VOT 74017

PROTEIN SECONDARY STRUCTURE PREDICTION FROM AMINO ACID SEQUENCE USING ARTIFICIAL INTELLIGENCE TECHNIQUE

(MERAMAL STRUKTUR PROTEIN SEKUNDER DARI JUJUKAN ASID AMINO MENGGUNAKAN TEKNIK KEPINTARAN BUATAN)

SAFAAI BIN DERIS
ROSLI BIN MD ILLIAS
SAHIDAN BIN SENAFI
SAAD OSMAN ABDALLA
SATYA NANDA VEL ARJUNAN

RESEARCH VOT NO:
74017

Jabatan Kejuruteraan Perisian
Fakulti Sains Komputer Dan Sistem Maklumat
Univerisit Teknologi Malaysia

2007

ABSTRACT

Large genome sequencing projects generate huge number of protein sequences in their primary structures that is difficult for conventional biological techniques to determine their corresponding 3D structures and then their functions. Protein secondary structure prediction is a prerequisite step in determining the 3D structure of a protein. In this research a method for prediction of protein secondary structure has been proposed and implemented together with other known accurate methods in this domain. The method has been discussed and presented in a comparative analysis progression to allow easy comparison and clear conclusions. A benchmark data set is exploited in training and testing the methods under the same hardware, platforms, and environments. The newly developed method utilizes the knowledge of the GORV information theory and the power of the neural network to classify a novel protein sequence in one of its three secondary structures classes. NN-GORV-I is developed and implemented to predict proteins secondary structure using the biological information conserved in neighboring residues and related sequences. The method is further improved by a filtering mechanism for the searched sequences to its advanced version NN-GORV-II. The newly developed method is rigorously tested together with the other methods and observed reaches the above 80% level of accuracy. The accuracy and quality of prediction of the newly developed method is superior to all the six methods developed or examined in this research work or that reported in this domain. The Mathews Correlation Coefficients (MCC) proved that NN-GORV-II secondary structure predicted states are highly related to the observed secondary structure states. The NN-GORV-II method is further tested using five DSSP reduction schemes and found stable and reliable in its prediction ability. An additional blind test of sequences that have not been used in the training and testing procedures is conducted and the experimental results show that the NN-GORV-II prediction is of high accuracy, quality, and stability. The Receiver Operating Characteristic (ROC) curve and the area under curve (AUC) are applied as novel procedures to assess a multi-class classifier with approximately 0.5 probability of one and only one class. The results of ROC and AUC prove that the NN-GOR-V-II successfully discriminates between two classes; coils and not-coils.

ABSTRAK

Projek-projek *genome* yang berskala besar telah menghasilkan jujukan-jujukan protein dalam bentuk struktur pertama yang sangat banyak bilangannya telah menyebabkan teknik-teknik biasa biologi sukar untuk menuntukan struktur 3D dan fungsinya. Peramalan struktur kedua protein diperlukan bagi menentukan struktur 3D protein dan fungsinya. Dalam tesis ini, satu kaedah untuk meramalkan struktur kedua protein telah dicadangkan dan dilaksanakan bersama-sama dengan kaedah-kaedah lain yang berkaitan. Kaedah itu telah dibincangkan dan ditunjukkan di dalam satu analisis perbandingan. Tujuh algoritma dan kaedah bagi peramalan struktur kedua protein telah dibangunkan dan dilaksanakan. Satu set data perbandingan digunakan untuk melatih dan menguji kaedah tersebut. Kaedah yang baru dibangunkan itu adalah menggunakan pengetahuan Teori Maklumat GORV dan Rangkaian Neural untuk mengkelaskan satu jujukan protein baru kepada salah satu daripada 3 kelas stuktur keduanya. NN-GORV-I dibangunkan dan diimplemenkan bagi meramal struktur kedua protein menggunakan maklumat biologi yang disimpan dalam bentuk keladak yang berhampiran dan jujukan-jujukan yang berkaitan. Seterusnya kaedah itu telah diuji dengan kaedah-kaedah lain dan telah mencapai lebih 80% ketepatan. Ketepatan dan kualiti peramalan bagi kaedah itu adalah melebihi 6 kaedah- kaedah lain yang juga telah dibangunkan dan diperiksa dalam penyelidikan ini. Pekali Korelasi Mathews (PKM) telah membuktikan struktur kedua yang telah diramalkan oleh NN-GORV-II adalah sangat berkait rapat dengan keadaan struktur kedua yang telah dicerapkan. Kaedah NN-GORV-II seterusnya diuji dengan menggunakan lima skema potongan DSSP dan disahkan kestabilannya dan boleh dipercayai kebolehannya untuk kerja peramalan tersebut. Satu penambahan ujian bagi jujukan-jujukan yang tidak digunakan dalam prosedur melatih dan menguji dijalankan dan hasil-hasil eksperimennya menunjukkan bahawa peramalan NN-GORV-II adalah berketepatan tinggi, berkualiti dan stabil. Lengkungan *Receiver Operating Characteristic* (ROC) dan *area under curve* (AUC) itu telah diaplikasikan sebagai satu prosedur baru bagi menilai pengkelas pelbagai kelas dengan anggaran kebarangkalian adalah 0.5 bagi satu dan hanya satu kelas. Hasil-hasil bagi ROC dan AUC membuktikan bahawa NN-GOR-V berjaya memisahkan 2 kelas; lingkaran dan bukan lingkaran.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	ABSTRACT	ii
	ABSTRAK	iii
	TABLE OF CONTENTS	iv
	LIST OF TABLES	x
	LIST OF FIGURES	xii
	LIST OF ABBREVIATIONS	xv
	LIST OF APPENDICES	xviii
1	**INTRODUCTION**	**1**
	1.1 Introduction	1
	1.2 Protein Structure Prediction	2
	1.3 Prediction Methods	3
	1.4 The Problem	6
	1.5 Objectives of the Research	7
	1.6 The Scope of the Research	8
	1.7 Organization and Overview of the Report	8
	1.8 Summary	10
2	**PROTEIN, SEQUENCES, AND SEQUENCE ALIGNMENT**	**11**
	2.1 Introduction	11
	2.2 Proteins	11
	2.2.1 Protein Primary Structure	15
	2.2.2 Secondary Structure	15

2.2.3 Tertiary Structure 17

2.2.4 Quaternary Structure 18

2.3 Methods of Determining Protein Structure 18

2.4 Characteristics of Protein Structures 20

2.5 Protein Homology 21

2.5.1 Types of Homologies 22

2.5.2 Homologues versus Analogues 22

2.6 Molecular Interactions of Proteins 23

2.7 Sequence Alignment Methods 24

2.7.1 Threading Methods 24

2.7.2 Hidden Markov Models 25

2.7.3 Types of Alignment Methods 26

2.7.3.1 Pairwise Alignment Methods 27

2.7.3.2 Profile Alignment Methods 29

2.7.3.3 Multiple Alignment Methods 30

2.7.4 Comparative Modelling 32

2.7.5 Overview of Alignment Methods and Programs 33

2.8 Summary 35

3 REVIEW OF PROTEIN SECONDARY STRUCTURE PREDICTION: PRINCIPLES, METHODS, AND EVALUATION 36

3.1 Introduction 36

3.2 Protein Secondary Structure Prediction 38

3.3 Methods Used In Protein Structure Prediction 40

3.4 Artificial Neural Networks 47

3.4.1 Inside the Neural Networks 47

3.4.2 Feedforward Networks 49

3.4.3 Training the Networks 51

3.4.4 Optimization of Networks 52

3.5 Information Theory 54

3.5.1 Mutual Information and Entropy 55

3.5.2 Application of Information Theory to Protein 57

Folding Problem
3.5.3 GOR Method for Protein Secondary Structure Prediction 59
3.6 Data Used In Protein Structure Prediction 61
3.7 Prediction Performance (Accuracy) Evaluation 63
3.7.1 Average Performance Accuracy (Q3) 64
3.7.2 Segment Overlap Measure (SOV) 65
3.7.3 Correlation 65
3.7.4 Receiver Operating Characteristic (ROC) 66
3.7.5 Analysis of Variance Procedure (ANOVA) 67
3.8 Summary 68

4 METHODOLOGY 70

4.1 Introduction 70
4.2 General Research Framework 70
4.3 Experimental Data Set 74
4.4 Hardware and Software Used 75
4.5 Summary 76

5 A METHOD FOR PROTEIN SECONDARY STRUCTURE PREDICTION USING NEURAL NETWORKS AND GOR-V 77

5.1 Introduction 77
5.2 Proposed Prediction Method – NN-GORV-I 78
5.2.1 NN-I 78
5.2.2 GOR-IV 78
5.2.3 Multiple Sequence Alignments Generation 79
5.2.4 Neural Networks (NN-II) 81
5.2.4.1 Mathematical Representation of Neural Networks 81
5.2.4.2 Generating the Networks 86
5.2.4.3 Networks Optimization 88
5.2.4.4 Training and Testing the Network 89
5.2.5 GOR-V 91

5.2.6 NN-GORV-I 94
5.2.7 Enhancement of Proposed Prediction Method - N-GORV-II 100
5.2.8 PROF 102
5.3 Reduction of DSSP Secondary Structure States 103
5.4 Assessment of Prediction Accuracies of the Methods 105
5.4.1 Measure of Performance (Q_H, Q_E, Q_C, and Q_3) 105
5.4.2 Segment Overlap (SOV) Measure 106
5.4.3 Matthews Correlation Coefficient (MCC) 106
5.4.4 Receiver Operating Characteristic (ROC) 107
5.4.4.1 Threshold Value 109
5.4.4.2 Predictive Value 109
5.4.4.3 Plotting ROC Curve 110
5.4.4.4 Area Under Curve (AUC) 110
5.4.5 Reliability Index 112
5.4.6 Test of Statistical Significance 112
5.4.6.1 The Confidence Level (P-Value) 113
5.4.6.2 Analysis of Variance (ANOVA) Procedure 114
5.5 Summary 114

6 ASSESSMENT OF THE PREDICTION METHODS 116

6.1 Introduction 116
6.2 Data Set Composition 117
6.3 Assessment of GOR IV Method 118
6.4 Assessment of NN-I Method 122
6.5 Assessment of GOR-V Method 123
6.6 Assessment of NN-II Method 126
6.7 Assessment of PROF Method 128
6.7.1 Three States Performance of PROF Method 130
6.7.2 Overall Performance and Quality of PROF Method 132
6.8 Assessment of NN-GORV-I Method 134
6.8.1 Three States Quality (SOV) of NN-GORV-I 136

Method
6.8.2 Overall Performance and Quality of NN-GORV-I Method 139
6.9 Assessment of NN-GORV-II Method 140
6.9.1 Distributions and Statistical Description of NN-GORV-II Prediction 140
6.9.2 Comparison of NN-GORV-II Performance with Other Methods 143
6.9.3 Comparison of NN-GORV-II Quality with Other Methods 148
6.9.4 Improvement of NN-GORV-II Performance over Other Methods 151
6.9.5 Improvement of NN-GORV-II Quality over Other Methods 155
6.9.6 Improvement of NN-GORV-II Correlation over Other Methods 156
6.10 Summary 158

