ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCE

MASTER THESIS

İrem ERSÖZ

Secondary Structure Prediction Of Hemoglobin

By Using Combined Neural Networks

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

ADANA, 2003

ÇUKUROVA ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

SECONDARY STRUCTURE PREDICTION OF HEMOGLOBIN BY USING COMBINED NEURAL NETWORKS

İrem ERSÖZ

YÜKSEK LİSANS TEZİ

ELEKTRİK-ELEKTRONİK ANABİLİM DALI

Bu tez 14/08/2003 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından Oybirliği/~~Oyçokluğu~~ İle Kabul Edilmiştir.

Yrd. Doç. Dr. Turgay İBRİKÇİ	Prof. Dr. Seyhan TÜKEL	~~Yrd.~~ Doç.Dr. Ulus ÇEVİK
DANIŞMAN	ÜYE	ÜYE

Bu tez Enstitümüz Elektrik-Elektronik Anabilim Dalında hazırlanmıştır.

Kod No 2196

Prof. Dr Fikri AKDENİZ
Enstitü Müdürü

Bu Çalışma Çukurova Üniversitesi Bilimsel Araştırma Projeleri Fonu ve kısmen TÜBİTAK-EEEAG Tarafından Desteklenmiştir.

Proje No: FBE2002YL328 / TÜBİTAK-EEEAG100E037

CONTENTS

ÖZ I
ABSTRACT II
ACKNOWLEDGEMENTS III
NOTATIONS IV
LIST of FIGURES V
LIST of TABLES VIII

1. INTRODUCTION
    1.1. Proteins 1
    1.2. Hemoglobin 5
    1.3. Artificial Neural Networks 7
2. RELATED WORKS 10
3. MATERIAL and METHODS
    3.1. Material
        3.1.1. Data Collection and Preprocessing 13
    3.2. Methods
        3.2.1. General Regression Neural Networks (GRNN) 15
        3.2.2. Probabilistic Neural Networks (PNN) 19
        3.2.3. Backpropagation (BP) 22
        3.2.4. Combined Neural Networks (CNN) 25
4. RESULTS and DISCUSSION 27
5. CONCLUSION 50

REFERANCES 51
BIOGRAPHY 58
APPENDIX 59

ÖZ

YÜKSEK LİSANS TEZİ

BİRLEŞTİRİLMİŞ YAPAY SİNİR AĞLARI KULLANILARAK HEMOGLOBİNİN İKİNCİL YAPISININ TAHMİN EDİLMESİ

İrem ERSÖZ

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK ANA BİLİM DALI

Danışman : Yard. Doç. Dr. Turgay İBRİKÇİ
Yıl: 2003 Sayfa: 60
Jüri : Yard. Doç. Dr. Turgay İBRİKÇİ
Prof. Dr. Seyhan TÜKEL
Yard. Doç. Dr. Ulus ÇEVİK

Proteinler, hayati önem taşıyan görevleri nedeniyle organizmaların en önemli kısımlarıdır. Bu bakımdan, bir organizmanın yaşam sürecini anlayabilmek için ilk olarak, fonksiyonuyla yakından ilgili olan yapısının bilinmesi gerekmektedir. Protein yapıları, Birincil, İkincil, Üçüncül, Dördüncül olmak üzere dört ana hiyerarşik düzeye sahiptirler. Birincil yapı amino asit dizilimidir. Polipeptit zincirlerinin bölgesel şekillenmeleri, ikincil yapıyı oluşturur, üçüncül yapı ise son şeklin oluşumunda ikincil yapıların ne şekilde bir araya gelmeleri gerektiğini belirler. Bir veya daha fazla polipeptit zincir arasındaki etkileşimler de dördüncül yapıyı oluşturur. Yapay sinir ağları, proteinlerin ikincil yapılarının tahmininde başarıyla kullanılabilen algoritmalardandır.

Bu tezde, genel regresyon sinir ağları (GRNN), probabilistik sinir ağları (PNN) ve geri yayılım algoritması (BP) hemoglobinin, amino asit dizilişlerinin farklı pencere boyutları elde edilmiş birincil yapısına uygulanarak ikincil yapısı tahmin edilmeye çalışılmıştır. Daha sonra tüm ağların sonuçları, GRNN ile birleştirilmiştir. 20 alfa-141 ve 20 beta-146 hemoglobin zincirleri içeren veri seti, Protein Veri Bankasından oluşturulmuştur. GRNN'nin genel başarı oranı, beta zincirleri için 90.2-91.7% ve alfa zincirleri için 85.9-87.3% dir. PNN, beta için 91.2-92%, alfa için 85.4-86.5% genel başarıya ulaşmıştır. BP'nin genel başarısı ise beta için 89.9-92.5% ve alfa için 86.5-90.3% dir. CNN sonuçlarına göre tahmin başarısında önemli bir artış olmamıştır.

Anahtar Kelimeler: Genel Regresyon Sinir Ağları, Probabilistik Sinir Ağları, Geri Yayılım Algoritması, Birleştirilmiş Sinir Ağları, Hemoglobin.

ABSTRACT

MSc THESIS

SECONDARY STRUCTURE PREDICTION OF HEMOGLOBIN BY USING COMBINED NEURAL NETWORKS

İrem ERSÖZ

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING
INSTITUTE OF NATURAL AND APPLIED SCIENCES
UNIVERSITY OF ÇUKUROVA

Supervisor : Asst. Prof. Dr. Turgay İBRİKÇİ
Year: 2003 Pages: 60
Jury : Asst. Prof. Dr. Turgay İBRİKÇİ
Prof. Dr. Seyhan TÜKEL
Asst. Prof. Dr. Ulus ÇEVİK

Proteins are one of the most important parts of an organism because of its vital important tasks. In this respect, to understand the life process of an organism, it is necessary to first know the protein's structure that is closely related to its function. Protein structures are described through four main hierarchical levels; Primary, Secondary, Tertiary, Quaternary. The primary structure is simply the sequence of amino acids, the secondary structure refers to the local conformations of the polypeptide chains, the tertiary structure describes how the secondary structure elements are arranged to form the overall shape of the chains and the interactions between one or more polypeptide chains gives the quaternary structure. Artificial Neural Networks are useful toolbox for secondary structures prediction of proteins.

In this thesis a generalized regression neural network (GRNN), probabilistic neural network (PNN) and backpropagation algorithm (BP) were applied to the hemoglobin primary structure with different window sizes of amino acid sequences to predict the secondary structure that has helix and coil. Then all results of the networks are combined with GRNN. The data set is prepared with 20 alpha-141 and 20 beta-146 hemoglobin chains from Protein Data Bank. The overall success rate of GRNN is around 90.2-91.7% for beta chains and 85.9-87.3% for alpha chains. The PNN achieved between 91.2-92% overall accuracy for beta and 85.4-86.5% for alpha. BP has the overall success rate of 89.9-92.5% for beta, 86.5-90.3% for alpha. There is no significant improvement in prediction accuracy with CNN.

Keywords: Generalized Regression Neural Network, Probabilistic Neural Network, Backpropagation, Combined Neural Networks, Hemoglobin.

ACKNOWLEDGEMENTS

I would like to express my respects and gratitude to my supervisor Asst. Prof. Dr. Turgay İbrikçi. Many thanks for all of his supports, guidance, patience, cooperates, suggestions on initiating, improving and completing this study.

I would like to express my appreciation to Prof. Dr. Süleyman Güngör, the head of the department of Electrical and Electronics Engineering, and all my colleagues in the Electrical Electronics Department for their moral supports.

I thank Prof. Dr. Kıymet Aksoy and Erdinç A. Yalin who are colleagues at the Medical School of Cukurova University for understanding the hemoglobin structures and discussions and also thank to Prof. Dr. Seyhan Tükel who is colleague at the Faculty of Sciences and Letters for understanding the protein structures.

I would like to thank to my committee member Asst. Prof. Dr. Ulus Çevik for his supports and very valuable discussions.

I would like to thank my family, my father Muhlis, my mother Gönül and my sister Ceren. They always encouraged and supported me with their loves and inspirations. Thanks my fiancé Cumhur, my best friend, for his love, tolerance, helps, patience and morale support.

I would like to express my best feelings to my lovely friend Ayça Çakmak for her great encouragement, help and support from all aspects. Finally, I thank all my friends, great people that I couldn't mention their names here for their good wishes and encourages.

NOTATIONS

w	Weight
f(net)	Sigmoid transfer function
Q_3	Accuracy factor
W	Window size of the kernel
r	The number of amino acid residue
$\boldsymbol{X}$	Independent vector random variable
Y	Dependent scalar random variable
n	Pattern size
$g(\boldsymbol{X}, Y)$	The joint continuous probability density function
x_i	Sample values for X random variable (input)
y_i	Sample values for Y random variable (output)
$\hat{g}(x,y)$	The estimate of joint probability function
σ	The window width of the kernel (smoothing factor)
$\hat{y}(x)$	The expected value of output
m_i	The number of training patterns that belong to class i
x_{ij}	The j^{th} training pattern of i^{th} class
f_i	The computed Parzen's pdf for i^{th} class
p_i	The prior probability of i^{th} class
h_i	The hidden units
o_k	The output responses
t_k	The target output
J	The criterion function
Δw	The value of weight update
η	The learning rate
δ	The sensitivity of a unit

FIGURES **PAGE**

Figure 1.1. The general structure of an amino acid 2

Figure 1.2. The structure of α-helix conformations 3

Figure 1.3. The structure of β-bridge conformations 3

Figure 1.4. The structure of parallel β-sheet conformation 4

Figure 1.5. The four levels of normal adult hemoglobin structure 6

Figure 1.6. The general model of learning in ANN 8

Figure 1.7. A simple multilayer perceptron (MLP) 8

Figure 3.1. The network of GRNN 15

Figure 3.2. The Network of PNN 19

Figure 3.3. The Network of BP 21

Figure 3.3. The Network of CNN 26

Figure 4.1. Overall results of 141 alpha set for each window size 28

Figure 4.2. Helix results of 141 alpha set for each window size 29

Figure 4.3. Coil results of 141 alpha set for each window size 29

Figure 4.4. Overall results of 146 alpha set for each window size 31

Figure 4.5. Helix results of 146 beta set for each window size 31

Figure 4.6. Coil results of 146 beta set for each window size 32

Figure 4.6. Overall results of 141 alpha set for each window size 34

Figure 4.7. Helix results of 141 alpha set for each window size 34

Figure 4.8. Coil results of 141 alpha set for each window size 35

Figure 4.9. Overall results of 146 beta set for each window size 36

Figure 4.10. Helix results of 146 beta set for each window size 37

Figure 4.11. Coil results of 146 beta set for each window size 37

Figure 4.12. Overall results of 141 alpha set for each window size 38

Figure 4.13. Helix results of 141 alpha set for each window size 40

Figure 4.14. Coil results of 141 alpha set for each window size 40

Figure 4.15. Overall results of 146 alpha set for each window size 42

Figure 4.16. Helix results of 146 beta set for each window size 42

Figure 4.17. Coil results of 146 beta set for each window size 43

Figure 4.18. Overall results of 141 alpha set for each window size 44

Figure 4.19. Helix results of 141 alpha set for each window size 45

Figure 4.20. Coil results of 141 alpha set for each window size 45

Figure 4.21. Overall results of 146 beta set for each window size 47

Figure 4.22. Helix results of 146 beta set for each window size 47

Figure 4.23. Coil results of 146 beta set for each window size 48

Figure 4.24. Overall results of 141 alpha set for each window size 48

Figure 4.25. Overall results of 146 beta set for each window size 49

TABLES **PAGE**

Table 3.1. DSSP assignment of the secondary structures 13

Table 4.1. Self-consistency and independent-dataset results for each window size 28

Table 4.2. Self-consistency and independent-dataset results for each window size 30

Table 4.3. Self-consistency and independent-dataset results for each window size 33

Table 4.4. Self-consistency and independent-dataset results for each window size 36

Table 4.5. Self-consistency and independent-dataset results for each window size 39

Table 4.6. Self-consistency and independent-dataset results for each window size 41

Table 4.7. Self-consistency and independent-dataset results for each window size 44

Table 4.8. Self-consistency and independent-dataset results for each window size 46

1. INTRODUCTION

Proteins are the main workhorse of an organism as a whole. Having the vital important tasks in a living organisms such as storage and transport of the required substances, defense against alien substances, etc., allow the protein to gain a fame of being one of the most essential part of molecular biology. The protein molecule represents much of the bulk of an organism and accomplishes almost all of the biochemical activities of these organisms. To understand the life process of an organism, it is necessary to first know the protein's structure that is closely related to its function.

