A COMPUTATIONAL APPROACH TO PREDICT CONTACT POTENTIAL AND DISULFIDE BOND OF PROTEINS

by

ELANUR ŞİRELİ

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
July 2004

A COMPUTATIONAL APPROACH TO PREDICT CONTACT POTENTIAL AND DISULFIDE BOND OF PROTEINS

APPROVED BY:

Asst. Prof. O. Uğur Sezerman ………………………….
(Thesis Supervisor)

Asst. Prof. BerrinYanıkoğlu ………………………….

Prof. Aytül Erçil ………………………….

DATE OF APPROVAL: ………………………….

ABSTRACT

Contact map and disulfide bond information of a protein give crucial clues about 3-dimensional structure and function of a protein. In this study, we represent a computational approach to predict both contact maps and disulfide bonds of the residues inside of a protein and these studies are two of the essential steps of protein folding problem.

In the first study, we predicted contacting residues of proteins using physical (ordering, length and volume), chemical (hydrophobicity), evolutionary (neighboring) and structural (secondary structure) information by implementing classification techniques, Neural Networks (NNs) and Support Vector Machines (SVMs). As a result, our method predicts 14% of the contacting residues with 0.6% false positive ratio and it performs 9 times better than a random predictor.

In the second study, using the same parameters we predicted cysteine residues forming. In this study, we used SVMs, we obtained 63.76% accuracy in disulfide bond prediction.

ÖZET

Bir proteinin temas eden amino asit ve ikili-sülfat bağı bilgileri proteinin 3 boyutlu yapısı ve fonksiyonu ile ilgili önemli ipucları vermektedir. Bu çalışmada, bilgilerin tahmini üzerine işlemsel yaklaşım sunulmaktadır ve bu çalışmaların her ikisi de protein katlanması probleminin önemli adımlarını oluşturmaktadır.

İlk çalışmada, proteinlerin temas matrikslerinin tahmini üzerine çalıştık. Tahmin işlemi için, Sinir Ağları ve Destek Vektör Makinaları tekniklerini uygulayarak, proteinlerin fiziksel (sıralanma, hacim ölçüleri), kimyasal, evrimsel (komşu bilgileri) ve yapısal (ikincil yapı) bilgileri kullanıldı. Çalışmanın sonunda, %0.6'lık temas dışı hata oranı ile temas örneklerinin %14'ünü tahmin edebildik ve bu tahmin, raslantısal tahminden 9 kat daha iyidir.

İkinci çalışmada, aynı parametreleri kullanarak, sistin amino asitlerinin bağlanıp ikili sülfat bağı oluşturabilirliğini tahmin ettik. Bu çalışmada SVM kullandık ve ikili-sülfat bağı tahmininde %63.76 doğruluğa ulaştık.

To my family with all my heart

ACKNOWLEDGEMENTS

I wish to thank my supervisor, Assist. Prof. O. Uğur Sezerman, for his guidance, encouragement and support throughout this study. I am also grateful to Assist. Prof. A. Berrin Yanıkoğlu and Prof. Aytül Erçil, who served on my advisory committee and examining committee, for their advice and useful contributions to my work.

I would like to thank to Beste Koral, Demet Dinç, Dilek Özüzümcü, Sevcan Karadadaş, Vildan Pınar, Batuhan Çayırlı, Murat Hacıömeroğlu, Burcu Dartan, Burcu Kaplan, Filiz Dede, Zehra Özaydın and F. Baran Eliş for their support and friendship.

TABLE OF CONTENTS

TABLE OF CONTENTS..viii

1 INTRODUCTION 1

2 OVERVIEW 3

2.1 Biological Background 3

2.1.1 Amino Acids 3

2.1.2 Volume.. 5

2.1.3 Hydrophobicity 5

2.1.4 Contact Definition.. 7

2.1.5 Disulfide Bond 7

2.1.6 The Peptide Bond.. 8

2.1.7 Proteins 8

2.1.8 Protein Folding 8

2.1.9 Levels of Protein Structure 9

2.1.10 Classification of Protein Structures 11

2.2 Prediction of Contact Map and Disulfide Bond.. 13

2.2.1 Role of Contact Map and Disulfide Bond 13

2.2.2 Contact Map Prediction in Literature 14

2.2.3 Disulfide Bond Prediction in Literature 15

2.3 Methods 16

2.3.1 Neural Networks 16

2.3.2 Neural Network Topology 17

2.3.3 Training 17

2.3.4 Testing 18

2.3.5 Global and Local Minima of Energy Function 18

2.3.6 Error Function and Back-Propagated Value 19

2.3.7 Summation Function 19

2.3.8 Transfer Function 20

2.3.9 Output Function 22

2.3.10 A NNs Tool, EasyNN 22

2.4 Support Vector Machines 23

2.4.1 SVM Hyperplane 24

2.4.2 SVM Training Rule 24

2.4.3 Linear SVMs 25

2.4.3.1 Classification of Linearly Separable Data 25

2.4.3.2 Classification of Nonlinearly Separable Data 26

2.4.4 Nonlinear SVMs 26

2.4.5 A SVM Tool BSVM 27

2.4.6 Performance Evaluation Metrics 28

2.4.7 Source of Data ... 28

3 RESULTS AND DISCUSSIONS ... 29

3.1 Contact Map Prediction Study ... 29

3.1.1 Experiment 1 ... 30

3.1.2 Experiment 2 ... 32

3.1.3 Experiment 3 ... 33

3.1.4 Experiment 4 ... 34

3.1.5 Experiment 5 ... 37

3.1.6 Experiment 6 ... 38

3.1.7 Experiment 7 ... 40

3.1.8 Experiment 8 ... 40

3.1.9 Experiment 9 ... 41

3.1.10 Experiment 10 ... 42

3.1.11 Experiment 11 ... 44

3.1.12 Experiment 12 ... 45

3.2 Disulfide Bond Prediction Study ... 47

3.2.1 Experiment 13 ... 47

3.2.2 Experiment 14 ... 48

3.2.3 Experiment 15 ... 49

3.2.4 Experiment 16 ... 49

3.2.5 Experiment 17 ... 50

4 CONCLUSION 52

5 REFERENCES 55

APPENDIX 60

ABBREVIATIONS

PDB: Protein Data Bank

NN: Neural Networks

CATH: A Hierarchic Classification of Protein Domain Structures

SCOP:Structural Classification of Proteins

SVM: Support Vector Machines

RBF: Radial Basis Function

KKT :Karush-Kühn-Tucker

C_α: alpha - carbon atom

C_β: beta - carbon atom

LIST OF FIGURES

Figure 2.1 Atomic Structure of Amino Acid 3

Figure 2.2 Common Atomic Structure of Residues 7

Figure 2.3 Peptide Bond 8

Figure 2.4 Forming of an Alpha Helix 10

Figure 2.5 Beta Sheet Structure 10

Figure 2.6 Steps of 3D Structure Prediction 13

Figure 2.7 Structure of a Biological Neuron and NN Neuron. 16

Figure 2.8 Topology of NN 17

Figure 2.9 Local Minima of Error Function Surface 19

Figure 2.10 SVM Classification by Separating Hyperplane 23

Figure 3.1 Distribution of the Prediction in Experiment 2 32

Figure 3.2 Desired Prediction Distribution 33

Figure 3.3 Representation of New Approach 37

LIST OF TABLES

Table 2.1 Volume and Hydrophobicity Scales of Residues 6

Table 2.2 Sample Transfer Functions 21

Table 2.3 Kernel Functions.. 27

Table 2.4 Used BSVM Parameters 27

Table 3.1 Cluster Information of Residues 30

Table 3.2 Results for Experiment 1 31

Table 3.3 Results for Experiment 4 36

Table 3.4 Results for Experiment 5 38

Table 3.5 Results for Optimization Process... 39

Table 3.6 Results for Experiment 7 40

Table 3.7 Results for Experiment 8 41

Table 3.8 Results for Experiment 9 42

Table 3.9 Superfamily Information 43

Table 3.10 Results for Experiment 10 43

Table 3.11 Results for Experiment 11 44

Table 3.12 Results for Experiment 12 45

Table 3.13 Results for Experiment 13 48

Table 3.14 Results for Experiment 14 48

Table 3.15 Results for Experiment 15 49

Table 3.16 Results for Experiment 16 50

Table 3.17 Results for Experiment 17 50

A COMPUTATIONAL APPROACH TO PREDICT CONTACT POTENTIAL AND DISULFIDE BOND OF PROTEINS

by

ELANUR ŞİRELİ

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
July 2004

A COMPUTATIONAL APPROACH TO PREDICT CONTACT POTENTIAL AND DISULFIDE BOND OF PROTEINS

APPROVED BY:

Asst. Prof. O. Uğur Sezerman ………………………….
(Thesis Supervisor)

Asst. Prof. BerrinYanıkoğlu ………………………….

Prof. Aytül Erçil ………………………….

DATE OF APPROVAL: ………………………….

ABSTRACT

Contact map and disulfide bond information of a protein give crucial clues about 3-dimensional structure and function of a protein. In this study, we represent a computational approach to predict both contact maps and disulfide bonds of the residues inside of a protein and these studies are two of the essential steps of protein folding problem.

In the first study, we predicted contacting residues of proteins using physical (ordering, length and volume), chemical (hydrophobicity), evolutionary (neighboring) and structural (secondary structure) information by implementing classification techniques, Neural Networks (NNs) and Support Vector Machines (SVMs). As a result, our method predicts 14% of the contacting residues with 0.6% false positive ratio and it performs 9 times better than a random predictor.

In the second study, using the same parameters we predicted cysteine residues forming. In this study, we used SVMs, we obtained 63.76% accuracy in disulfide bond prediction.

ÖZET

Bir proteinin temas eden amino asit ve ikili-sülfat bağı bilgileri proteinin 3 boyutlu yapısı ve fonksiyonu ile ilgili önemli ipucları vermektedir. Bu çalışmada, bilgilerin tahmini üzerine işlemsel yaklaşım sunulmaktadır ve bu çalışmaların her ikisi de protein katlanması probleminin önemli adımlarını oluşturmaktadır.

İlk çalışmada, proteinlerin temas matrikslerinin tahmini üzerine çalıştık. Tahmin işlemi için, Sinir Ağları ve Destek Vektör Makinaları tekniklerini uygulayarak, proteinlerin fiziksel (sıralanma, hacim ölçüleri), kimyasal, evrimsel (komşu bilgileri) ve yapısal (ikincil yapı) bilgileri kullanıldı. Çalışmanın sonunda, %0.6'lık temas dışı hata oranı ile temas örneklerinin %14'ünü tahmin edebildik ve bu tahmin, raslantısal tahminden 9 kat daha iyidir.

İkinci çalışmada, aynı parametreleri kullanarak, sistin amino asitlerinin bağlanıp ikili sülfat bağı oluşturabilirliğini tahmin ettik. Bu çalışmada SVM kullandık ve ikili-sülfat bağı tahmininde %63.76 doğruluğa ulaştık.

To my family with all my heart

ACKNOWLEDGEMENTS

I wish to thank my supervisor, Assist. Prof. O. Uğur Sezerman, for his guidance, encouragement and support throughout this study. I am also grateful to Assist. Prof. A. Berrin Yanıkoğlu and Prof. Aytül Erçil, who served on my advisory committee and examining committee, for their advice and useful contributions to my work.

I would like to thank to Beste Koral, Demet Dinç, Dilek Özüzümcü, Sevcan Karadadaş, Vildan Pınar, Batuhan Çayırlı, Murat Hacıömeroğlu, Burcu Dartan, Burcu Kaplan, Filiz Dede, Zehra Özaydın and F. Baran Eliş for their support and friendship.

TABLE OF CONTENTS

TABLE OF CONTENTS..viii

1 INTRODUCTION 1

2 OVERVIEW 3

2.1 Biological Background 3

2.1.1 Amino Acids 3

2.1.2 Volume... 5

2.1.3 Hydrophobicity 5

2.1.4 Contact Definition... 7

2.1.5 Disulfide Bond 7

2.1.6 The Peptide Bond.. 8

2.1.7 Proteins 8

2.1.8 Protein Folding 8

2.1.9 Levels of Protein Structure 9

2.1.10 Classification of Protein Structures 11

2.2 Prediction of Contact Map and Disulfide Bond.................................... 13

2.2.1 Role of Contact Map and Disulfide Bond 13

2.2.2 Contact Map Prediction in Literature 14

2.2.3 Disulfide Bond Prediction in Literature 15

2.3 Methods 16

2.3.1 Neural Networks 16

2.3.2 Neural Network Topology 17

2.3.3 Training 17

2.3.4 Testing 18

2.3.5 Global and Local Minima of Energy Function 18

2.3.6 Error Function and Back-Propagated Value 19

2.3.7 Summation Function 19

2.3.8 Transfer Function 20

2.3.9 Output Function 22

2.3.10 A NNs Tool, EasyNN 22

2.4 Support Vector Machines 23

2.4.1 SVM Hyperplane 24

2.4.2 SVM Training Rule 24

2.4.3 Linear SVMs 25

2.4.3.1 Classification of Linearly Separable Data 25

2.4.3.2 Classification of Nonlinearly Separable Data 26

2.4.4 Nonlinear SVMs 26

2.4.5 A SVM Tool BSVM 27

2.4.6 Performance Evaluation Metrics 28

2.4.7 Source of Data 28

3 RESULTS AND DISCUSSIONS 29

3.1 Contact Map Prediction Study 29

3.1.1 Experiment 1 30

3.1.2 Experiment 2 32

3.1.3 Experiment 3 33

3.1.4 Experiment 4 34

3.1.5 Experiment 5 37

3.1.6 Experiment 6 38

3.1.7 Experiment 7 40

3.1.8 Experiment 8 40

3.1.9 Experiment 9 41

3.1.10 Experiment 10 42

3.1.11 Experiment 11 44

3.1.12 Experiment 12 45

3.2 Disulfide Bond Prediction Study 47

3.2.1 Experiment 13 47

3.2.2 Experiment 14 48

3.2.3 Experiment 15 49

3.2.4 Experiment 16 49

3.2.5 Experiment 17 50

4 CONCLUSION 52

5 REFERENCES 55

APPENDIX 60

ABBREVIATIONS

PDB: Protein Data Bank

NN: Neural Networks

CATH: A Hierarchic Classification of Protein Domain Structures

SCOP:Structural Classification of Proteins

SVM: Support Vector Machines

RBF: Radial Basis Function

KKT :Karush-Kühn-Tucker

C_α: alpha - carbon atom

C_β: beta - carbon atom

LIST OF FIGURES

Figure 2.1 Atomic Structure of Amino Acid 3

Figure 2.2 Common Atomic Structure of Residues... 7

Figure 2.3 Peptide Bond 8

Figure 2.4 Forming of an Alpha Helix 10

Figure 2.5 Beta Sheet Structure 10

Figure 2.6 Steps of 3D Structure Prediction 13

Figure 2.7 Structure of a Biological Neuron and NN Neuron. 16

Figure 2.8 Topology of NN 17

Figure 2.9 Local Minima of Error Function Surface 19

Figure 2.10 SVM Classification by Separating Hyperplane.. 23

Figure 3.1 Distribution of the Prediction in Experiment 2 32

Figure 3.2 Desired Prediction Distribution.. 33

Figure 3.3 Representation of New Approach 37

LIST OF TABLES

Table 2.1 Volume and Hydrophobicity Scales of Residues 6

Table 2.2 Sample Transfer Functions 21

Table 2.3 Kernel Functions... 27

Table 2.4 Used BSVM Parameters 27

Table 3.1 Cluster Information of Residues 30

Table 3.2 Results for Experiment 1 31

Table 3.3 Results for Experiment 4 36

Table 3.4 Results for Experiment 5 38

Table 3.5 Results for Optimization Process.. 39

Table 3.6 Results for Experiment 7 40

Table 3.7 Results for Experiment 8 41

Table 3.8 Results for Experiment 9 42

Table 3.9 Superfamily Information 43

Table 3.10 Results for Experiment 10 43

Table 3.11 Results for Experiment 11 44

Table 3.12 Results for Experiment 12 45

Table 3.13 Results for Experiment 13 48

Table 3.14 Results for Experiment 14 48

Table 3.15 Results for Experiment 15 49

Table 3.16 Results for Experiment 16 50

Table 3.17 Results for Experiment 17 50

1 INTRODUCTION

Proteins are the biochemical molecules that make up cells, organs and organisms so they are the building stones of living organisms. Each protein has its own fold and as a result of this fold it has its own function and three-dimensional structure. This fold occurs to provide the native conformation of lowest available free energy in given environmental conditions. To predict the native fold of a protein from the primary sequence of residues is referred to as the protein-folding problem [1].