7 THE EFFECT OF DIFFERENT REDUCTION METHODS 160

7.1 Introduction 160
7.2 Effect of Reduction Methods on Dataset and Prediction 161
7.2.1 Distribution of Predictions 162
7.2.2 Effect of Reduction Methods on Performance 166
7.2.3 Effect of Reduction Methods on SOV 169
7.2.4 Effect of Reduction Methods on Matthews's Correlation Coefficients 171
7.3 Summary 173

8 PERFORMANCE OF BLIND TEST 174

8.1 Introduction 174
8.2 Distribution of CASP Targets Predictions 175
8.3 Performance and Quality of CASP Targets Predictions 179
8.4 Summary 183

9 RECEIVER OPERATING CHARACTERISTIC (ROC) TEST 184

9.1 Introduction 184

9.2 Binary Classes and Multiple Classes 185
9.3 Assessment of NN-GORV-II 189
9.4 Summary 193

10 CONCLUSION 194

10.1 Introduction 194
10.2 Summary of the Research 195
10.3 Conclusions 197
10.4 Contributions of the Research 199
10.5 Recommendations for Further Work 199
10.6 Summary 201

REFERENCES 202

APPENDIX A (PROTEIN STRUCTURES) 230
APPENDIX B (CUFF AND BARTON'S 513 PROTEIN DATA SET) 233
APPENDIX C (DESCRIPTIVE STATISTICS) 244
APPENDIX D (SELECTED PUBLICATIONS) 246

LIST OF TABLES

TABLE NO.	TITLE	PAGE
2.1	The twenty types of amino acids that forms the proteins	12
2.2	The standard genetic code	14
3.1	Well established protein secondary structure prediction methods with their reported accuracies and remarks briefly describing each method.	46
5.1	The contingency table or confusion table for ROC curve	108
5.2	ANOVA table based on individual observations (One way ANOVA)	114
6.1	Total number of secondary structures states in the data base	118
6.2	The percentages of prediction accuracies with the standard deviations of the seven methods	120
6.3	The SOV of prediction accuracies with the standard deviations of the seven methods	121
6.4	The Mathew's correlation coefficients of predictions of the seven methods	122
6.5	Descriptive Statistics of the prediction accuracies of NN-GORV-II method	142
6.6	Descriptive Statistics of the prediction of SOV measure for NN-GORV-II method	142
6.7	Percentage Improvement of NN-GORV-II method over the other six prediction methods	152
6.8	SOV percentage improvement of NN-GORV-II method over the other prediction methods	155

6.9	Matthews Correlation Coefficients improvement of NN-GORV-II method over the other six prediction methods	157
7.1	Percentage of secondary structure state for the five reduction methods of DSSP definition (83392 residues)	162
7.2	The analysis of variance procedure (ANOVA) of the Q_3 for the five reduction methods	163
7.3	The analysis of variance procedure (ANOVA) of SOV for the five reduction methods	164
7.4	The effect of the five reduction methods on the performance accuracy of prediction (Q_3) the of NN-GORV-II prediction method	167
7.5	The effect of the five reduction methods on the segment overlap measure (SOV) of the NN-GORV-II prediction method	169
7.6	The effect of reduction methods on Matthews's correlation coefficients using NN-GORV-II prediction method	172
8.1	Percentages of prediction accuracies for the 42 CASP3 proteins targets	180
8.2	Percentages of SOV measures for the 42 CASP3 proteins targets	181
8.3	The mean of Q3 and SOV with and standard deviation, and Mathew's Correlation Coefficients (MCC) of CASP	182
9.1	The contingency table or confusion matrix for coil states prediction	187
9.2	The cut scores for the NN-GORV-II algorithm considering coil only prediction	189
9.3	The cut scores, true positive rate (TPR), false positive rate (FPR), and area under ROC (AUC) for the NN-GORV-II prediction algorithm considering coil state only prediction	191

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
3.1	Basic graphical representations of a block diagram of a single neuron artificial neural networks.	48
3.2	Representation of multilayer perceptron artificial neural networks.	50
4.1	General framework for protein secondary structure prediction method	72
4.2	An example of a flat file of CB513 data base used in this research, 1ptx-1-AS.all file.	75
5.1	Basic representation of multilayer perceptron artificial neural network	82
5.2	The sigmoidal functions usually used in the feedforward Artificial Network. (a) Hyperbolic tangent sigmoid transfer function or bipolar function (b) Log sigmoid transfer function or uniploar function	83
5.3	A general model for the newly developed protein secondary structure prediction method.	95
5.4	A detailed representation for the first version of the newly developed protein secondary structure prediction method (NN-GORV-I)	96
5.5	A detailed representation for the second version of the newly developed protein secondary structure prediction method (NN-GORV-II)	101
5.6	The 1ptx-1-AS.all file converted into a FASTA format (zptAS.fasta) readable by the computer programs.	104
5.7	The 1ptx-1-AS.all file parsed into a format readable by	105

	the Q3 and SOV program	
5.8	A typical example of area under curve (AUC) for training data, test data, and chance performance or random guess	111
6.1	The performance of the GOR-IV prediction method with respect to Q3 and SOV prediction measures	119
6.2	The performance of the NN-I prediction method with respect to Q3 and SOV prediction measures	123
6.3	The performance of the GOR-V prediction method with respect to Q_3 and SOV prediction measures	124
6.4	The performance of the NN-II prediction method with respect to Q3 and SOV prediction measures	127
6.5	The performance of the PROF prediction method with respect to Q3 and SOV prediction measures	129
6.6	The α helices performance (Q_H) of the seven prediction methods	130
6.7	The β strands performance (Q_E) of the seven prediction methods	130
6.8	The coils performance (Q_C) of the seven prediction methods	132
6.9	The performance of the NN-GORV-I prediction method with respect to Q3 and SOV prediction measures	135
6.10	The helices segment overlap measure (SOV_H) of the seven prediction methods	137
6.11	The strands segment overlap measure (SOV_E) of the seven prediction methods	137
6.12	The coils segment overlap measure (SOV_C) of the seven prediction methods	138
6.13	The performance of the NN-GORV-II prediction method with respect to Q_3 and SOV prediction measures	141

6.14	Histogram showing the Q_3 performance of the seven prediction methods	144
6.15	A graph line chart for the Q_3 performance of the seven prediction methods.	147
6.16	Histogram showing the SOV measure of the seven prediction methods	148
6.17	A graph line chart for the SOV measure of the seven prediction methods	150
7.1	Five histograms showing the Q3 distribution of the test proteins with respect to the five reduction methods	165
7.2	Five histograms showing the SOV distribution of the test proteins with respect to the five reduction methods	166
7.3	The performance accuracy (Q_3) of the five reduction method on the test proteins	168
7.4	The SOV measure of the five reduction method on the 480 proteins using NN-GORV-II prediction method	171
8.1	The distribution of prediction actuaries of the of the 42 Casp targets blind test for the secondary structure states.	176
8.2	The performance of the 42 CASP targets with respect to Q3 and SOV prediction measures	177
8.3	The distribution of SOV measure of the of the 42 Casp targets blind test for the secondary structure states.	178
9.1	An idealized curve showing the (TP, TN, FP, and FN) numbers of a hypothetical normal and Not normal observations	188
9.2	The cut scores of the coils and not coils secondary structure states predicted by the NN-GORV-II algorithm using Method V reduction scheme.	190
9.3	The area under ROC (AUC) for the NN-GORV-II prediction algorithm considering coil only prediction.	192

LIST OF ABBREVIATIONS

1D	-	One Dimensional Protein Structure
3D	-	Three Dimensional Protein Structure
HGP	-	Human Genome Project
GenBank	-	Gene Bank
PDB	-	Protein Data Bank
EMBL	-	European Molecular Biology Laboratory
DNA	-	Deoxyribonucleic Acid
RNA	-	Ribonucleic Acid
mRNA	-	Messenger RNA
NMR	-	Nuclear Magnetic Resonance
GOR	-	Garnier-Osguthorpe-Robson
BLAST	-	Basic Local Alignment Search Tool
PSIBLAST	-	Position Specific Iterated Blast
ROC	-	Receiver Operating Characteristic
AUC	-	Area Under Curve
NN-GORV-I	-	Neural Network GOR V Version 1 Prediction Method
NN-GORV-II	-	Neural Network GOR V Version 2 Prediction Method
Q_3	-	Prediction Accuracy of Helices, Strands, And Coils
Q_H	-	Prediction Accuracy of Helix State
Q_E	-	Prediction Accuracy of Strand State
Q_C	-	Prediction Accuracy of Coil State
SOV_3	-	Segment Overlap Measure Of Helices, Strands, And Coils
SOV_H	-	Segment Overlap Measure Of Helix State
SOV_E	-	Segment Overlap Measure Of Strand State
SOV_C	-	Segment Overlap Measure Of Coil State
MCC	-	Matthews Correlations Coefficient
NN	-	Neural Network
CASP	-	Critical Assessment Of Techniques For Protein Structure Prediction

RF	-	Radio Frequency Pulses
CE	-	Combinatorial Extension
FSSP	-	Database F Families Of Structurally Similar Proteins
SCOP	-	Structural Classification Of Proteins
HMMs	-	Hidden Markov Models
FASTA	-	Fast Alignment
GenThreader	-	Genomic Sequences Threading Method
MSA	-	Multidimensional Sequence Alignments
PRINTS	-	Protein Fingerprints
PRODOM	-	Protein Domain
PROF	-	Profile Alignment
PSSM	-	Position Specific Scoring Matrix
PRRP	-	Prolactin-Releasing Peptide
SCANPS	-	Protein Sequence Scanning Package
PHD	-	Profile Network From Heidelberg
DSSP	-	Dictionary Of Protein Secondary Structure Prediction
SAM	-	Sequence Alignment Method
MULTALIGN	-	Multiple Alignment
MULTAL	-	Multiple Alignment
HMMT	-	Hidden Markov Model Training For Biological Sequences
BAliBASE	-	Benchmark Alignments Database
PIM	-	Protein Interaction Maps
ITERALIGN	-	Iteration Alignment
MLP	-	Multi-Layer Perceptron
MI	-	Mutual Information
H	-	α Helix
E	-	β Strand
C	-	Coil
CPU	-	Central Processing Unit
RCSB	-	Research Collaboratory For Structural Bioinformatics
PDB	-	Protein Data Bank
NNSSP	-	Nearest-Neighbor Secondary Structure Prediction
DSC	-	Discrimination Of Protein Secondary Structure Class

3Dee	-	Database Of Domain Definitions (DDD)
CB513	-	Cuff And Barton 513 Proteins
TP	-	True Positive
TN	-	True Negative
FP	-	False Positive
FN	-	False Negative
ANOVA	-	Analysis Of Variance
nr	-	Non Redundant Database
PERL	-	Practical Extraction And Reporting Language
RES	-	Residues
LMS	-	Least Mean Square
SNNS	-	Stuttgart University Neural Network Simulator
ANSI	-	American National Standards Institute
RI	-	Reliability Index
FTP	-	File Transfer Protocol
SPSS	-	Statistical Package For Social Sciences
SAS	-	Statistical Analysis Software
SE	-	Standard Error
PIR	-	Protein Information Resource

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Protein Structures	230
B	Cuff and Barton's 513 Protein Data Set	233
C	Descriptive Statistics	244
D	Selected Publications	246

CHAPTER 1

INTRODUCTION

1.1 Introduction

Advances in molecular biology in the last few decades, and the availability of equipment in this field have allowed the increasingly rapid sequencing of considerable genomes of several species. In fact, to date, several bacterial genomes, as well as those of some simple eukaryotic organisms (e.g. yeast) have been completely sequenced. The Human Genome Project (HGP), aimed to sequence all of the human chromosomes, is almost completed with a rough draft announced in the year 2000 (Heilig *et al.*, 2003). Known sequencing databases projects, such as GenBank, PDB, and EMBL, have been growing significantly. This surge and overflow of data and information have imposed the rational storage, organization and indexing of sequence information.