Routinely used and reliable methods to determine the structure of a protein are laborious techniques such as X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR). Since biological techniques are quite expensive, laborious and time consuming, in recent years scientists have concerned computer and statistical modeling. In this study, the aim is to test the success on predicting the secondary structure of the protein that is the first step for determination of the overall structure by combined neural networks.

1.1. Proteins

All proteins are composed of linear unbranched combinations of twenty different amino acids (Append. Table 1.). The structural and functional properties of the proteins primarily depend on this linear sequence of amino acids (i.e. their number and positions). Protein structures are described through four main hierarchical levels; *Primary, Secondary, Tertiary, Quaternary.*

In the general structure, amino acids contain an alpha carbon (C_α) that is attached to a carboxyl group (-COOH), an amino group ($-NH_2$), a hydrogen atom (-H) and a side chain group (-R) that gives each amino acid its distinguished properties [Solomons, 1995]. Figure 1.1. illustrates the general structure of an amino acid.

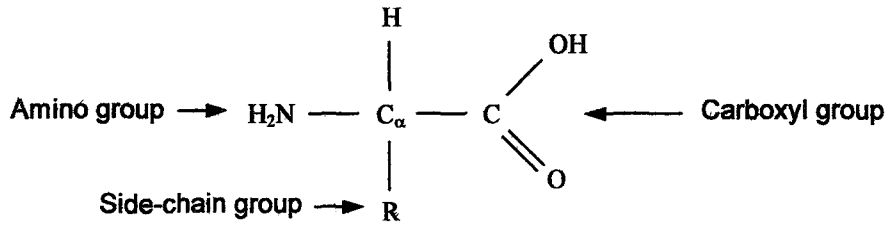

Figure 1.1. The general structure of an amino acid

When the carboxyl group of one amino acid is joined to the amino group of another amino acid by a peptide bond, this forms a dipeptide. The molecule constitute *n*-sequence of amino acid residues are termed a polypeptide. This sequence of amino acids that make up protein is called the *primary structure* [Denniston et al., 2001]. Each protein has distinct primary sequence with different amino acids which are placed differently along the chain.

Protein molecules are not rigid because the bonds that join C_α to amino group (N-C_α) and to carboxyl group (C_α-C) can rotate in a great flexibility with the angles of PHI (ϕ) and PSI (ψ), respectively. Then the protein folds into a three-dimensional configuration according to the association between amino acid residues that can stabilize certain conformations above others to conserve the most energy by this rotation of its segments along the ϕ and ψ angles. These local conformations are known as the *secondary structures* that is the result of hydrogen bonding between the hydrogen of amino group and the oxygen of carboxyl group [Denniston et al., 2001]. There are eight distinct classes of secondary structures;

1. **α-helix** is an helical configuration of residues with 3.6 amino acid per turn and formed by hydrogen bonding between the backbone CO of one residue and the backbone NH of fourth residue along the chain (Figure 1.2).

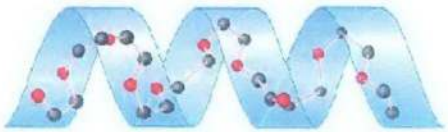

Figure 1.2. The structure of α-helix conformations [Denniston et al., 2001]

2. **π-helix** has similar conformation with α-helix but 4.4 residues per turn with hydrogen bonding between the atoms of i^{th} and $i + 5^{th}$ residues.
3. **3_{10} helix** is also similar to α-helix but formed by hydrogen bonds between the atoms of i^{th} and $i + 3^{rd}$ amino acids.
4. **Isolated β-bridge** has the configuration of a fashion that is made up of a hydrogen binding between a pair of three adjacent amino acids by bridging the i-1, i, i+1 and the j-1, j, j+1 residues. Two kinds of bridges are seen; the parallel bridges that is formed when the chains run in the same direction, the antiparallel bridges that is formed when the conformation of the chains is in the opposite direction. These orientations are shown in Figure 1.3. [Denniston et al. 2001].

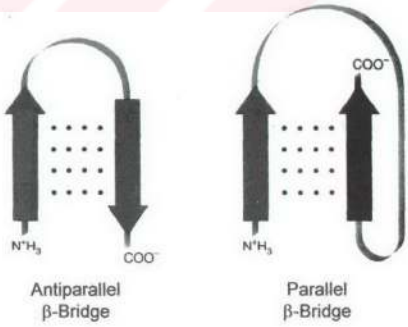

Figure 1.3. The structure of β-bridge conformations

5. **Extended β-strand** is the combination of two or more consecutive β-bridges. A beta-sheet is composed of several individual beta-strands arranged in a plane. There are also two kinds of β-sheet; parallel and antiparallel β-sheets that are formed by strands linked of parallel bridges and anti-parallel bridges, respectively. Figure 1.4. illustrates an example of parallel β-sheet structure.

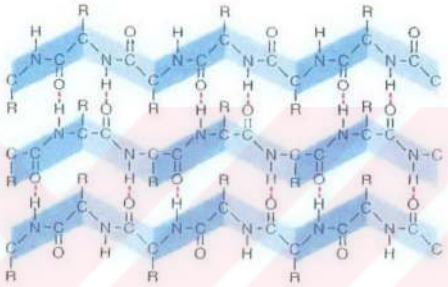

Figure 1.4. The structure of parallel β-sheet conformation [Denniston et al., 2001]

6. **Turn** occurs when a hydrogen bonding exists between the CO atom of i^{th} residue and the NH atom of $i + n^{th}$ residue.
7. **Bend** occurs when the angle between atoms $C_{\alpha(i-2)}$, $C_{\alpha(i)}$ and $C_{\alpha(i+2)}$
8. **Coil (Loop)** is a region that can not be included into any other seven structures. The term is used for the segments of a chain that do not form a regular secondary structure.

α-helices, β-strands and coils are the most common types that are coincided with abundantly in secondary structures. The polypeptide chain with its regions of the secondary structure is arranged to form its the overall folding pattern. Thus the protein takes on its final three-dimensional shape. This completion is termed the *tertiary structure*. For many proteins, the functional form is composed of several polypeptides. The arrangement of these polypeptides defines the *quaternary*

structure of a protein [Solomons, 1995].

1.2. Hemoglobin

Hemoglobin is one of the most important protein and commonly illustrated one (i.e. more is known about the chemical structure and function than any other protein molecules) because the various types of disorders in its structure cause lots of diseases that are encountered in the humans and the oxygen, carried via the hemoglobin protein, is the main element of the human and animal physiology. All vertebrate and many invertebrates use the hemoglobin as an oxygen carrier. Thus the hemoglobin, the vital importance for a human body, is found in the red corpuscles (referred to as erythrocyte) of the blood. The major function of the hemoglobin is to collect oxygen that diffuses into the plasma of the blood from the lungs, then to deliver it through the arteries to the tissues for maintaining the viability of cells and to transport carbon dioxide back to the lungs through the veins [Rinsho, 1998].

A hemoglobin that carries a full load of four oxygen atoms, is called oxyhemoglobin. Reversibly, the hemoglobin with no oxygen is known as deoxyhemoglobin. The hemoglobin molecule is made up of four polypeptide subunits (i.e. minor proteins) that are known as globins and four heme groups for each globin chain, as its name states [Pauling, 1952]. The four globins of normal adult hemoglobin refer to as two identical alpha chains, each with 141 amino acids and two beta chains, each with 146 amino acids [Shirasawa et al., 2003] (Append. Table 2.). These words of alpha and beta are irrelevant with the concept of α-helix and β-sheet secondary structures. Even in fact these chains contain α-helices, but no β-sheets in their secondary structures.

The α and the β chains fold up to form a globular tertiary structure. The heme groups are encased in the folded helices like a packaging. At the center of these heme complexes there is an iron atom that is actually responsible for holding one oxygen atom and for supplying red color to blood. The hemoglobin protein is shown in Figure 1.5. [Denniston et al., 2001].

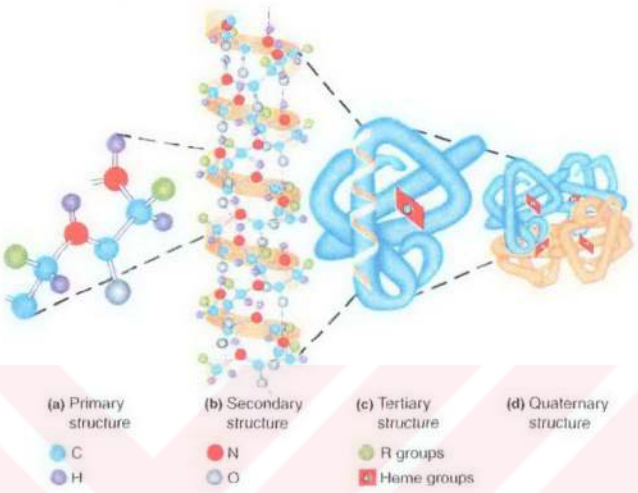

Figure 1.5. The four levels of normal adult hemoglobin structure

Hemoglobin molecule has different abilities about the oxygen carrying associated with the development stage from embryo to adult. Embryonic hemoglobin, found only in the first three or four months of intrauterine life, is made from one alpha (α) and one gamma (γ) chain. This initial hemoglobin dimer has no function about blood circulation. The functional type of the fetus hemoglobin is formed with the combination of two α and two γ chains and predominates during the rest of fetal life. This is called fetal hemoglobin (HbF) that makes up all the blood until about two months before birth [Pauling, 1952]. This is completely replaced by adult hemoglobin within a certain months after birth.

All adult hemoglobin throughout the world has the same structure. However, sometimes some defects can occur in the genes code for hemoglobin and cause abnormalities in the hemoglobin. An abnormal hemoglobin appears in three basic conditions:

Structural defects is occurred via the alteration in the gene for one of the two

globins. This is called mutation that occasionally disturbs the function of the hemoglobin molecule and produces a disease state. The sickle cell disease is an example of this type of mutations' outcome.

Deficiency in production of one of the two subunits is termed "thalassemias". The mutation of hemoglobin chain imbalance produce the disease of anemia by destroying the red cells.

Abnormal combinations of normal subunits are the association with β-globins (or α-globins) into groups of four (tetramers). In severe α-thalassemia, that is the combination of β-globins, the mutations of these tetramers result in a loss of function [Anonymous/a].

In conclusion, these are the reasons for the importance of finding the hemoglobin's structure that is exactly related with the implementation of its function.

1.3. Artificial Neural Networks

ANNs are computational paradigm that has an amazing capacity to actually learn from input data, inspired by the way the biological nervous system in brain process information [Stergiou]. A neural network is composed of highly interconnected processing units or artificial neurons. These units are joined together with weighted connections that are similar to synapses of the brain. The neural network architecture consists of three or more layers of neurons that fully connect to consecutive layer neurons (i.e. each connection are associated with numerical weights) [Sarle, 1994]. The input exemplars from previous neurons are processed through the weighted sum and then transmitted to another neuron via a transfer or activation function as an input (i.e. each neuron contains the weighted sum of its inputs filtered by the function) [Jordan and Bishop, 1996].

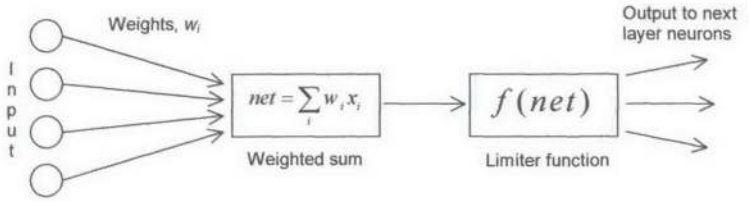

Figure 1.6. The general model of learning in ANN

This process is repeated after the modification of the weights until the error between the predicted and expected output values is minimized [Lippmann, 1987]. This iterative modification known as a training, characterize the learning ability of ANNs. The goal of the training is to reach an optimal solution based on the performance measurement.

The most common architecture of neural networks is the multilayer perceptron (MLP). The Figure 1.7. illustrates a simple MLP. The first layer, named the input layer, presents a pattern to the network. The last layer that is the output layer consists of the output response to given input. The other layers between the input and output layer are the hidden layers (pattern layers) that perform the processing tasks of the network.

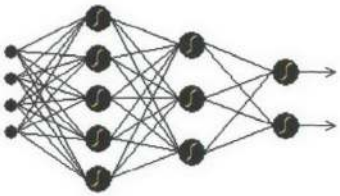

Figure 1.7. A simple multilayer perceptron (MLP)

Neural Networks are useful toolbox for prediction of secondary structures of proteins because of the capability of analyzing the overwhelmingly big data sets and mapping the relationships between the protein sequences [İbrikçi, 2000]. For this study GRNN, PNN, BP and the network of their combination were chosen to predict the secondary structure of hemoglobin.

