argos, 1996) calculates the potential of hydrogen bonded amino acids. This algorithm examines protein sequence in eight structural classes instead three by looking at also non-local interactions. The parameter for it is only the assignment type. Here, STRIDE was selected. SOPMA (Geourjon and Deleage, 1994) and SIMPA96 (Levin et al. 1986) are the algorithms which uses sequence homology information. Whereas SOPMA uses sequence similarity of protein families, SIMPA96 takes into account that 7 length local sequence similarity decides the structure of the window. This is automated by the nearest neighbor method. For SOPMA, the number of conformational states was set to 4, helix, sheet, turn, coil; similarity threshold was set to 8; the window length was set to 17.

In order to compare prediction results of these algorithms with STRIDE assignment results, two different decomposition methods were applied. STRIDE assignment provides eight different states of secondary structure elements, however secondary structure prediction algorithms perform prediction in three main state of elements; helix (H), sheet (E) and coil (C/T). Usage of decomposition methods allow us to reduce STRIDE alphabet and predictions into same level.

Eventhough there are several decomposition techniques are currently available, in this study simple reduction technique and CB1999 (Cuff and Barton, 1999) were used. Other decomposition techniques (Rost and Sander, 1993) and (Frishman and Argos, 1996) were used in training of PHD and PREDATOR. Therefore, using these decomposition techniques would create a bias on the evaluation of algorithms. In simple decomposition method, E and B characters were converted to E; G, H, I were converted to H; and rest were converted to C. On the other hand in CB1999 decomposition technique, E and B characters were converted to E; G and H were converted to H; and rest were converted to C.

2.2.4. Protein Ontology

In order to discriminate proteins into their sub cellular locations and functions in that location, gene ontology information was used. Basically, gene ontology provides a unified representation of gene and gene products on their annotations and functions with a controlled vocabulary. (http://www.geneontology.org) Mainly, three domains

“Molecular function”, “Cellular Component”, and “Biological Process” are covered in gene ontology. Molecular function covers the activities which are carried out in the cell such as enzymatic activities. Biological function describes the series of molecular functions performed connectively in the cell. Finally, cellular component refers two meanings. The first one is the cellular unit where gene or product of interest is located. Second meaning is that gene or product of interest is part of a cellular unit as a structural component.

While attempts to parse ontology information, it has been realized that some proteins do not have some domains of gene ontology. This would cause entries with null values in the database. As a caution, “Ontologies” table was created. Each of parsed ontology information of protein as in Figure 2.1 was inserted into database “ProteinOntologies” table shown in Figure 2.2 with protein identifier, corresponding ontology domain id, and the value for that ontology domain.

2.2.5. Protein Homology and Needleman-Wunsch Algorithm

In previously proposed databases, the homology of the proteins was restricted to 25-30% similarity. However, in this study, to group proteins into their locations with their functions was aimed considering the high inter group similarities. Therefore, global sequence similarity was computed by using Needleman-Wunsch algorithm. (Needleman and Wunsch, 1970) The general basis of the algorithm is to determine homology level of two input nucleotide or amino acid sequences by using dynamic programming. The alignment is scored for match, mismatch and gaps. For amino acid sequences, the scores for matches and mismatches are assigned through using a suitable substitution matrix.

Here, two step global sequence similarity calculations have been done. In order to select suitable substitution matrix type, first of all, global identity was computed by using identity matrix. The feature of the matrix is built on 1-0 scores, 1 for identical matches, 0 for mismatches. Self- maximum score and global identity were computed for the first and second sequences by using identity matrix. Later, similarity percentage was calculated by normalizing the global identity of two sequences according to the maximum self-score of the first sequence. Similarity percentage is shown in Equation 2.1.

$$Percentage\ Similarity = \left(\frac{AlignmentScore+minOfAlignment}{maxOfAlignment+minOfAlignment}\right)x100$$

This computation was conducted to select suitable substitution matrix type precisely. Alignment was performed for one sequence against the rest of the proteins in the database. This procedure was repeated until all sequences are aligned each other. If the similarity varies between 0-30 %, BLOSUM30 would be suitable. If the calculated similarity is between 30-62 %, BLOSUM62 matrix can be selected. If the similarity between two sequences is higher than 62%, BLOSUM90 can be selected. Aligned protein identifiers, the global identity, and similarity, minimum score of the alignment, and matrix type were stored in the table "Alignments" shown in Figure 2.2. Furthermore, in order to normalize each alignment, these scores were computed for self-alignment. The similarity between two sequences for corresponding substitution matrix was calculated by normalizing the score. Since the global similarity scores for each pair of sequences, they need to be normalized according to the maximum and minimum scores of self- alignment for certain substitution matrix. The normalization formula was given in Equation 2.2

$$Percentage\ Similarity = \left(\frac{AlignmentScore+minOfAlignment}{maxOfAlignment+minOfAlignment}\right)x100 \quad (2.2)$$

CHAPTER 3

RESULTS AND DISCUSSION

3.1. Data Acquistion from PDB

The data was obtained from PDB derived S2C database. The reason has been mentioned earlier that parsing PDB file format was fruitless because of the initial format design. Each chain in S2C file of a protein was considered as a protein itself with PDBID (PDB Identifier) -Chain as name. If the sequence of different chains were exactly same, one of them was kept and the rest was discarded while parsing the files. Therefore, starting with 64872 files, we collected 80980 proteins with their PDB identifiers, chain names, UniProt accession numbers, experimental techniques that were used to verify structures, resolutions of the experiments and isoelectric points to describe the biochemical properties of each protein. Data distribution according to experimental technique was shown in the Figure 3.1.