Finding the fold of a protein is important, because the structure determines the function of proteins in organisms and their impact on biological reactions, task in cell and role in diseases such as cancer. In addition, if we discover why and how a protein achieves its fold, it is possible to design drug and artificial proteins to perform some desired functions.

By the genome project, millions of proteins have been identified from different organisms [2]. However, their folded structures and their functions are still mostly unknown. Thus, prediction of the structure and function of proteins, based on their residue sequences, is the major challenge in computational biology [3].

The three-dimensional structure of the protein molecule can be represented in a convenient way as a two dimensional map of the contacts, called contact map, between residues [4]. In the first part of the study we represent a computational approach to generate contact map of any given protein sequence. As a fundamental intermediary step, the contact map of a protein gives crucial hints about three-dimensional structure of this protein. There are many approaches developed to predict contact map such as finding correlated inter changes in multiple sequence alignment [5]; likelihood matrix methods [6]; and training NNs with encodings of multiple sequence alignments [7, 8, 9, 10].

In our approach, we divide the primary sequence of a protein into N-size windows and analyze it by using some pattern recognition techniques (Neural Networks (NNs) and Support Vector Machines (SVMs)). Thus, we may theoretically find contacting residues, which would help us determine the fold of proteins by computational methods without a great deal of experimentation and in a more robust way.

In training, we used some chemical and physical characteristics of contacting residues and, in addition to these, we used some characteristics of neighbors such as hydrophobicity, secondary structure patterns, volume etc.

In the second part of the study, we have predicted disulfide bond, which is formed by side chain sulfide atoms of cysteine residues. This bond is crucial for protein folding problem because it is the strongest bond in protein structure and introduces extra stability to the structure. Hence, disulfide bond makes a major contribution to three-dimensional structure of protein.

Because of the importance of finding disulfide bonds, many researchers have tried to predict the characteristics of disulfide bond formation in proteins using statistical studies [11], NNs studies [12, 13] etc.

Disulfide bond prediction is similar to contact map prediction by nature. In this case, we tried to predict contacts between cysteine residues only rather than any two residues. Therefore, we used similar information, physical (ordering, length, volume) and chemical characteristics (hydrophobicity scales) of cysteine residues and neighboring residues. Same as the previous study, the window approach is used in this phase of the study. However, in this study, only one pattern recognition technique, SVM, is used.

2 OVERVIEW

2.1 Biological Background

2.1.1 Amino Acids

Amino acid is an organic compound containing an amino group (-NH2), a carboxyl group (-COOH) and a side chain that distinguish one amino acid from another. [14]

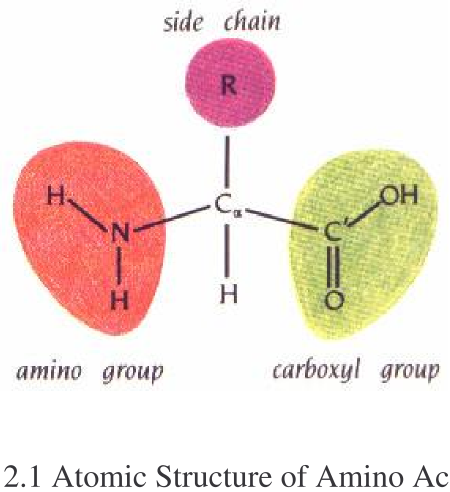


Figure 2.1 Atomic Structure of Amino Acid [15]

Amino acids fall into several naturally occurring groups. However, usually they are grouped into three different classes with using their side chains [16]. Those classes are as follows:

Hydrophobic amino acids do not want to interact with water. They tend to cluster on the inside of the molecule. Thus, core of the protein structure, stabilized by numerous van der Waals interactions, which is a non-covalent force that result from the attraction of one atoms nucleus for the electrons of another atom in a non-covalent form [16], is composed of hydrophobic residues. Hydrophobicity gives them an important role to play in determining the three-dimensional structure of proteins. This class comprises those Alanine, Proline (they are weakly hydrophobic and have small, nonpolar side chains), Valine, Leucine, Isoleucine, Phenylalanine and Methionine (they are strongly hydrophobic and have larger side chains) [14].

Charged amino acids are normally found on the surface of the protein where they interact with water and with other biological molecules. Thus, these amino acids are important in the determining of oppositely charged groups on molecules that interact with proteins. The acids of this class are Aspartic acid, Glutamic acid (they have carboxyl groups on their side chains so they are naturally charged), Lysine and Arginine (they have side chains with amino groups so they are positively charged) [14].

Polar amino acids exist both interior as well as on the surface of the protein. They form hydrogen bonds with water or with other polar residues. This class comprises those with polar side chains Serine, Threonine (they have hydroxyl groups on their side chains and extraordinarily important in the regulation of the activity of different proteins), Asparagine, Glutamine (they cannot be ionized and therefore, they are uncharged), Histidine (it is either uncharged or positively charged, depending on local environment. These states make it important, in the catalytic mechanism of enzymes and explain why it is often found in the active site.), Tyrosine (it is weakly acidic and can be chemically modified by combining with a peptide chain), Tryptophan (it tends to be found buried inside of protein structure), Glycine (it has a single hydrogen atom as its side chain and it is the simplest amino acid) and Cysteine (it can provide a bond with another cysteine via the sulfur atoms to form a covalent disulfide bridge. This bond is important in determining the three-dimensional structure of many proteins) [14].

2.1.2 Volume

Volume is a size measure to define the space that residue occupy. It is an important property to determine contact tendency of residues, because the residue substitution probability is inversely related with the difference of residue sizes. This feature is also important for contact potentials. Big residues would have higher probability to contact another big residue if surrounded by small residues. This probability would decrease if big residues surround them. Also big residues by nature would make contact easily.

In the experiments, we used the volume scales, which are taken from an implementation study of the method Lee and Richards [17, 18] and Baysal et al. study.

2.1.3 Hydrophobicity

Hydrophobicity is a non-covalent bond and has a central role in determining the shape of a protein. In order to minimize the deteriorating effect on the hydrogen-bonded network of water molecules, hydrophobic molecules tend to be forced together in an aqueous environment. Therefore, an important factor controlling the folding of protein is the distribution of its polar and nonpolar residues. The hydrophobic side chains tend to cluster in the inside of the molecule core. This provides them to avoid making contact with the water molecules that surround them inside of a cell. On the contrary, polar side chains want to take place near the surface of the molecule, where they form hydrogen bonds with water or with other polar or charged residue. When polar residues are embedded within the protein, generally, they make hydrogen bond with other polar residues or with the polyprotein backbone. This explains how hydrophobic effect is important as one of the contacting forces inside of a protein [16]. Therefore, we use hydrophobic values of a window of residues in our experiments.

Hydrophobic value of a residue has been measured experimentally in different ways such as using the free residues, residues with the amino and carboxyl groups blocked and side-chain analogues with the backbone replaced by a hydrogen atom. In contact potential experiments, we have just use ROSEF's hydrophobicity scale [19, 20]. In disulfide-bond experiments, we have used three of hydrophobicity scales, ROSEF, Eisenberg and Hopp-Woods [21].

Residue Type	**Volume**	**Hydrophobicity**		
		ROSEF	**Eisenberg**	**Hopp-Woods**
Alanine	107.95	0.50	0.25	-0.5
Arginine	238.76	-2.01	-1.8	3
Asparagine	143.94	-2.26	-0.64	0.2
Aspartic acid	140.39	-2.51	-0.72	3
Cysteine	134.28	4.77	0.04	-1
Glutamine	178.50	-2.51	-0.69	0.2
Glutamic Acid	172.25	-2.51	-0.62	3
Glycine	80.10	0	0.16	0
Histidine	182.88	1.51	-0.4	-0.5
Isoleucine	175.12	4.02	0.73	-1.8
Leucine	178.63	3.27	0.53	-1.8
Lysine	200.81	-5.03	-1.1	3
Methionine	194.15	3.27	0.26	-1.3
Phenylalanine	199.48	4.02	0.61	-2.5
Proline	136.13	-2.01	-0.07	0
Serine	116.50	-1.51	-0.26	0.3
Threonine	139.27	-0.5	-0.18	-0.4
Tryptophan	249.36	3.27	0.37	-3.4
Tyrosine	212.76	1.01	0.02	-2.3
Valine	151.44	3.52	0.54	-1.5

Table 2.1 Volume and Hydrophobicity Scales of Residues

2.1.4 Contact Definition

A residue is any molecule that contains amino and carboxylic acid functional groups and a side chain as illustrated in Figure 2.1. In side chain region, there is a β carbon atom. When two residues' β carbon (α carbon for gylcine) are closer than 7Å, that means, these residues are in contact. There are other methods which use different contact definitions, but we use just C_β atoms (C_α for glycine) to determine the contact relation between two residues.

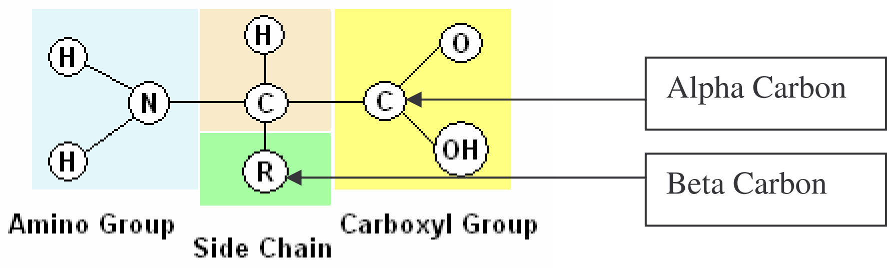


Figure 2.2 Common Atomic Structure of Residues

2.1.5 Disulfide Bond

It is a single covalent bond between the sulfur atoms in cysteine residues. By forming these covalent bonds, very distant fragments of a protein sequence may be forced to make bond. Thus, the location of these bonds is a very informative constraint on understanding some characteristics of the protein such as the folding, structure and function of proteins [22].

By existing such bonds, the conformational stability of the protein is increased both by lowering the entropy of the folded state and by forming stabilizing interaction in the native state. However, the disulfide bonds can be considered as part of the primary structure of a protein. In addition, they are very important in determining the tertiary structure of proteins and the quaternary structure of some proteins by having function to stabilize the tertiary and/or quaternary structures of proteins [23].

2.1.6 The Peptide Bond

When amino acids are joined together, peptide bonds are generated. The carboxyl group of the first amino acid is attached to the amino group of the next amino acid to eliminate water then they form peptide bond.

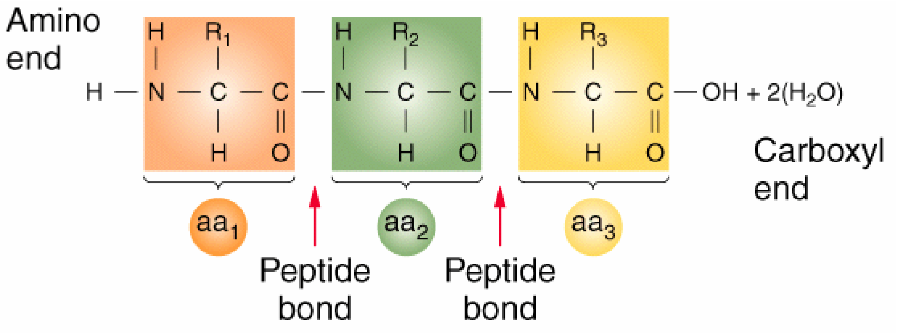


Figure 2.3 Peptide Bond [24]

2.1.7 Proteins

Proteins have a crucial role in living organisms by executing nearly all the functions in the cell. Without proteins, growth or development is not possible. They are made of 20 different building blocks, called residues or amino acids, which give distant structure backbone side chain. Each protein has a unique residue sequence.

2.1.8 Protein Folding

Proteins cannot be described exactly by just using their residue sequence. Even though, they can be denatured by high temperature or pH as soon as the natural conditions are introduced, they fold into their nature form. Three-dimensional structures are determined by its sequence. Each protein has its own robust fold and this event is not coincidental. It is robust. The final folded structure is generally the one in which the free energy is minimized.

Many different weak non-covalent bonds, between different chains, force the folding of a protein chain [25]. Although these bonds are 30-300 times weaker than the typical covalent bonds that create biological molecules. Many weak bonds are able to act together to hold two different regions of a protein chain together. Therefore, the merged force of large numbers of these non-covalent bonds help to determine the stability of a structure. Because of all these interaction forces, each protein has a particular three-dimensional structure.

2.1.9 Levels of Protein Structure

There are four levels in the protein structure organization [26]; *Primary structure* is the first level of this organization. The amino acid sequence by connected peptide bonds is called the primary structure of a protein.

Secondary structure is the conformation of residues in localized regions of a polypeptide. By stabilizing folding patterns, hydrogen bonds play an important role in secondary structure. The two main and the most stable secondary structures are the alpha helix and the beta sheet. Both types are characterized by having the main chain amino and carboxyl groups participating in hydrogen bonds to each other.

Alpha helix has a clockwise spiral form in which each peptide bond is in the trans conformation. There are 3.6 residues in an alpha helix turn. The amino group places generally upward and parallel to the axis of the helix; inversely, the carboxyl group places downward as illustrated in Figure 2.4.