2. RELATED WORKS

Studies on the protein secondary structure prediction by the computational and statistical methods started with Krigbaum and Kuntton [Krigbaum and Kuntton, 1973: from Cai et al., 2003]. They used the multiple linear regression algorithm to predict the amino acid composition of a protein. This attempt continued with Chou-Fasman method that had achieved a three-state (Q_3) accuracy of 52% [Chou and Fasman, 1978]. This method is well-known empirical statistical algorithm that is based on the frequencies of the secondary structure types.

Within the field of protein secondary structure prediction the idea of combining different prediction methods is also an old one. Schultz et al. combined 11 methods, and Argos et al. combined five different methods [Schultz et al., 1974] [Argos et al., 1976]. Sternberg stressed the merits and limitations of different prediction methods for different types of secondary structure (multi-strategy learning) [Sternberg, 1983]. Nishikawa and Ooi combined methods using a voting strategy [Nishikawa and Ooi, 1986]. In 1988, Biou et al. used a complicated biased method to combine three methods (including GOR 111) [Biou et al., 1988].

At the end of 1980's, researches about the secondary structure prediction have developed considerably with starting to use neural network methods. The average success reached 64.3% on three types of secondary structure conformation (α-helix, β-sheet and coil) of non-homologous data set by using neural network models [Qian and Sejnowski,1988]. Qian and Sejnowski applied the sliding window strategy to preprocess the data like Chou and Fasman. Zhang et al. used a neural network (which they found to be better than voting) to combine predictors obtained from neural network, nearest neighbour and naive Bayes algorithms. They achieved more accurate level than individual experts level with 66.4% accuracy [Zhang et al., 1992].

In the last decade of 20^{th} century, accuracy percentage level rose over 70%. During this period, Rost et al. studied in several works about the prediction of secondary structure. They found a method called PHD (profile network prediction) that has the same structure as Qian and Sejnowski with 40 hidden units, window size

of 13 amino acids and addition of the alignment feature (i.e. amino acid frequency vector for each position) in the form of a profile [Rost and Sander, 1993a; Rost and Sander, 1993b]. The success of the network that is trained a set of profiles of multiple alignment sequences, is about 70%. For instance in 1993, they predicted the secondary structure of proteins at 70.8 % accuracy with two-layered feed forward neural network. To verify the success rate of the performance 7-fold cross validation test was used in their study. Rost and Sander also used an ensemble of networks, i.e. combine predictions from 10 distinct neural network systems. They observed over 1% improvement in accuracy rate with 72% Q_3 [Rost and Sander, 1994a; Rost et al., 1994b].

In 1996, Mehta et al. applied several types of statistical, neural network and artificial intelligence (AI) based methods to predict the secondary structure of proteins according to their amino acid sequence, and obtained around 60%-75% accuracy [Mehta et al., 1996]. They used Chou-Fasman, Microgenie for statistical methods, Fuzzy Artmap for AI methods and GRNN, PNN for NN methods in their studies. Krogh and Riis made another study on protein structure prediction. They applied structured neural networks and multiple sequence alignment with an accuracy of 80% [Krogh and Riis, 1996].

Salamov and Solovyev applied their NNSSP (nearest neighbor secondary structure prediction) method and PHD method on non-membrane non-homogenous proteins with 71%-72% of overall three-state accuracy [Salamov and Solovyev, 1997]. NNSP is based on nearest neighbor rule, which states a classification for a test sample according to the classifications of neighbor training examples from a database of known structure.

In 1998, one of the combine methods, Jpred was applied by Cuff [Cuff et al., 1998]. This method combined the outputs of several certain methods such as PHD, NSSP, etc. by using a consensus method and achieved 72.9%.

By using currently available accurate methods, Chandonia and Karplus obtained 74.8% success rate on non-homologous protein data that is considerably larger than the other data sets [Chandonia and Karplus, 1999]. Zhang et al. proposed multiple linear regression with 80% correct protein prediction [Zhang et al., 1999]. In

2000, Siermala, Juhola and Vihinen studied on pre-processing of protein secondary structures for neural networks [Siermala, 2000]. Petersen et al. achieved 77.2%-80.2% accuracy by structure filtering neural networks [Petersen, 2000]. Another combined method Jnet that was used in Cuff and Barton's study with 76.4% accuracy, consisted of HMM (Hidden Markov Model) and PSI-BLAST. These methods are combined with a neural network model instead of a consensus [Cuff et al., 2000]. Rost et al. also used Jnet by combining with some other methods (i.e. PHD, NNSP, Predator, Mulpred, DSC and Zpred). They obtained 74.8% success rate with slightly improvement [Rost et al., 2001]. In recent years three and eight classes of protein secondary structures is predicted with the performance of about 78% by using recurrent neural networks and profiles in the study of Pollastri et al. [Pollastri et al., 2002].

3. MATERIAL AND METHODS

3.1. Material

3.1.1. Data Collection and Preprocessing

In this study, the dataset has taken from Protein Data Bank, USA and exists of 20 alpha-141 and beta-146 hemoglobin chains which are 1A3N:A/:B, 1A4F:A/:B, 1A0U:A/:B, 1BIJ:A/:B, 1DXV:A/:B, 1CG5:A/:B, 1FHJ:A/:B, 1FSX:A/:B, 1G0B:A/:B, 1GBV:A/:B, 1HBS:A/:B, 1HBR:A/:B, 1IRD:A/:B, 1J7Y:A/:B, 1HDS:A/:B, 1ITH:A/:B, 1NIH:A/:B, 1QPW:A/:B, 1O1K:A/:B, 2HBD:A/:B [Protein Data Bank]. These hemoglobin dataset were divided into two parts; 2/3 as train and 1/3 as test sets. The networks used in this study were trained to predict a three-category (HEC) secondary structure assignment reduced from the eight-category assignment produced by DSSP program [Kabsch and Sander, 1983].

DSSP class	Abbrevation	Reduced Class
α - helix	H	H
3_{10} helix	G	H
π helix	I	H
Extended Strand	E	E
Isolated β Bridge	B	C
Turn	T	C
Bend	S	C
Coil and rest	-	C

Table 3.1. DSSP assignment of the secondary structures

In this dataset, there is no Strand (E) structure; only Helix (H) and Coil (C) represent the secondary structure. Training files contain a primary structure (i.e. a list of amino acids) and its corresponding secondary structure. These structures were manipulated by determining certain size of slide windows that consist of contiguous amino acid residues. Since there are 20 known amino acids in the nature, each of them was coded as a 20-bit strings with all, but one bit turned on, and each of outputs was coded as a 1-bit string that can be 1 for helix and 2 for coil;

Alanine (A): **1**0000000000000000000 (20-bit string coded for input)
Helix (H) : **1** (1-bit string coded for output)

Window (input)	ENLKGFL	Primary Structure (Amino Acid Sequence)
Output	HHCHH	Secondary Structure (Helix or Coil)

In this study, each dataset derives from eight different sizes of windows:

11-13-15-17-19-21-23-25

Each window size is chosen according to following formula with **r** number of amino acid residues before and after center element of the windows:

$$window\ size\ (W) = 2*r+1 \quad (1)$$

Center technique appears from the thought of that the central amino acid had a large influence in structure classification of that window [Qian and Sejnowski, 1988]. Prediction is applied by labeling the input pattern (i.e. amino acid sequence window) with the secondary structure, i.e. output that belongs to the central element of the amino acid sequence. For example, if the window size were chosen as 11, five amino acid residues before and five residues after would be considered, and at the opposite of the sixth element is the secondary structure, helix or coil. Starting from the first residue the window turns by sliding one residue to the right until the end of sequence, then all residues are classified.

{ENLKGFLVNKG} EELFLV... *input before coded with 20-bit strings code*

HHCHCHCCCHH... *output before coded with 1-bit string code*

Hence, each input pattern is coded as 20-bit string with *W x 20* binary input elements per window with an output 1 or 2.

For combined network, all the outputs that obtained from the other three networks were gathered in a single file as an input dataset and for the output dataset the same output files of the networks were used. The outputs from self-consistency dataset test (testing on train dataset) constituted dataset for training and the outputs from independent dataset test (testing on test dataset) formed the test dataset of the combined network.

3.2. Methods

3.2.1. General Regression Neural Networks (GRNN)

GRNN introduced by Donald Specht in 1990 is memory-based feed forward neural networks based on the estimation of Probability Density Function (pdf) from observed samples using Parzen-window estimation [Specht, 1991]. It approximates any arbitrary function either linear or non-linear relationships between input and output vectors, drawing the appropriate function estimate directly from the training data. This approach removes the necessity to specify a functional form of the estimation.

The method is a kind of regression network that utilizes a probabilistic model between an independent vector random variable $\boldsymbol{X}$ (i.e. input) and a dependent scalar random variable Y (i.e. output). Let x and y are the particular measured values of $\boldsymbol{X}$ and Y random variables, respectively, and $g(\boldsymbol{X}, Y)$ is the joint continuous probability density function of these random variables.

A good choice for non-parametric estimate of the probability density function g (as noted in Specht) is using the Parzen window estimator that is proposed by Parzen and performed for multidimensional cases by Cacoullos . Given a sample of n real d dimensional x_i values for a scalar y_i values, the estimate of joint probability density is:

$$\hat{g}(x,y)=\frac{1}{(2\pi)^{(d+1)/2}\sigma^{(d+1)}}\frac{1}{n}\sum_{i=1}^{n}\left[\exp\left(-\frac{(x-x_i)^T(x-x_i)}{2\sigma^2}\right)\exp\left(-\frac{(y-y_i)^2}{2\sigma^2}\right)\right] \quad (2)$$

where σ is the window width of a sample probability and called smoothing factor of the kernel [Yeung and Chow, 2002].

The expected value of Y given x (the regression of Y on x) can be given as:

$$E[Y/x]=\frac{\int_{-\infty}^{+\infty} Y.g(x,Y)dY}{\int_{-\infty}^{+\infty} g(x,Y)dY} \quad (3)$$

Substituting the joint probability estimate $\hat{g}(x, y)$ into the definition of the conditional mean, $E[Y/x]$ yields:

$$\hat{y}(x)=E[Y/x]=\frac{\sum_{i=1}^{n}[y_i\exp(d_i)]}{\sum_{i=1}^{n}\exp(d_i)} \quad (4)$$

where d_i is the distance function between the input vectors and the centers recorded in the pattern nodes, which is given by

$$d_i=-\frac{(x-x_i)^T(x-x_i)}{2\sigma^2} \quad (5)$$

The estimate $\hat{y}(x)$ is thus a weighted average of all the observed y_i valves where each weight is exponentially proportional to its Euclidean distance from x.

The structure of the GRNN consists of 4 layers; the input layer, the hidden (pattern) layer, the summation layer and the output layer (Figure 3.1.).

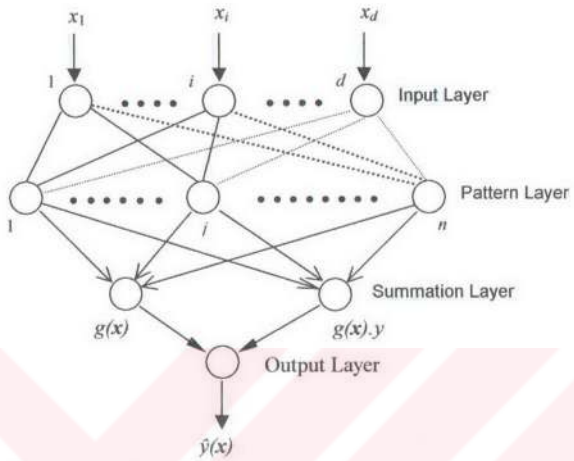

Figure 3.1. The network of GRNN

As a preprocessing step, all input variables of the training data are scaled roughly. Then, they are simply copied as the weights into the pattern units [Hu et al., 2002]. This means that each exemplar of the training data is dedicated to one neuron of the pattern layer that is called cluster center. Input units are the connection points for the elements of test exemplars. When a new input variable of the test data is entered into the network, it is subtracted from each cluster center vector. Thus, the distances between the input vectors and the centers are computed. A nonlinear function that is usually the exponential activation function is applied to the distances, and then these values are summed as shown in Equation 4 [Rzempoluck, 1997]. Two neurons of the summation layers can be denoted as the numerator and the denominator of the equation. Lastly, the division of $y_* g(x)$ by $g(x)$ gives the result of expected value of $\hat{y}(x)$. If y and $\hat{y}$ are the vector variables, in this case, one extra summation unit is added for each component of $\hat{y}$ in the unit [Hyun and Nam].