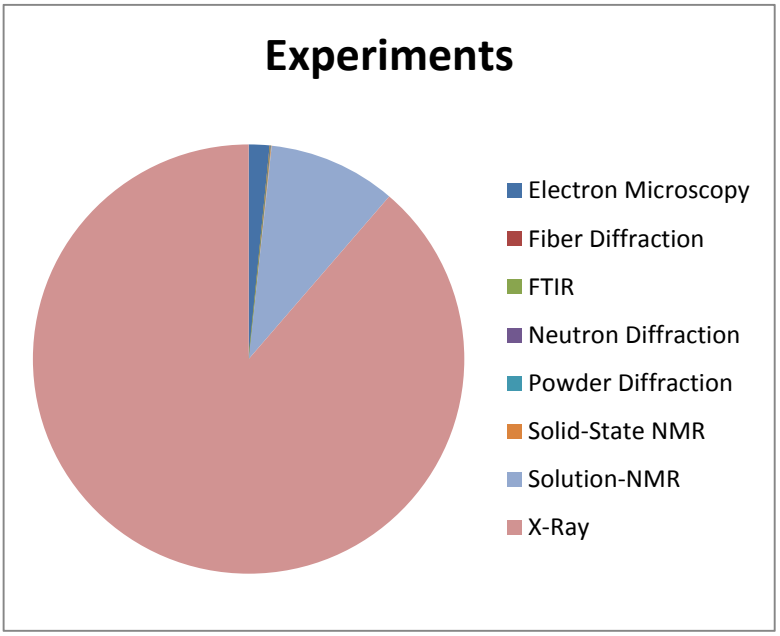

Figure 3.1. The data extracted from PDB was verified by Electron microscopy, Fiber Diffraction, FTIR, Neutron Diffraction, Powder diffraction, Solid and Solution states NMR and X-Ray crystallography. The most of the data was verified via X-Ray with 71821 entries, then 7775 solution-NMR and 1263 electron microscopy.

However, some proteins with unique PDBID-Chain-UniProt accession number had the same amino acid sequence. When these entries were examined, it was seen that same protein could be named with different identifiers when it was analyzed in different conditions. Therefore, the first filtering through table Proteins was performed in such a way that all amino acids sequences are seen as unique with the highest resolution. 43179 entries were were inserted into UniqueProteins table. Nevertheless, it was found out that there were missing UniProt accession information for 2513 in UniqueProteins table. To overcome this problem, the entries with missing UniProt accession numbers were collected and it was checked that whether other entries of Proteins table for each sequence with missing information has UniProt accession number. Only the 234 entries in Proteins table have UniProt accession numbers for the same amino acid sequences. Therefore, those 234 rows in UniqueProteins table were updated. The rest missing rows were deleted.

Another issue was related with experimentally verified structures. Due to the problems occurred during experiments, some of entries in UniqueProteins had unsuitable STRIDE structures in UniqueStructures table. These entries had many X, B, and dash characters. Moreover, some of the entries are less than 20 amino acid sequence length. All of these entries were collected and removed from UniqueStructures and UniqueProteins tables. In the end, UniqueProteins and UniqueStructures table had 38417 entries.

Unlike RS126 dataset, in this database, the number of proteins verified by different experimental techniques is increased. Whereas RS126 has 126 proteins, we collected 38417 proteins.

3.2. Data Acquisition from UniProtKB and GeneOntology

The data acquisition from UniProtKB fetch was quite problematic. Some of the proteins are given individual accession numbers while some of them are given family accession number. This difference is not mentioned in PDB data. While extracting data from UniProt, this issue came to the stage. Therefore, only one protein with highest resolution for a certain UniProtID was selected from the same family. Another issue was removing redundancy. Since in some of different sequences or proteins, UniProt accession numbers are the same, UniqueProteins table was grouped by UniProtIDs.

Thus, 20345 unique UniProt accession numbers were queried in the UniProtKB fetch and Gene Ontology. Protein sub cellular localizations were defined in 37,083 given unique UniProt accession number. Some of the entries had several sub cellular localizations. 4469 entries had several different sub cellular locations. However, most of these entries had cellular unit and sub cellular unit, for instance, cytoplasm and melanosome. These entries were not considered. The distribution of proteins according to their sub cellular locations according to twelve main groups as cytoplasm, endoplasmic reticulum, nucleus, extracellular space, golgi apparatus, mitochondria, peroxisome, vacuole, cytoskeleton, nucleolus, ribosomes, nuclear matrix was shown in Figure 3.2.