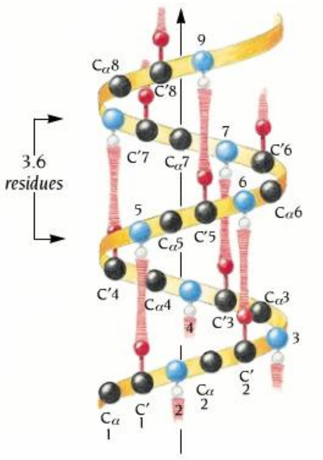


Figure 2.4 Forming of an Alpha Helix [27]

The beta sheet is the second major pattern in secondary structure, which consists of extended polypeptide chains with neighboring chains. It is stabilized by hydrogen bonds between the amino groups of one chain and the carboxyl groups of neighboring chain. The two strands, which form a beta sheet, can be either parallel (Figure 2.4 (a)), when successive strands have same biochemical direction, or anti-parallel (Figure 2.4 (b)), in the case of having opposite biochemical direction.

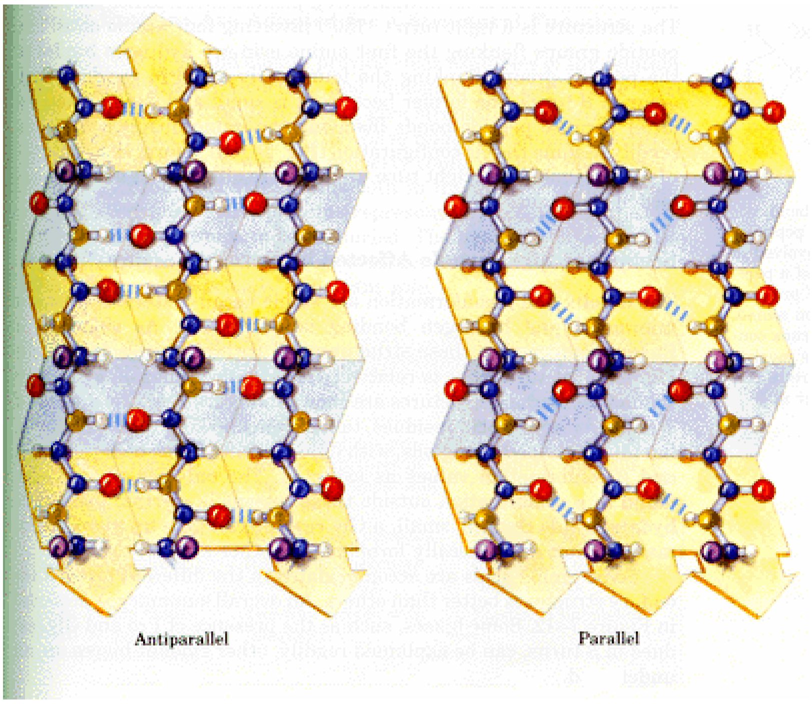


Figure 2.5 Beta Sheet Structure [28]

These patterns generate strict chains in proteins and this chain structure provides energy integrity. Each residue, which is in α helix or β sheet, is affected by this energy integrity, because neighbors and their properties, chemically, have effects on it. Thus, it may behave according to this evolutionary information, which is provided by neighbors [29, 39]. In order to use this evolutionary information, we have used α helix, β sheet or coil (neither α helix nor β sheet) information in our study. Secondary structure information of all the residues within the window are given as input.

Tertiary structure is the three-dimensional arrangement of the atoms within a single polypeptide chain. It is usually formed by disulfide bonds. When a polypeptide includes a single folding pattern (i.e. an alpha helix), the secondary and tertiary structure will be same. Similarly, when a protein is consisted single polypeptide molecule, tertiary structure and quaternary structure can be considered as the same.

Quaternary structure describes protein, which is composed of multiple polypeptides. Hydrophobic force is the main stabilizing force in this structure. When a single monomer folds into a three-dimensional shape to expose its polar side chains to an aqueous environment and to shield its nonpolar side chains, there are still some hydrophobic sections on the exposed surface. Two or more monomers will assemble so that their exposed hydrophobic sections are in contact.

2.1.10 Classification of Protein Structures

During evolution, a protein had evolved by folding up into a stable structure with useful properties, so its conformation could be mutated to make it possible for performing new functions. Genetic mechanisms have accelerated this process by producing duplicated copies of genes and by allowing one gene copy to evolve independently to perform a new function.

Such evolutions have occurred frequently in the past and because of this process, many of today’s proteins can be clustered into subgroups. Member of each subgroup has a sequence and a three-dimensional conformation that shows similarity with the members of the same subgroup. There are some kind of standards to group proteins such that CATH [30], SCOP [31] (Structural Classification of Proteins). In our study, we used SCOP database. SCOP clusters proteins into family, superfamily and fold subclusters. Similarity rises from folds through family.

2.2 Prediction of Contact Map and Disulfide Bond

2.2.1 Role of Contact Map and Disulfide Bond

A fundamental problem in molecular biology is the prediction of the three-dimensional structure of a protein from its sequence because of complexity of the task of searching possible conformations. Unfortunately, the experimental prediction of protein structure is time consuming and expensive. By using simple physical laws, machine-learning techniques have proven to be very useful for prediction of protein secondary structure from the amino acid sequence. They cannot manage to predict exact fold of a protein so far, but they achieve limited success. In order to improve structure prediction, some preliminary information such as contact map and disulfide map be used.

The contact map is a matrix, which has a binary format. Instead of the exact distances between residues, the contact map only contains ones for contacting interactions and zeroes for non-contacting interactions, respectively. Disulfide bond information includes the bond information of cysteine residues in protein sequence. Similar to contact map matrix, it has either zero (for non-contacting interactions) or one (for contacting interactions).

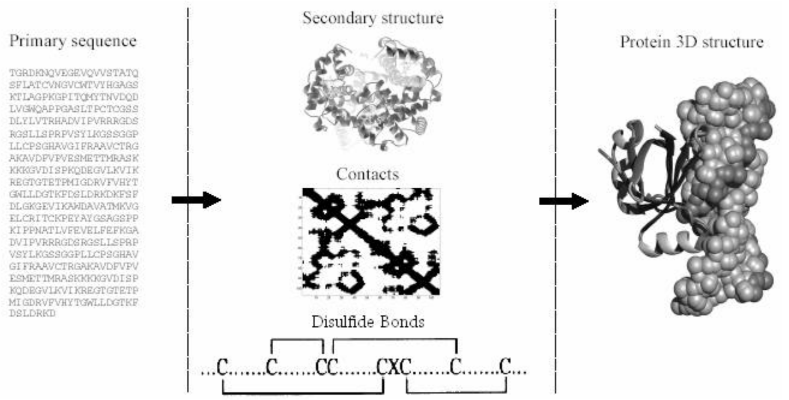


Figure 2.6 Steps of 3D Structure Prediction [32]

Proteins have very similar three-dimensional structure when they have homologous sequences or similar conserved regions. Therefore, a new sequence can be predicted by comparing known sequences. There are around 37 million reported protein sequences [33]. By comparing or using pattern recognition techniques, it is possible to predict unknown structures. In order to compare known and unknown structures, we use secondary structure, contact map and disulfide bond. So, as an intermediate step, contact map prediction and disulfide bond prediction are essential steps in the way of prediction of protein structure. For example, when contact map of an unknown protein would be predicted, its structure could be determined by using i.e. graph-matching algorithm [34]. Therefore, contact potential and disulfide bond of a protein is crucial for deriving constraints useful in modeling protein structure and protein folding.

2.2.2 Contact Map Prediction in Literature

Ying, Z. and Karypis, G. [35] present a contact-map prediction algorithm, which combine a set of features such as sequence profiles and conservation, physicochemical properties (i.e. hydrophobicity scale) and secondary structure (alpha helix and beta sheet), by using SVMs. They used three data set which is extracted from different families of CATH. Their predictor achieved a correctly predicted contact samples accuracy of 0.2238 by improving a random predictor of a factor 11.7.

In Akan, P. and Sezerman, U. [36] study, they tried to predict contact potentials of proteins with using NN. They used physical (volume), chemical (hydrophobicity, charge) and structural (secondary structure) characteristics of residues with the same sliding window approach of ours. In this study, a dataset, which was used by Casadio et al. [37], composed of 608 proteins is used. They correctly predict 11% of the contacting residues with a false positive ratio of 2%. This predictor performs 7 times better than a random predictor.

In Casadio et al. [37] study, they also tried to improve contact map prediction problem by implementing NN. In this study several numbers of network architecture is examined by using many different input vectors. As a result of these experiments, they saw that hydrophobicity and evolutionary information are the most useful characteristics of residues for this problem. The sliding window approach was also used in this study and presented as a useful technique for prediction performance. HSSP files [38] are used for sequence alignment encoding. The predictor is 6 times better than a random predictor.

2.2.3 Disulfide Bond Prediction in Literature

Martelli *et al.* [39] have published the best accuracy in disulfide prediction. They implement a new hybrid system that combines a NN and a hidden Markov model (HMM) by using 4136 containing cysteine residues, which extracted from 969 cysteine rich proteins. They have advantage both of local and global characteristics of the protein chains. A feed-forward NN captures local characteristics of protein chains with a sliding window. Output of the first stage is used in a four-state HMM as emission probabilities by defining global rules. By applying 20-fold cross-validation, obtained accuracies are 88% for cysteine basis study and 84% for protein basis study, respectively. These results are the best among previously described methods for prediction of disulfide bond task.

2.3 Methods

2.3.1 Neural Networks

Unlike traditional computing models, NN is a system modeled by the way biological nervous system, which has a structure and operation that resembles that of the mammal brain.

NNs are composed by a series of interconnected neurons that operate in parallel. These elements are called neurons. Similar to biological neurons, each neuron is linked to another neuron with connectivity weights that represent how strength this connection is. These links determine the flow of information between neurons. In Figure 2.2, the similarity between biological neuron and NN neuron is illustrated.

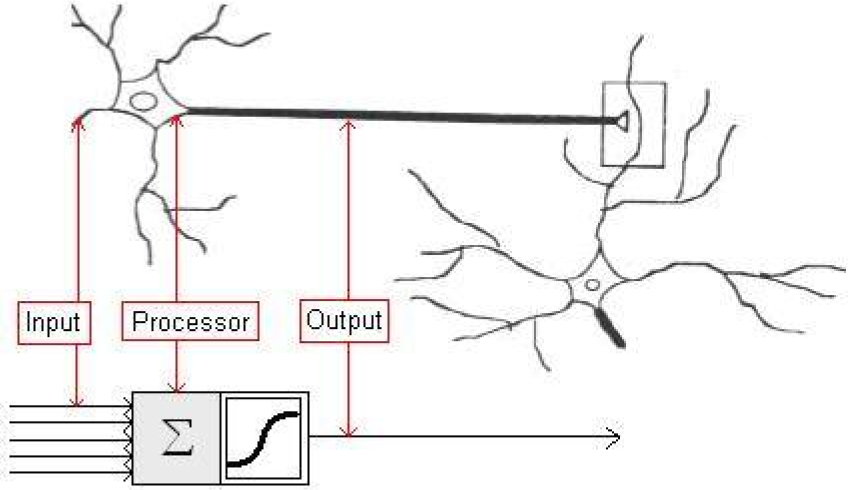


Figure 2.7 Structure of a Biological Neuron and NN Neuron.

Each neuron has an activation function, which is a simplistic representation and causes the signal integration and threshold firing behavior of it by means of mathematical equations [40].

Simply, the behavior of a single neuron can be determined as follows: First, the neuron collects the received signals from other interconnected neurons in the network by taking into account weight of each link. This signal is transmitted through a weighted connection, which is typically described as being analogous to a synapse. Second, it applies its activation function over this total signal to compute output signal. Third, it sends this output signal to other interconnected neurons in the network.

2.3.2 Neural Network Topology

The network is constructed using layers. The network requires at least two layers, an input layer and an output layer and possibly, it has one or more hidden layers. An example of a typical network is as follows:

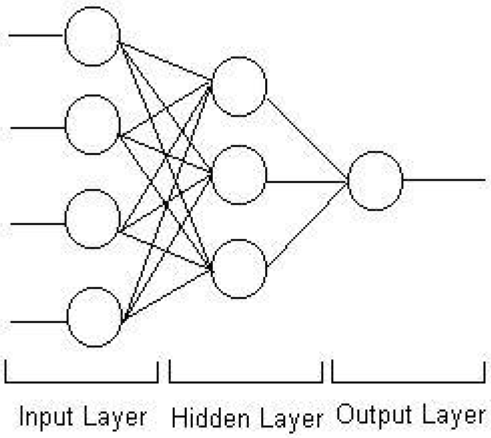


Figure 2.8 Topology of NN

2.3.3 Training

In biological systems, training involves adjustments to the synaptic connections that exist between the neurons. It is generated by adjusting these weights to reach the appropriate results for overall network.

NNs, like a human brain, learn by given examples. First, a network has been structured for a particular application, which is varying according to applications such as pattern recognition or data classification. Before this process, the weights are initialized randomly. Then, the training begins.

While training process, a set of samples, train set, is presented to the network. At the beginning of the training process, the network predicts the output for each example. However, as training goes on, the network updates strength of the connections between neurons, by using the following formula, until it reaches a stable stage at which prediction performance reaches a satisfactory level by taking into consideration the difference between actual and produced outputs, namely error criteria.

$$\Delta w_i^j = -\alpha . \frac{\partial E}{\partial w_i^j}(W)$$
$$w_i^{j,new} = w_i^j + \Delta w_i^j$$

where α is learning rate.

2.3.4 Testing

At the testing process, the network receives an input signal and produces an output signal. If the network trained correctly, generalization should be done. For this purpose, network can produce similar output with actual one, which is almost as good as the ones produced in the training stage for similar inputs.

2.3.5 Global and Local Minima of Energy Function

Mostly, training of an NN is based on numerical optimization of a usually nonlinear function. There is not the unique and the best method for nonlinear optimization for all cases. It is necessary to choose a method based on the characteristics of the problem, in hand. These methods find local optima in error surface such as in Figure 2.4.

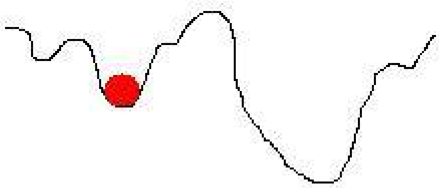


Figure 2.9 Local Minima of Error Function Surface

2.3.6 Error Function and Back-Propagated Value

The difference between the produced output and the desired output is determines error of the prediction. By the error function, this raw error is transformed to match particular network architecture. This error is used directly but other paradigms are used to modify this raw error to fit topologies' specific purposes.

$$E = \frac{1}{N}\sum_{t=1}^{N}(f(x_t, w) - y_t)^2$$

The error is propagated backwards to a previous layer. In order to update weights of connections before the next training cycle, back-propagated value is multiplied against each of the incoming connection weights.