The only adjustable parameter is the smoothing factor for the kernel function. The critical point is to decide an optimum value for σ. The larger values of this factor cause the density to be smooth and converges $\hat{y}(x)$ to the value of the sample mean of

the observed y_i. On the other hand, when σ is chosen very small (i.e. σ goes to zero), the density is forced to have non-gaussian shapes. Therefore the wild points have great effects on the estimate. In this case, $\hat{y}(x)$ considers only the value of y_i concerning with the observation(s) closest to x. All values of y_i are taken into account where the points closer to x are given heavier weights, if the optimum value of σ is selected [Specht, 1991].

There are several ways to provide an optimal value for σ. One of the useful method is *k*-fold cross validation [El-Solh et al., 1999]. The basic idea of cross validation is to remove some of the training data for using it to test the performance of the learned model on new data. *k*-fold cross validation uses *k* subsets for validation [Anonymous/b]. In Equation 5, a single common sigma (σ) is used for all the input variables and pattern units. However, it can be necessary to assign an independent sigma for each of the variables when each stored pattern has its own importance or the centers are spaced uniform with each other [Baker and Richards, 1999].

The main advantage of GRNN according to other techniques is fast learning. It is a one-pass training algorithm, and does not require an iterative process. The training time is just the loading time of the training matrix. Also, it can handle both linear and non-linear data because the regression surface is instantly described everywhere, even just one sample is enough for this. Thus, other existing pattern nodes tolerate faulty samples. Another advantage is the fact that adding new samples to the training set does not require re-calibration of the model. As the sample size increases, the estimate surface converges to the optimal regression surface. Thus, it requires many training samples to span the variation in the data and all these to be stored for the future use. Solely, this causes a trouble of an increase in the amount of the computation to evaluate new points. In the course of time, highly improvements in the speed of the computer's processing power prevent this being a major problem. Furthermore, this also can be overcome by applying the various clustering techniques for grouping samples that each center is represented by this group of samples [Specht, 1991]. However, there is only one disadvantage that there is no intuitive method for choosing the optimal smoothing factor.

3.2.2. Probabilistic Neural Networks (PNN)

The Probabilistic Neural Network (PNN), also introduced by Donald Specht is feed-forward and one-pass training algorithm primarily based on Bayes-Parzen Classification Technique which is theoretically used in many classical pattern recognition problems [Specht, 1990]. Parzen's method is used for the estimating probability density function (pdf) of random variables form a set of training examples similar as in GRNN. Bayes' theory that allows the determination of the classes for the unknown class pattern vectors, is included into the PNN for the classification decisions [Walsh, 2000].

To understand Bayesian technique, consider a pattern vector x with dimension d taken from a set of examples that belong to k number of distinct classes (1,2,....,k). Let p_k is the prior probability for the accuracy of k^{th} class. For all classification classes, assume that $F_1(x)$, $F_2(x)$, ... , $F_k(x)$ are true probability density functions of that sample. Bayes' decision rule then favors a class for the unknown sample [Goh, 2002]. According to this strategy, x belongs to the i^{th} class if:

$$p_i F_i > p_j F_j \tag{6}$$

for all categories $j \neq i$.

Typically, the class densities of an unknown input vector are calculated from all the training pattern vectors in the PNN by using the multivariate case of Parzen's pdf estimator as for GRNN. However contrarily, PNN requires one pdf for each class in usage of the estimator. The estimate of the probability density function can be expressed as [Niwa, 2003]:

$$\hat{f}_i(x) = \frac{1}{(2\pi)^{(d+1)/2}\sigma^{(d+1)}} \frac{1}{m_i} \sum_{j=1}^{m_i} \exp\left(-\frac{(x-x_{ij})^T (x-x_{ij})}{2\sigma^2} \right) \tag{7}$$

where m_i is the number of training patterns that belong to class i, x_{ij} is the j^{th} training

pattern of this class and x is the i^{th} test vector. The other probability density functions for each category are similarly defined by Equation 8. PNN has the only one training parameter called the smoothing factor and can be find by . k-fold cross-validation method like in GRNN [Anonymous/c].

There are two differences between the architecture of GRNN and PNN [Hajmeer and Basheer, 2002].

1. The summation layer that contains one neuron for each class.
2. The output layer that simply retains the maximum of the summation nodes .

The architecture of PNN is shown in Figure 3.2.

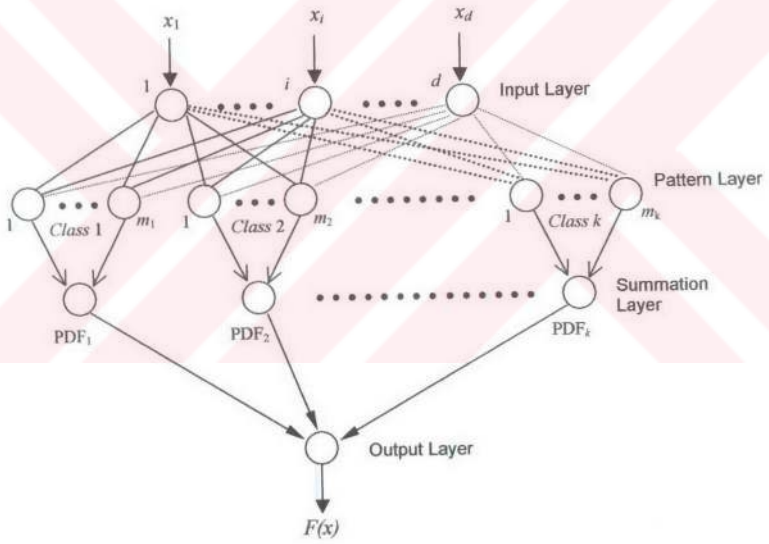

Figure 3.2. The Network of PNN

Input units are the distribution units for the elements of test samples to all the pattern units. After computing the distance measure, and then the activation functions for each class are subjected to the corresponding summation layer neuron which adds all

the outputs in a particular class together. The constant multiplier of the Parzen's estimator equation is then applied to the activation of each summation neuron to get the estimated pdf value of each category. However, common variables of this constant for each class can be neglected in the computation of PNN. Here, the only distinct factor is $1/m_i$. The summation neurons calculate the posterior probability of each class according to the Bayes' theorem by combining these outcoming values of pdfs with their relative frequencies (i.e. the prior probabilities) [Raykar];

$$F_i(x) = \frac{p_i f_i(x)}{\sum_{i=1}^{k} p_i f_i(x)} \tag{8}$$

f_i is the computed Parzen's pdf for i^{th} class, p_i is the prior probability. Both f_i and p_i have the factor $1/m_i$ and m_i in their contents, respectively. These constants simplifies each other; this yields:

$$F_i(x) = \frac{f_i(x)}{\sum_{i=1}^{k} f_i(x)} \tag{9}$$

Subsequently, the output neuron performs a classification of the unknown input vector according to the highest pdf. This choice is decided from Bayes' decision rule stated as before in Equation 7 [Low and Togneri, 1998].

Training for both GRNN and PNN is proceeded in two parts. In the first part, the network is trained with the data of the training set. The second part of the training can be regarded as finding the best smoothing factor by using a validation set of examples to maximize the success.

In practice, PNN is often an excellent pattern classifier, outperforming other classifiers because it derives an estimate of pdf from the training examples rather than just assume normal distribution. Although PNN and GRNN have totally different algorithms, their pros. and cons. show similar qualities about speed of learning, data variation, adding new samples, choosing smoothing factor, etc.

3.2.3. Backpropagation (BP)

The Backpropagation introduced by Werbos in 1974 [Werbos, 1974: from Cai et al., 2002], is the most popular learning algorithm for multilayered structure in neural networks. BP algorithm is processed in two distinct phases that are forward motion signals and error back propagation. The feed forward phase is to propagate input signal from the input layer to the output layer. During this process, the basic arrangement of NN structure and properties are used.

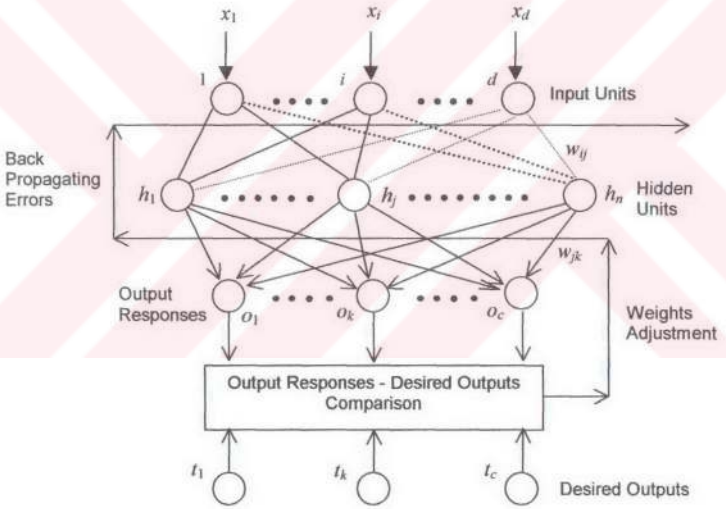

Figure 3.3. The Network of BP

The process starts with the random values for weights (w). An activation function the sigmoid function (Eq.11.) is appropriate because of its easy differentiable feature. The output of this function gives a value between 1 and 0.

$$f(net) = \frac{1}{1+e^{-net}} \tag{10}$$

The calculated output, o_k is then compared to the target output, t_k and the backward phase begins with the computation of least mean squares (LMS) for weight adjustment [Duda et al., 2000]. This yields the criterion function, J;

$$J(w) = \tfrac{1}{2}\sum_{k=1}^{c}(t_k - o_k)^2 \tag{11}$$

BP algorithm updates the weights by using the gradient descent rule starting with hidden-to-output weights;

$$\Delta w_{kj} = -\eta \frac{\partial J}{\partial w_{kj}} \tag{12}$$

where η is the learning rate that controls the learning time of the network (i.e. the weight change with relative size). The chain rule is used for differentiation;

$$\frac{\partial J}{\partial w_{kj}} = \frac{\partial J}{\partial net_k}\frac{\partial net_k}{\partial w_{kj}} = -\delta_k \frac{\partial net_k}{\partial w_{kj}} \tag{13}$$

δ_k, the sensitivity of unit k, that gives the overall error changes over the net activation of the unit is simply;

$$\delta_k = \frac{\partial J}{\partial o_k}\frac{\partial o_k}{\partial net_k} = (t_k - o_k) f'(net_k) \tag{14}$$

and

$$\frac{\partial net_k}{\partial w_{kj}} = h_j \tag{15}$$

Substituting these results into the Equation 13 gives the learning rule otherwise known as for the weight update for hidden-to-output weights;

$$\Delta w_{kj} = \eta \delta_k h_j = \eta (t_k - o_k) f'(net_k) h_j \tag{16}$$

The learning rule for the input-to-hidden units can be found with the similar way;

$$\frac{\partial J}{\partial w_{ji}} = \frac{\partial J}{\partial h_j} \frac{\partial h_j}{\partial net_j} \frac{\partial net_j}{\partial w_{ji}} \tag{17}$$

Calculation of the first term on the right hand side is subtle, because it involves the hidden-to-output weights, w_{kj} [Bishop, 1995]. When the criterion function, given in Equation 12, is substituted into the term, this yields;

$$\frac{\partial J}{\partial h_j} = \frac{\partial}{\partial h_j} \left[\frac{1}{2} \sum_{k=1}^{c} (t_k - o_k)^2 \right]$$

$$= -\sum_{k=1}^{c} (t_k - o_k) \frac{\partial o_k}{\partial h_j}$$

$$= -\sum_{k=1}^{c} (t_k - o_k) \frac{\partial o_k}{\partial net_k} \frac{\partial net_k}{h_j}$$

$$= -\sum_{k=1}^{c} (t_k - o_k) f'(net_k) w_{kj} \tag{18}$$

Applying the Equation 15, the sensitivity for a hidden unit is defined;

$$\delta_j = f'(net_k) \sum_{k=1}^{c} w_{kj} \delta_k \tag{19}$$

Thus the learning rule for the input-to-hidden weights can be represented as;

$$\Delta w_{ji} = \eta \delta_k x_i = \eta \left[\sum_{k=1}^{c} w_{kj} \delta_k \right] f'(net_j) x_i \tag{20}$$

Finally, the weights are updated as follows;

$$w(new) = w(old) + \Delta w \tag{21}$$

This iterative algorithm of the forward and backward phases is repeated until the overall network error is less than a pre-defined threshold value. This causes a disadvantage of long training time.

3.2.4. The Combined Neural Networks (CNN)

The CNN is the combination of several neural network methods with a combiner. This combiner can be a neural network, statistical method or some other prediction algorithm like consensus method. Figure 3.4. illustrates the network schema of a CNN.