Explaining the tasks undertaken in Bioinformatics field in details might be far beyond this introductory chapter. However, they fall in the creation and maintenance of databases of biological information with nucleic acid or protein sequences cover the majority of such databases. Storage and organization of millions of nucleotides is essential portion in these databases. Designing, developing, and implementing databases access and exchange information between researchers in this field is progressing significantly.

The most fundamental tasks in bioinformatics include the analysis of sequence information which involves the following the prediction of the 3D structure

of a protein using algorithms that have been derived from the knowledge of physics, chemistry and from the analysis of other proteins with similar amino acid sequences. Some researchers refer to this area with the name Computational Biology.

1.2 Protein Structure Prediction

Protein structure prediction is categorized under Bioinformatics which is a broad field that combines many other fields and disciplines like biology, biochemistry, physics, statistics, and mathematics. Proteins are series of amino acids known as polymers linked together into contiguous chains. In a living cell the DNA of an organism encodes its proteins into a sequence of nucleotides (transcribed), namely: adenine, cytosine, guanine and thymine that are copied to the mRNA which are then translated into protein (Branden and Tooze, 1991)

Protein has three main structures: primary structure which is essentially the linear amino acid sequence and usually represented by a one letter notation. Alpha helices, beta sheets, and loops are formed when the sequences of primary structures tend to arrange themselves into regular conformations; these units are known as secondary structure (Pauling and Corey, 1951; Kendrew, 1960). Protein folding is the process that results in a compact structure in which secondary structure elements are packed against each other in a stable configuration. This three-dimensional structure of the protein is known as the protein tertiary structure. However, loops usually serve as connection points between alpha-helices and beta-sheets, they do not have uniform patterns like alpha-helices and beta-sheets and they could be any other part of the protein structure rather than helices or strands (Appendix A).

In the molecular biology laboratory, protein secondary structure is determined experimentally by two lengthy methods: X-ray crystallography method and Nuclear Magnetic Resonance (NMR) spectroscopy method.

Since Anfinsen (1973) concluded that the amino acid sequence is the only source of information to survive the denaturing process, and hence the structured

information must be somehow specified by the primary protein sequence, researchers have been trying to predict secondary structure from protein sequence. Anfinsen's hypothesis suggests that an ideal theoretical model of predicting protein secondary structure from its sequence should exist anyhow.

1.3 Prediction Methods

There are two main different approaches in determining protein structure: a molecular mechanics approach based on the assumption that a correctly folded protein occupies a minimum energy conformation, most likely a conformation near the global minimum of free energy. Potential energy is obtained by summing the terms due to bonded and non-bonded components estimated from these force field parameters and then can be minimized as a function of atomic coordinates in order to reach the nearest local minimum (Weiner and Kollman, 1981; Weiner *et al*., 1984). This approach is very sensitive to the protein conformation of the molecules at the beginning of the simulation.

One way to address this problem is to use molecular dynamics to simulate the way the molecule would move away from that initial state. Newton's laws and Monte Carlo methods were used to reach to a global energy minima. The approach of molecular mechanics is faced by problems of inaccurate force field parameters, unrealistic treatment of solvent, and spectrum of multiple minima (Stephen *et al*., 1990).

The second approach of predicting protein structures from sequence alone is based on the data sets of known protein structures and sequences. This approach attempts to find common features in these data sets which can be generalized to provide structural models of other proteins. Many statistical methods used the different frequencies of amino acid types: helices, strands, and loops in sequences to predict their location. (Chou and Fasman, 1974b; Garnier *et al*., 1978; Lim, 1974b; Blundell *et al*., 1983; Greer, 1981; Warme *et al*., 1974). The main idea is that a

segment or motif of a target protein that has a sequence similar to a segment or motif with known structure is assumed to have the same structure. Unfortunately, for many proteins there is no enough homology to any protein sequence or of known structure to allow application of this technique.

The previous review leads us to the fact that the approach of deriving general rules for predicting protein structure from the existing data sets or databases and then applies them to sequences of unknown structure appears to be promising. Several methods have utilized this approach (Richardson, 1981; Chou and Fasman, 1974a; Krigbaum and Knutton, 1973; Qian and Sejwaski, 1988; Crick, 1989).

Artificial Neural networks have great opportunities in the prediction of proteins secondary structures. These methods are based on the analogy of operation of synaptic connections in neurons of the brain, where input is processed over several levels or phases and then converted to a final output. Since the neural network can be trained to map specific input signals or patterns to a desired output, information from the central amino acid of each input value is modified by a weighting factor, grouped together then sent to a second level (hidden layer) where the signal is clustered into an appropriate class.

Artificial Neural Networks are trained by adjusting the values of the weights that modify the signals using a training set of sequences with known structure. The neural network algorithm adjusts the weight values until the algorithm has been optimized to correctly predict most residues in the training set.

Feedforward neural networks are powerful tools. They have the ability to learn from example, they are extremely robust, or fault tolerant, the process of training is the same regardless of the problem, thus few if any assumptions concerning the shapes of underlying statistical distributions are required. The most promising is that programming artificial neural networks is fairly easy (Haykin, 1999).

Thus, neural networks and specially feedforward networks have a fair chance to well suite the empirical approach to protein structure prediction. In the process of protein folding, which is effectively finding the most stable structure given all the competing interactions within a polymer of amino acids, neural networks explore input information in parallel style.

The GOR method was first proposed by (Garnie *et al.*, 1978) and named after its authors Garnier-Osguthorpe-Robson. The GOR method attempts to include information about a slightly longer segment of the polypeptide chain. Instead of considering propensities for a single residue, position-dependent propensities have been calculated for all residue types. Thus the prediction will therefore be influenced not only by the actual residue at that position, but also to some extent by other neighbouring residues (Garnier and Robson, 1989). The propensity tables to some extent reflect the fact that positively charged residues are more often found in the C-terminal end of helices and that negatively charged residues are found in the N-terminal end.

The GOR method is based on the information theory and naive statistics. The mostly known GOR-IV version uses all possible pair frequencies within a window of 17 amino acid residues with a cross-validation on a database of 267 proteins (Garnier *et al.*, 1996). The GOR-IV program output gives the probability values for each secondary structure at each amino acid position. The GOR method is well suited for programming and has been a standard method for many years.

The recent version GORV gains significant improvement over the previous versions of GOR algorithms by combining the PSIBLAST multiple sequence alignments with the GOR method (Kloczkowski *et al*., 2002). The accuracy of the prediction for the GOR-V method with multiple sequence alignments is nearly as good as neural network predictions. This demonstrates that the GOR information theory based approach is still feasible and one of the most considerable secondary structure prediction methods.

1.4 The Problem

The problem of this research focuses on the protein folding dilemma. The question is how protein folds up to its three dimensional structure (3D) from linear sequences of amino acids? The 3D structure protein is the protein that interacts with each other 3D protein and then produces or reflects functions. By solving the protein folding problem we can syntheses and design fully functioning proteins on a computational machine, a task that may requires several years in the molecular biology labs. A first step towards that is to predict protein secondary structures (helices, strands, and loops). At the time of writing this chapter, the prediction level of protein secondary structures is still at its slightly above the 70% range (Frishman, and Argos, 1997; Rost, 2001; Rost, 2003).

Prediction can not be completely accurate due to the facts that the assignment of secondary structure may vary up to 12% between different crystals of the same protein. In addition, β-strand formation is more dependent on long-range interactions than α-helices, and there should be a general tendency towards a lower prediction accuracy of β-strands than α-helices (Cline *et al.*, 2002).

To solve the above mentioned problems, or in other words to increase the accuracy of protein secondary structure prediction, the hypothesis of this research can be stated as: "construction and designing advanced well organized artificial neural networks architecture combined with the information theory to extract more information from neighbouring amino acids, boosted with well designed filtering methods using the distant information in protein sequences can increase the accuracy of prediction of protein secondary structure".

1.5 Objectives of the Research

The goal of this research is to develop and implement accurate, reliable, and high performing method to predict secondary structure of a protein from its primary

amino acid sequence. However, the specific objectives of this research can be stated in the following points:

a. To analyse and study existing methods developed in the domain of protein secondary structure prediction to help in the development and implementation of a new prediction method.

b. To develop and implement a new accurate, robust, and reliable method to predict protein secondary structure from amino acid sequences.

c. To assess the performance accuracy of the method developed in this research and to compare the performance of the newly developed method with the other methods studied and implemented in this research work.

d. To study the differences between the secondary structure reduction methods and the effects of these methods on the performance of the newly developed prediction method.

e. To carry out blind test on the newly developed method. That is to analyse the output of the newly developed method with respect to an independent data set.

f. To study the performance of the coil prediction of the newly developed method using the ROC curve. This is also to examine the ability of ROC analysis to discriminate between two classes in a multi-class prediction classifier.

1.6 The Scope of This Research

Following the goal and objectives of this study is its scope. Since Bioinformatics is a multi-disciplinary science, the scope of each study must be stated

clearly. The protein sequence data is obtained from the Cuff and Barton (1999) 513 protein database. The data is prepared from the Protein Data Bank (PDB) by Barton's Group and considered as a benchmark sample that represents most PDB proteins. This study focuses on the neural networks and information theory since they are found to be effective for the prediction of protein secondary structure. The output results of the prediction methods are analysed and tested for performance, reliability, and accuracy. The limitation of this research work is the nature of the biological data which needs a great effort of pre-processing before the training and testing stages.

1.7 Organization and Overview of the Report

The organization and the flow of the contents of this report may be described as follows:

- The report begins with Chapter 1 which we are reading now. The chapter explains key concepts, introducing the problem of this research, list the objectives, and determine the scope of this work.

- Chapter 2 reviews and explains the proteins, sequences, and sequence alignments. It also examines amino acids and proteins in terms of their nature, formation, and their importance. The chapter reviews protein homology and homology detection and types of homologies proteins and then explains sequence alignment methods, pair-wise alignment, multiple alignments, as well as profile generation methods.

- The following is Chapter 3 which discusses and overviews protein structure prediction. The generation of profiles that uses the evolutionary information in similar sequences and the multiple sequence alignment methods are thoroughly reviewed in this chapter. This chapter describes the benchmark data sets conventionally used to predict protein structure as well. The chapter also reviews the artificial neural networks and the information theory for prediction of

protein secondary structure with special emphasis to GOR theory. The tools and techniques used in this research as well as prediction performance evaluation procedures are introduced in this chapter..

- Chapter 4 represents a brief and comprehensive methodology of this research. The chapter outlines and represents the framework followed in this research to implement the method proposed and developed in this research.

- Chapter 5 represents and explains the modelling of the methodology and algorithms used to develop the new method NN-GORV-I and its advanced version NN-GORV-II. The data set for training and testing the newly developed methods beside the other methods that are implemented in this work was described. The implementation of PSIBLAST program search of the *nr* database to generate multiple sequences which in turns are aligned by the CLUSTALW program is demonstrated in this chapter. The reduction methods used for the secondary structure data and the different statistical analysis and performance tests are demonstrated in this chapter.