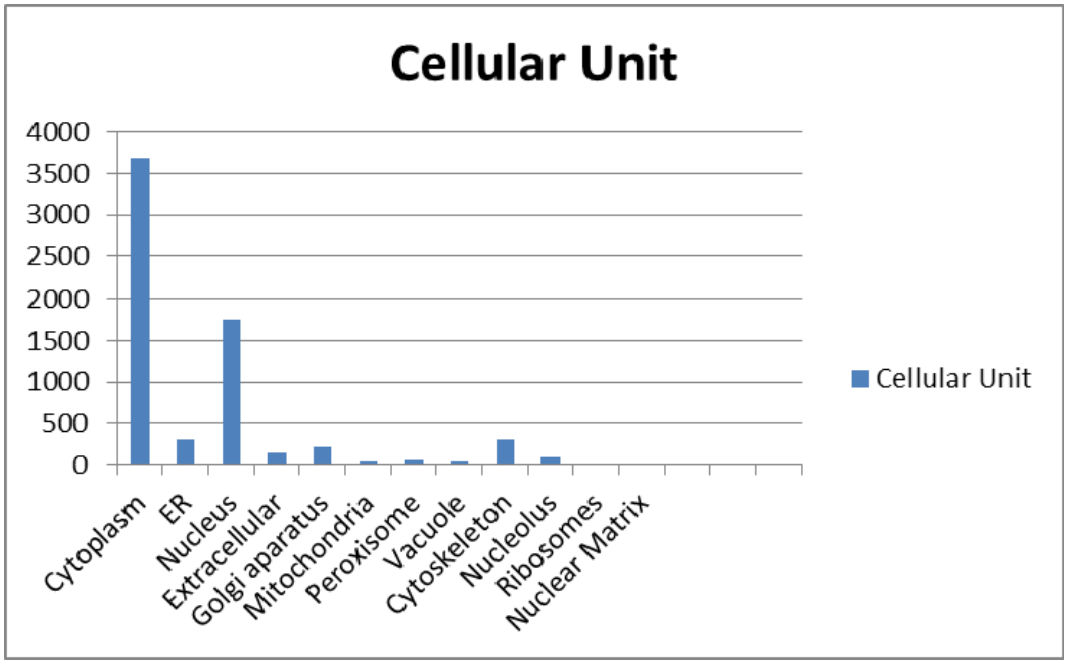

Figure 3.2. Sub cellular localizations of proteins are shown.

In addition, 19.774 of proteins in the database were eukaryotic and they were grouped in 731 functional groups in general. 14.021 of proteins were bacterial and they beloned to 479 functional groups. 2.383 of them were viral proteins and they belonged to 147 functional groups. 1.783 proteins were archaeal proteins and they belonged to 101 functional groups. 456 proteins were not identified with any taxa information. Thus, they were removed.

3.3. Global Sequence Alignment of Proteins

38417 sequences were aligned with global pairwise alignment. Therefore, 737.952.153 alignments were performed. In order to prevent the redundancy, sequences aligned with sequences that are not aligned previously.

3.4. Construction of Benchmark Datasets and Evaluation of The Algorithms

Here, in order to evaluate the performance of seven predictors, two sample benchmark datasets derived from benchmark database were introduced. For the comparision, RS126 dataset was also analyzed and results were presented.

When RS126 dataset was further examined in respect to protein function types and structural classes, it was seen that 50 different types of proteins were included and most of them were eukaryotic proteins with similar average length. 7 proteins were verified by NMR and the rest were verified by X-Ray crystallography. The resolution was assigned as -1 to NMR solved proteins. The variation in the pI was expected since the proteins were localized in different subcellular localizations with different function types. The pI and resolution values of proteins were shown in Figure 3.3.

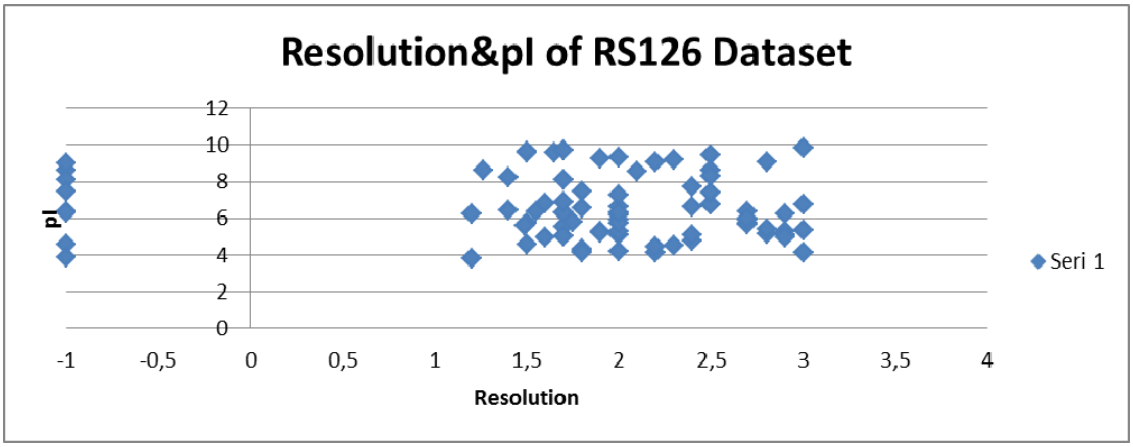

Figure 3.3. Resolution (x-axis) and pI (y-axis) values of proteins in RS126 dataset are shown. Proteins that were verified by NMR are shown as -1 in x-axis. The rest of the proteins that were by X-Ray crystallography are shown on the right hand side with their resolution values. The distribution in the pI is related to the different subcellular localizations and functionalities of proteins.

When the structural classes of the proteins were analyzed, it was seen that most of the proteins belonged to all alpha proteins, all beta proteins and alpha/beta proteins with 20%. The rest of the proteins were belonged to groups respectively alpha and beta, multi domain small proteins. There were also few membrane proteins and small peptides in the dataset. The overall percentages of each structural group were given in the Table 3.1.

Table 3.1. Structural classes and percentage of the proteins

Class Definition	**RS126 set (%)**
Alpha and beta (a/b)	25 proteins (20%)
Alpha and beta (a+b)	17 proteins (13%)
All alpha	27 proteins (21%)
All beta	38 proteins (20%)
Multi domain	3 proteins (2%)
Small proteins	18 proteins (14%)
Membrane	1 protein (< 1)
Peptides	1 protein (< 1)

The algorithms were run on the proteins in RS126 dataset and the prediction results and STRIDE assignments of proteins were converted into three structural element classes in order to make them comparable. The accuracies of the algorithms were computed according to two different evaluation techniques.