2.3.7 Summation Function

The first step of the training process is to compute the weighted sum of all of the received inputs. When input vector is $(A_1, A_2, ..., A_n)$ and weight vector is $(w_1, w_2, ..., w_n)$, summation of the inner product of these two vector will be ;

$$\sum_{i=1}^{N} w_{ij} Ai + \theta_j$$

where θ_j is bais for connection.

By multiplying each component of the A vector by the corresponding component of the w vector and then adding up all the products, we compute weighted sum of inputs.

In addition to this method, the summation function can be depending on different algorithms such as the minimum, maximum, majority, product, or several normalizing algorithms. In this way, the input and weighting coefficients can be combined in many different ways before passing on to the transfer function. We pick a specific algorithm to combine inputs by considering the chosen network architecture.

2.3.8 Transfer Function

The result of the summation function is received by the neuron and inside of each neuron, there is a transfer function to transform the signal to a working output through an algorithmic process known as the transfer function or activation function. If *f* is the transfer function, A_j is the computed output for current neuron and the formula is as in the following,

$$A_j = f\left[\sum_{i=1}^{N} w_{ij} Ai + \theta_j\right]$$

This function is used to compare summation total with some threshold to generate output signal. If the sum is greater than the specified threshold value, a signal is generated. Otherwise, no signal is generated.

There are several different kinds of transfer functions, see Table 2.2 for sample transfer functions. The transfer function is generally non-linear. However, linear functions are limited, the output depends to the input. As investigated in the former researches [41], linear transformation functions are so strict that they are not very useful.

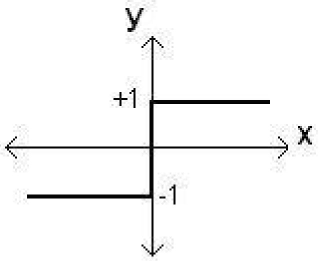
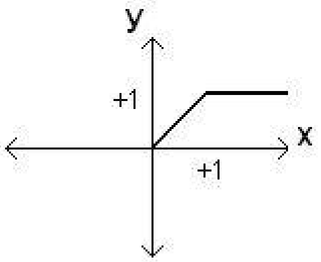
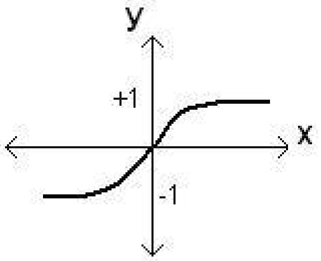
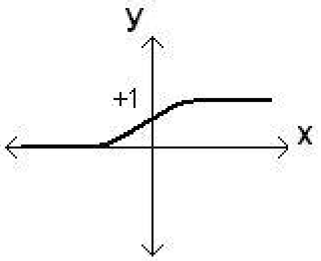
Transfer Function	x-y Graph	Formula
Hard Limiter		$x < 0, y = -1$ $x \ge 0, y = 1$
Ramping Function		$x < 0, y = 0$ $0 \le x \le 1, y = x$ $x > 1, y = 1$
Sigmoid Function 1		$x \ge 0, y = 1 - 1/(1 + x)$ $x < 0, y = -1 + 1/(1 - x)$
Sigmoid Function 2		$y = 1/(1 + e^{-x})$

Table 2.2 Sample Transfer Functions

The transfer function defines summation function either positive/one/one or negative/zero/minus one, respectively. "Hard limiter" transfer function can be used for such a desired response.

Another type of transfer function has a curve within a given range and still act as a hard limiter outside that range. However, outside of the range, it behaves as a linear function, inside of the range, as a non-linear function. That curve approaches a minimum and maximum value at the asymptotes.

Sigmoid is the most used transfer function between non-linear ones, because curve derivatives of sigmoid function are continuous. Thus, it works fairly well and is often preferable as transfer function. If it has a curve, it ranges between 0 and 1. When it ranges between -1 and 1, it has a hyperbolic tangent, respectively.

2.3.9 Output Function

Each neuron has inputs and produces an output. Generally, this output is produced by the transfer function. However, in some network topologies, neurons are allowed to compete with other neurons. In this purpose, the output is modified to include competition among interconnected neurons. This process may appear in two levels. In the first level, competition is used to determine the neuron, which will provide an output. In the second level, competitive inputs determine the neuron, which will participate in the training among all interconnected neurons.

2.3.10 A NNs Tool, EasyNN

In our experiments, we used EasyNN plus 4.0 tool to build NN by Stephen Wolstenholme. The release version of EasyNN can be downloaded from the following web site: http://www.easynn.com/.

2.4 Support Vector Machines

The Support Vector Machine (SVM) is a supervised training technique purposed by Vladimir Vapnik in 1979. It is designed for efficient multidimensional function approximation and for creating functions from a labeled training data. It nonlinearly maps N dimensional input space into a high dimensional feature space. In this high dimensional feature space, a linear classifier is constructed.

SVM is based on a training algorithm, which has some simple ideas and provides a clear intuition of what training from examples is about. It provides high performance in practical application with constructing models that are complex enough. It can be shown to correspond to a linear method in a high-dimensional feature space nonlinearly related to input space. However, it is easy to be analyzed mathematically.

SVM operates by finding a hypersurface in the space of possible inputs. This hypersurface divides input space into two or more subspace (depending to number of classes). The split will be chosen to have the largest distance from the hypersurface to the nearest of the positive and negative examples as illustrated in the Figure 2.5. Intuitively, this makes the classification correct for testing data that is near, but not identical to the training data. Thus, it prevents memorizing by maintaining generalization idea.

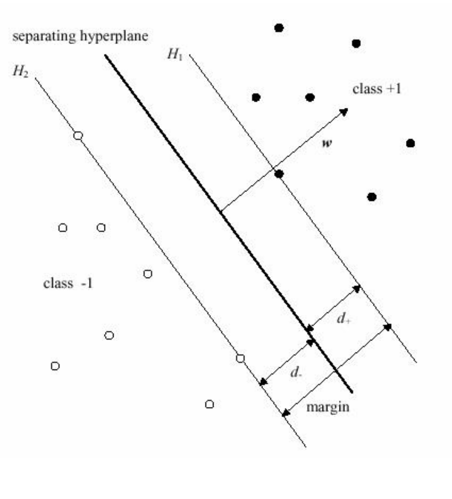


Figure 2.10 SVM Classification by Separating Hyperplane

2.4.1 SVM Hyperplane

For m-dimensional input vector $x = [x_1,...,x_m]^T \in X \subset R^m$, a one-dimensional output $y \in \{-1,1\}$ and label the training data $\{x_i, y_i\}$ where $i = 1,...,n$, suppose we have a hyperplane, which separates the positive from the negative examples. The hyperplane is designed performing a linear separation of the training data is described by

$$w^T x + b = 0 \qquad (1)$$

where $w = [w_1,...,w_m]^T, w \in W \subset R^m$. w is the normal to the hyperplane. In order to find a vector w and scalar b such that the points in each class are correctly classified and the following inequalities are satisfied:

$$w.x + b > 0 \text{ , for all i such that } y_i = 1$$

$$w.x + b < 0 \text{ , for all i such that } y_i = -1 \qquad (2)$$

The distance d between x_i and the hyperplane is

$$d(w,b;x_i) = \frac{| w^T x_i + b |}{|| w ||} \qquad (3)$$

2.4.2 SVM Training Rule

In SVM training, w_0, b_0 (2) is minimized. For such a problem Langrange multipliers is well suited for nonlinear constraints such as in (2). Thus, the Lagrangian is implemented

$$\mathrm{L(w,b,}\alpha) = \frac{1}{2} || w ||^2 - \sum_{i=1}^{n} \alpha_{\mathrm{i}} (y_i [w^T x_{\mathrm{i}} + b] - 1) \qquad (4)$$

where α_i are the Lagrange multipliers and $\alpha_i > 0$.

Here, (w_0, b_0) parameters specify the properties of the optimal hyperplane. From the given Lagrange multipliers, we can calculate the weight vector directly in terms of the training vectors. The training vectors are called *support vectors*.

2.4.3 Linear SVMs

2.4.3.1 Classification of Linearly Separable Data

A SVM can be defined as

$$f(x) = \operatorname{sgn}\{w^T x + b\} \tag{5}$$

where w, b are found from the training set. Hence, (5) may be written as

$$f(x) = \operatorname{sgn}\left\{\sum_{i \subset S} \alpha_{0i} y_i (x_i^T x) + b_0\right\} \tag{6}$$

where b_0 is found as

$$b_0 = \frac{1}{2}(w_0^T x_i^+ + w_0^T x_i^-) \tag{7}$$

where x_i^+ and x_i^- are any input training vector examples from two different classes.

2.4.3.2 Classification of Nonlinearly Separable Data

Here, the data is nonlinearly separable and we can extend the above approach to find a hyperplane, which minimizes the number of errors on the training set. For this purpose, we try to get

$$y_i\left[w^T x_i + b\right] \geq 1 - \xi_i \qquad (8)$$

where $\xi_i > 0$, $i = 1,...,n$.

2.4.4 Nonlinear SVMs

In most case, linear separation in input space is a too restrictive hypothesis to be of practical use. Fortunately, the theory can be extended to nonlinear separating surfaces by mapping the input points into feature points. The classifier is obtained by $\mathbf{x}_i^T\mathbf{x}$ where i ⊂ S. However, it is not necessary to use the input data to form the classifier. Instead, all that is needed is to use these inner products between the support vectors and the vectors of the feature space.

That is, by defining the kernel

$$K(x_i, x) = x_i^T x \qquad (9)$$

the non-linear classifier can be obtained as

$$f(x) = \text{sgn}\left\{\sum_{i \subset S} \alpha_{0i} y_i K(x_i, x) + b_0\right\} \qquad (10)$$

There are number of kernels that can be used in SVM models. Some of them is as

Kernel Function Type	$K(x_i, x_j)$
Linear	$x_i^T x_j$
Polynomial	$(\gamma < x_i, x_j > +b)^p$
Radial Basis	$\exp(-\gamma \|x_i - x_j\|^2)$
Sigmoid	$\tanh(\gamma < x_i, x_j > +b)$

Table 2.3 Kernel Functions

2.4.5 A SVM Tool BSVM

In order to implement SVM algorithm we used BSVM 2.05 by Chih-Wei Hsu and Chih-Jen Lin (2002). This is a freeware software for academic use and freely downloadable from the web site: http://www.csie.ntu.edu.tw/~cjlin/bsvm/. It is essentially used to solve binary classification problems.

The explanations of parameters in BSVM, which we optimized in our experiments, are in the following table.

-c cost	Set the parameter C of SVM (default 1)
-g gamma	Set gamma in kernel function (default 1/k)
-t kernel type	Set type of kernel function (default 2) 0 -- for linear kernel function 1 -- for polynomial kernel function 2 -- for radial basis kernel function 3 -- for sigmoid kernel function

Table 2.4 Used BSVM Parameters

2.4.6 Performance Evaluation Metrics

In prediction of contact map experiments, many different sets of information have been used. It is not possible to compare results and find the better one or more preferable one. Thus, a measurement is found to show how good the prediction is. By this method, it is possible to compare different results of different studies. While there are N-number of non-contacts, false positive (FP) ratio will be

$$FP = {N_c}/{N}$$

where N_c is the number of non-contact predicted as contact.Accuracy of the contact prediction is;

$$A = {C_c}/{C}$$

The number of residue pairs (N_P) is

$$N_P = (L-4)*(L-3)/2$$

By using N_P , we can calculate testing performance by the following formula.

$$A_r = {C}/{N_P}$$

Finally, improvement over a random predictor is

$$R = {A}/{A_r}$$

2.4.7 Source of Data

In both phases of the study, we used Protein Data Bank (PDB) [42]. PDB is an archive of experimentally determined three-dimensional structures of proteins, serving a global biology community of researchers, educators and students.

3 RESULTS AND DISCUSSIONS

3.1 Contact Map Prediction Study

In nature, proteins tend to have about 5 non-contacts for 1 contact. Thus, in the beginning, we collect data set by picking 5 non-contacts for 1 contact among whole residue interactions in a protein and we tried to respect this ratio in our experiments. However, for some experiments, we used different combinations for this ratio (i.e. 1 to 3) to be able to predict more contact samples. This approach will be called “contact / non-contact ratio” in the following parts.

In training, we used some chemical, physical and structural properties of not only contacting residues but also residues, which are neighbors of contacting residues. Therefore, information of both contacting residue and its environment will be captured. For this purpose, we generated a sliding window approach which slides on protein backbone. The contacting residues are located in the center of the windows.

According to some chemical and physical characteristics of residues such as polarity, charge and volume properties, 20 residues were clustered into 11 groups as shown in Table 3.1. However, if we used 20 residues one by one, training will be too specific. As a generalization and performance point of view, it is better to use smaller and compact feature set as much as possible. In addition, these clusters would make system learn how similar or different these residues are, which are from the same cluster.

Cluster #	Residue(s)
1	VAL,ILE,LEU,MET
2	TYR,PHE
3	GLN,ASN
4	GLU,ASP
5	TRP
6	CYS
7	SER,THR
8	ALA
9	GLY
10	LYS,HIS,ARG
11	PRO

Table 3.1 Cluster Information of Residues

As machine learning algorithm, two most popular and powerful method is used, SVM and NN. At the beginning, we built a feed-forward NN architecture, which uses sigmoid kernel function because it effectively finds the most stable structure given all the competing interactions within a protein of residues. It had three layers and 5 to 20 hidden nodes. As a binary classifier, SVM was used as well. According to the settings, SVM used either sigmoid or radial basis function as kernel function.

3.1.1 Experiment 1

In this experiment, we started with 7 residue wide-sliding window. The cluster information of contacting residue, which are located in the center of this window structure, was added to the feature set by using 11-digit vectors. Hydrophobicity of each residue in the window and average volume of three residues in the middle of the window were used in feature set as well. PDB codes of the proteins that we used in this experiment are 1bhg, 1dfx, 1ivt, 1l4i, 1obs and 2mcm. Data set was generated by picking 5 non-contact samples for 1 contact sample then we randomly selected two sets for both training and testing. In training set, there are around 12,400 residue interactions by including 10,000 non-contacting residue interactions and 2,400 contacting residue interactions. In test set, there are 1,200 residue interactions, which comprise 1,000 non-contacting residue interactions and 200 contacting residue interactions. We choose 10% of the train set as validation set.

In this experiment, we implemented NN Algorithm to predict contact potential of proteins. There were 20 hidden nodes in NN architecture. Learning rate (α) was 0.2 and momentum constant was 0.9. Learning rate and momentum values were optimized during training by the tool (EasyNN). At the end of the training, learning rate was 0.6 and momentum was 0.8. Sigmoid kernel function was picked as the transfer function of neurons in the network. Stopping criteria of training was either "Stop when the average error is less than 0.005" or "Stop when error for validation set starts to increase". If the output of the network is greater than 0.5, it is classified as contact, otherwise it is classified as non-contact, respectively. Therefore, decision threshold was 0.5. The result of this experiment is given in the Table 3.2.