- Chapter 6 discusses the results of the seven different prediction methods developed or studied in this research. The Q_3 , the segment overlap (SOV) measure and the Matthews correlations coefficients MCC are discussed and examined in this chapter.

- Chapter 7 discuses the effect of the five eight-to-three secondary structure reduction methods on the newly developed method in this research and trying to judge the argument that the eight-to-three state reduction scheme can alter the prediction accuracy of an algorithm.

- Chapter 8 explores the performance of an independent data set test on the NN-GORV-II method. Few protein targets of CASP3 are

predicted by the newly developed method to judge its performance and quality.

- Chapter 9 introduces the Receiver Operating Characteristics (ROC) analysis and area under curve (AUC) to the newly method which is a multi-class classifier to estimate the prediction accuracy of the coil states.

- Chapter 10 concludes and summarizes this research, highlights the contributions and findings of this work, and suggests some recommendations to further extend work.

1.8 Summary

This chapter introduces the problem of predicting protein secondary structure which is the core concern of this research. The chapter presents a brief introduction to bioinformatics, proteins, sequences, protein structure prediction. Known methods and algorithms in this domain are briefly introduced and presented. The problem of this research is clearly stated in this chapter and the objectives and scope of this research are thoroughly explained. The chapter ends with a description and overview of the organization of the report.

CHAPTER 2

PROTEIN, SEQUENCES, AND SEQUENCE ALIGNMENT

2.1 Introduction

To grasp a better understanding to this research, a molecular biology introductory concepts and facts are inevitable. This chapter reviews in a comprehensive style the protein definition, nature, and it's important to life. The chapter also explains the composition of proteins and its building blocks, the amino acids. The sequences and their alignments are discussed thoroughly in this review chapter. The different structures of proteins, methods of determining protein structure, and methods for generating homologue sequences and sequence alignment methods are presented in this chapter.

2.2 Proteins

Proteins are composed of individual units called amino acids. Amino acids share a similar structure. The difference between them is the 'R' group which is the cluster of atoms that give an amino acid its particular characteristics. Amino acids are grouped together in particular sequences that naturally fold up into a specific structure. While an amino acid is a letter in the sequence of the protein, in the structure each amino acid letter is actually a piece of a 3D structural object. Appendix A illustrates the different structures of protein.

The importance of sequence data can be used to make predictions of the functions of newly identified genes, estimate evolutionary distance in phylogeny, determine the active sites of enzymes, construct novel mutations and characterize alleles of genetic diseases. Sequence data also facilitates the analysis of the organization of genes and genomes and their evolution with respect to species and the identification of mutations that cause the diseases.

Multiple alignments of protein sequences are important tools in studying proteins. The basic information they provide is the identification of conserved sequence regions. This is very useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins, and in identifying new members of protein families (Durbin *et al*., 2002).

Proteins can be considered as series of amino acids linked together into contiguous chains. The 20 amino acids are shown in Table 2.1 with their respective three letter and one letter codes conventionally used in molecular biology.

Table 2.1: The twenty types of amino acids that forms the proteins

No.	Amino acid name	Three letter code	One letter code
1	Alanine	Ala	A
2	Arginine	Arg	R
3	Asparagine	Asn	N
4	Aspartic acid	Asp	D
5	Cysteine	Cys	C
6	Glutamic acid	Glu	E
7	Glutamine	Gin	Q
8	Glycine	Gly	G
9	Histidine	His	H
10	Isoleucine	Ile	I
11	Leucine	Leu	L
12	Lysine	Lys	K
13	Methionine	Met	M
14	Phenylalanine	Phe	F
15	Proline	Pro	P
16	Serine	Ser	S
17	Threonine	Tht	T
18	Tryptophan	Trp	W
19	Tyrosine	Tyr	Y
20	Valine	Val	V

In Bioinformatics research the one letter code is more commonly used than the three letter code. The training and testing protein sequences data used in this research adopts the one letter coding scheme.

The production of proteins in a cell is governed by codes and information transferred to the DNA, and RNA of the organism. Proteins are synthesized in the cells of living organisms, Prokaryotes (single cell) or Eukaryotes (high order) by a structured mechanism. The DNA of an organism encodes its proteins in a sequence of nucleotides, namely: adenine, cytosine, guanine and thymine. These nucleotides considered as information which is copied to the mRNA (messenger RNA) that serves as an intermediate medium, which is then processed during protein synthesis.

The codon (a non-overlapping triplet of nucleotides), specifies a corresponding subunit, or residue, to be added to the always growing polypeptide chain. The genetic code shown in Table 2.2 resembles the correspondence between the sequence of nucleotides of the codon and the amino acids which is constant in almost all organisms (Brian, 1998).

Amino acids consist of a carbon as a central atom linked to hydrogen. The bonding of carbon and oxygen forms what is known as Carboxyl group, while the bonding of carbon with hydrogen forms what is known as Amino group. Molecules of amino acids connect with each other through a side chain. Table 2.2 shows the standard genetic code of living organisms, where there are 64 different amino acids but only twenty different types of amino acids work as basic building units of a protein as shown in Table 2.1.

Table 2.2: The standard genetic code

		Second Position					
		T	**C**	**A**	**G**		
First Position	**T**	TTT Phe (F) TTC Phe (F) TTA Leu (L) TTG Leu (L)	TCT Ser (S) TCC Ser (S) TCA Ser (S) TCG Ser (S)	TAT Tyr (Y) TAC Tyr (Y) TAA *Ter* (end) TAG *Ter* (end)	TGT Cys (C) TGC Cys (C) TGA *Ter* (end) TGG Trp (W)	**T** **C** **A** **G**	**Third Position**
	C	CTT Leu (L) CTC Leu (L) CTA Leu (L) CTG Leu (L)	CCT Pro (P) CCC Pro (P) CCA Pro (P) CCG Pro (P)	CAT His (H) CAC His (H) CAA Gln (Q) CAG Gln (Q)	CGT Arg (R) CGC Arg (R) CGA Arg (R) CGG Arg (R)	**T** **C** **A** **G**	
	A	ATT Ile (I) ATC Ile (I) ATA Ile (I) ATG Met (M)	ACT Thr (T) ACC Thr (T) ACA Thr (T) ACG Thr (T)	AAT Asn (N) AAC Asn (N) AAA Lys (K) AAG Lys (K)	AGT Ser (S) AGC Ser (S) AGA Arg (R) AGG Arg (R)	**T** **C** **A** **G**	
	G	GTT Val (V) GTC Val (V) GTA Val (V) GTG Val (V)	GCT Ala (A) GCC Ala (A) GCA Ala (A) GCG Ala (A)	GAT Asp (D) GAC Asp (D) GAA Glu (E) GAG Glu (E)	GGT Gly (G) GGC Gly (G) GGA Gly (G) GGG Gly (G)	**T** **C** **A** **G**	

With the exception of *proline*, the amino acids described in Table 2.1 share the common feature of an amino and carboxyl group joined by a single carbon atom from which different side-chains are attached. However, *glycine* has no side-chain. Each type of amino acid has different side chain which gives it its distinguished characteristics. The peptide bond does not rotate freely, but the other two backbone bonds can rotate, allowing the polypeptide chain to fold in almost any direction.

The sequence of amino acids in a protein chain forms the protein structure. Protein structures may be classified into four levels or classes: primary, secondary, tertiary, and quaternary structure.

2.2.1 Protein Primary Structure

The amino acid sequence is the primary structure of a protein. It is usually represented by the one letter notation of the amino acids. Amino acids combine to form a protein through polypeptide bonds and here the protein could be considered as polypeptide chain and the amino acids as residues (Table 2.1). Anyhow the reaction here is complex and lengthy to be mentioned in detail. A protein could be formed out of 2000 amino acids or residues although short chain proteins are not unusual. Shorter chains are called peptides. The different physical and chemical properties of the side-chains determine both the local and global conformations adopted by polypeptide chains. Anyhow the sequence direction is very important and usually represented from the amino, (N) terminus to the carboxyl (C) terminus.

2.2.2 Secondary Structure

The three-dimensional structure of proteins is potentially determined by its primary structure (Anfinsen, 1973), although the folding process can be aided by other molecules (Hartl, 1996). Most proteins always fold into the same configuration (Branden and Tooze, 1991).

Pauling and Corey (1951) predicted the existence of sheet-like structures of non-covalently cross-linked strands of extended polypeptide chain which they called beta-sheet and a helical arrangement. Studying the structures of myoglobin, Kendrew (1960) confirmed the existence of a regular helical arrangement, called alpha-helix. Alpha-helices and beta-sheets are the most common form of secondary structure in proteins.

When the sequences of primary structures tend to arrange themselves into regular formations, these units are referred to as secondary structure. The angles and hydrogen bond patterns between the backbone atoms are determinant factors in protein secondary structure. Secondary structure is subdivided into three parts: alpha-helix, beta-sheet, and loop.

Alpha-helix is spiral turns of amino acids while a beta-sheet is flat segments or strands of amino acids formed usually by a series of hydrogen bonds. As the polypeptide chain coils in, the CO and NH groups of residues form hydrogen bonds which stabilize the helix. Most of the residues in a helix are bonded in this way, making it somewhat a rigid unit of structure with a little free space in its core. A helix and can have 4 - 50 residues and makes a whole turn every 3.6 residues.

Beta-strands are the most regular form of extended polypeptide chain in protein structures. Like alpha-helices, beta-sheets are stabilized by hydrogen bonds between CO and NH groups, but they are distantly separated along the chain. Because of the geometry of the peptide backbone, the amino acid side chains of beta-strands alternate on either side of the sheet.

Loops usually serve as connection points between alpha-helices and beta-sheets, they do not have even patterns like alpha-helices and beta-sheets and they could be any other part of the protein structure. They are recognized as random coil and not classified as protein secondary structure. When the polypeptide chain makes very sharp changes in direction using as few as four residues by means of hydrogen bond, it forms turns. These secondary structures commonly contain proline or glycine or both residues (Hutchinson and Thornton, 1994).

However, many researchers refer to anything which is not helix or strand as coil or random coil which is known as loop, and of course ignoring the existence of beta-turns. Anyhow, Chothia *et al.* (1989) proved that some protein structures (antibodies) have well defined conformations in a number of loops.

2.2.3 Tertiary Structure

The three-dimensional structure of the protein, which is formed from the secondary structures as subunits elements, is known as the protein's tertiary structure. Protein folding is the process that results in a compact structure in which secondary structure elements are packed against each other in a stable configuration. Dill (1990) reported that, the tendency for the burial of hydrophobic side-chains in the core of proteins has been observed in almost all structures discovered. It is believed this tendency is the driving force of tertiary structure formation.

Hydrogen bonds, van der Waals forces, and oppositely charged amino acid side-chains are other interactions that help to stabilize the fold. Folds are considered as sets of connected secondary structure elements, so they are known as *topologies*. Longer polypeptide chains that are usually clearly distinguished by a naked eye as self-contained units of structure, and have distinct hydrophobic cores, are known as domains. Swindells (1995), Islam *et al*. *(*1995), Siddiqui and Barton (1995) argued that the definition of domains is in this way is unreliable. A covalent linkage made during the folding process between sulphur atoms from cysteine residues is known as the disulphide bond (Freedman, 1995). Examples of proteins that exhibit disulphide are snake and scorpion toxins.