The first technique is three-state prediction accuracy measurement (Q3). It is a measure of the overall percentage of predicted residues to observed residues for each structural state. The average in other words Q3-All measurement is the percentage of correctly predicted residues to all residues. The formulation of Q3-All measurement is given in Equation 3.1.

$$Q3 = \frac{\text{number of residues correctly predicted}}{\text{number of all residues}} * 100 \quad (3.1)$$

The correctness of a prediction does not rely on the correct prediction of single amino acid residues. Therefore, correct prediction of portion of secondary structure segments gains importance. The second technique, segment overlap measure (SOV) (Venclovas et al. 1999) emerged to calculate the correctness of segment prediction in order to assess the quality in sense. Although per-state segment overlap measure could be computed, the overall segment overlap measure could be performed. The formulation of overall SOV is given in Equation 3.2.

$$SOV = \frac{1}{N} \sum_{i} \frac{\sum_{S(i)} \sum_{S(i)} MINOV(S1;S2) + DELTA(S1;S2)}{\sum_{S(i)} \sum_{S(i)} MAXOV(S1;S2)} * LEN(S1) \quad (3.2)$$

The accuracy results of algorithms on RS126 dataset that were computed according to Q3-All and Q3-per state and SOV-All and SOV-per state formulations were given in Table 3.2 and Table 3.3.

Table 3.2. Q3 accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in Simple and CB1999 decomposition methods on RS126

	Q3 All		**Q3 Helix**		**Q3 Sheet**		**Q3 Coil**	
	Sim	C&B	Sim	C&B	Sim	C&B	Sim	C&B
GOR	51,3	51,3	66,73	66,80	61,78	61,78	39,27	39,22
GOR4	64,23	64,23	58,43	58,49	55,70	53,70	75,27	75,21
HNN	66,82	66,77	65,37	63,35	54,31	54,31	77,66	77,59
PHD	75,42	75,17	71,27	70,99	64,83	64,20	78,48	78,27
PREDATOR	78,19	78,14	66,49	66,47	59,11	59,11	91,79	91,70
SIMPA96	69,65	69,55	65,40	65,12	54,45	53,53	79,78	79,83
SOPMA	69,05	69,01	73,27	73,27	59,53	59,53	72,95	72,90

According to Q3 results, PREDATOR and PHD showed high accuracy in compare to other predictors both in simple decompostion and in CB1999 decompostion methods.

In order to obtain more reliable information on the accuracy, SOV measurements were performed for decomposition by simple reduction and CB1999 reduction of results of each algorithm.

Table 3.3. SOV accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in Simple and CB1999 decomposition methods on RS126

	SOV All		**SOV Helix**		**SOV Sheet**		**SOV Coil**	
	Sim	C&B	Sim	C&B	Sim	C&B	Sim	C&B
GOR	47,30	47,32	58,04	58,07	54,20	54,20	39,74	39,75
GOR4	58,34	58,32	60,44	60,46	57,05	57,05	60,49	60,38
HNN	59,07	59,02	66,74	66,72	53,78	53,78	61,81	61,61
PHD	70,26	69,75	71,27	70,73	63,46	64,00	69,6	68,93
PREDATOR	66,19	66,13	68,99	68,96	59,65	59,65	65,60	65,48
SIMPA96	63,43	63,67	67,69	67,72	59,11	58,11	64,68	65,09
SOPMA	64,52	64,48	72,76	72,75	62,11	62,11	63,22	63,06

According to these results, it was seen that PHD and PREDATOR are more accurate than other algorithms in all structural states. Notice that PREDATOR and PHD have been trained on RS126 dataset. Therefore, the advantage of being trained by RS126 dataset, PREDATOR and PHD gave higher accuracy over other algorithms. It can be concluded that usage of proteins that are part of training set results a tendency on the accuracy results.

To assess the the impact of clustering proteins according to their sub cellular localizations and functions in that cellular location, three different benchmark datasets were generated from the new benchmark database.

The first dataset contains 31 eukaryotic glycoproteins. These proteins are found in the cell membrane. The resolution values and pI values are shown in Figure 3.4.

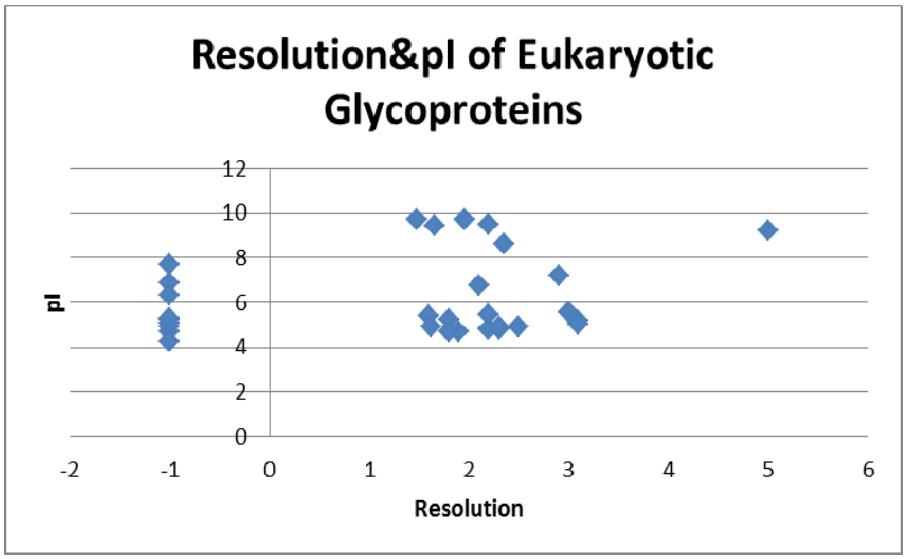

Figure 3.4. Resolution (x-axis) and pI (y-axis) values of proteins in Eukaryotic Glycoprotein dataset are shown. Proteins that were verified by NMR are shown as -1 in x-axis. The rest of the proteins that were by X-Ray crystallography are shown on the right hand side with their resolution values.