	C_C	N_N
# of Occurrence	**22**	**937**
Accuracy	**11%**	**93.7%**
Overall Accuracy	**79.91%**	

Table 3.2 Results for Experiment 1

where C_C is the accuracy of correctly predicted contacting residue interactions and N_N is the accuracy of correctly predicted non-contacting residue interactions.

As discussion of this result, we may say either "good" or "bad". They are both true for some aspects which are explained in the following part.

The result of the experiment may look like poor and unsuccessful. However, this problem is not easy to solve. Generally, predictors tend to classify all contacting residue interactions as non-contacting because it is a big issue to learn contacting residue pattern. For such problems, it is fare enough to make a better prediction than a random predictor does. For this purpose, we tried to improve performance then compared it with a random predictor.

For such a hard problem, predicting 79.91% of test data seems to be successful. Nevertheless, the main contribution of this problem is to have a false positive accuracy, which must approximate to zero. This is the desired case and if predictor fails in non-contact prediction, this error will give more damage to 3D structure prediction than any error in contact prediction. Therefore, our primary goal is to have minimum false positive, then maximum correctly predicted contact accuracy.

3.1.2 Experiment 2

The result of the previous experiment showed us that with using the data set and architecture that are explained in the above part, we could not reach a good false positive ratio. There is too much error in non-contact prediction. That may because of trying to predict all of the residues together. Therefore, to generate less specific feature vector, we tried to predict contacting interaction of residues from Cluster1. The rest of the feature vector was the same with the feature vector in previous experiment. The same NN architecture with the experiment 1 was used. The training and testing sets were same as well, but we just took contacting interactions of residues from Cluster1. After processing, we got a prediction distribution in Figure 3.1.

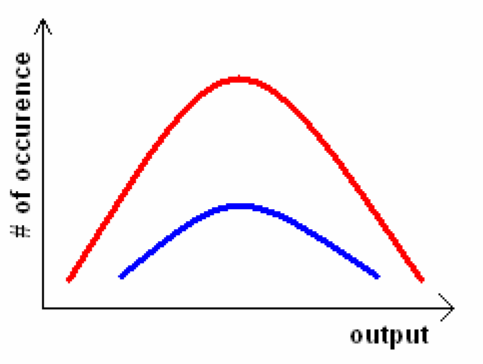


Figure 3.1 Distribution of the Prediction in Experiment 2

Our aim was to have a distribution as shown in the Figure 3.2 where red curve represents graph of non-contact class and blue curve represents contact class. When we compare the distribution of two classes, it is necessary to have divergent regions on both sides of these curves to be able to correctly classify some of contacting interactions. If we would have such a distribution in Figure 3.2, by looking at the produced output, which were on the left corner of graph, we could say that sample of this output were in the non-contact class. Respectively, if the produced output was in the right corner, the sample of this output was classified as contact. This was the ideal case but the result of this experiment was faraway than this shape (Figure 3.1). There is no divergent region on the curves to determine a classification threshold. Thus, we could not predict any of the contacting residue interactions.

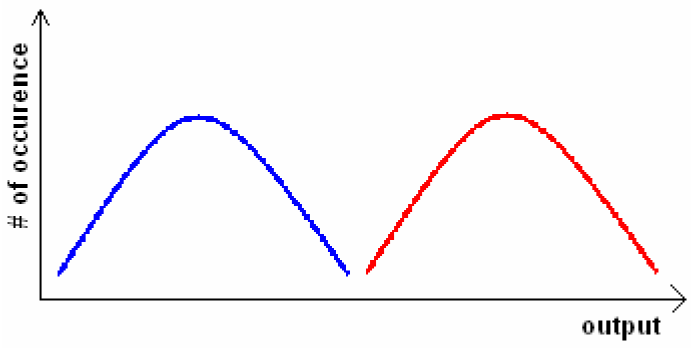


Figure 3.2 Desired Prediction Distribution

By comparing this experiment with the previous one, we can say that clustering information is an important feature in classification and it affects the classification performance by positively. By getting cluster information out, we lost advantage of using it. This indicates that cluster information is important to learn behavior of residues as a group, their similarity and their tendency to make a connection with a residue for example from the same cluster (or vice versa).

3.1.3 Experiment 3

In this experiment, we added the cluster information again to have advantage of hints that cluster information carries. In previous experiments, we mainly focused on environmental features such as using hydrophobicity of each residue in the windows. However, after that, we decided to use more information about contacting residues. Of course, neighboring information would be used as well, but it would be in a more compact manner. In order to use combination of hydrophobicity, we took average hydrophobicity of the three residues, which are located the center of the sliding windows. Cluster information of contacting residues, their hydrophobicity and volume, average hydrophobicity and average volume of three residues in the middle of the window were used in feature vector. The network architecture and data sets were same with the experiment 1.

In this study, all contacting interactions were predicted as non-contact because there was not any reasonable threshold in output distribution to separate contacting residues from non-contacting residues.

These unsuccessful results may depend on either feature vector, or classification technique or both. In this experiment, we might loose the advantage of using hydrophobicity of the residues by using its combination. Hydrophobicity scale of residues may be more useful when they are used individually. Namely, using average of the hydrophobicity may be not as effective as using hydrophobicity of neighboring residues. In addition, we may not need to use volume of the contacting residue when we take average of three middle residues. Therefore, in the next experiment, this feature is extracted from feature set. By this way, we may decrease the dimension of feature vector.

As another reason of these unsuccessful results, we gave too many non-contact samples that may cause just learning non-contact class instead of learning both contact and non-contact classes. Another reason may be that the classification technique or used kernel function may not be suitable for our problem.

3.1.4 Experiment 4

After previous two unsuccessful experiments, we changed the classification technique and feature set. First, instead of using NN algorithm, we applied SVM algorithm by using BSVM tool. This is an acceptable solution, because, in some cases, some algorithms can be more suitable than other ones. They use different techniques in training and according to our problem; these techniques can either be suitable or not. However, as kernel function, we still used sigmoid kernel function. The parameters of BSVM were set; gamma was 40 and cost was 100.

In this experiment, to decrease feature vector dimension, we used smaller sliding window by changing its size from 7 to 5. Cluster information was taken out from feature vector. Because we wanted to obtain a useful threshold to classify contact samples as well as non-contact samples and we tried to find more compact and more useful features. Therefore, we eliminate some of the features.

The normalized sequential distance between contacting residues was inserted to feature vector as a physical property. This feature is an important measure to give information about long or short distance contacting residues. In this way, predictor may learn long distance and short distance contact interactions and their properties separately.

Another difference from previous studies, as evolutionary information, secondary structures of contacting residues was added to the feature vector. This feature will give us a chance to evaluate each residue by its own structure. For example, predictor will be able to catch similar behaviors of residues, which are in alpha helix structure. By this way, predictor may extract secondary structure-specific properties. There is, of course, a relation between forces, which drive contact interactions and structural properties of proteins. Residues, which are in a more compact structure such as an alpha helix, will act together and combine their forces. Therefore, secondary structure will play a big role in contact interactions.

Nevertheless, some parts of feature vector were still kept the same such as hydrophobicity scale of contacting residues and the average volume of the contacting residues together with their first order neighbors.

In addition to 1:5 contact/non-contact ratio, 1:1 ratio were also used, because, in 1:5 ratio, there were too many non-contact residue interactions and contacting residue interactions cannot be analyzed and then classified correctly. By taking into consideration of those, we generated 5 different data sets;

1. 100 interactions from each protein with having 1:1 ratio
2. 200 interactions from each protein with having 1:1 ratio
3. 100 interactions from each protein with having 1:5 ratio
4. 200 interactions from each protein with having 1:5 ratio
5. All of the interactions from each protein with having 1:1 ratio.

10% of whole data set was used in testing. Result of this experiment is given the following table.

Data Set	Test Size	C_C %	N_N %	%
Set 1	60	20%	46%	33.3%
Set 2	120	55%	63%	59%
Set 3	72	9%	95%	48%
Set 4	36	17%	86%	75%
Set 5	232	57%	60%	58%

Table 3.3 Results for Experiment 4

By considering result of this experiment, we can say that, when we use bigger data sets in training, we achieve better prediction accuracies. That is why the performance of prediction in Set5 is better than the performance in Set1. Thus, it is better to use bigger train sets. Even contacting residue interactions were separated from non-contacting residue interactions, we got lower accuracies than the accuracy in experiment 1.

This experiment was also a test to see whether we could overcome threshold problem or not. We saw that, some of the contact interactions might be correctly predict by using SVM. Although, results are not sufficient, this experiment shows us that SVM can be more suitable than NN for this problem. Therefore, we started to use SVM technique.

3.1.5 Experiment 5

In order to get further improvement, we generated bigger train set with using different contacting/non-contacting ratios. Four different data sets were produced with 1:1, 1:2, 1:3 and 1:4 contact/non-contact ratios. Even though, there is a reasonable connection between this ratio and contact prediction performance, these ratios were generated to control the dependency to contact/non-contact ratio of prediction.

For each occurring N-M contact, we added 5 N-M non-contacts to the data set to allow the SVM to learn under what conditions N-M forms a contact and under what conditions they do not contact each other. For example, for Glutamine residue interactions of 23rd Alanine residue with 1:5 ratio, we picked 5 non-contacts and 1 contact information among 23rd Alanine and Glutamine residue interactions. The main point is; all the interactions were selected by using the same protein backbone. By this approach, predictor will catch any relation between a particular residue and a residue set from a particular type within the same protein backbone. This relation can be local as protein specific, or global as a general interaction behavior between these two residue-types. The schematic explanation of this approach is shown Figure 3.3.

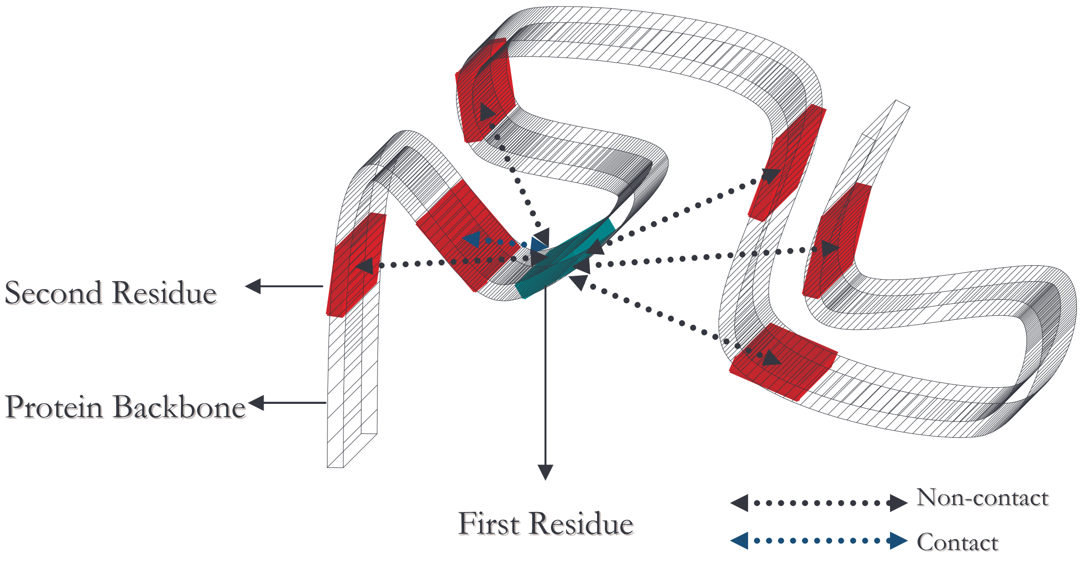


Figure 3.3 Representation of New Approach

After finding a better training algorithm, in order to see their effects on performance for current platform, we used some of previously used features, such as hydrophobicity scale of each residue, secondary structure information and volume scales of contacting residues and the normalized sequential distance between two contacting residues. The parameters of SVM were; Kernel function was sigmoid. Gamma coefficient was 40. Cost of train was 100. The result for this experiment is as following.

Ratio	Test size	C_C%	N_N%	%
1:1	1390	32%	51%	41.29%
1:2	1800	26%	53%	43.77%
1:3	2095	25%	56%	48.83%
1:4	2233	30%	82%	71.65%

Table 3.4 Results for Experiment 5

These results indicate that when we pick more non-contacts per one contact, predictor is getting better and starting to correctly predict more non-contact samples. Namely, false positive ratio reduces. By this aspect, 1:4 data set gave the best result for this experiment. Although this result is not perfect, in this step, we can use 1:4 data set to optimize parameters. This optimization makes us surer about the appropriateness of the parameters that we used in training. By this way, we may continue with using parameters that are more preferable.

3.1.6 Experiment 6

In this experiment, parameter optimization was done by just changing kernel function type, gamma constant and training cost parameters of BSVM. 1:4 data set taken from the previous experiment was used in optimization. Optimization results are as in the following Table 3.5.

Kernel (t)	Gamma (g)	Cost (c)	Accuracy
Sigmoid	400	100	70.8912%
Sigmoid	40	100	71.6525%
Sigmoid	4	100	71.2494%
Sigmoid	400	1	71.0255%
Sigmoid	40	10	71.7421%
Sigmoid	80	1	72.0555%
Sigmoid	70	1	72.1003%
Sigmoid	50	1	71.6525%
Sigmoid	40	1	72.0107%
Sigmoid	30	1	71.7868%
Sigmoid	4	1	71.4734%
Sigmoid	1	1	70.8464%
Radial Basis	0.1	100	76.8025%
Radial Basis	0.3	10	78.7730%
Radial Basis	0.2	10	79.0864%
Radial Basis	70	1	79.8925%
Radial Basis	0.9	1	80.8330%
Radial Basis	**0.1**	**10**	**82.4451%**
Radial Basis	0.1	1	81.5495%

Table 3.5 Results for Optimization Process

In these results, there is an important point; when we change kernel function type from sigmoid to radial basis, we got considerable improvement in prediction performance. When we used NN in our experiments, the kernel function was sigmoid and results of these experiments were unsuccessful. As we told before, some machine learning techniques are more suitable than some others. Likewise the above case, the most useful kernel function is changing according to the data set. In our case, data set has a semi positive defined matrix form. Sigmoid kernel function does not work properly for such matrices as experienced in our study and as known.

The combination of radial basis kernel function, 0.1 (gamma coefficient) and 10 (cost) gave the best accuracy for this set. Therefore, in next experiments, we used this parameters in training.

3.1.7 Experiment 7

By using the parameters that we find in previous experiment, we repeat the process for the data sets in experiment 5. A new data set with 1:5 ratio is used in this analysis as well. The result of the experiment is shown in the Table 3.6.