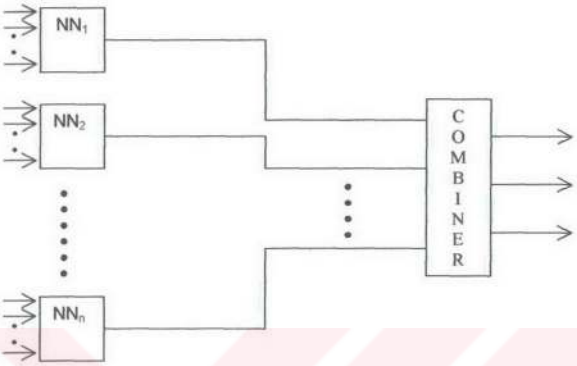

Figure 3.4. The Network of CNN

In this thesis, the three algorithms, GRNN, PNN and BP as the subunits were combined by the GRNN that was chosen as the combiner.

4. RESULTS AND DISCUSSION

The algorithms of GRNN, PNN and BP are applied to the datasets of 141 alpha and 146 beta chains constructed by slide windows for the prediction of the secondary structure of proteins. Self-consistency test (testing on train dataset) and independent-dataset test (testing on test dataset) were used with eight different window sizes to obtain success rates of secondary structure prediction for two classes namely helix and coil. Then all results were calculated for overall success rates via Q_2 method. This method was computed with the following formula;

$$Q_2 = \frac{H + C}{N} \tag{22}$$

where H and C stand for correctly predicted helix and coil, respectively. N shows the total number of predicted residues. After individual applications of the methods, all outputs are combined by the method GRNN, as a combiner and then helix, coil and overall predictions (Q_2) values are computed.

4.1. GRNN Results

4.1.1. 141 Alpha Set

According to the self-consistency and independent test results, stabilization of accuracy level is observed. While this level is quite good for helix with 99%, it is around 95.6% for coil because of its distribution in dataset.

The window size of 17 amino acid residues of independent dataset provides slightly better prediction of the secondary structure of protein in overall prediction with 87.3% (Table 4.1.). GRNN algorithm correctly-classified 93.4% of the helix classes, 69.6% of the coil classes for this window. If classes are examined individually, class of helix has the best result on window size 11; this is 19 and 21 for coil with 70.7%.

Table 4.1. Self-consistency and independent-dataset results for each window size for 141 alpha set

W	Helix		Coil		Overall	
	Train	Test	Train	Test	Train	Test
11	99.0	93.8	94.5	65.3	97.9	86.7
13	99.1	93.7	95.0	66.3	98.1	86.8
15	99.1	93.5	95.6	66.1	98.2	86.5
17	99.0	93.4	95.6	69.6	98.2	87.3
19	99.1	93.0	95.6	70.7	98.2	87.2
21	99.1	93.1	95.6	70.7	98.1	87.1
23	99.0	92.7	95.6	69.5	98.1	86.4
25	99.0	92.3	95.6	68.9	98.1	85.9

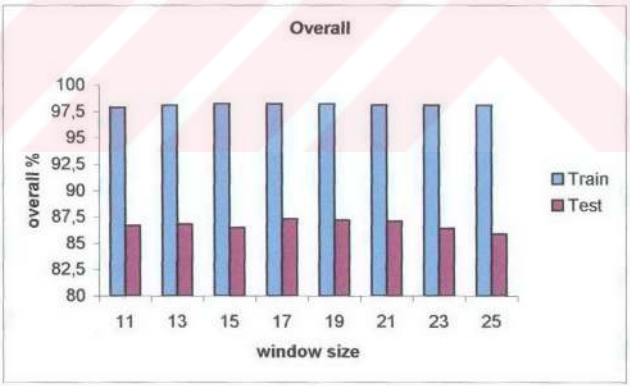

Figure 4.1. Overall results of 141 alpha set for each window size

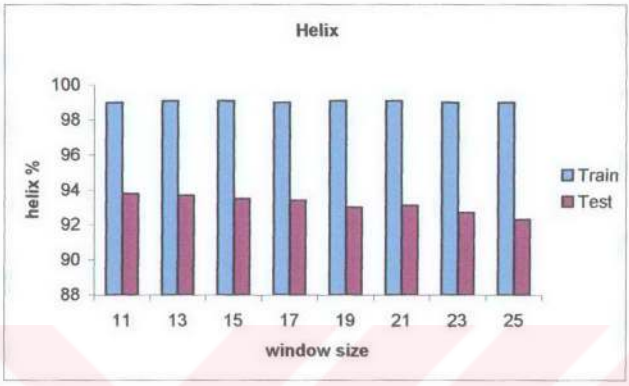

Figure 4.2. Helix results of 141 alpha set for each window size

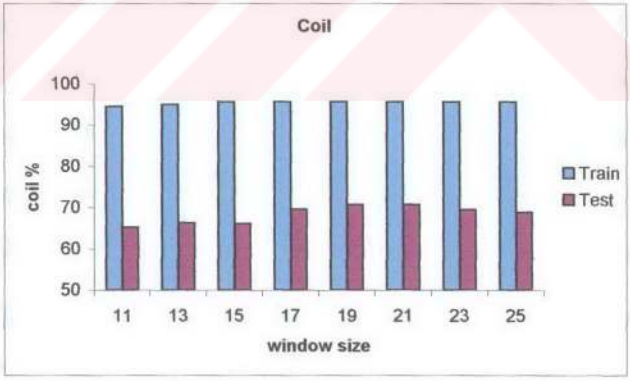

Figure 4.3. Coil results of 141 alpha set for each window size

4.1.2. 146 Beta Set

Window size of 13 has given the best overall prediction for independent data set with 91.7 with GRNN (Table 4.2.). This window also gives the best accuracy for coils with 76.1% and for helix with 95.8%. After window size of 13, size of 15 achieved the best result to classify helices with percentage of 95.8 and overall with 91.3. Based on the coils results from best to worst, window size scale is 13-11-19-23-25-17-15-21 (Figure 4.6.). According to these results, it can be said that the small window sizes give the better results for 146 beta test set in GRNN.

Table 4.2. Self-consistency and independent-dataset results for each window size for 146 beta set

W	Helix		Coil		Overall	
	Train	Test	Train	Test	Train	Test
11	99.1	95.1	96.1	75.1	98.5	91.0
13	99.1	95.8	96.1	76.1	98.5	91.7
15	99.3	95.7	96.4	74.3	98.7	91.3
17	99.4	95.5	96.6	74.5	98.8	91.0
19	99.5	94.4	96.8	75.0	99.0	90.2
21	99.6	93.9	97.1	74.1	99.0	89.6
23	99.6	94.2	97.3	74.8	99.1	89.9
25	99.5	93.4	98.3	74.8	99.2	89.3

Self-consistency test of 146 beta set gives the better results than 141 alpha set with 99.6 for helix, 97.3 for coil and 99.2 for overall.

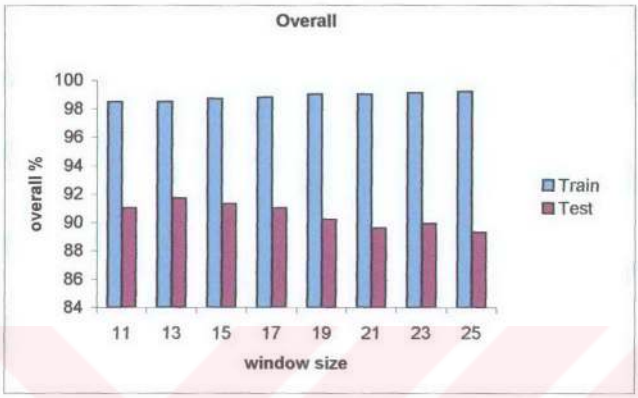

Figure 4.4. Overall results of 146 alpha set for each window size

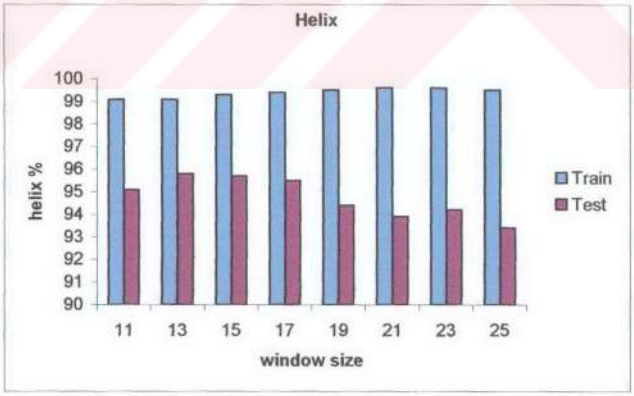

Figure 4.5. Helix results of 146 beta set for each window size

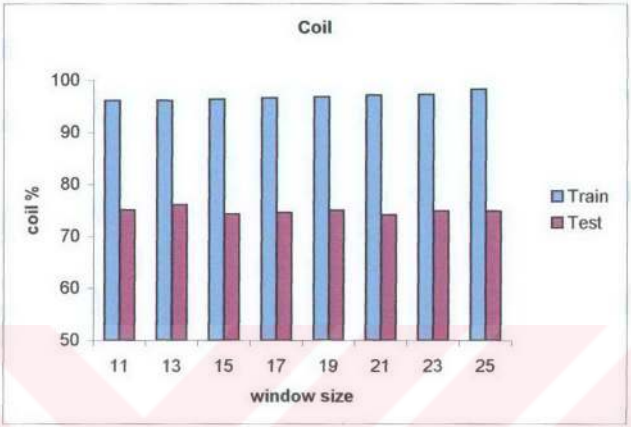

Figure 4.6. Coil results of 146 beta set for each window size

As shown in Figure 4.6., 146 beta set gives the result of stable window size values for coils in both self-consistency and independent-dataset tests. But there is a small difference on the success value of window size 13.

4.2. PNN Results

According to PNN results, the overall accuracy resembles to the GRNN results for both 141 alpha and 146 beta data test sets. But in detail, PNN has slightly better overall test results on 146 beta, worst results on 141 alpha. As for GRNN, there are also not considerably differences between the results in both self-consistency test set results.

While PNN has better results in helix prediction, it has given considerably worse results in prediction of coil comparing with GRNN. Despite of this approach, PNN obtained the best overall predictions, because of helix's weighted density.

The study of Mehta et al., the best success rate was obtained at the window

size 19 with around 60%-75%. In our study was also reached the maximum rate at the window size of 19 for 146 beta set with 92.1% and 141 alpha at window size 11 is better prediction than Mehta et al.'s prediction.

4.2.1. 141 Alpha Set

The success rate of the helix and overall test results are inversely proportional with the window sizes but opposite result is valid for coil. Thus, window size of 11 has given the best prediction with 94.0% and 86.5% for the helix and overall prediction, respectively. Coil prediction has the similar results between the window sizes, and gives the best result in the window sizes of 19, 21, 23 and 25 (Figure 4.8.).