None of the proteins in this group were used to train any of these algorithms in order to prevent bias on the accuracy results. As an example cross similarity between sequences of this group was shown here as a heatmap in Figure 3.5.

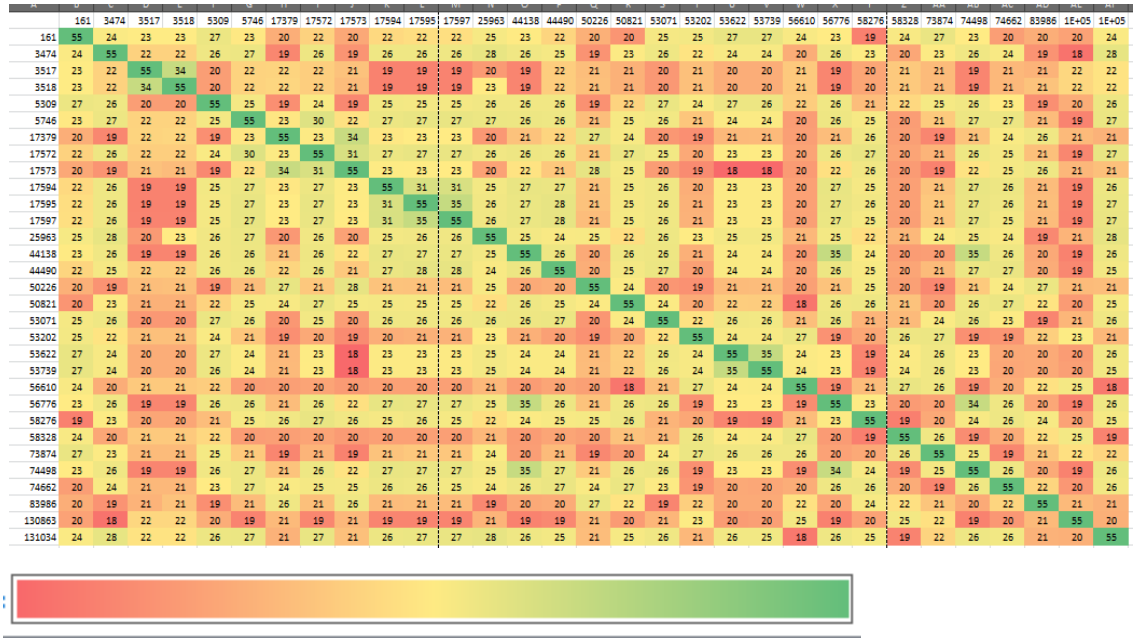

Figure 3.5. The cross similarities of proteins in eukaryotic glycoprotein group are shown as heatmap. Green color shows the maximum identity. Yellow tones show the similarity degree up to 80%. The red colors show the lowest similarities that are close to 50%.

According to the heatmap, it could be inferred that the inner group similarity is quite high. Since the proteins in this group locate in the same subcellular compartment, cell membrane, and play the same role in that location, it was expected to see the high sequence similarity within the group.

The evaluation of the algorithms in this group was computed as done in RS126 dataset. The results of Q3 measurement is shown in Table 3.4.

Table 3.4. Q3 accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in simple and CB1999 decomposition methods on Eukaryotic Membrane Glycoproteins

	Q3 All		**Q3 Helix**		**Q3 Sheet**		**Q3 Coil**	
	Sim	C&B	Sim	C&B	Sim	C&B	Sim	C&B
GOR	50,39	50,39	67,25	67,25	55,06	55,06	42,30	42,34
GOR4	60,25	60,25	41.41	41,41	47,66	47,66	73,14	74,95
HNN	63,12	63,12	60.16	60,16	41,15	41,15	75,12	75,16
PHD	72,44	72,44	48,74	48,74	66,05	66,05	75,06	77,26
PREDATOR	67,17	67,17	47,71	47,71	38,08	38,08	87,29	87,29
SIMPA96	63,38	63,38	55,57	55,57	43,61	43,61	77,06	77,06
SOPMA	67,94	67,94	64,46	64,46	53,18	53,18	78,51	76,5

Table 3.5. SOV accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in simple and CB1999 decomposition methods on Eukaryotic Membrane Glycoproteins

	SOV All		**SOV Helix**		**SOV Sheet**		**SOV Coil**	
	Sim	C&B	Sim	C&B	Sim	C&B	Sim	C&B
GOR	47,35	47,35	59,37	59,37	52,25	52,25	42,73	42,73
GOR4	59,98	52,98	43,78	43,78	50,72	50,72	56,10	56,10
HNN	48,60	48,60	58,01	58,01	41,71	41,71	55,47	55,47
PHD	64,02	64,02	47,96	47,96	65,67	65,67	63,83	63,83
PREDATOR	52,53	52,53	47,46	47,46	41,84	41,84	57,43	57,43
SIMPA96	54,56	54,56	54,29	54,29	47,27	47,27	58,9	58,92
SOPMA	58,53	58,53	59,27	59,27	56,16	56,16	61,92	61,92

Here, the results showed that PHD worked significantly well in eukaryotic membrane glycoproteins. Even though the number of proteins in this dataset was limited, the group specifity resulted in a considerable accuracy. Since the structure of membrane proteins differ from the other proteins and several algorithms have been developed to predict structure of these proteins specifically, it would be acceptable to obtain such an accuracy.