Ratio	Test Size	N_N %	C_C %	%
1:1	1390	70.64%	65.18%	67.9137%
1:2	1800	89.58%	49%	76.0556%
1:3	2095	88.65%	39.92%	76.42%
1:4	2233	**95.51%**	**26.28%**	**81.5943%**
1:5	2314	100%	0.389864%	77.9361%

Table 3.6 Results for Experiment 7

As you see, by using new parameters, we got big improvements in the prediction performance of 1:1, 1:2 and 1:3 data sets. Still the best accuracy among all results is obtained from the analysis of 1:4 data set. Predicting of 26.28% of contacting interactions is sufficient but false positive value (95.51%) is still high then desired.

3.1.8 Experiment 8

Before that, we generated test and train data by using whole proteins. After this point, some proteins were used only for testing and others were used only in training. Therefore, test data would be completely separate from train data. This approach was more appropriate for the case in real life. However, while predicting contact map of a protein, we would not know any of the contacting residues in this protein. For this purpose, among 16 proteins, we used 14 proteins in training and 2 proteins in testing.

In this experiment, we picked the proteins that are from the same superfamily, but from the different families. Furthermore, we would deduce bias of being in the same family. In this way, we would see how successfully we could predict contact potential of a protein, whose superfamily is known. For this purpose, we select 16 proteins from "Winged helix DNA-binding domain" SCOP superfamily. The PDB codes of these proteins are; 1bia, 1jhf, 1aoy, 1cgp, 1i1g, 1smt, 1mkm, 1lnw, 1ku9, 1hw1, 1bm9, 1ixc, 1b9m, 1i1s, 1hkq, 1in4. Obtained results are shown in the Table 3.7.

Ratio	Test Size	N_N%	C_C%	%
1:1	982	72.5051%	72.35%	72.4033%
1:2	1473	88.9002%	55.6911%	77.8004%
1:3	1965	97.0808%	35.1626%	81.6191%
1:4	2456	98.2688%	29.6748%	84.5621%
1:5	**2946**	**99.2671%**	**24.4399%**	**86.7957%**

Table 3.7 Results for Experiment 8

Our aim was to predict as many contacts as possible while keeping false positive value near to zero. For 1:5 data set, we got almost excellent performance. Because false positive is too low and correct contact ratio is high enough. Thus, we can say that we could successfully predict contact map of any protein from "Winged helix DNA-binding domain" SCOP superfamily. However, the point is which ratio will be more useful in a general manner. For this experiment, 1:5 ratio gave the best accuracy but to select more reliable and general ratio, we repeated this experiment for another SCOP superfamily in the next experiment.

3.1.9 Experiment 9

In this experiment, we used proteins from another SCOP superfamily class, “ARM repeat”. 12 proteins were picked from this superfamily (10 proteins for training, 2 proteins for testing). The PDB codes of these proteins are; 1jdh, 1gw5, 1b3u, 1oxj, 1h19, 1b89, 1c9l, 1lrv, 1e8z, 1m8z, 1ho8, 1n8v.

Feature set was the same with previous experiment and it was the combination of hydrophobicity scale of the residues in windows, secondary structure and volume scales of the contacting residues and the normalized sequential distance between two contacting residues. Results for this experiment are as in the Table 3.8.

Ratio	Test Size	N_N%	C_C%	%
1:1	3767	85.6081%	81.6879%	83.6431%
1:2	5651	83.4616%	42.9708%	83.7049%
1:3	7536	93.7887%	40.9019%	85.3769%
1:4	9419	94.8507%	18.4615%	85.9236%
1:5	**11298**	**99.9469%**	**6.47902%**	**84.3689%**

Table 3.8 Results for Experiment 9

As shown in the Table 3.8, 1:5 ratio, again, gave the best accuracy among all ratios. It gave a small false positive error, which is closest to zero for non-contact samples and a sufficient accuracy in contact prediction. Therefore, after that, we decided to collect data by using 1:5 ratio instead of other ratios. As we explained before, in nature, proteins tend to have an approximate 1:5 contact/non-contact ratio.

3.1.10 Experiment 10

In addition to the data sets in experiment 8 and 9, we repeated same study for 8 different kinds of SCOP superfamilies, which include 1 all alpha, 5 all beta and 2 alpha/beta superfamily (All alpha class proteins have alpha helix more than 15% and beta sheet less than 10%. All beta class proteins have alpha helix less than 15% and beta sheet more than 10%. Alpha & Beta class proteins have alpha helix more than 15% and beta sheet more than 10%), by using 1:5 ratio. Some of proteins were used only for testing and remaining part is used in training as written in Table 3.9. Detailed information about these super families is written in the following table.

Code	**Class**	**Superfamily Name**	**#of protein**
A1	All alpha	"Winged helix" DNA-binding domain	16 (2 for test)
A2	All alpha	ARM repeat	12 (2 for test)
A3	All alpha	EF-hand	9 (2 for test)
B1	All beta	E set domains	16 (2 for test)
B2	All beta	Galactose-binding domain-like	17 (2 for test)
B3	All beta (barrel)	Nucleic acid-binding proteins	10 (2 for test)
B4	All beta (barrel)	PH domain-like	6 (1 for test)
B5	All beta (barrel)	Composite domain of metallo-dependent hydrolases	7 (2 for test)
AB1	Alpha & Beta	Metallo-dependent hydrolases	13 (2 for test)
AB2	Alpha & Beta	P-loop containing nucleotide triphosphate hydrolases	20 (4 for test)

Table 3.9 Superfamily Information

In order to see whether results are good enough or not, we should compare results with random predictor by using random evaluation metrics.

Superfamily	Test Size	C%	N%	A%	FP	R
A1	2947	24.43%	99.26%	86.79%	0.0073	8.7496
A2	11299	6.479%	99.94%	84.36%	0.0005	4.3096
A3	2719	20.97%	99.24%	86.20%	0.0075	8.6667
B1	9007	0.066%	100%	83.34%	0	0.0473
B2	7951	0.075%	100%	83.34%	0	0.0445
B3	11599	9.415%	99.48%	84.47%	0.0052	8.8996
B4	217	2.777%	100%	83.79%	0	4.0532
B5	4567	18.00%	98.66%	85.21%	0.0134	20.8154
AB1	6559	19.48%	99.14%	85.86%	0.0086	10.0944
AB2	617	18.64%	98.67%	85.33%	0.0133	3.1756
Overall	**5748.2**	**12.03%**	**99.44%**	**84.86%**	**0.0058**	**6.88**

Table 3.10 Results for Experiment 10

By comparing our predictor with a random predictor, we have a chance to measure our prediction performance and the improvement that we provide for contact map prediction.

As shown in the Table 3.10, some of accuracies are good enough but some of them are not such as B1 and B2 data sets. These are all beta–class proteins and the long distance contacts of this class are difficult to predict, because, unlike all alpha or other classes, the distant fragments of the protein backbone make up these patterns. On the other hand, for example for B5 set, our predictor achieves 20 times better prediction than a random predictor. Having a better prediction than a random predictor is a considerable improvement, because this problem is hard to solve and has many possible solutions. Therefore, even small difference between our predictor and a random predictor indicates big improvement in prediction performance.

3.1.11 Experiment 11

In order to have further improvement, we tried to add useful information into feature set by using our experiences from previous experiments and our knowledge. Secondary structure caused improvement in performance because after adding secondary structure information to the feature vector, we got better accuracies. Therefore, in order to give more secondary structure information, we added secondary structure information of the residues in whole window to the feature vector.

Superfamily	Test Size	C%	N%	A%	FP	R
A1	2947	28.30%	98.81%	87.06%	0.0118	10.1349
A2	11299	0.11%	100%	83.35%	0	0.0706
A3	2719	19.20%	99.33%	85.98%	0.0066	7.9369
B1	9007	3.80%	99.50%	83.55%	0.0049	2.6942
B2	7951	0.91%	99.92%	83.42%	0.0008	0.5346
B3	11599	22.14%	96.96%	84.49%	0.0303	20.9287
B4	217	2.78%	100%	83.79%	0	4.0532
B5	4567	24.68%	97.70%	85.53%	0.0184	29.1719
AB1	6559	23.14%	98.70%	86.10%	0.013	11.99
AB2	617	36.14%	97.83%	87.55%	0.0217	6.1579
Overall	**5748.2**	**16.12%**	**98.875%**	**85.08%**	**0.0175**	**9.363**

Table 3.11 Results for Experiment 11

R determines how many times better performance we get than a random predictor does. For some superfamilies, we got sufficient performance (i.e. 29 times better prediction than a random predictor for BB3 superfamily set). Yet, for some superfamilies, we get insufficient prediction accuracy (i.e. B4 superfamily). However, for overall performance, adding secondary structure information of each residue in the windows makes performance better.

3.1.12 Experiment 12

In order to use more environmental information such as secondary structure, we increased window size to 7 and repeated process by using the same superfamilies and settings in experiment 11.

Superfamily	Test Size	C%	N%	A%	FP	R
A1	2947	26.48%	26.48%	86.64%	0.0134	9.4787
A2	11299	0.05%	0.05%	83.35%	0	0.0353
A3	2719	22.52%	22.52%	86.42%	0.0079	9.3054
B1	9007	2.40%	2.40%	83.48%	0.0031	1.7016
B2	7951	0.60%	0.60%	83.40%	0.0005	0.3564
B3	11599	21.31%	21.31%	84.25%	0.0317	20.1463
B4	217	2.78%	2.78%	83.80%	0	4.0532
B5	4567	22.22%	22.22%	85.29%	0.0209	25.2627
AB1	6559	23.88%	23.88%	86.37%	0.0113	12.3691
AB2	617	34.85%	34.85%	87.32%	0.022	5.937
Overall	**5748.2**	**15.71%**	**98.89%**	**85.03%**	**0.011**	**8.86**

Table 3.12 Results for Experiment 12

When we compare these results with results in experiment 10, as expected, we improved the performance (i.e. for B1 and B3 sets) as shown in the Table 3.14. Nevertheless, the improvement is not as big as the improvement that we got in previous experiment. This indicates that when a protein has many helix and sheet inside of the sequence, it can be predicted more accurate. However, when we started to use 7-residue wide sliding window, normally, dimension of the feature vector is increased and this may cause a confusion and noise in training.

3.2 Disulfide Bond Prediction Study

In this phase of the study, we tried to predict disulfide bond interactions. This study has not finished yet. However, the results and discussions about completed parts will be presented in this section.

As described before, disulfide bond is a kind of contact interaction between cysteine residues. Therefore, we used same cysteine rich proteins in [43] study. The list of the proteins and their chain information are written in Appendix. In this data set, there were 737 proteins with 624 cysteines forming disulfide bridges. PDB were used to extract secondary structure information, protein sequence and disulfide bond information.

The contact definition of disulfide bond is different from the residue contacting interaction definition that we used in the previous phase. Disulfide bond is a type of the bond, which is formed by sulfide atoms in cysteine residues. In more chemical point of view, cysteine atoms has the particular side chain which includes the thiol group (-SH) and oxidation of the thiol group yields a disulfide (S-S) bond. This is the reason of that disulfide bond is called also “SS bond”. These bonds are found experimentally and the SS bond information was taken from PDB files.

In training, we used cross validation by applying 5-fold cross validation because number of contacting interactions in train set was small. In addition to ROSEF hydrophobicity scales, we used two different hydrophobicity scales as well, Hopp-Woods and Eisenberg.

3.2.1 Experiment 13

As a starting point, we used simple features. Although, we know the important features; what they are and how they affect. However, the important issue is to find the most useful combination of them. For the purpose, as the first feature vector, we used only hydrophobicity and volume scales of the residues in windows. Window size was 5.

For this study, of course, contacting residues are always cysteine residues. Thus, we did not use the connecting cysteine residues' hydrophobicity and volume because they are always same. BSVM parameters are defined; gamma as 40, cost as1000, kernel type as sigmoid, validation fold as 5. For the set with ROSEF hydrophobicity, we obtained the results as follows;

Ratio	N_N %	C_C %	%
1:1	57.28%	39.04%	48.16%
1:2	99.84%	2.56%	67.41%
1:3	99.84%	2.88%	75.6%

Table 3.13 Results for Experiment 13

For the other two hydrophobicity scales, we got exactly same results. That means this classification is not meaningful. The error probably occurs because of using the sigmoid kernel function in training.

3.2.2 Experiment 14

After failure in the previous experiment, we changed kernel function to radial basis kernel function and repeated the same analysis in experiment 13. Results are as in the following table.

Ratio	Hydrophobicity Type	N_N %	C_C %	%
1:1	Eisenberg	53.6%	47.68%	50.64%
1:2	Eisenberg	77.47%	35.68%	62.06%
1:3	Eisenberg	74.24%	29.28%	63.00%
1:1	Hopp-Woods	51.84%	50.56%	51.2%
1:2	Hopp-Woods	66.00%	34.24%	55.41%
1:3	Hopp-Woods	75.36%	24.44%	62.88%
1:1	ROSEF	50.56%	49.92%	50.24%
1:2	ROSEF	65.36%	33.44%	54.72%
1:3	**ROSEF**	**75.68%**	**28.00%**	**63.76%**

Table 3.14 Results for Experiment 14

This time we get reliable results than previous results. Therefore, we fixed kernel function to radial basis. If we compare results, the set, which includes the ROSEF hydrophobicity and has 1:3 ratio, gives the best result among all data sets.

3.2.3 Experiment 15

The results of previous experiment were the starting point of this study, so we should try to find more useful information to improve the performance. For this purpose and using more environmental information, we increased the window size by setting it to 9. Feature vector was same by including volume and hydrophobicity scales of the neighboring residues of contacting residues.

Ratio	Hydrophobicity Type	N_N %	C_C %	%
1:1	Eisenberg	50.08%	46.08%	48.08%
1:2	Eisenberg	65.2%	22.56%	50.98%
1:3	Eisenberg	74.77%	19.2%	60.88%
1:1	Hopp-Woods	47.84%	49.6%	48.72%
1:2	Hopp-Woods	68.4%	32.48%	56.42%
1:3	Hopp-Woods	77.33%	21.12%	63.28%
1:1	ROSEF	50.08%	50.24%	50.16%
1:2	ROSEF	68.56%	30.24%	55.78%
1:3	ROSEF	76.69%	20.16%	62.56%

Table 3.15 Results for Experiment 15

This time, the data set including Hopp-Woods hydrophobicity scale and 1:3 ratio was predicted with the best accuracy among all data sets. Nevertheless, very similar results were obtained with experiment 2. Moreover, there is a decrease in performance.

3.2.4 Experiment 16

As we know, secondary structure information gives important clues about combined forces of residues. Therefore, in addition to volume and hydrophobicity scales, the secondary structure information of the neighboring residues was used in to feature vector. Window size was redefined as 5 because using bigger window caused a recession in performance.