Table 4.3. Self-consistency and independent-dataset results for each window size

W	Helix		Coil		Overall	
	Train	Test	Train	Test	Train	Test
11	98.8	94.0	93.3	64.1	97.4	86.5
13	98.7	93.8	93.7	64.5	97.4	86.5
15	98.6	93.7	93.5	64.9	97.3	86.4
17	98.5	93.6	93.9	64.9	97.3	86.2
19	98.7	93.4	94.1	65.3	97.5	86.1
21	98.6	93.3	93.5	65.3	97.3	85.9
23	98.7	93.1	93.9	65.3	97.4	85.6
25	98.6	93.0	93.9	65.3	97.3	85.4

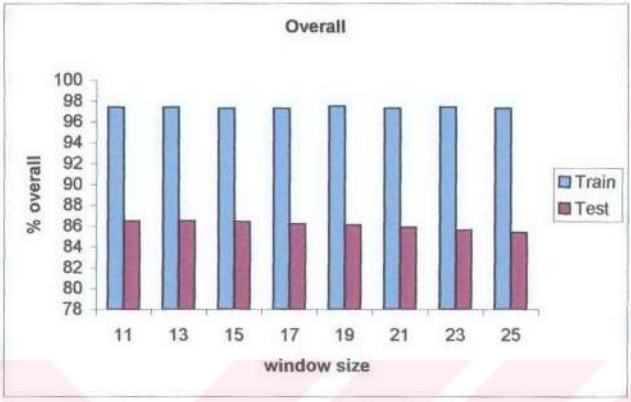

Figure 4.6. Overall results of 141 alpha set for each window size

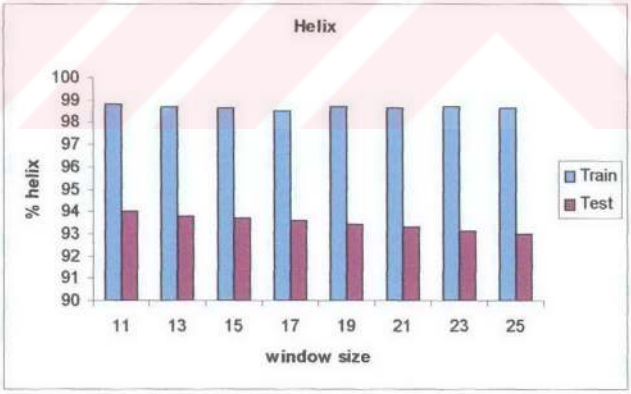

Figure 4.7. Helix results of 141 alpha set for each window size

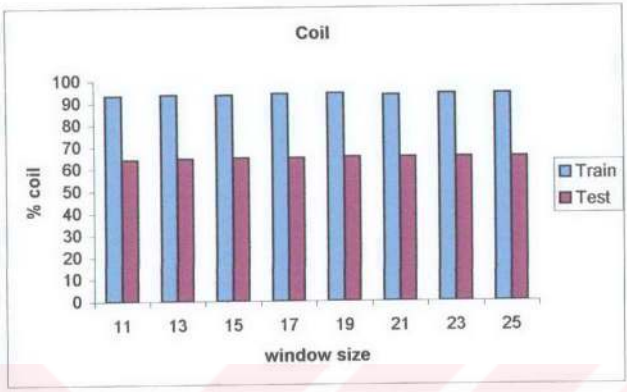

Figure 4.8. Coil results of 141 alpha set for each window size

4.2.2. 146 Beta Set

While there are no significant differences among the overall and helix accuracy levels of windows in test set results, it can be seen that window size 17 gives the best result for coil class (Table 4.4.). The results of coil from the best to the worst, window size scale is 17-19-13-11-15-21-23-25 (Figure 4.11.).

Table 4.4. Self-consistency and independent-dataset results for each window size

W	Helix		Coil		Overall	
	Train	Test	Train	Test	Train	Test
11	99.0	97.4	95.7	71.4	98.3	92.0
13	99.0	97.3	95.7	71.7	98.3	92.0
15	99.0	97.3	95.9	71.3	98.4	91.9
17	99.0	97.2	96.1	72.3	98.3	91.9
19	99.0	97.6	96.1	72.1	98.4	92.1
21	99.2	97.8	96.1	71.1	98.5	92.0
23	99.2	97.9	96.1	71.1	98.5	92.0
25	99.2	97.2	96.3	70.4	98.6	91.2

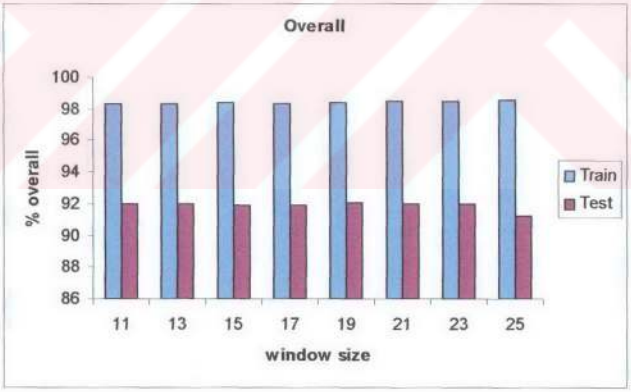

Figure 4.9. Overall results of 146 beta set for each window size

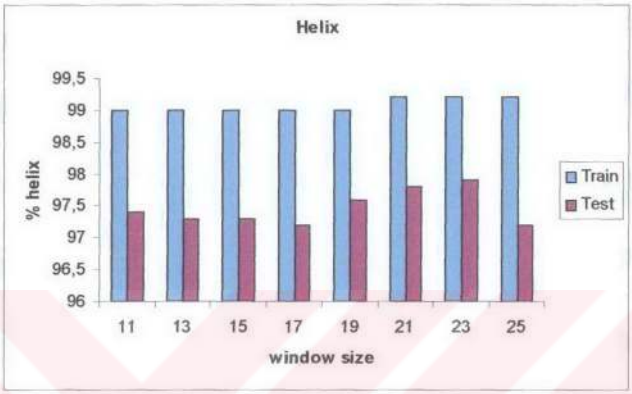

Figure 4.10. Helix results of 146 beta set for each window size

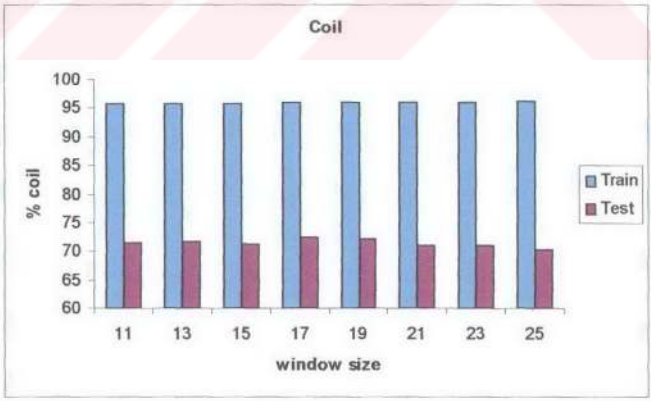

Figure 4.11. Coil results of 146 beta set for each window size

4.3. BP Results

When examined the BP results, it can be observed that the results for different window sizes are close to each other. Since BP learning depends on the weights, in another training, results can be successful for any different size of window. Thus, choosing a certain size of window as the best has no influence on prediction via BP. However, according to our data set, it may be said that small sizes of windows can give better results.

BP has more accurate results on coil prediction than the other methods, GRNN and PNN, but the overall results are worse because of the lower success in helix prediction.

The success rate of our study is better than Rost and Sander's study at the window size of 13 with 87.2% for 141 alpha set and with 91.7% for 146 beta set. They predicted the secondary structure of proteins at 70.8 % accuracy with two-layered feed forward neural network at the window size of 13.

4.3.1. 141 Alpha Set

The success levels of self-consistency test results are quite good for helix with 98.7%, for coil with 98.5% and for overall 98.7%. This is better from PNN, but worse from GRNN.

Table 4.5. Self-consistency and independent-dataset results for each window size

W	Helix		Coil		Overall	
	Train	Test	Train	Test	Train	Test
11	97.6	89.4	96.9	78.6	97.4	87.1
13	97.7	90.0	96.9	77.6	97.5	87.2
15	98.3	89.7	97.2	77.9	98.0	87.0
17	98.4	89.5	97.2	77.9	98.1	86.8
19	98.2	89.6	97.4	78.0	98.0	86.9
21	98.1	89.1	97.4	78.8	97.9	86.7
23	98.2	88.8	97.4	77.7	98.0	86.1
25	98.2	89.1	97.4	78.7	98.0	86.5

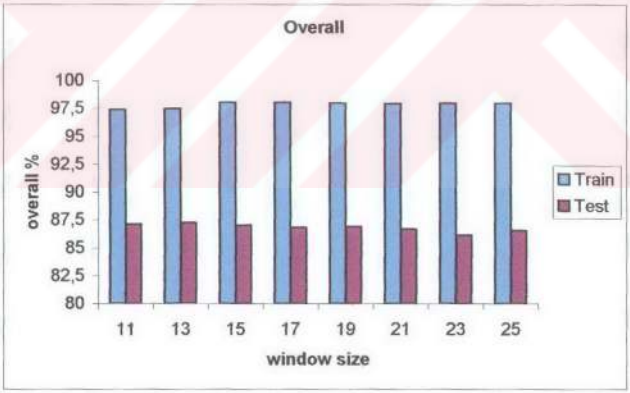

Figure 4.12. Overall results of 141 alpha set for each window size

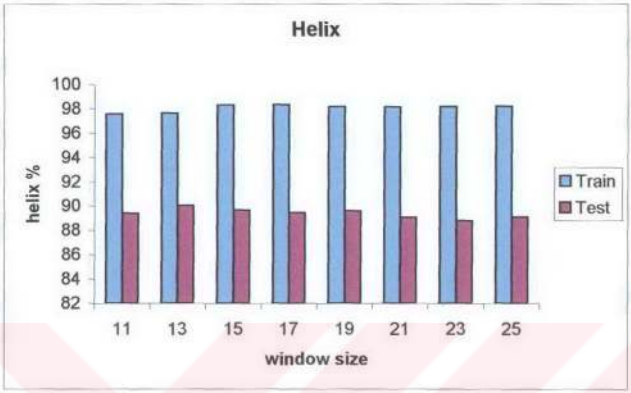

Figure 4.13. Helix results of 141 alpha set for each window size

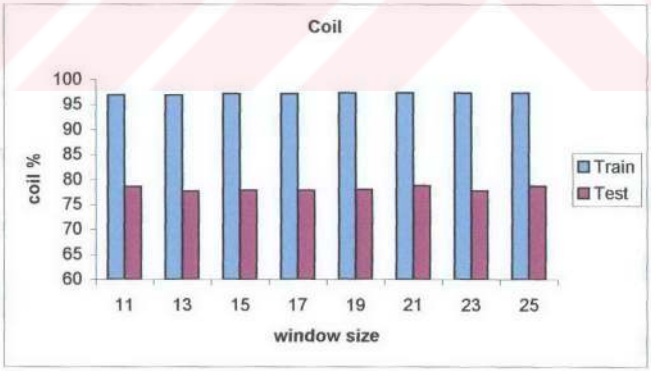

Figure 4.14. Coil results of 141 alpha set for each window size

The residue of independent dataset with the window size of 13 amino acids provides slightly better prediction of the secondary structure of protein in overall and helix prediction with 87.2% and 90%, respectively (Table 4.5.). The best prediction of coil occurs with the window size of 21. After window size of 13, size of 11 achieved the best result for overall prediction with percentage of 87.1 and then followed by window sizes of 15 (Figure 4.12.). The result shows the small sizes of windows are preferable.

4.3.2. 146 Beta Set

The Window size of 11 has given the best overall prediction for independent data set with 92%. The accuracy of this window is 82.2% for coil and 94.5% for helix. The best of the helix prediction is 85.6% with the 17-window size. This is 13 for coil with 93.7% (Table 4.6.). Like in 141 alpha results, small window sizes give better prediction successes.

Table 4.6. Self-consistency and independent-dataset results for each window size

W	Helix		Coil		Overall	
	Train	Test	Train	Test	Train	Test
11	98.6	94.5	95.6	82.2	97.9	92.0
13	98.0	93.7	98.0	83.3	98.0	91.7
15	98.7	93.4	97.1	80.3	98.4	90.8
17	98.5	93.0	98.2	85.6	98.6	91.6
19	98.5	92.7	98.5	85.3	98.5	91.4
21	99.0	92.9	97.0	84.0	98.6	91.2
23	98.7	93.3	98.2	82.4	98.6	91.1
25	98.7	92.4	98.5	83.9	98.7	90.8

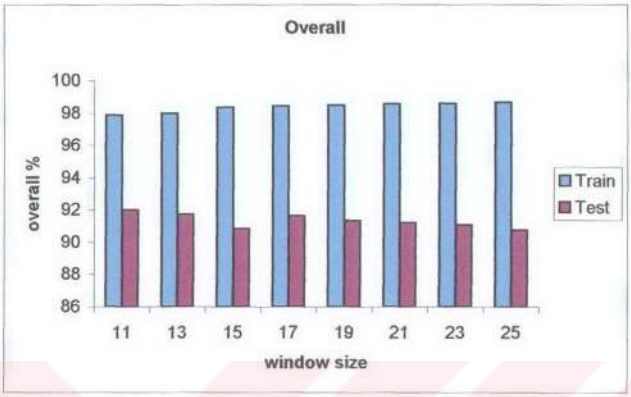

Figure 4.15. Overall results of 146 alpha set for each window size

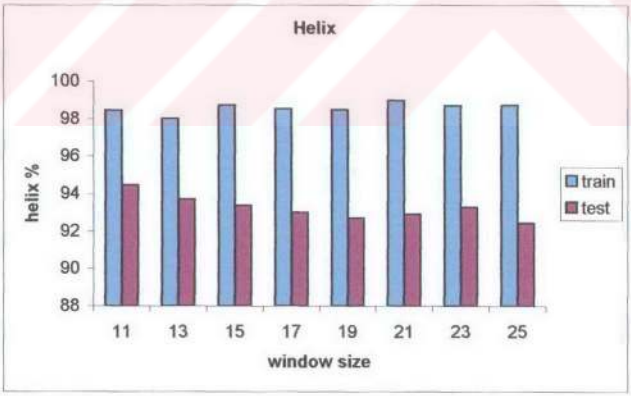

Figure 4.16. Helix results of 146 beta set for each window size

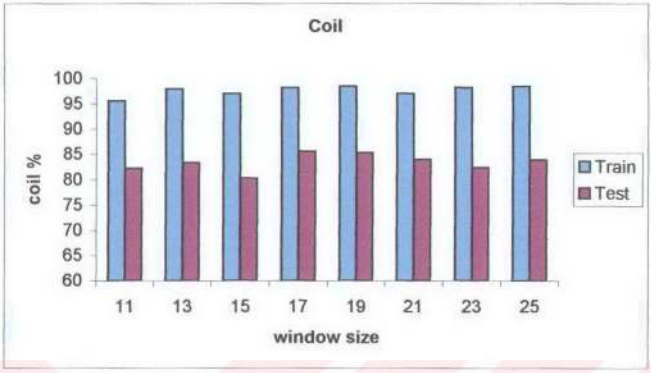

Figure 4.17. Coil results of 146 beta set for each window size

4.4. CNN Results

Generally, CNN has not given any improvement in comparing with other individual methods' accuracy results. The method gives the same result on 141 alpha set test as GRNN that has the best overall test result of 141 alpha set. In 146 beta set test, there is an unimportant improvement that can be neglected, with 0.5% for overall accuracy results. Rost et al. said that the individual methods were more successful than the combined one. They observed around 1%-2% improvement on accuracy rate in their studies.