Another example is given for eukaryotic proteins that play role in translation. The resolutions and pI values of proteins in this group are shown in the Figure 3.6.

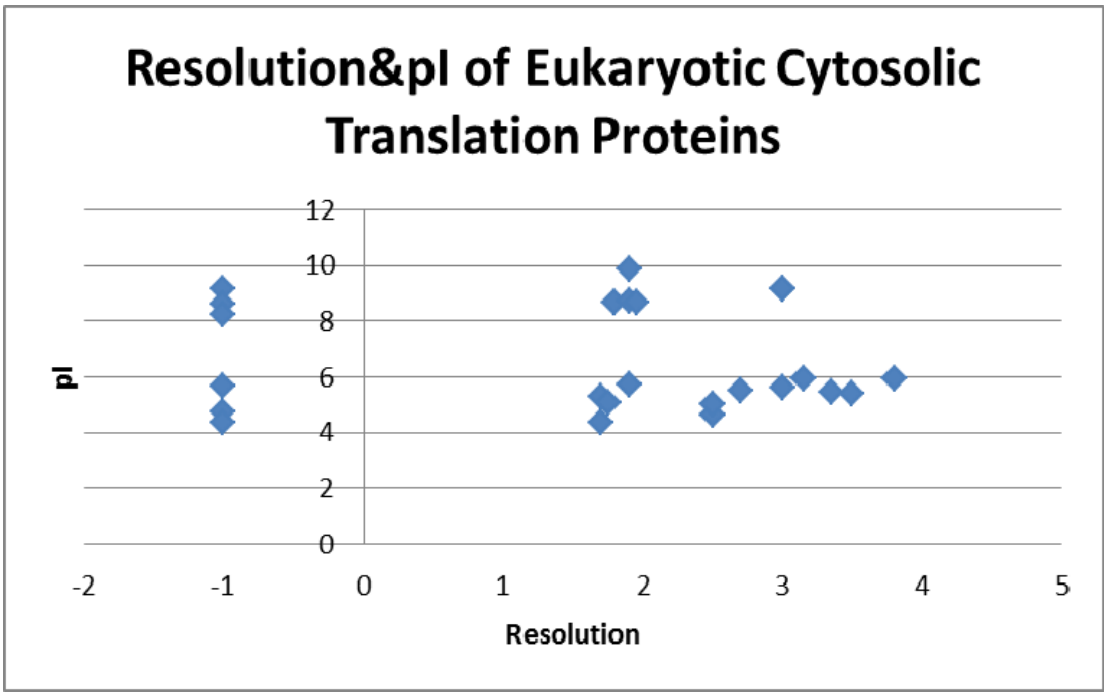

Figure 3.6. Resolution (x-axis) and pI (y-axis) values of proteins in Eukaryotic Cytolic Translation protein dataset are shown. Proteins that were verified by NMR are shown as -1 in x-axis. The rest of the proteins that were by X-Ray crystallography are shown on the right hand side with their resolution values.

The Q3 accuracy results of the algorithms in this dataset are given in Table 3.6.

Table 3.6. Q3 accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in simple and CB1999 decomposition methods on Eukaryotic Cytosolic Translation Proteins

	Q3 All		**Q3 Helix**		**Q3 Sheet**		**Q3 Coil**	
	Sim	C&B	Sim	C&B	Sim	C&B	Sim	C&B
GOR	48,51	48,51	77,35	77,35	63,09	63,09	28,90	28,90
GOR4	61,51	61,51	67,35	67,35	49,27	49,27	69,38	69,38
HNN	66,76	66,78	76,36	76,36	55,38	55,38	70,88	70,88
PHD	75,68	75,68	86,67	86,68	70,78	70,78	71,46	71,46
PREDATOR	64,85	64,85	59,86	59,86	47,66	47,66	78,85	78,85
SIMPA96	65,76	65,43	74,54	75.08	54,40	53.57	70,22	70,18
SOPMA	70,81	70,81	85,91	85,91	62,28	62,28	69,58	69,58

The accuracy results of the algorithms in this dataset according to SOV are given in Table 3.7.

Table 3.7. SOV accuracy results for each structural state; helix, sheet and coil are given for GOR, GOR4, HNN, PHD, PREDATOR, SOPMA and SIMPA96 in simple and CB1999 decomposition methods on Eukaryotic Cytosolic Translation proteins

	SOV All		**SOV Helix**		**SOV Sheet**		**SOV Coil**	
	Sim	C&B	Sim	C&B	Sim	C&B	Sim	C&B
GOR	46,42	46,42	64,39	64,39	59.2	59.2	32.44	32.44
GOR4	58,23	58,23	72.85	72.85	57.26	57.26	56.78	58.78
HNN	64,13	64,13	76.99	76.99	59.99	59.99	63.58	64.05
PHD	70,71	70,71	84.58	84.58	74.73	74.53	63.05	63.05
PREDATOR	60,06	60,06	65.83	65.83	56.03	56.03	61.61	61.61
SIMPA96	64,35	63.36	76.69	77.45	63.80	62.78	61.86	61.97
SOPMA	70,09	70,08	81.41	81.41	73.78	73.78	65.83	65.83

According to the results, SOPMA and PHD had higher accuracy over other algorithms. In compare to RS126 dataset results, it is explicit that SOPMA has a high performance as well as PHD. However, PREDATOR, which is seen as an accurate algorithm according to results on RS126, did not produce accurate results.