Levitt and Chothia (1976) grouped proteins into naturally four classes based upon the gross secondary structural content of their tertiary structures. These classes were: mainly-alpha, mainly-beta, alternating alpha-beta, and alpha and beta (not alternating). However, with the construction of a classified database of domains an automated approach to classification was developed (Michie *et al.*, 1996).

Different folds that often possess similar arrangements of a two to four consecutive recurring units of secondary structures are called super-secondary structures (Rao and Rossmann, 1973) and (Richardson, 1981; Richardson, 1986) or motifs (Sternberg and Thornton, 1976).

2.2.4 Quaternary Structure

An individual protein that its independent fold or substructures form a three dimensional structure of the protein is known as quaternary structure. This is true for some proteins because they do not work in isolation; haemoglobin and RNA polymerase are examples of such proteins.

2.3 Methods of Determining Protein Structure

Three-dimensional structures of a protein can be determined by describing the relative position of a single atom within the protein using two laboratory methods: (i) X-ray crystallography and (ii) Nuclear Magnetic Resonance (NMR) spectroscopy. X-ray crystallography is the most popular method of protein structure determination. X-ray beams are applied to a crystal of proteins that has been grown by purifying a protein sample. The structure of the protein is then determined by studying the diffraction pattern of X-ray. Anyhow X-ray crystallography is a lengthy and complicated process; it requires a high level of technical ability in the laboratory reach to an inference of the x-ray diffraction patterns (Branden and Tooze, 1991).

Nuclear Magnetic Resonance (NMR) spectroscopy requires a highly concentrated and purified and a lowered pH sample of a protein. The protein is then put in a strong magnetic field, and subjected to radio frequency (RF) pulses. This will force the protein to emit RF radiation. Then information of protein structure can be inferred from the frequencies and intensities of the emitted radiation. Practically, this process is not as easy as been described and there are many biochemical constraints in this process (Branden and Tooze, 1991).

Protein structure determination methods mentioned above require several months or even years of laboratory work, and they are not viable for some proteins. This why introducing procedures or processes of protein sequence prediction can save a considerable amount of time and effort.

As far as hydrophobicity is concerned, many researchers identified the amino acids that commonly substitute with each other and categorized them with regard to their properties or structures and found that the most common clusters of a single column amino acid profiles were mostly hydrophobic or polar in nature (Han and Baker, 1995; Fiser *et al.*, 1996; Ladunga and Smith, 1997).

The scale to measure hydrophobicity is not standardized and since it depends on the physico-chemical properties of amino acids, it was opened to subjective interpretations. However, Nakai *et al.* (1988), and Tomii and Kanehisa (1996) constructed a database of reported amino acids that shows their hydrophobicity scales and substitution matrices.

The distribution of disulphide bonds in cysteine residues stabilizes this amino acid and encodes important structural information since these bonds are mostly well conserved (Carrington and Boothroyd, 1996), while the distribution of cysteine residues does not encode important structural information in intracellular proteins interaction. However, pairwise interactions between distant homologues are not very well conserved (Russell and Barton, 1994).

The hydrophobic core residues of proteins are more conserved than non-core residues (Taylor, 1997). Patterns of hydrophobicity and sequence conservation are widely used to predict secondary structure. This prediction typically encodes important information to fold recognition but cannot contain further information than is already available in multiple sequences (Taylor and Thornton, 1984; Fischer and Eisenberg, 1996; Defay and Cohen, 1996; Hubbard and Park, 1995; Rice and Eisenberg, 1997; Rost *et al.*, 1997).

2.4 Characteristics of Protein Structures

A protein could be subjected to denaturing forces like high temperature or low pH which force the protein to loose its original structure. Proteins tend to revert to their original structure, after the denaturing forces are removed. Anfinsen (1973) showed that the amino acid sequence is the only source of information to survive the denaturing process, so the structured information must be somehow specified by the sequence.

Many proteins exist in an aqueous solution within the cell, and certain amino acid side chains tend to interact with the water molecules. These amino acids are known as hydrophilic which are polar. Their interaction with water often involves forming hydrogen bonds (Pace *et al*., 1996). On the other hand, hydrophobic amino acids, lack the atomic structure that enables them make hydrogen bonds with water. Protein folding is significantly affected by hydrophobic forces (Dill, 1990).

Patterns of amino acids interaction of a protein is another characteristics of a protein. Pairwise interaction and disulfide bonds play a great role in protein stability. Natural or induced mutations turn a protein to unstable condition. Proteins interact with each other through only certain portions of them. This portion is known as the functional site, and residues within the functional site are called functional residues. Protein function usually depends on the three-dimensional structure of its functional site. Anyhow mutation has an adverse affects on protein function. However, recently, Lise and Jones (2005) investigated two databases, one of disordered proteins and the other of globular proteins, in order to extract simple sequence patterns of amino acid properties that characterize disordered segments and concluded that the derived patterns provide some insights into the physical reasons for disordered structures. They are expected to be helpful in improving currently available prediction methods.

2.5 Protein Homology

Proteins of the same family are known as homologous proteins or homologs. Proteins change conservatively through evolution and similar proteins express similar functions (Jacob, 1977). Comparing two different proteins homologs, one of the three states occurs: substitution which is the replacement of one or more residues, deletion which the removal of one or more residues, insertion which is the addition of one or more residues. This is known as protein sequence alignment.

Sequence alignment is performed when different protein sequences are put in rows while columns represent regions of match or mismatch. When aligning two sequences, regions of mismatch in the other sequence are deleted and represented by dashes. These deleted regions are called gaps.

Alignments that contain two protein sequences are known as Pairwise alignment, while those contain many sequences are known as multiple alignments. Researchers (Burkhard, 1999; Sander and Schneider, 1991) showed that similar protein sequences usually reflect similar functions. Although there are exceptions of the previous conclusion, it has been proved that two proteins may have very different structures but almost identical function (Gilbrat *et al*., 1996). However, Lichtarge *et al.*, 1996 showed that functional regions residues are conserved within the same protein subfamilies but between different subfamilies.

The terms *homology* and *similarity* should not be confused. Sequences either have or do not have a common ancestor. Thus, sequences can either be homologous or not, but they cannot be 75% homologous, for instance. However, sequences can be similar by different degree and therefore be 75% similar. Moreover, that is not informative enough unless we know what the significance of this similarity is. Proteins that have significant sequence similarity are most often homologous. The next sections explain homology in more detail.

2.5.1 Types of Homologies

Gilbrat *et al*. (1996), Liisa and Chris (1996), and Hubbard (1997) enumerated instances of proteins with very similar structures but no or few sequence homology. These types of instances are known as structural homologs, on the other hand when these sequence similarities are week, such protein is referred to as remote homologs. Homology is estimated by percent identity (Burkhard, 1999; Julie *et al*., 1999).

There are several systems that make Pairwise structural alignments or organize proteins structures into families and classes Examples of these systems are: Yale aligner (Mark and Michael, 1998), CE (Shindyalov and Bourne, 1998), FSSP (Liisa and Chris, 1996), VAST (Gilbrat *et al*., 1996), CATH (Orengo *et al*., 1997), SCOP database (Hubbard *et al*., 1997; Andreeva *et al*., 2004), and CASP2 which uses individual human knowledge (Michael, 1997). Anyhow, the number of distinct folds in proteins is very small compared to the huge number of proteins (Chothia, 1992).

Remote homologies were able to be detected by dynamic programming alignments methods using a 3x3 substitution matrix derived from database counts (Fischer and Eisenberg, 1996; Defay and Cohen, 1996; Hubbard and Park, 1995; Rice and Eisenberg, 1997; Rost *et al.*, 1997). Most of these methods have included secondary structure prediction information.

2.5.2 Homologues versus Analogues

In classification of proteins, two main types or classes of pairs of protein structures could be distinguished: Homologues and analogues. Homologues are the pairs of proteins that have the same fold, more or less the same function, and common ancestry while analogous are the pairs of proteins that have the same fold, different functions, and unknown ancestry.

Doolittle (1981) and Sander and Schneider (1991) reported that some successfully aligned homologues shared sequence identity as less as up to 25% . This zone of sequence similarity is known as *twilight zone.* It also refer to pairs of analogues align with very low sequence identity. However, Homologues and analogues and protein folds have been used in the study evolution process of proteins and then species through million of years.

2.6 Molecular Interactions of Proteins

A protein function is highly affected by interaction occurring at the interface between solvent (typically water) and protein. The shapes of the protein, hydrophobic forces, and electrostatic attractive forces are among the most factors that affect protein functions although Chothia and Janin (1975) disagreed with that.

Hydrophobicity of a folding chain is one of the major forces in *ligand* (other molecules rather than water) recognition. When two molecules come together there is an increase in the entropy of the system as the solvent molecules become disordered (Chothia and Janin, 1975; Jones and Thornton, 1996). Hydrogen bonds and van der Waals forces provide attractive forces between molecules. However, hydrogen bonds are considered conferring *specificity* to interactions because they depend on the location of participating atoms (Fersht, 1984; Fersht, 1987).

Complementarity of two proteins interfaces is seen in electrostatic distributions and in three-dimensional shape. A computer generated methods have been developed to quantify Shape complementarity (Lawrence and Colman, 1993; Norel *et al.*, 1994) Predicting the location, orientation and conformation of protein molecules in their physiological interactions with proteins using knowledge of protein surfaces and interactions is known as docking technique.

2.7 Sequence Alignment Methods

Needleman and Wunsch (1970) introduced the concepts and algorithms of dynamic programming to biological sequence alignment. Since this algorithm needs to include the termini of both or all sequences, it is known as global alignment. A modified type of this algorithm was developed by Smith and Waterman (1981) to locate the best local alignments between two sequences.

The superposition methods which use iterative application of least-squares fitting techniques to optimize the definitions of residue equivalences between structures was then developed (Chothia and Lesk, 1986; Johnson *et al.*, 1990; Russell and Barton, 1992; May and Johnson, 1994; May and Johnson, 1995; May, 1996)

Other algorithms and methods of alignments include Falicov and Cohen method which uses a dynamic programming algorithm to generate the minimum soap-film area (Schulz, 1977) between arbitrarily superposed carbon-alpha backbones. Holm and Sander (1993) developed the DALI program which uses simulated annealing to generate alignments of structural fragments. DALI also can find alignments involving chain reversals and different topologies. The following section explores briefly some of the alignment methods.

2.7.1 Threading Methods

Jones *et al*. (1992) applied the double dynamic programming algorithm of Taylor and Orengo (1989) to solve the problem of misalignment of sequences when defining them in structural environments or residue classes. This is known as threading methods. A low level alignment is used to score the pairwise residue interactions (Sippl, 1990).

Jones et *al*. (1992) alignment threading methods has a serious problem that the number of all possible sub-alignments at each equivalence is exponential with

respect to sequence length (Lathrop, 1994). However, the frozen approximation method of (Flockner et *al.*, 1995) could speed up the alignment process by testing the suitability of pairwise distances between query residue *k* and library residues *l*.