Ratio	Hydrophobicity Type	N_N %	C_C %	%
1:1	Eisenberg	51.2%	45.92%	48.56%
1:2	Eisenberg	63.04%	39.52%	55.2%
1:3	Eisenberg	70.02%	27.84%	59.48%
1:1	Hopp-Woods	49.12%	45.76%	47.44%
1:2	Hopp-Woods	62.72%	34.08%	53.17%
1:3	Hopp-Woods	69.92%	27.36%	59.28%
1:1	ROSEF	51.36%	46.56%	48.96%
1:2	ROSEF	63.28%	36.32%	54.29%
1:3	ROSEF	69.38%	28.16%	59.08%

Table 3.16 Results for Experiment 16

This time, performance reduced. This might cause using small window because secondary structure gives a chance to use environmental information about residue contact interactions. Thus, in order to use more environmental information, window size was increased in the next experiment.

3.2.5 Experiment 17

Experiment 16 was repeated for 9-residue wide sliding window to have more neighboring residue information in training to use more environmental information.

Ratio	Hydrophobicity Type	N_N %	C_C %	%
1:1	Eisenberg	51.2%	55.84%	53.52%
1:2	Eisenberg	64.56%	36.8%	55.30%
1:3	Eisenberg	70.50%	31.84%	60.84%
1:1	Hopp-Woods	51.68%	52.48%	52.08%
1:2	Hopp-Woods	61.28%	38.24%	53.6%
1:3	Hopp-Woods	70.08%	25.28%	58.88%
1:1	ROSEF	49.12%	56.16%	52.64%
1:2	ROSEF	61.68%	37.76%	53.70%
1:3	ROSEF	73.06%	30.4%	62.45%

Table 3.17 Results for Experiment 17

Again overall accuracy was similar with the best results in experiment 2. While false positive accuracy was increasing, we got remarkable improvement in correct contact accuracy. This indicates that secondary structure information is not useful in disulfide bond prediction.

These results are as successful as the other studies in literature [13]. Because of the time constraint, we stopped here.

4 CONCLUSION

The main contribution of this study is the obligation of having small (even ~zero) false positive by predicting almost all of the non-contact samples correctly because the any error in non-contact prediction causes fatal error in 3D structure of protein. While keeping this balance, our secondary aim is to predict contact samples as well. The error in prediction of contact samples is harmless because it can be repaired by some methods such as graph matching. For this purpose, results are evaluated primarily false positive and then correct contact predictions. In order to overcome this difficulty, we followed an optimization strategy by changing all parameters with considering results of the experiments. First, we used feed-forward NN architecture in training. Although, NN is widely used and a powerful classification technique for most of the problems in bioinformation, in our case, the feed-forward NN was not good enough. This may occur because of using raw architecture of NN. After this phase, we started to use SVM in training. SVM is very powerful and the one of the most efficient method in many real-world applications and we may expect that SVM to become a standard tool for bioinformaticians. Results of our experiment showed us that SVM is a more useful classification technique than NN by transforming data into a high dimensional space.

As well as picking the right classification technique, choosing the most useful kernel function is an important issue for training performance. In our experiments, we always started with using sigmoid function as a kernel function. Nevertheless, we saw that any of the data set that we generated was not suitable to use in sigmoid kernel function. For such a complex and big feature vectors radial basis give more reliable and better results. When feed-forward NN structure was implemented, sigmoid kernel function was used. The problem in the experiment 1,2 and 3 most probably occurs because of using sigmoid function in feed-forward NN structure. As a future work, radial basis based NN structure may be used by using the same data sets in the successfully resulted experiments.

Another difficulty in this study is to find the best combination of features. In literature, there are many of the important properties of residues used. However, we have to know how we use to get better performance. Even if all the important features are added into the feature vector this may cause confusion in learning. Therefore, feature vector should be compact and it should carry useful information as much as possible. In order to find the best combination of features, we used our knowledge and experiences of the experiments in hand. Different features of a window of residue were considered in analysis such as hydrophobicity, volume, ordering distance and secondary structure information.

Our study showed that secondary structure information plays an important role in, especially, contact map prediction. Combined forces within a secondary structure pattern determine how a residue, which is in this pattern, is willing to make a contact with another residue. However, this is not true for disulfide bond prediction. Because disulfide bridge structure is different from a contact definition and as a covalent bond, disulfide bond probably is not too much affected by the environmental forces. Nevertheless, this is not true for disulfide bond prediction. These two problems, prediction of contact potential and prediction of disulfide bridge, are different from each other.

Contact potential prediction studies also suggest that predictor learned separately for different protein fold superfamilies may achieve better performance than a unified predictor. Thus, the contact potential of proteins having unknown fold based on known superfamily can be easily predict. We probably catch important structural properties by using the proteins from same superfamily in training and this gave important improvement in training performance.

For future improvements in contact map prediction, new classifier architectures can be built (such as presented in Pollastri *et al.* study [7]) by having same feature sets in the successful experiments of this study. In this way, we will use architecture that is more complex and after this step, we may change feature vector for further improvement. Some of the important features were not used in this study such as charge scale. These can be considered to use in feature set. In addition, in the first experiments, we used cluster information and these experiments showed that the cluster information causes improvement and it is useful information in prediction. Therefore, we may add cluster information of residues.

Our prediction of disulfide bridge study has not finished yet. For this problem, we need to find better feature set, which carries information about the driving forces for building SS Bridge because it is a chemical bond and this problem is different from a normal contact between residues. As a future work, these features may be explored more deeply. In addition, the dataset, in hand, may be growth. In this way, predictor may learn more information from more samples.

5 REFERENCES

1. Branden, C. and Tooze, J., Introduction to Protein Structure. Second Edition ed. New York: Garland Publishing,1999.

2. http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml.

3. Micheal, J.E., Protein Structure Prediction: Principles and Approaches.1996,New York: Oxford University Press. 1-26.

4. Vendruscolo, M., Kussell, E. and Domany, E., Recovery of protein structure from contact maps. Structure Fold. Des., 1997.2:p.941-948.

5. Göbel, U., Sander, C., Scheider, R. and Valencia, A., Correlated mutations and residue contacts in proteins. Proteins,1994.18:p.309-317.

6. Singer, MS., Vriend, G. and Bywater, R.P., Prediction of Protein Residue Contacts with a PDBderived Likelihood matrix. Protein Eng,2002.15:p.721-725.

7. Pollastri, G., Baldi, P., Fariselli, P. and Casadio, R., Improved Prediction of the Number of Residue Contacts in Proteins by Recurrent Neural Networks. Bioinformatics,2001. 17 Suppl. 1,:p.234-242.

8. Fariselli, P., Olmea, O., Valencia, A. and Casadio, R., Prediction of Contact Maps with Neural Networks and Correlated Mutations. Protein Eng, 2001. 14:p.835-843.

9. Fariselli, P., Olmea, O., Valencia, A. and Casadio, R., Progress in Predicting Inter-residue Contacts of Proteins with Neural Networks and Correlated Mutations. Proteins, 2001.5:p.157-162.

10. Thomas, D., Casari, G. and Sander, C., The prediction of protein contacts from multiple sequence alignments. Protein Engineering, 1996. 9(11):p. 941-948.

11. Fiser, A. and Simon, I., Predicting the oxidation state of cysteines by multiple sequence alignment. Bioinformatics, 2000. vol. 16, no. 3, p. 251-256.

12. Muskal, S.M., Holbrook, R.S. and Kim, S.H., Prediction of the disulfide-bond state of cysteine in proteins. Protein Eng., 1990. 3:p.667-672.

13. Fariselli, P., Riccobelli, P. and Casadio, R., Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins, 1999. 36:p.340-346.

14. Branden, C. and Tooze, J., Introduction to Protein Structure. New York: Garland Publishing, 1991. P:62-63.

15. Branden, C. and Tooze, J., Introduction to Protein Structure. New York: Garland Publishing, 1991. P.4.

16. Johnson, A., Raff, L., W. Roberts, Molecular Biology of The Cell. Fourth Edition ed. New York: Garland Publishing, 2002. P.54-55.

17. http://www.biochem.ucl.ac.uk/bsm/sidechains/.

18. Baysal, C. and Atilgan, A.R., Coordination topology and stability for the native and binding conformers of chymotrypsin inhibitor 2. Proteins, 2001. 45: p. 62-70.

19. http://pref.etfos.hr/scacor/.

20. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. and Zehfus, M.H., Hydrophobicity of Amino Acid Residue in Globular Proteins. Science, 1985. 229:p.834-838.

21. http://molvis.chem.indiana.edu/C687_S99/hydrophob_scale.html.

22. Abkevich, V.I. and Shakhnovich, E.I., What can disulfide bonds tell us about protein energetics, function and folding: Simulations and bioinformatics analysis. J. Mol. Biol., 2000. 300: p. 975-985.

23. Wedemeyer, W.J., Welker, E., Narayan, M. and Scheraga, H.A., Disulfide bonds and protein folding. Biochemistry, 2000. 39:p.4207-4215.

24. http://www.sbs.utexas.edu/genetics/Supplements/Ch9.htm.

25. Gromiha, M.M., Saraboji, K., Ahmad, S., Ponnuswamy, M.N. and Suwa, M., Role of non-covalent interactions for determining the folding rate of two-state proteins. Biophys Chem., 2004. 107(3):p. 263-72.

26. http://pps98.man.poznan.pl/ppscore/section7/interact.html.

27. Branden, C. and Tooze, J., Introduction to Protein Structure. New York: Garland Publishing, 1991. P.13.

28. http://www.agsci.ubc.ca/courses/fnh/301/protein/protprin.htm.

29. Alessandro, V. and Frasconi, P., Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics, 2004. 20(5):p.653-659.

30. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M., CATH- A Hierarchic Classification of Protein Domain Structures. Structure, 1997. Vol 5. No 8. p.1093-1108.

31. Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol., 1995. 247:p.536-540.

32. Jacobsson, A., Christian, G., Prediction of the Number of Residue Contacts in Proteins Using LSTM Neural Networks.Halmstad University, 2003. P.9.

33. ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

34. Akutsu, T., Protein structure alignment using a graph matching technique. Genome Informatics, 1995. 6:p.1-8.

35. Ying, Z. and Karypis, G., Prediction of Contact Maps Using Support Vector Machines. BIBE, 2003. P:26-.

36. Akan, P. and Sezerman, U., Computational Approaches To Understanding The Protein Structure.Sabancý University, 2002.

37. Fariselli, P. and Casadio, R., A Neural Network Based Predictor of Residue Contacts in Proteins. Protein Eng., 1999. 12: P. 15-21.

38. Sander, C. and Schneider, R., Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alingment. Proteins, 1991. 9: P. 56-68.

36. Martelli, P.L., Fariselli, P., Malaguti, L. and Casadio, R., Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy.Protein Science, 2002. 11:p.2735-2739.

37. Minsky M. and Papert, S., Perceptrons. MIT Press, 1969.

38. Vapnik, V., The Nature of Statistical Learning Theory. Springer, 1995, New York.

39. Bernstein, F., Koetzle, T.F., Williams, G.J., Meyer, E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., The protein data bank: a computer-based archival file for macromolecular structures. JMB, 1977. 112: p. 535-542.

40. Rost, B., Liu, J., Przybylski, D., Nair, R., Wrzeszczynski, K.O., Bigelow, H. and Ofran, Y.,Predicting protein structure and function through evolutionary information, Chemoinformatics, Wiley, 2002.

APPENDIX

PDBC	C	PDBC	C	PDBC	C	PDBC	C	PDBC	C	PDBC	C	PDBC	C	PDBC	C
10mh	A	1aq0	A	1bgv	A	1c7n	A	1d0n	A	1ds0	A	1ejf	A	1f2l	D
153l	_	1aqb	_	1bhe	_	1c8d	A	1d0q	A	1dsb	A	1ejj	A	1f2n	C
1a12	A	1arb	_	1bhg	A	1c8k	A	1d1z	A	1dtw	B	1ekj	G	1f2v	A
1a28	B	1aru	_	1bhh	A	1cc8	A	1d2k	A	1du3	G	1ekv	A	1f37	B
1a2w	A	1aso	A	1bhs	_	1ccz	A	1d2r	A	1dug	A	1el4	A	1f39	A
1a2z	_	1atl	A	1bht	A	1cdh	_	1d2v	C	1duj	A	1en7	A	1f52	A
1a44	_	1au1	A	1bhu	_	1cdq	_	1d3s	A	1duw	A	1enw	A	1f5m	B
1a4i	B	1aua	_	1bj7	_	1cem	_	1d4b	A	1dwm	A	1eo9	B	1f5s	A
1a4u	A	1aui	B	1bk5	A	1cew	I	1d4o	A	1dys	A	1ep0	B	1f5v	A
1a6d	A	1auz	_	1bkp	B	1cf2	0	1d6b	A	1dz3	A	1ep3	B	1f5x	A
1a7g	E	1avk	_	1bm0	A	1cfb	_	1d7c	A	1dz4	B	1ep9	A	1f5y	A
1a8e	_	1avp	_	1bn6	A	1cfe	_	1d7l	A	1dz7	A	1epf	B	1f6k	A
1a8h	_	1aw8	B	1boe	A	1cfr	_	1d7q	A	1dzf	A	1ept	C	1f7s	A
1a8l	_	1awe	_	1bou	B	1chc	_	1dbf	A	1e1a	A	1eqf	A	1f82	A
1a8p	_	1axn	_	1bov	A	1chd	_	1dbs	_	1e2t	A	1eqr	B	1f8m	A
1a8q	_	1ay2	_	1boy	_	1chm	A	1dce	B	1e3u	B	1erd	_	1f8v	A
1a99	A	1ayf	A	1bpo	B	1cid	_	1dci	A	1e4m	M	1erz	A	1faz	A
1aa7	A	1ayo	A	1bqv	_	1cjc	A	1ddb	A	1e4u	A	1esc	_	1fbr	_
1aaz	A	1az9	_	1br9	_	1cku	A	1ddl	A	1e5d	A	1esg	B	1fbx	A
1ad6	_	1b0p	A	1bs0	A	1cl7	L	1ddz	A	1e5l	A	1esl	_	1fc9	A
1ade	A	1b10	A	1bs2	A	1cle	A	1de3	A	1e5m	A	1ete	A	1fcd	A
1adn	_	1b1a	_	1bsl	B	1cli	A	1deo	A	1e5w	A	1etp	A	1fce	_
1ado	A	1b2p	A	1btn	_	1cmi	A	1dev	A	1e6u	A	1eua	A	1fd7	D
1aew	_	1b35	A	1bu7	A	1cmk	E	1dfx	_	1e6v	A	1euc	A	1ffy	A
1af7	_	1b3a	A	1buo	A	1cnz	A	1dgn	A	1e6y	E	1euh	A	1fgj	A
1afr	A	1b3r	A	1bvp	1	1co4	A	1dgs	A	1e8u	B	1euu	_	1fgp	_
1afw	B	1b3u	A	1bvz	A	1cp2	A	1dii	A	1eaj	A	1evx	A	1fgu	A
1ahj	A	1b4b	A	1bw3	_	1cpn	_	1dj0	A	1ecf	B	1ew4	A	1fi2	A
1ahk	_	1b5e	A	1by1	A	1cpo	_	1djn	A	1ecs	A	1ewi	A	1fiq	B
1ahl	_	1b5q	A	1by4	B	1cpq	_	1dk8	A	1ecy	_	1eww	A	1fj2	A
1ahs	A	1b71	A	1bya	_	1cq3	A	1dl6	A	1ed1	A	1ex1	A	1fjr	A
1aij	H	1b74	A	1byf	B	1cqx	A	1dli	A	1ed8	A	1ex2	A	1fl2	A
1air	_	1b8p	A	1bzh	A	1cqz	B	1dmr	_	1edg	_	1exg	_	1flk	A
1ajy	A	1b8t	A	1bzy	A	1cs6	A	1dor	A	1edq	A	1exk	A	1fn9	A
1ajz	_	1b9h	A	1c01	A	1css	_	1dov	A	1ee6	A	1ext	A	1fnc	_
1ako	_	1b9w	A	1c1k	A	1cuo	A	1dp0	A	1ee8	A	1eyq	A	1fo1	B
1alv	A	1b9y	A	1c1z	A	1cvr	A	1dp4	C	1eej	A	1eys	C	1fo5	A
1amm	_	1bbi	_	1c3q	A	1cw5	A	1dq3	A	1efv	A	1ezg	A	1foa	A
1amp	_	1bea	_	1c3y	A	1cwv	A	1dqb	A	1eg7	A	1ezw	A	1fod	1
1amy	_	1beb	A	1c4z	A	1cx1	A	1dqe	A	1eg9	A	1f00	I	1fp0	A
1aoc	A	1bet	_	1c52	_	1cz1	A	1dqg	A	1ehk	B	1f08	A	1fp2	A
1apj	_	1bf2	_	1c5m	D	1czf	A	1dqq	A	1ei9	A	1f16	A	1fp3	A
1apq	_	1bfd	_	1c75	A	1czp	A	1dqt	A	1eix	C	1f2d	A	1fps	_
1apy	A	1bfs	_	1c7k	A	1czt	A	1dr9	A	1ej2	A	1f2h	A	1fro	A