4.4.1. 141 Alpha Set

The best result is obtained at the window size of 21 for the overall and the coil prediction. For helix, the best prediction occurs in the size of 11 and the accuracy is inversely proportional with the window size (figure 4.19.).

Table 4.7. Self-consistency and independent-dataset results for each window size

W	Helix	Coil	Overall
	Test	Test	Test
11	93.8	65.3	86.7
13	93.7	66.3	86.8
15	93.5	66.1	86.5
17	93.4	66.7	86.5
19	93.0	67.7	86.4
21	93.1	71.3	87.3
23	92.7	70.1	85.5
25	92.7	67.1	85.7

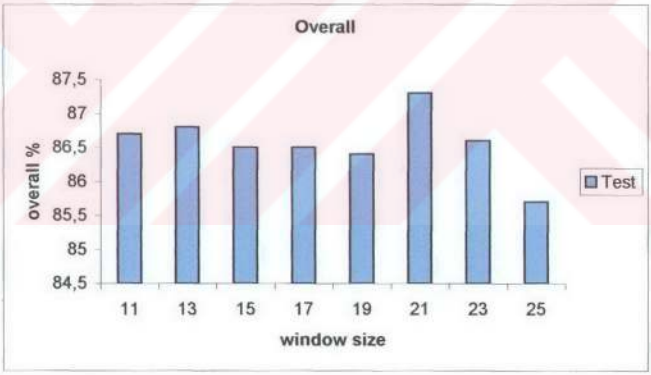

Figure 4.18. Overall results of 141 alpha set for each window size

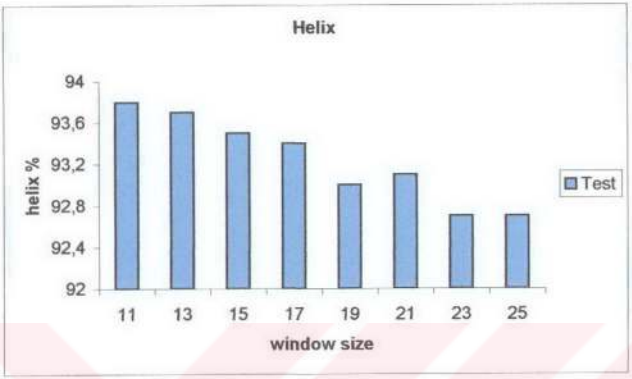

Figure 4.19. Helix results of 141 alpha set for each window size

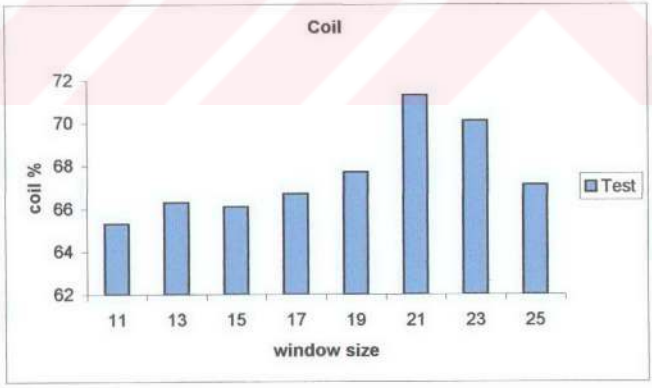

Figure 4.20. Coil results of 141 alpha set for each window size

4.4.2. 146 Beta Set

The overall results of 146 beta set test gives the similar distribution with the helix accuracy distribution as shown in Figure 4.21. and Figure 4.22. Both of them have the best prediction accuracy on 13-window size and the worse prediction is observed at the window size of 21.

The success rate at the window size of 19 is also very low for helix and overall accuracy prediction. Contrarily, coil prediction has the best result on the window size of 19.

Table 4.8. Self-consistency and independent-dataset results for each window size

W	Helix	Coil	Overall
	Test	Test	Test
11	97.2	74.3	92.5
13	97.3	74.6	92.6
15	96.9	72.8	91.9
17	97.2	73.0	92.1
19	94.4	75.0	90.2
21	93.9	74.1	89.6
23	96.9	73.3	91.7
25	96.9	72.6	91.2

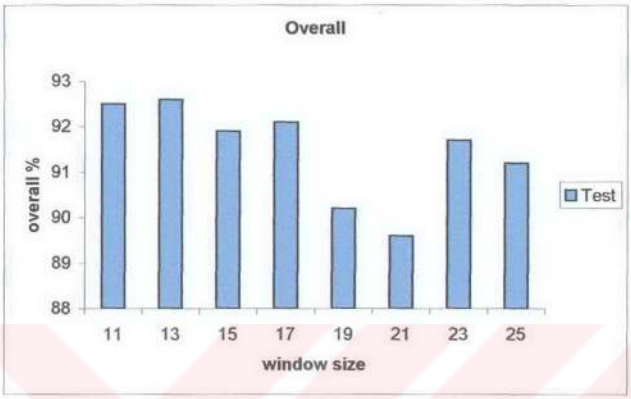

Figure 4.21. Overall results of 146 beta set for each window size

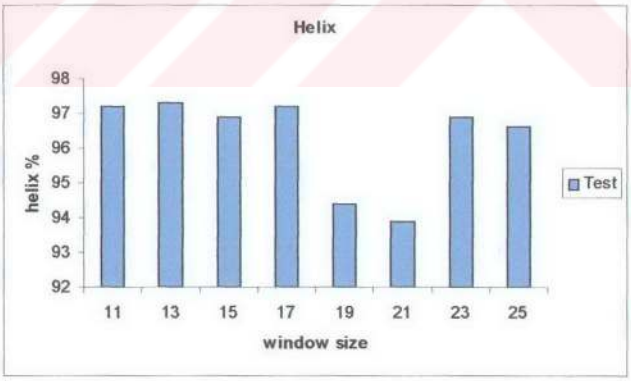

Figure 4.22. Helix results of 146 beta set for each window size

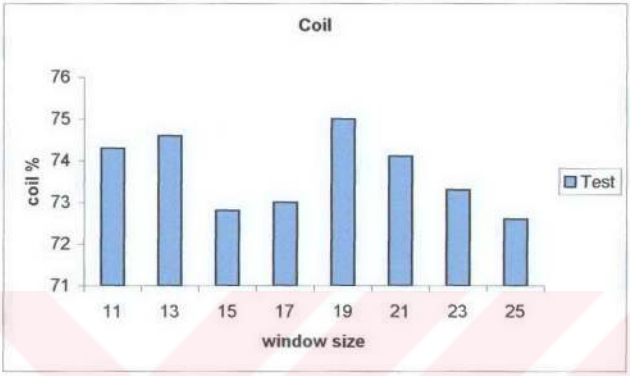

Figure 4.23. Coil results of 146 beta set for each window size

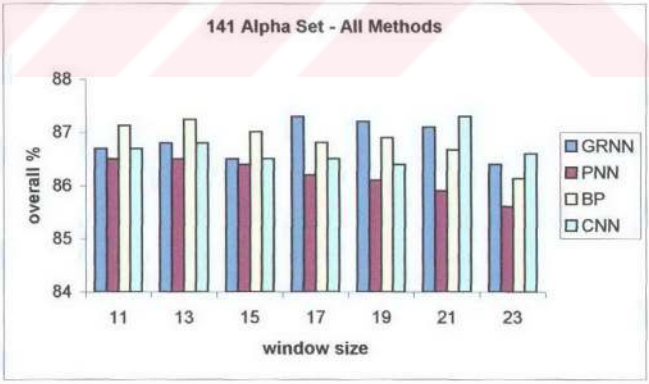

Figure 4.24. Overall results of 141 alpha set for each window size of all methods.

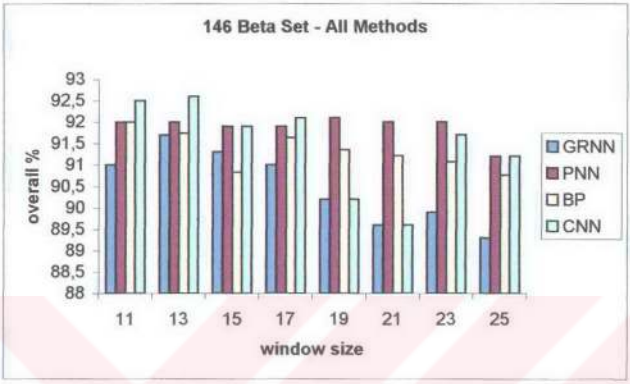

Figure 4.25. Overall results of 146 beta set for each window size of all methods.

As seen in Figure 4.24., GRNN and CNN give the best results in prediction of 141 alpha dataset at the window sizes of 17 and 21, respectively. For the window sizes of 11, 13 and 15, BP has the best prediction.

The dataset of 146 beta prediction has the best result at the window size of 13 with the CNN method. The CNN also wins at the window size of 11, 17 and 15. At the window size of 19, 21 and 23, the success rate of the PNN is the best (Figure 4.25.).

5. CONCLUSION

The problem of prediction of hemoglobin secondary structure is successfully handled by the networks, which have already applied to many daily problems.

All the self-consistency results that are obtained from the algorithms, GRNN, PNN, BP and CNN, are quite successful and generally stable results. Therefore, it cannot be easily said anything about which window size is better. In the independent dataset test results, the accuracies of the window sizes are slightly differ from each other. For 141 alpha set tests, the minimum overall result is 85.4% and the maximum one is 90.3%. This is min. 89.3% and max. 92.6% overall accuracy level for 146 beta set. In this respect, it is not logical to interpret about the window size accuracies. However, when comparing success of window sizes of independent data set results, it is seemed that the small sizes give the better results. According to the CNN results, there is no significant improvement comparing with GRNN, PNN and BP results. There is only 0.5% increase for overall accuracy in 146 beta set tests.

Because of their one-pass learning algorithm, GRNN and PNN have the advantageous of fast training time when it is compared with BP training time. Another advantage of the GRNN and the PNN is the fact that adding new samples to the training set does not require re-calibrating the model. Also erroneous samples are tolerated by other existing pattern nodes, this causes the good success. Another advantage that while GRNN and PNN have only one parameter, BP has lots of parameters. However, the main disadvantage of GRNN and PNN is that there is no intuitive method for selecting the optimal smoothing parameter. In this thesis, the value of σ is chosen as 0.3 by looking at the references [Niwa, 2003] [El-Solh et al., 1999]. There are no any advantages for CNN because it did not give any significant improvement in results. It also causes the loosing time.

REFERANCES

ANONYMOUS/a, 2002. http://sickle.bwh.harvard.edu/hemoglobin.html

ANONYMOUS/b, Cross Validation. www-2.cs.cmu.edu/~schneide/tut5/node42.html

ANONYMOUS/c, Alaska Project: Artificial Neural Networks. www.ccg.leeds.ac.uk/alaska/ann.html

ARGOS, P., SCHWARTZ, J., 1976. Biochim. Biophys. Acta., 439, 261–273.

BAKER, B. D., RICHARDS, C. E., 1999. A Comparison of Linear Regression and Neural Network Methods for Forecasting Educational Spending. Economics of Educational Review, International Journal of Economics Literature Citation (Report of Original Empirical Research), 4, 18, 405-416.

BISHOP, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press Inc., New York, 482.