CHAPTER 4

CONCLUSION

In this study we report the advantages of grouping proteins into their subcellular locations and their functions in that location over evaluating secondary structure prediction algorithms. This would put the advantage of seeing the varying behaviours of the algorithms in respect to protein function, sequence similarity and the subcellular localization. In addition to that increasing the number of experimentally verified proteins in the datasets with a certain sequence similarity would lead more precise evaluation of the algorithms.

In the future, it is aimed to include newly verified structures of proteins into benchmark database. This would allow us to remove possible tendencies in the accuracy results of algorithms over each other. Since newly verified proteins would not be used in the training sets of currently available algorithms, it would be easy to assess the actual performances of algorithms for different taxas, different subcellular localizations, and functions.

We also aim to increase the number of algorithms to present more informative results to users. All benchmark datasets that can be generated from our database and the comparision results of algorithms in different states are planned to be published in a user-friendly web-interface.

REFERENCES

Allen, F., Almasi, G., Andreoni, W., Beece, D., Berne, B. J., Bright, A., Brunheroto, J., et al. (2001). Blue Gene: A vision for protein science using a petaflop supercomputer. *IBM Systems Journal*, *40*(2), 310-327. IBM Corporation.

Babaei, S., Geranmayeh, A., & Seyyedsalehi, S. A. (2010). Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks. *Computer methods and programs in biomedicine*, *100*(3), 237-47. Elsevier Ireland Ltd.

Baker, E. N., Arcus, V. L., & Lott, J. S. (2003). Protein structure prediction and analysis as a tool for functional genomics. *Applied bioinformatics*, *2*(3 Suppl), S3-10.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Lautrup, B., Norskov, L., Olsen, O. H., et al. (1988). Protein secondary structure and homology by neural networks. *FEBS Letters*, *241*, 223-228.

Bork, P., & Koonin, E. V. (1998). Predicting functions from protein sequences--where are the bottlenecks? *Nature Genetics*, *18*(4), 313-318.

Brünger, A. T., & Nilges, M. (1993). Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Quarterly Reviews of Biophysics*, *26*(1), 49-125.

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *the The European Molecular Biology Organization Journal*, *5*(4), 823-826. Nature Publishing Group.

Cole, C., Barber, J. D., & Barton, Geoffrey J. (2008). The Jpred 3 secondary structure prediction server. *Nucleic acids research*, *36*(Web Server issue), W197-201.

Cuff, J., & Barton, G J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, *34*(4), 508-19.

Cuff, J. a, & Barton, G J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, *40*(3), 502-11.

Fernández, C., & Wider, G. (2003). TROSY in NMR studies of the structure and function of large biological macromolecules. *Current Opinion in Structural Biology*, *13*(5), 570-580.

Fiser, A. (2004). Protein structure modeling in the proteomics era. *Expert review of proteomics*, *1*(1), 97-110.

Floudas, C., Fung, H., Mcallister, S., Monnigmann, M., & Rajgaria, R. (2006). Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, *61*(3), 966-988.

Frishman, D., & Argos, P. (1996). PREDATOR: Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Engineering*, *9*, 133-142.

Garnier, J., Gibrat, J.-F., & Robson, B. (1996). GOR secondary structure prediction method version IV. *Methods in Enzymology*, *266*, 540-553.

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.

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

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.

Guermeur, Y., & Gallinari, P. (n.d.). Combinaison de Classifieurs Statistiques, Application à la Prédiction de la Structure Secondaire des Protéines = Statistical Classifier Combination, Application to Protein Secondary Structure Prediction.

Van Gunsteren, W. F. (1993). Molecular dynamics studies of proteins. *Current Opinion in Structural Biology*, *3*(2), 277-281.

Han, K. F., & Baker, D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(12), 5814-8.

Heinig, M., & Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, *32*(Web Server issue), W500-W502. Oxford University Press.

Hua, S., & Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of molecular biology*, *308*(2), 397-407.

Huang, L., Hung, L., Odell, M., Yokota, H., Kim, R., & Kim, S.-H. (2002). Structure-based experimental confirmation of biochemical function to a methyltransferase, MJ0882, from hyperthermophile Methanococcus jannaschii. *Journal of Structural and Functional Genomics*, *2*(3), 121-127.

Nyugen J.C., & Rajapakse M. N. (2005). Two-stage multi-class support vector machines to protein secondary structure prediction. *Pacific Symposium on Biocomputing*, *357*, 346-357.

Jung, J.-W., & Lee, W. (2004). Structure-based functional discovery of proteins: structural proteomics. *Journal of Biochemistry and Molecular Biology*, *37*(1), 28-34.

Kabat, E. A., & Wu, T. T. (1973). The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in proteins: comparison of predicted and experimental determination of -sheets in concanavalin A. *Proceedings of the National Academy of Sciences of the United States of America*, *70*(5), 1473-1477.

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

Kabsch, W., & Sander, C. (1983b). How good are predictions of protein secondary structure? *FEBS Letters*, *155*(2), 179-182.