Table A. Cysteine Rich Proteins (1)

* PDBC: PDB code, C: Chain type

PDBC	C	PDBC	C	PDBC	C	PDBC	C	PDBC	C	PDBC	C	PDBC	C	PDBC	C
1fs7	A	1gat	A	1i0d	A	1lam	_	1pmp	A	1qr2	A	1thv	_	2bnh	_
1fsz	_	1gcb	_	1i0r	B	1lbe	A	1pnb	B	1qrj	B	1tib	_	2bpa	1
1ft1	A	1gcu	A	1i17	A	1leh	A	1png	_	1qsa	A	1tii	D	2cmd	_
1ft5	A	1gd0	A	1i1i	P	1lki	_	1poa	_	1qst	A	1tki	A	2cpl	_
1ftr	A	1gd5	A	1i39	A	1lmk	A	1poi	A	1qsv	A	1tlf	A	2ctb	_
1fua	_	1gen	_	1i3j	A	1lrv	_	1ppn	_	1qtr	A	1tlk	_	2dkb	_
1fui	A	1gg6	B	1i50	C	1mhl	C	1prt	B	1qts	A	1tpf	A	2dln	_
1fup	A	1gh9	A	1i5p	A	1mho	_	1psr	B	1qtw	A	1tpm	_	2eia	A
1fva	B	1gia	_	1i71	A	1mka	A	1ptq	_	1qu1	F	1trk	A	2ezm	_
1fvl	_	1gnc	_	1i7q	B	1mkn	A	1pud	_	1qu5	A	1ttq	B	2fcb	A
1fwl	A	1gnd	_	1iab	_	1mla	_	1pvc	1	1qu6	A	1tul	_	2fcp	A
1fwq	A	1gof	_	1iat	A	1mml	_	1qaz	A	1quu	A	1tvs	_	2fdn	_
1fxj	A	1gpc	_	1ib2	A	1msk	_	1qb0	A	1qvb	A	1tyf	A	2gf1	_
1fxr	A	1gpe	A	1ice	_	1mty	D	1qba	_	1rbl	A	1udh	_	2gmf	A
1fyb	A	1h2r	L	1icj	A	1mug	A	1qcc	A	1rcb	_	1uok	_	2hgs	A
1fzc	B	1h5q	A	1ig0	B	1muy	A	1qck	A	1rgf	A	1uro	A	2hpa	A
1fzd	A	1h7w	D	1ig8	A	1mwp	A	1qcx	A	1rkm	_	1vca	A	2hrv	A
1fzq	A	1h8u	A	1ilr	2	1nba	A	1qd1	B	1rla	A	1vhh	_	2hvm	_
1g0h	A	1h8v	A	1im3	A	1nfa	_	1qdp	_	1rmd	_	1vhi	A	2i1b	_
1g12	A	1hbk	A	1iml	_	1ngl	A	1qex	A	1rmg	_	1vhr	A	2if1	_
1g1b	A	1hcz	_	1in1	A	1ngr	_	1qfs	A	1rpl	_	1vid	_	2jhb	A
1g40	A	1he7	A	1iq3	A	1nkr	_	1qft	A	1rpx	A	1vmo	A	2kau	C
1g5b	B	1het	A	1isu	A	1nmt	A	1qg3	B	1rth	A	1vpn	B	2mcm	_
1g5c	A	1hf8	A	1ixx	A	1nse	A	1qgi	A	1rtm	1	1vsg	A	2mhr	_
1g5t	A	1hfe	L	1jb0	F	1nsf	_	1qgj	A	1rzl	_	1vsr	A	2min	B
1g5v	A	1hfh	_	1jb3	A	1obw	A	1qgk	A	1sac	A	1wab	_	2mnr	_
1g61	A	1hgf	A	1jc5	B	1onr	A	1qgo	A	1sft	A	1waj	_	2ms2	A
1g63	B	1hh7	C	1jdb	F	1opm	A	1qh4	A	1skf	_	1wdc	C	2mss	A
1g66	A	1hhs	A	1jdw	_	1orb	_	1qh5	A	1sll	_	1wer	_	2nac	A
1g6e	A	1hi7	A	1jf9	A	1ord	A	1qhd	A	1smd	_	1wjb	A	2paw	_
1g6n	B	1hjc	A	1jfr	A	1oro	A	1qhv	A	1sml	A	1xik	B	2pfl	A
1g6s	A	1hjr	A	1jhb	_	1oun	A	1qhw	A	1sqc	_	1xpa	_	2pgd	_
1g71	A	1hp4	A	1jj2	2	1pam	A	1qi9	A	1sra	_	1xva	A	2pia	_
1g72	A	1hr6	A	1jkm	B	1pbn	_	1qjd	A	1sry	A	1xwl	_	2pol	A
1g73	D	1hre	_	1jly	A	1pbv	_	1qjv	A	1stm	A	1yac	A	2psp	A
1g7o	A	1hsb	A	1joe	A	1pbw	A	1qk8	A	1svb	_	1yge	_	2pva	A
1g8k	B	1hsk	A	1jvr	_	1pce	_	1qks	A	1tap	_	1zin	_	2rel	_
1g8l	A	1htr	A	1kjs	_	1pcn	_	1ql0	A	1tbc	_	1zpd	A	2rgf	_
1g8q	A	1hul	_	1klo	_	1pcz	A	1qmu	A	1tbd	_	2a39	A	2rn2	_
1g93	A	1hux	A	1koe	_	1phn	A	1qnf	_	1tca	_	2aai	B	2scu	B
1g96	A	1hw7	A	1kp6	A	1phr	_	1qnr	A	1tcr	B	2alc	A	2sil	_
1g99	A	1hyj	A	1kpf	_	1pkm	_	1qnx	A	1tf4	A	2baa	_	2sn3	_
1ga3	A	1hyn	Q	1ksa	A	1plq	_	1qqf	A	1tfe	_	2bbk	L	2tgi	_
1gak	A	1hzt	A	1kve	A	1pmi	_	1qqj	A	1tfi	_	2bid	A	2tnf	A

Table B. Cysteine Rich Proteins (2)

* PDBC: PDB code, C: Chain type

PDBC	**C**
2tpt	_
2trx	B
2utg	_
2vpf	H
2vsg	A
3cyr	_
3daa	A
3ebx	_
3ezm	A
3grs	_
3hsc	_
3lzm	_
3mdd	A
3msp	A
3pah	_
3pmg	A
3pte	_
3rub	L
3ssi	_
4fgf	_
4lzt	_
4sbv	C
4wbc	A
5eat	_
5pti	_
6at1	B
6taa	_
7fd1	A
7rsa	_
7yas	A

Table C. Cysteine Rich Proteins (3)

* PDBC: PDB code, C: Chain type

5 REFERENCES

1. Branden, C. and Tooze, J., Introduction to Protein Structure. Second Edition ed. New York: Garland Publishing,1999.

2. http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml.

3. Micheal, J.E., Protein Structure Prediction: Principles and Approaches.1996,New York: Oxford University Press. 1-26.

4. Vendruscolo, M., Kussell, E. and Domany, E., Recovery of protein structure from contact maps. Structure Fold. Des., 1997.2:p.941-948.

5. Göbel, U., Sander, C., Scheider, R. and Valencia, A., Correlated mutations and residue contacts in proteins. Proteins,1994.18:p.309-317.

6. Singer, MS., Vriend, G. and Bywater, R.P., Prediction of Protein Residue Contacts with a PDBderived Likelihood matrix. Protein Eng,2002.15:p.721-725.

7. Pollastri, G., Baldi, P., Fariselli, P. and Casadio, R., Improved Prediction of the Number of Residue Contacts in Proteins by Recurrent Neural Networks. Bioinformatics,2001. 17 Suppl. 1,:p.234-242.

8. Fariselli, P., Olmea, O., Valencia, A. and Casadio, R., Prediction of Contact Maps with Neural Networks and Correlated Mutations. Protein Eng, 2001. 14:p.835-843.

9. Fariselli, P., Olmea, O., Valencia, A. and Casadio, R., Progress in Predicting Inter-residue Contacts of Proteins with Neural Networks and Correlated Mutations. Proteins, 2001.5:p.157-162.

10. Thomas, D., Casari, G. and Sander, C., The prediction of protein contacts from multiple sequence alignments. Protein Engineering, 1996. 9(11):p. 941-948.

11. Fiser, A. and Simon, I., Predicting the oxidation state of cysteines by multiple sequence alignment. Bioinformatics, 2000. vol. 16, no. 3, p. 251-256.

12. Muskal, S.M., Holbrook, R.S. and Kim, S.H., Prediction of the disulfide-bond state of cysteine in proteins. Protein Eng., 1990. 3:p.667-672.

13. Fariselli, P., Riccobelli, P. and Casadio, R., Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins, 1999. 36:p.340-346.

14. Branden, C. and Tooze, J., Introduction to Protein Structure. New York: Garland Publishing, 1991. P:62-63.

15. Branden, C. and Tooze, J., Introduction to Protein Structure. New York: Garland Publishing, 1991. P.4.

16. Johnson, A., Raff, L., W. Roberts, Molecular Biology of The Cell. Fourth Edition ed. New York: Garland Publishing, 2002. P.54-55.

17. http://www.biochem.ucl.ac.uk/bsm/sidechains/.

18. Baysal, C. and Atilgan, A.R., Coordination topology and stability for the native and binding conformers of chymotrypsin inhibitor 2. Proteins, 2001. 45: p. 62-70.

19. http://pref.etfos.hr/scacor/.

20. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. and Zehfus, M.H., Hydrophobicity of Amino Acid Residue in Globular Proteins. Science, 1985. 229:p.834-838.

21. http://molvis.chem.indiana.edu/C687_S99/hydrophob_scale.html.

22. Abkevich, V.I. and Shakhnovich, E.I., What can disulfide bonds tell us about protein energetics, function and folding: Simulations and bioinformatics analysis. J. Mol. Biol., 2000. 300: p. 975-985.

23. Wedemeyer, W.J., Welker, E., Narayan, M. and Scheraga, H.A., Disulfide bonds and protein folding. Biochemistry, 2000. 39:p.4207-4215.

24. http://www.sbs.utexas.edu/genetics/Supplements/Ch9.htm.

25. Gromiha, M.M., Saraboji, K., Ahmad, S., Ponnuswamy, M.N. and Suwa, M., Role of non-covalent interactions for determining the folding rate of two-state proteins. Biophys Chem., 2004. 107(3):p. 263-72.

26. http://pps98.man.poznan.pl/ppscore/section7/interact.html.

27. Branden, C. and Tooze, J., Introduction to Protein Structure. New York: Garland Publishing, 1991. P.13.

28. http://www.agsci.ubc.ca/courses/fnh/301/protein/protprin.htm.

29. Alessandro, V. and Frasconi, P., Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics, 2004. 20(5):p.653-659.

30. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M., CATH- A Hierarchic Classification of Protein Domain Structures. Structure, 1997. Vol 5. No 8. p.1093-1108.

31. Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol., 1995. 247:p.536-540.

32. Jacobsson, A., Christian, G., Prediction of the Number of Residue Contacts in Proteins Using LSTM Neural Networks.Halmstad University, 2003. P.9.

33. ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

34. Akutsu, T., Protein structure alignment using a graph matching technique. Genome Informatics, 1995. 6:p.1-8.

35. Ying, Z. and Karypis, G., Prediction of Contact Maps Using Support Vector Machines. BIBE, 2003. P:26-.

36. Akan, P. and Sezerman, U., Computational Approaches To Understanding The Protein Structure.Sabancý University, 2002.

37. Fariselli, P. and Casadio, R., A Neural Network Based Predictor of Residue Contacts in Proteins. Protein Eng., 1999. 12: P. 15-21.

38. Sander, C. and Schneider, R., Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alingment. Proteins, 1991. 9: P. 56-68.

36. Martelli, P.L., Fariselli, P., Malaguti, L. and Casadio, R., Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy.Protein Science, 2002. 11:p.2735-2739.

37. Minsky M. and Papert, S., Perceptrons. MIT Press, 1969.

38. Vapnik, V., The Nature of Statistical Learning Theory. Springer, 1995, New York.

39. Bernstein, F., Koetzle, T.F., Williams, G.J., Meyer, E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., The protein data bank: a computer-based archival file for macromolecular structures. JMB, 1977. 112: p. 535-542.

40. Rost, B., Liu, J., Przybylski, D., Nair, R., Wrzeszczynski, K.O., Bigelow, H. and Ofran, Y.,Predicting protein structure and function through evolutionary information, Chemoinformatics, Wiley, 2002.