Branch-and-bound search (Lathrop and Smith, 1996), Monte Carlo (Madej *et al.*, 1995) and exhaustive searches using heuristics (Russell *et al*., 1996).are among several methods that searches for the best alignment. The statistics of threading scores has been studied by (Bryant and Altschul, 1995), and (Jones and Thornton, 1996). However, Russell and Barton (1994) and Russell *et al.* (1997) showed that pairwise interactions are poorly conserved across large evolutionary distances

The alignment of biological sequences occupies a central role in modern molecular biology. Fundamental to biological sequence alignment is the incorporation of gaps, which represent insertions or deletions of sequence characters as mentioned in this chapter. In an experiment to evaluate the type and quality of an alignment, Zachariah *et al*. (2005) reported that Evaluation of the alignment quality revealed that the generalized affine model aligns fewer residue pairs than the traditional affine model but achieves significantly higher per residue accuracy. They then concluded that generalized affine gap costs should be used when alignment accuracy carries more importance than aligned sequence length.

2.7.2 Hidden Markov Models

Hidden Markov models (HMMs) are statistical models that have been used in speech recognition problems HMMs construct a general profile of each word, in which the more salient or known characteristics are expected with high probability. Then, when a a person pronounces a word, the word is recognized by comparing its sequence of frames against the HMMs for various words to look for the best match. HMMs were first used in computational biology by (Krogh *et al*., 1994) and in for sequence analysis by (Baldi *et al*., 1994; Eddy *et al*., 1995).

In proteins sequence prediction, members of a protein family share certain characteristics, such as the presence of conserved motifs; there could be clear differences between members of the same family in this aspect. HMMs model each protein family in such a way that the distinguishing characteristics are expected with high probability while variation is permitted. So, when a new homologous sequence is presented or introduced, the model estimates the likelihood that the sequence is a new homolog.

HMMs have been used successfully in different applications of protein sequence prediction (Kulp *et al*., 1996) used them in recognizing human genes in DNA, Grundy *et al*. (1997) in protein families detection, Francesco *et al*. (1997) in secondary sequence and protein topology. HMMs have been used effectively in protein structure prediction experiments in CASP (Kevin *et al*., 1997; Kevin *et al*., 1999) and CASP2 (Bystroff and Baker, 1997). However, comprehensive and useful reviews of HMMs can be found in Eddy (1996) and Eddy (1998).

2.7.3 Types of Alignment Methods

Many threading methods use the dynamic programming algorithm in various forms, including local alignment (Jones *et al*., 1992), global alignment (Bowie *et al.*, 1990; Matsuo and Nishikawa, 1995), and the so-called global-local alignment (Fischer and Eisenberg, 1996; Rice and Eisenberg, 1997). These protocols basically differ in the scoring of terminal gaps and the extent of the alignment (Zachariah *et al.*, 2005). The processing of scores in fold recognition is something of a black art. Theoretical proof exists to show that the scores from local alignments follow a Poisson-like distribution from which reasonable estimates of biological significance can be drawn (Henikoff, 1996; Bryant and Altschul, 1995).

The widely used sequence database searching methods BLAST (Altschul *et al.*, 1990), FASTA (Pearson, 1990) and Smith and Waterman's algorithm (Smith and Waterman, 1981) all use local alignments. However, the distributions of global alignment scores are less well understood. Some methods use the global method to

generate the alignment, and then calculate an energy score based on mounting the query sequence onto the library structure(Matsuo and Nishikawa, 1995), thus avoiding the direct use of the global score. Using global or local scores, Z-scores for each query-library pair can be calculated independently using scores from the alignments of randomised sequences (Rice and Eisenberg, 1997).

The global alignment algorithm (Needleman and Wunsch, 1970) gave the best results when combined with a simple score normalisation step. Before the calculation of Z-scores, the dynamic programming score is divided by the sum of the lengths of the two protein sequences. Without this correction, longer alignments (from longer library sequences) rank higher than they should.

Sequence alignment methods are divided into two categories: pairwise methods, which use only two sequences, and multiple sequence methods, which can use more than two sequences. Moreover, multiple sequences methods are subdivided into two categories: profile methods and multiple alignment estimation methods. In his paper “the art of matchmaking”, Smith (1999) presented sequence alignment of proteins and discussed their implications. However, Apostolico and Giancarlo (1998), Eddy (1998), and Gotoh (1999) presented detailed review and discussion about sequence alignment methods.

2.7.3.1 Pairwise Alignment Methods

The famous Needleman -Wunsch (Smith, 1999) and Smith-Waterman (Smith and Waterman, 1981) algorithms are used in pairwise alignments. The Needleman-Wunsch algorithm uses dynamic programming to find the lowest-cost *global* alignment of two sequences, while the Smith-Waterman algorithm (Smith and Waterman, 1981) finds the optimal *local* alignment of two sequences. The alignment is allowed to start and end in the middle of the sequences by deleting low-scoring regions.

As briefly discussed above, a well established method (Feng, 1985; Barton, and Sternberg, 1987) to measure the similarity between two protein sequences *x* and *y* is to align the proteins by a standard dynamic programming algorithm (Needleman and Wunsch, 1970) and obtain the score for the alignment . The order of amino acids in each protein sequence is then randomised and a dynamic programming alignment of the randomised sequences. This procedure is repeated typically several times and the mean and standard deviation of the scores for comparison of the randomised sequences is calculated. The standard deviation of the scores is better than the percentage identity since it corrects for bias due to the length and composition of the sequences.

The most widely used FASTA (Pearson and Lipman, 1988) and BLAST (Stephen *et al*., 1990) use heuristic algorithms, which offer higher efficiency of pairwise alignments. However, when applied to the complete proteomes of some organisms (Fleischmann *et al*., 1995; Fraser *et al.*, 1995; Bult *et al.*, 1996), these methods find similar sequences between only 58% - 78% of the sequences. Increasing the coverage of Smith-Waterman sequence search methods will increase the accuracy of prediction (Brenner, 1996; Hubbard, 1997).

Henikoff and Henikoff (1997) showed that simple embedding of consensus sequences from conserved regions of a multiple sequence alignment into a single representative sequence improves BLAST and FASTA searches. In order to align whole sequences, gap penalties can also be calculated on a position specific basis (Gribskov *et al.*, 1990). Hidden Markov models (HMMs) similarly deal with position specific substitutions and gap penalties in the alignment of multiple sequences (Krogh *et al*., 1994; Eddy, 1996).

Sequence database clustering requires high speed pairwise comparisons (Van-Heel, 1991; Wu *et al.*, 1992). Ferran *et al.* (1994) and Hanke *et al*. (1996) have used non-linear mappings of sequence composition data to cluster large sets of sequences.

2.7.3.2 Profile Alignment Methods

Profile alignment methods algorithms are more complex than the previous pairwise alignment algorithms. They were first used by Gribskov *et al.* (1987). This algorithm constructs a profile of the alignment under consideration. The profile consists of gap costs and a set of costs for aligning each of the twenty amino acids to each alignment column. The costs are derived from the amino acid probability distribution in each column. Sequence are given weights generally range between 0 and 1, and is that due to the fact that biological databases are skewed toward the proteins most heavily studied (Sjolander *et al.*, 1996; Smith, 1999.).

Examples of systems that use profile information include TOPITS (Rost, 1995), PSI-BLAST (Jones, 1999a; Altschul, 1997), GenThreader (Jones, 1999b), SAM-T98 (Kevin *et al*., 1998) and CLUSTALW (Julie *et al*., 1994; Higgins *et al*., 1996; Durbin *et al*., 2002).

Abagyan *et al.* (1994) calculated profiles based on the side-chain modelling energies of alternate amino acid substitutions in the library structure. Ponder and Richards (1987) were among the first researchers that conducted a side-chain replacement for fold recognition

However, there is a considerable number of reported methods that use or encode 3D structural information into strings of symbols or profiles against which 1D strings derived from the query sequence are aligned (Bowie *et al*., 1990; Bowie *et al.*, 1991; Abagyan *et al.*, 1994; Matsuo and Nishikawa, 1995; Hubbard and Park, 1995; Fischer and Eisenberg, 1996; Defay and Cohen, 1996; Taylor, 1997; Rost *et al.*, 1997; Rice and Eisenberg, 1997)

2.7.3.3 Multiple Alignment Methods

A more complicated estimation derived by several methods is the multiple alignment estimation methods which search for an alignment to maximize the overall homology in a pool of sequences. Multiple alignment methods use two-dimensional dynamic programming algorithms. The more complex method which is the K-

dimensional dynamic programming algorithms that seek to align *K* sequences simultaneously. The computational complexity of this task is proportional to K $(2L)^K$, where *K* is the number of sequences to align and *L* is the length of the alignment. Because of the computational complexity, 3-4 sequences are used (Gotoh, 1996; Gotoh, 1999); however, MSA (Lipman *et al*., 1989) which uses approximations can use up to 10 sequences only.

BLOCKS (Henikoff and Henikoff, 1994), PRINTS (Attwood *et al.*, 1997; Attwood *et al.*, 2003), PRODOM (Sonnhammer and Kahn, 1994), PROFILES (Gribskov *et al.*, 1987), PROSITE patterns (Bairoch *et al.*, 1997) and (Barton, 1990; Krogh *et al.*, 1994) are examples of multiple sequence alignment methods.

It has been shown recently that simple embedding of consensus sequences from conserved regions of a multiple sequence alignment into a single representative sequence improves BLAST and FASTA searches, and outperforms PSSM based methods (Henikoff and Henikoff, 1997). In order to align whole sequences, gap penalties can also be calculated on a position specific basis (Gribskov *et al.*, 1990). Hidden Markov models (HMMs) similarly deal with position specific substitutions and gap penalties in the alignment of multiple sequences (Krogh *et al*., 1994; Eddy, 1996).

Other methods are the progressive methods which calculate the alignment in a progressive mode, starting by aligning two sequences. Then, either profile methods are used to align a third sequence to the pair, or two other sequences are aligned. The process continues repeating until all sequences are aligned. Anyhow, the disadvantage of progressive method is that it can not correct mistakes made at earlier stages and so continue repeating aligning on incorrect estimations. This disadvantage suggested a need of refinement methods. However, iterative refinement methods generate high quality alignments, but require more computing resources than their predecessor progressive methods. Examples of progressive methods are CLUSTALW and PRRP (Notredame *et al*., 1998).

Stochastic alignment methods modify parts of the alignment according to a probability function, and then assessing the value of the modifications according to an objective function. The disadvantage of stochastic alignment methods is that they do not guarantee an optimal solution. However, they can build high quality alignments. The genetic algorithm for estimating multiple alignments SAGA-COFFEE (Notredame *et al*., 1998) is an example of stochastic methods. However, Hidden Markov models (HHM) for multiple alignment estimation are other examples of stochastic methods. It is worthy to mention that researchers reported that many of the best alignment results they achieved were supported significantly by involving manual refinements methods (Bates and Sternberg, 1999; Koretke *et al*., 1999; and Kevin *et al*., 1999).

As far as practically generating the multiple sequence alignments for large numbers of proteins is concerned, researchers simplify this process by developing automatic procedures for that. Some researchers perform a BLAST (Altschul *et al.*, 1990)database search of the OWL or nr databases (Cuff and Barton, 2000) The BLAST output is then screened by SCANPS, an implementation of the Smith Waterman dynamic programming algorithm(Smith and Waterman, 1981; Barton,1993) Sequences are rejected if their SCANPS probability score is higher than $1x10^{-4}$. Sequences are also rejected if they do not fit a length cut-off of 1.5. If sequences exceed the length criterion determine by SCANPS , they are truncated by removing end residues until the length of the sequence satisfies the cut-off value. Sequences that are shorter than the lower length limit are discarded. Although this method removes very long, very short and unrelated sequences, it allows sequences that are longer than the query, and are related, to be included after truncation. The sequence similar proteins selected by this method are then aligned by CLUSTALW (Thompson, 1994), with default or adjusted parameters.