Kakumani, R., Devabhaktuni, V., & Ahmad, M. O. (2008). A Two-Stage Neural Network Based Technique for Protein Secondary Structure Prediction. *30th Annual International IEEE EMBS Conference* (pp. 1355-1358). Vancouver: IEEE.

Karp, P. D. (1998). What we do not know about sequence analysis and sequence databases. *Bioinformatics*.

Kim, H., & Park, H. (2003). Protein secondary structure prediction based on an improved support vector machines approach. *Protein Engineering Design and Selection*, *16*(8), 553-560.

Kopp, J., & Schwede, T. (2004). Automated protein structure homology modeling: a progress report. *Pharmacogenomics*, *5*(4), 405-416.

Lee, D., Redfern, O., & Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature reviews. Molecular cell biology*, *8*(12), 995-1005.

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.

Liu, H.-L., & Hsu, J.-P. (2005). Recent developments in structural proteomics for protein structure determination. *Proteomics*, *5*(8), 2056-68.

Madera, M., Calmus, R., Thiltgen, G., Karplus, K., & Gough, J. (2010). Improving protein secondary structure prediction using a simple k-mer model. *Bioinformatics (Oxford, England)*, *26*(5), 596-602.

Mariani, V., Kiefer, F., Schmidt, T., Haas, J., & Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics*

Mika, S., & Rost, Burkhard. (2003). UniqueProt: creating representative protein sequence sets. *Nucleic Acids Research*, *31*(13), 3789-3791. Oxford University Press.

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology*, *15*(3), 285-9.

Noguchi, T., & Akiyama, Y. (2003). PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Research*, *31*(1), 492-493. Oxford University Press.

Palopoli, L., Rombo, S. E., Terracina, G., Tradigo, G., & Veltri, P. (2009). Improving protein secondary structure predictions by prediction fusion. *Information Fusion*, *10*(3), 217-232. Elsevier B.V.

Pauling, L., & Corey, R. B. (1951a). The structure of synthetic polypeptides. *Proceedings of the National Academy of Sciences of the United States of America*, *37*(5), 241-250.

Pauling, L., & Corey, R. B. (1951b). Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*, *37*(5), 235-240. National Academy of Sciences.

Pauling, L., & Corey, R. B. (1951c). Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds. *Proceedings of the National Academy of Sciences of the United States of America*, *37*(11), 729-740. National Acad Sciences.

Petrey, D., & Honig, B. (2005). Protein Structure Prediction□: Inroads to Biology. *Structure*, *20*, 811-819.

Phizicky, E., Bastiaens, P. I. H., Zhu, H., Snyder, M., & Fields, S. (2003). Protein analysis on a proteomic scale. *Nature*, *422*(6928), 208-15.

Pirovano, W., & Heringa, J. (2011). Protein Secondary Structure Prediction. *Methods Mol Bio*, *6*(1).

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.

Qu, W., Sui, H., Yang, B., & Qian, W. (2011). Improving protein secondary structure prediction using a multi-modal BP method. *Computers in biology and medicine*, *41*(10), 946-59.

Rost, B, & Sander, C. (1993, July 20). Prediction of protein secondary structure at better than 70% accuracy. *Journal of molecular biology*.

Rost, B, & Sander, C. (2000). Third generation prediction of secondary structures. *Methods In Molecular Biology Clifton Nj*, *143*, 71-95. Springer.

Rost, Burkhard, & O'Donoghue, S. (1997). Sisyphus and prediction of protein structure. *Comput Appl Biosci* , *13*(4), 345-356.

Szent-Gyorgyi, A. G., & Cohen, C. (1957). Role of proline in polypeptide chain configuration of proteins. *Science*, *126*(3276), 697-698.

Schneider, R., & Sander, C. (1996). The HSSP database of protein structure-sequence alignments. *Nucleic Acids Research*, *24*(1), 201-205.

Shoyaib, M., Baker, S. M., Jabid, T., Anwar, F., & Khan, H. (2007). Protein secondary structure prediction with high accuracy using Support Vector Machine. *2007 10th International Conference on Computer and Information Technology*, 1-4. Ieee.

Sreerama, N., Venyaminov, S. Y., & Woody, R. W. (1999). Estimation of the number of alpha-helical and beta-strand segments in proteins using circular dichroism spectroscopy. *Protein Science*, *8*(2), 370-380. Cold Spring Harbor Laboratory Press.

Ubarretxena-Belandia, I., & Stokes, D. L. (2010). Present and future of membrane protein structure determination by electron crystallography. *Advances in protein chemistry and structural biology*, *81*(10), 33-60. Elsevier Inc.

Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. (1999). Some measures of comparative performance in the three CASPs. *Proteins: Structure Function and Genetics*, 34, 220-223

Ward, J. J., McGuffin, L. J., Buxton, B. F., & Jones, D. T. (2003). Secondary structure prediction with support vector machines. *Bioinformatics*, *19*(13), 1650-1655.

Watson, J. D., Laskowski, R. a, & Thornton, J. M. (2005). Predicting protein function from sequence and structural data. *Current opinion in structural biology*, *15*(3), 275-84.

Whisstock, J. C., & Les, A. M. (2003). Prediction of protein function from protein sequence and structure. *Quarterly Reviews of Biophysics*, *36*(3), 307-340.

Wistow, G., & Piatigorsky, J. (1987). Recruitment of enzymes as lens structural proteins. *Scienceence*, *236*, 1554-1556.

Xu, Jinbo Peng, Jian Zhao, F. (2009). Template-based and free modeling by RAPTOR++ in CASP8. *Proteins*, *77*(Suppl 9), 133-137.

Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 342-348.

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*, 430.