BIOU, V., GIBRAT, J., LEVIN, J.M., ROBSON. B., GARNIER. J., 1988. Secondary Structure Prediction: Combination Of Three Different Methods. Protein Eng., 2, 185–191.

CAI, Y., LIU, X., XU, X., CHOU, K., 2002. Artificial Neural Network for Predicting Protein Secondary Structure Content. Computers and Chemistry, 26, 347-350.

CHANDONIA, J.M., KARPLUS, M., 1999. New Methods for Accurate Prediction of Protein Secondary Structure. Proteins, 35, 293-306.

CHOU, P.Y., FASMAN, G.D., 1978. Empirical Predictions of Protein Conformation. Annual Review Biochemistry, Vol. 47, pp.251-276.

CUFF, J.A., CLAMP, M.E., BARTON, G.J., 1998. JPred: A Consensus Secondary Structure Prediction Server. Bioinformatics, 14, 892-893.

CUFF, J.A., BARTON, G.J., 2000. Application of Enhanced Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction. Proteins: Structure, Function and Genetics, 40, 502-511.

DENNISTON, K.J., TOPPING, J.J., CARET, R.L., 2001. General, Organic and Biochemistry. Third International Edition, McGraw-Hill Companies Inc., New York.

DUDA, R.O., HART, P.E., STORK, D.G., 2000. Pattern Classification. Wiley, Second Edition, 654.

EISENHABER, F., PERSON, B., ARGOS, P., 1995. Protein Structure Prediction: Recognition of primary, secondary and tertiary structural features from amino acid sequences. Crit. Rev. Biochem. Mol. Biol., 30, 1-94.

EL-SOLH, A., HSIAO, C., GOODNOUGH, S., SERGHANI, J., GRANT, B.J., 1999. Predicting Active Pulmonary Tuberculosis Using an Artificial Neural Network. Chest., 4, 116, 968-73.

GOH, A.T.C., 2002. Probabilistic Neural Network For Evaluating Seismic Liquefaction Potential. Canadian Geotechnical Journal, Vol. 39, pp. 219-232. www.nrc.ca/cisti/journals/sample/t01-073.pdf

HAJMEER, M., BASHEER, I., 2002. A probabilistic Neural Network Approach for Modeling and Classification of Bacterial Growth / No-Growth Data. Journal of Microbiological Methods, 51, 217-226.

HAYKIN, S., 1994. Neural Networks. Prentice-Hall Inc., New Jersey, 696.

HU, Y., ZHANG, H., WIELICKI, B., STACKHOUSE, P., 2002. A Neural Network MODIS-CERES Narrowband to Broadband Conversion, IGARSS 2002 Conference Proceedings.
asd-www.larc.nasa.gov/~yhu/c_pub/I2C06_1592.PDF

HYUN, B., NAM, K., Faults Diagnoses of Rotating Machines by Using Neural Nets: GRNN and BPN. www.postech.ac.kr/ee/cmd/publication/iecon1995hyun.pdf

IBRIKCI, T., 2000. Neural Network Models for Secondary Protein Structure. PhD Thesis.

JORDAN, M.I., BISHOP, C.M., 1996. Neural Networks. CRC Handbook of Computer Science.

KABSCH, W., SANDER, C., 1983. Dictionary of Protein Secondary Structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22: 12, pp. 2577-637.

LIPPMANN, R.P., 1987. An Introduction to Computing with Neural Nets. Institute of Electrical and Electronics Engineers ASSP Magazine, 4-22.

LOW, R., TOGNERI, R., 1998. Speech Recognition Using the Probabilistic Neural Network. Proceedings of ICSLP98 (SST Student Day), Sydney, Australia, Paper No. 645.
ciips.ee.uwa.edu.au/~roberto/research/ speech/local/papers/tr98-01.pdf

MEHTA, B.V., MEHTA, S.B., RABELO, L.C., KOPCHICK, J.J., 1996. Artificial Intelligence Techniques for Analyzing the 3-D Structure of Proteins: Designing New Proteins. Journal of Biomedical Engineering Applications, Basis & Communications, Vol. 8., No. 6, pp. 556.

NISHIKAWA, K., OOI, T., 1986. Biochem. Biophys. Acta., 871, 45–54.

NIWA, T., 2003. Using General Regression and Probabilistic Neural Networks to Predict Human Intestinal Absorption with Topological Descriptors Derived From Two-Dimensional Chemical Structures. J. Chem. Inf. Comput. Sci., 43, 113-119.

PAULING, L., 1952. The Hemoglobin Molecule in Health and Disease. Proceedings of The American Philosophical Society, Vol. 96, No: 5, October, 555-565.

PETERSEN, T.N., LUNDEGAARD, C., NIELSEN, M., BOHR, H., BOHR, J., BRUNAK, S., GIPPERT, G.P., LUND. O., 2000. Prediction of Protein Secondary Structure at % 80 Accuracy. Proteins: Structure. Funct., and Genetics, Vol. 41, 17-20.

POLLASTRI, G., PRZYBYLSKI, D., ROST, B., BALDI, P., 2002. Improving The Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. Proteins, 47, 228-235.

PROTEIN DATA BANK, Chemistry Department, Brookhaven National Laboratory. Upton, NY 11973 USA, www.rcsb.org/pdb/

QIAN, N., SEJNOWSKI, T.J., 1988. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. J. Mol. Biol., Vol.202, pp. 865-884.

RAYKAR, C. V., Probability Density Function Estimation by Different Methods. cvl.umiacs.umd.edu/users/vikas/projects/enee739q_2.pdf

RIIS, S.K., KROGH, A., 1996. Improving Prediction of Protein Secondary Structure Using Structured Neural Networks and Multiple Sequence Alignments. J. Comp. Biol., 1, 3, 163-183.

RINSHO, N., 1998. Human Hemoglobin Structure and Respiratory Transport. Department of Human Genetics. National Institute of Genetics, 9, 54, 2320-2325.

ROST, B., SANDER, C., 1993a. Improved Prediction Of Protein Secondary Structure by Use of Sequence Profiles and Neural Networks. Proc. Natl. Acad. Sci., U S A, 90, 7558-62.

ROST, B., SANDER, C., 1993b. Prediction of Protein Secondary Structure at Better Than 70% Accuracy. J. Mol. Biol., 232, 584-599.

ROST, B., SANDER, C., 1994a. Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure. Proteins, 19, 55-72.

ROST, B., SANDER, C., SCHNEIDER, R., 1994b. Redefining The Goals of Protein Secondary Structure Prediction. J. Mol. Biol., 235, 13-26.

ROST, B., BALDI, P., BARTON, G., CUFF, J., EYRICH, V.A., JONES, D., KARPLUS, K., KING, R., POLLASTRI, G., PRZYBYLSKI, D., 2001. Simple Jury Predicts Protein Secondary Structure Best. New York: Columbia University, Preprint Cubic_2001_10.

RÖGNVALDSSON, T. S., Overfitting and Generalization. http://www.hh.se/staff/denni/SLS_stuff/lecture_09_generalization.pdf

RZEMPOLUCK, E.J., 1997. Neural Network Classification of EEG During Camouflaged Object Identification. International Journal of Medical Informatics, 44, 169-175.

SALAMOV, A.A., SOLOVYEV, V.V., 1997. Protein Secondary Structure Prediction Using Local Alignments. J. Mol. Biol., 268, 31-6.

SARLE, W.S., 1994. Neural Networks and Statistical Models. Proceedings of the Nineteenth Annual SAS Users Group International Conference.

SHIRASAWA, T., IZUMIZAKI, M., SUZUKI, Y., ISHIHARA, A., SHIMIZU, T., TAMAKI, M., HUANG, F., KOIZUMI, K., IWASE, M., SAKAI, H., TSUCHIDA, E., UESHIMA, K., INOUE, H., KOSEKI, H., SENDA, T., KURIYAMA, T., HOMMA, I., 2003. Oxygen Affinity of Hemoglobin Regulates O_2 Consumption. Metabolism and Physical Activity, The Journal of Biological Chemistry, Vol. 278, No: 7, pp. 5035-5043.

SOLOMONS, T.W.G., 1995. Organic Chemistry. Fifth Edition. John Wiley & Sons Inc, New York.

SPECHT, D., 1990. Probabilistic Neural Networks. Neural Networks, Vol. 3, pp. 109-118.

SPECHT, D., 1991. A General Regression Neural Network. IEEE Trans. Neural Networks, Vol. 2, No 6, pp. 568-576.

STERGIOU, C., Neural Networks: The Human Brain and Learning. www.doc.ic.ac.uk/~nd/surprise_96/journal/vol2/cs11/article2.html.

STERNBERG, M.J.E., 1983. Computing in Biological Science. Elsevier Biomedical Press.

WALSH, B., 2000. Introduction to Bayesian Analysis. nitro.biosci.arizona.edu/courses/EEB600A/download/Bayesian.pdf

YEUNG, D., CHOW, C., 2002. Parzen Window Network Intrusion Detectors. Proceedings of the Sixteenth International Conference on Pattern Recognition. Quebec City, Canada, Vol.4, pp.385-388. www.cs.ust.hk/~dyyeung/paper/pdf/yeung.icpr2002.pdf

ZHANG, X., MESIROV, J.P., WALTZ, D.L., 1992. Hybrid System For Protein Secondary Structure Prediction. J. Mol. Biol., 225, 1049–1063.

ZHANG, Z., SUN, Z., ZHANG, C., 2001. A New Approach to Predict the Helix/Strand Content of Globular Proteins. J. Theor. Biol, 208, 65-78.

BIOGRAPHY

I was born on Nov. 19th. 1975 in Ankara, TÜRKİYE. After completing my primary (1982-1987), secondary (1987-1990) and high school (1990-1993) education at T.E.D. Ankara College. I received my B.S. degree from Hacettepe University in the field of Physics Engineering.

After that I worked for T.E.D. Ankara College as a Science Education Project Assistant. Then I started as a master student at the Electrical and Electronics Engineering Department of Çukurova University. Meanwhile I have been working as a Preparing and Controlling Data Operator at the Student's Affairs Office of the Çukurova University.

APPENDIX

The amino acids names can be abbreviated by the three-letter and the one-letter code. These abbreviations are shown in Append. Table 1.:

Amino Acid	Three-Letter Abbreviation	One-Letter Abbreviation
ALANINE	ALA	A
ARGININE	ARG	R
ASPARAGINE	ASN	N
ASPARTIC ACID	ASP	D
CYSTEINE	CYS	C
GLUTAMINE	GLN	Q
GLUTAMIC ACID	GLU	E
GLYCINE	GLY	G
HISTIDINE	HIS	H
ISOLEUCINE	ILE	I
LEUCINE	LEU	L
LYSINE	LYS	K
METHIONINE	MET	M
PHENYLALANINE	PHE	F
PROLINE	PRO	P
SERINE	SER	S
THREONINE	THR	T
TRYPTOPHAN	TRP	W
TYROSINE	TYR	Y
VALINE	VAL	V

Append. Table 1. Three-letter and one-letter abbreviations of the α-amino acids names

In Append. Table 2., the upper lines of the chains show the primary and the bottom lines of the chains show the secondary structures:

α-Chain, HbA

```
  1 V L S P A D K T N V K A A W G K V G A H A G E Y G A E A L E R M F L S F P T T K T Y F P H F D L S H
      H H H H H H H H H H H H H H G G G H H H H H H H H H H H H H H H   G G G G G G T T T S     S T

 51 G S A Q V K G H G K K V A D A L T N A V A H V D D M P N A L S A L S D L H A H K L R V D P V N F K L
    T   H H H H H H H H H H H H H H H H H H H H T T T S H H H H T H H H H H H H H H T T       T H H H H H

101 L S H C L L V T L A A H L P A E F T P A V H A S L D K F L A S V S T V L T S K Y R
    H H H H H H H H H H H H T T T T T T H H H H H H H H H H H H H H H H H H T T T
```

β-Chain, HbA

```
  1 M H L T P E E K S A V T A L W G K V N V D E V G G E A L G R L L V V Y P W T Q R F F E S F G D L S T
        H H H H H H H H H H T T     H H H H H H H H H H H H H H H S G G G G G G T T T S       S S

 51 P D A V M G N P K V K A H G K K V L G A F S D G L A H L D N L K G T F A T L S E L H C D K L H V D P
    H H H H H T H H H H H H H H H H H H H H H H H H T T T T T   H H H H T H H H H H H H H H T S       T

101 E N F R L L G N V L V C V L A H H F G K E F T P P V Q A A Y Q K V V A G V A N A L A H K Y H
    H H H H H H H H H H H H H H H H H H G G G S   H H H H H H H H H H H H H H H H H H H H T T
```

Append. Table 2. Globin sequence of normal adult hemoglobin