Gaps in aligned sequences must be carefully observed since they can affect alignment and hence accuracy of prediction significantly. In several methods, the multiple sequence alignments are modified so that they do not contain gaps in the query sequence. The PHD (Rost and Sander, 1993; Rost and Sander, 1994; Rost *et*

al., 1994) uses a slightly different method whereby gaps at the end of the target sequence are removed.

The reference secondary structure for the data set is usually defined using DSSP (Kabsch, and Sander, 1983), STRIDE (Frishman, and Argos, 1995) or DEFINE (Richards, and Kundrot, 1988) where all definitions are then reduced to 3 state helix, strand, and coil. Care must be taken when using alternative reduction methods for the DSSP or other methods since this affect the prediction accuracies of different algorithms.

2.7.4 Comparative Modelling

Using either sequence-only or structure-based fold recognition techniques, one or more sequences of known structure are found to be related to a novel sequence under investigation

Comparative modelling is building a model of the newly introduced protein sequence based upon known (parent) structures. The major steps of this model are: alignment of the newly introduced sequence with the parents and other homologous sequences, copying the core from the parent to the model, building the non-core regions into the model, and refining the side-chain geometry and packing (Sanchez and Sali, 1997).

2.7.5 Overview of Alignment Methods and Programs

Needleman-Wunsch pairwise alignment, CLUSTALW multiple alignment, and PRRP multiple alignment methods were compared according to their performance (Gotoh,1996). The test set consisted of about 50 protein families, each consisting of two to ten sequences. Gotoh found that PRRP performed better than CLUSTALW and Needleman-Wunsch, and CLUSTALW performed better than

Needleman-Wunsch. Anyhow the gap penalties significantly affect the performance of each method.

Notredame *et al*. (1998) compared CLUSTALW and PRRP together with SAM, PILEUP, SAGA-COFFEE and SAGA-MSA methods using their default parameters. The test sets were selected with each having at least five sequences, and a consensus length of 50 or greater. Methods were scored according to the proportion of residue pairs in columns that they aligned accurately. Although all methods were close in score, PRRP and SAGA-COFFEE performed the best and in ten out of the eleven cases; SAM had the worst performance among all other methods while CLUSTALW, SAGA-MSA, and PILEUP showed similar performance estimates in most cases.

Julie *et al.* (1999) compared CLUSTALX a CLUSTAL with X windows interface, PILEUP, PRRP, and SAGA-COFFEE with MULTALIGN (Barton and Sternberg, 1987), MULTAL (Taylor, 1998), PIMA (Smith and Smith, 1992) DIALIGN (Morgenstern *et al*., 1998) and HMMT (Sean, 1995) methods. The BAliBASE alignment benchmark set (Julie *et al*., 1999) database was used for this test which is divided into five subsets, with each subset representing a distinct class of alignment test. In this experiment, global methods generally performed better than local methods. PRRP, CLUSTALW, and SAGA-COFFEE achieved the best performance. Anyhow, PRRP performed better than the other two. In general, this test showed that iterative and stochastic refinement methods outperformed most progressive alignment methods.

Briffeuil *et al*. (1998) compared the performance of MATCH-BOX server, a method they developed which uses a local multiple sequence alignment method (Depiereux *et al.,* 1997) with CLUSTALW, MSA (Lipman *et al*., 1989), PIM (Smith and. Smith, 1992), MAP (Huang, 1994), Block Maker (Henikoff *et al*., 1995), and MEME (Timothy *et al*., 1994) servers. All methods were tested on their each own server using 20 families, each family included at least three sequences of well known structure.

Specificity which is the number of correctly predicted residue pairs compared to the number predicted, and sensitivity which is the number of correctly predicted residue compared the number of correct pairs were used in the scoring of this test. Results suggested that there were differences in specificity and sensitivity of local aligners and global aligners. However, among global aligners, MAP performed better and among local aligners, MATCH-BOX showed very high specificity and low sensitivity (Briffeuil *et al.,* 1998).

Hudak and McClure (1999) compared SAM (Richard and Anders, 1996), MATCH-BOX (Depiereux *et al*., 1997), PIMA (Smith and. Smith, 1992), Block Maker (Henikoff, *et al*., 1995), and MEME (Timothy, *et al*., 1994), ITERALIGN (Brocchieri and Karlin 1998), and PROBE (Neuwald *et al*., 1997). In contrary to a previous experiment conducted by Hudak and McClure(1999), who concluded that global alignment methods often perform better than local alignment methods (Marcella, 1994) and SAM performed much better (Marcella, 1996). Hudak and McClure (1999) found that while all methods could detect the conserved Motif IV, only ITERALIGN, MEME, SAM, and PROBE could detect the entire series of motifs, with PROBE outperformed all of them.

Sauder, *et al*. (2000) studied the profile alignment methods in their work on homology modelling experiments (Dunbrack, 1999). They used SCOP (Hubbard *et al*., 1997) and CE (Shindyalov and Bourne, 1998) structures, and BLAST, PSI-BLAST, CLUSTALW sequence alignment methods. In summary, the results showed that BLAST performed better with 28% sensitivity and PSI-BLAST did better with 40% sensitivity. Although CLUSTALW aligned 100% of all structure pair, it was concluded that the results obtained in this range were not very good because CLUSTALW has no fold recognition component.

2.8 Summary

This chapter begins with a molecular biology definition and description proteins and amino acids with a brief review to the 20 amino acids that form proteins

and the standard genetic map of living entities. The different structures of proteins; primary, secondary, tertiary, and quaternary structures and the known methods of determine these structures are explained in details. Protein homology, the types of homology, and the difference between protein homology, analogy, and similarity are reviewed in this chapter. This chapter also reviews and discuses the different sequence alignment methods and the ways and procedures of automating the generation of multiple sequence alignments for large number of proteins. The generation of protein profiles to get the maximum possible distant biological information from related sequences is reviewed in this chapter. The chapter ends with an overview of the known alignment methods and programs. Some of these alignment methods and programs are used in this research to generate the necessary aligned sequences as discussed in the modelling of the methods in Chapter 5.

CHAPTER 3

REVIEW OF PROTEIN SECONDARY STRUCTURE PREDICTION: PRINCIPLES, METHODS, AND EVALUATION

3.1 Introduction

Protein secondary structure prediction essentially means the prediction of the formation of regular local structures such as α helices and β strands within a single protein sequence; of course the remaining non regular structures are coils. This is an essential intermediate step on the way to predicting the 3D structure of a protein. If the secondary structure of a protein is known, it is possible to derive a quite small number of 3D structures using knowledge about the ways that secondary structural states formed. A good number of prediction methods and algorithms have been developed using the advances in algorithms and computational power and storage ability.

Most probably solving the protein folding problem will pave the way to rapid progress in the fields of protein engineering and drug design. Moreover, since the number of protein sequences is growing much faster than our ability to solve their structures experimentally in the molecular biology laboratories; this will widen the gap between sequence and structure. The need for alternative methods to solve the protein folding problem becomes crucial.

Artificial neural networks method is inspired from the mechanism of *synaptic* connections of neurons of the brain, where input is processed on several levels and

mapped to a final output. In protein secondary structure prediction, information from the central amino acid of each input is modified by a weighting factor and sent to another level of the network until it is passed to the output layer. The output layer then decides whatever this amino acid or residue will fold to helix, strand, or coil. The work of Qian and Sejnowski (1988) sparked the implementation of neural networks in the domain of protein secondary structure prediction.

The information theory is a naive statistical method that is based on the conditional probabilities of variables. Garnier *et al.* (1978) implemented this approach to protein secondary structure prediction problem. This method calculates probability values for a specific amino acid based on the adjacent amino acids up to eight residues away using principles of the information theory mentioned above. The GOR method which is named after the first letters of its authors' name was first developed in 1978 and has been updated many times since then until it reached a comparatively high accuracy of prediction.

In this chapter the problem of predicting the secondary structure of a novel protein from its primary sequence will be addressed and reviewed. The different methodologies and algorithm used in this domain, collaborative programs and utilities, and data set exercised in this filed are presented and explained. The chapter also briefly reviews the contribution of many researchers and the advances in the domain of protein secondary structure prediction.

The chapter introduces and presents the artificial neural networks, its concepts, applications, and implementation. The chapter also reviews the information theory with special reference to the GOR implementation of this approach to calculate propensities of proteins. Both artificial neural networks and information theory (GOR-V) constitute the basis of this research. The evaluation and assessment of such prediction methods and programs are presented and explained briefly. Full description and explanation of the protein secondary structure prediction accuracy assessment methods are presented in the methodology chapter.

3.2 Protein Secondary Structure Prediction

The prediction of protein structure from amino acid sequence has become the target of of scientist since Anfinsen (1973) showed that the information necessary for protein folding resides completely within the primary structure. Researchers have then been considerate with the possibility of obtaining a complete three-dimensional structure of a protein by applying the proper algorithm to a known amino acid sequence.

The appearance of rapid methods of DNA sequencing and the translation of the genetic code into protein sequences has boosted the need for automated methods of interpreting these linear sequences into terms of two or three-dimensional structure (Stephen *et al.*, 1990).

Although the development of advanced molecular biology laboratory techniques reduced the amount of time necessary to determine a protein structure by X-ray crystallography, a crystal structure determination may still require many months if not years. NMR techniques helped in determining protein structure, but NMR is also costly, time-consuming, requires large amounts of protein of high solubility and is severely limited by protein size (Stephen *et al.*, 1990). The conclusion is that current experimental methods of determining protein structure will not suffice the present and future need for protein structure determination.

There are two different approaches in determining protein structure. A molecular mechanics approach based on the assumption that a correctly folded protein occupies a minimum energy conformation, most likely a conformation near the global minimum of free energy. In this approach, predictions are based on a force field of energy parameters derived from a variety of sources including ab initio and experimental observations of amino acids. Potential energy is obtained by summing the terms due to bonded and non-bonded components estimated from these force field parameters and then can be minimized as a function of atomic coordinates in order to reach the nearest local minimum (Weiner and Kollman, 1981, Weiner, *et al.*, 1984) However, this approach is very sensitive to the protein conformation of the molecules at the beginning of the simulation.

One way to address this problem is use molecular dynamics to simulate the way the molecule would move away from that initial state. Newton's laws and Monte Carlo methods were used to reach to a global energy minima. The approach of molecular mechanics is faced by problems of inaccurate force field parameters, unrealistic treatment of solvent, and spectrum of multiple minima (Stephen *et al.*, 1990).

The second approach of predicting protein structures from sequence alone is an empirical one, based on the data sets of known protein structures and sequences. This approach attempts to find common features in these data sets which can be generalized to provide structural models of other proteins.

Many statistically based methods use the different frequencies of amino acid types in sequences to predict their location in the secondary structure conformations: helices, strands, and coils (Chou and Fasman, 1974a; Chou and Fasman, 1974b; Garnier, *et al.*, 1978; Lim, 1974a; Lim, 1974b, Blundell, *et al.*, 1983; Greer, 1981; Warme, *et al.*, 1974). The basic idea is that a segment or motif of a target protein that has a sequence similar to a segment or motif with known structure is assumed to have the same structure. Unfortunately, for many proteins there is not enough homology to any protein sequence or of known structure to allow application of this technique.

Thus, the approach of deriving general rules for protein structure from the existing data sets or databases and then applies them to sequences of unknown structure appears to be promising for protein structure prediction. Various methods have been used for extracting rules from structural databases. Examples of these methods are: visual inspection of protein structures (Richardson, 1981), multivariate analyses methods (Chou and Fasman, 1974a; Krigbaum and Knutton, 1973), and artificial neural networks (Qian and Sejwaski, 1988; Crick, 1989).

3.3 Methods Used In Protein Structure Prediction

Organizing proteins into classes and families made the protein structure prediction a viable process. In addition, the growth in precise, fast, computerized structure prediction algorithms turned predicted structures good alternatives to obtain actual structures. Researchers distinguish between two categories of protein structure prediction methods: fold recognition methods which assume that a given protein is similar in structure to known protein structure; ab-initio which is a term indicates first principles or basic facts. Ab-initio methods search for a conformation that brings biochemical and biophysical forces to minimum. Comparing these two methods, the fold recognition methods outperformed the ab-initio methods (Alexey, 1999), moreover ab-initio methods require complex computations and they work better in short proteins sequences (Moult *et al.*, 1999). Fold recognition methods predict the structure of a protein by searching the protein structure databases for a fold family that best fits the protein, and then figure out which portions of the protein will adopt or match which portions of the fold (Daniel *et al*., 1999; Kevin *et al*., 1999). Ab-initio methods focus on predicting the novel structure of a sequence from basic facts or principles.

Homology modelling methods are usually applied to fairly close homologs, for which an accurate alignment can be predicted with high confidence (Srinivasan *et al.*, 1996). Docking prediction algorithms study the protein under observation and the nucleic acid or proteins with which it interacts, and then predict the functional site of the protein, and predict the nature of the interaction. Eisenhaber *et al.*, (1996) developed a secondary structural content prediction algorithm known as SSCP, which can indirectly be used to predict structural class defined using secondary structure composition cut-offs (Nakashima *et al.*, 1986).

The prediction of a protein tertiary or 3D structure however, begins with the prediction of its secondary structure elements as mentioned before. The reported accuracy of these methods is around 70-80% using differently constructed datasets with varying degrees of cross-validation. However, some researchers reported accuracy of nearly 100% (Zhou *et al.*, 1992; Chou and Zhang, 1994; Chou and Zhang 1995), but their method had been criticized of neglecting the memorization effect of weighted vectors they have used.

The first experiments to predict secondary structure of proteins were restricted by the few numbers of available structures and limited computing resources available. Using simple statistical and mathematical estimates of helix and strand, predictions of 60-65% Q_3 accuracy were reported (Periti *et al.*, 1967; Ptitsyn, 1969; Nagano, 1973; Chou and Fasman, 1974a; Garnier *et al.*, 1978; Lim, 1974a; Lim, 1974b). Many researchers have used the increased availability of structural information in the analysis of sequence or structure correlations for pairs of amino acids (Gibrat *et al.*, 1987; Rooman and Wodak, 1991; Han and Baker, 1995; Han and Baker, 1996). However, their prediction was not of significant improvement to the overall accuracy of secondary structure prediction.

Garnier *et al.* 1978 used their own algorithm to show that aligned protein sequences could provide valuable evolutionary information relevant to secondary structure prediction. However, their work was not of practical use until recently when databases of sequences were built (Zvelebil *et al.*, 1987). A linear discrimination function is used to determine the relative contributions of each sequence-based attribute to the final prediction (Weiss and Kulikowski, 1991; Michie *et al.*, 1994), which is 70% accurate (Q_3). Some researchers performed secondary structure predictions with a manual analysis of patterns of conservation and residue types (Benner and Gerloff, 1991; Benner *et al.*, 1994).

The nearest neighbour methods (Yi and Lander, 1993; Salamov and Solovyev, 1995; Salamov and Solovyev,1997; Frishman and Argos, 1996) have around 70% Q3 accuracy although there is redundancy in the mapping between local sequence and structure (Kabsch and Sander, 1984) For short fragments of query sequence, these methods search a database of sequences with known structure and allocate secondary structure according to that of the nearest neighbours. Many researchers reported that long-range contacts cannot be usefully predicted using statistics based methods (Thomas *et al.*, 1996; Gobel *et al.*, 1994; Olmea and Valencia, 1997).

Using SWISS-PROT (Bairoch and Boeckmann, 1991; Bairoch and Apweller, 1997) sequence database, Frishman and Argos (1997) tested their PREDATOR

program which uses amino acid pair statistics to predict hydrogen bonds between neighbouring strands and other residues. They expected an increase in Q3 of 5-10% given a ten-fold increase in sequence database size. Since PREDATOR uses pairwise local alignments (Russell and Barton, 1993), there is expected further improvement of the accuracy of this method.

As far as further improvements in secondary structure prediction are concerned, many researchers reported that may require more attention to specific sequential and structural motifs and turns (Han and Baker, 1996; Hutchinson and Thornton, 1994; Yang *et al.*, 1996), termini of beta-sheets and alpha-helices (Jimenez *et al.*, 1994; Aurora *et al.*, 1994; Donnelly *et al.*, 1994; Elmasry and Fersht, 1994) and super-secondary structures of proteins (Taylor and Thornton, 1984).

If we would like to simulate the folding process in detail in tertiary structure prediction, that might be impossible for the time being. However, Dill (1990) attempted to reduce the search space by using a simplified polypeptide representation and restrain atom or residue positions to a lattice (Dill *et al.*, 1995). Folding or conformational search experiments are hard to succeed, even for small proteins. However, theoretical experiments using these algorithms may be informative (Thomas and Dill, 1996).

Critical Assessment of Structure Prediction (CASP) is meeting sessions for evaluating prediction methods in a competitive environment. The first meeting experiment (CASP1) was held in 1994 and then being held every two years to compare between protein structures that are suggested by prediction methods and that are determined by X-ray crystallography or NMR spectroscopy The main benefit of this coordination is the evaluation of prediction results on the same targets using the same criteria (Lattman, 1995; Dunbrack *et al.*, 1997; Marchler-Bauer and Bryant, 1997).

Nakashima *et al*. (1986) conducted experiment to predict structural classes of proteins from amino acid composition with small dataset. The results reported showed accuracies of around 70-80% (Nakashima *et al.*, 1986; Klein and Delisi,

1986; Chou, 1989). Several researchers reported that the size and makeup of the dataset crucially affected the prediction accuracies; large and comprehensive datasets gave accuracies as low as 57% for three classes (helices, strands, and coils) implementing the jack-knifed method (Nakashima *et al*., 1986).

The whole sequences or a collection of genes of an organism is known as the genome. The aim of sequencing a genome is to identify the genes and the proteins that they code for. Gene prediction systems can predict which sections of DNA code for genes with over 90% accuracy (Kulp *et al*., 1996). After genes have been predicted and identified, then the proteins that might be expressed or produced could be identified and characterized.

Neural network models have the advantage of making complex decisions based on the unbiased selection of the most important factors from a large number of competing variables. This is particularly important in the area of protein structure determination, where the principles governing protein folding are complex and not yet fully understood (Stephen *et al*., 1990).

At present, the largest application of feedforward neural networks has been used in the prediction of protein secondary structure. As secondary structures (alpha-helices, beta-strands, and coils) are by definition the regions of protein structure that have ordered, locally symmetric backbone structures. Many researchers have sought to predict secondary structure from the sequence of contributing amino acids (Bohr *et al*., 1988).

Qian and Sejnowski (1988), Holley and Karplus (1989), Bohr *et al*. (1990), and McGregor *et al*. (1989) have applied neural network models to extract secondary structural information from local amino acid sequences and have achieved improved secondary structure prediction levels over that derived by statistical analysis (Chou and Fasman,1974a; Chou and Fasman, 1974b).

Qian and Sejnowski (1988) used a fully connected multilayer perceptron with a single hidden layer of 40 units for this purpose. A sliding window consisting of 13 consecutive residues was used as the input to the network to predict the secondary

structure of the residue in the middle of the window. The window is used to incorporate neighbourhood influence into the prediction. The network employed three output nodes, each representing a class of the secondary structure. The 20 distinct residues were represented using what is termed orthogonal encoding in which each residue is assigned a unique binary vector (100, 011,001, for alpha, beta, and coil, respectively).Therefore, for the network, the input dimension was of size (20 binary bits) × (13 residues) = 260. After training the network with the standard back-propagation algorithm, it scored 64.3% correct predictions. The Qian and Sejnowski's work pioneered the work of artificial neural networks in predicting protein secondary structure and now become almost the standard method in this domain.

Maclin and Shalvik (1994) for example, combined the Chou and Fasman (1978) residue statistics into the design of their Artificial Neural Networks to improve the prediction accuracy. However, Rost and Sander (1993) incorporated distant or what they called evolutionary information into their neural network. It was the very first work that introduced the long range effects using a profile of evolutionary information (Rost and Sander, 1996).

Baldi and co-researchers designed a bidirectional recurrent neural networks (BRNN) in different architectures to intelligibly utilize evolutionary information without over-fitting by rolling them along the multiple aligned sequences in both directions (i.e like wheels) until they reach the residue under consideration. The final prediction is computed by using a simple averaging scheme to form an ensemble of all the networks (Baldi *et al.,* 1999; Baldi *et al.,* 2001)

Rost and Sander's Artificial Neural Networks reached 71.9% accuracy and then being called the PHD (Profile network from HeiDelberg), (Rost and Sander, 1994). However, it has been reported recently that the latest version of the SSpro server has an accuracy of 74.5% (Pollastri *et al*., 2002).

Because the neural networks are effective they have produced the most accurate secondary structure predictions for the majority of the past few years. However, criticism to neural networks falls in that they are *black boxes*. Neural

networks may be effective classifiers but they cannot explain why a given pattern has been classified as *a* rather than *b*. People defend this criticism by proving that many things in our life give good results without explanations and we do not reject these results.

The content of secondary structural classes can be estimated experimentally by spectroscopy (Woody, 1995), or secondary structure predictions, from which the class can be derived (Rost and Sander, 1994; Eisenhaber *et al.*, 1996). Nishikawa and Ooi, (1982), Nishikawa *et al.*, (1983), and Nakashima *et al.*, 1986 reported that the amino acid composition of a protein is correlated with the structural class. Artificial neural networks have also been used to predict structural classes by representing proteins in 20 dimensional amino acid composition space (Muskal and Kim, 1992; Metfessel *et al.*, 1993; Rost and Sander, 1994). Variations on distance measures and multivariate analysis methods have used for the same prediction too (Nakashima *et al.*, 1986; Chou, 1989; Metfessel *et al.*, 1993; Klein and Delisi, 1986; Chou and Zhang, 1995; Boberg *et al.*, 1995; Eisenhaber *et al.*, 1996).

Methods of protein secondary structure prediction improved significantly in the past few years through the use of information contained in neighbouring residues and accumulated databases. Recently, the evolutionary information resulting from improved searches and larger databases has boosted prediction accuracy to the 77% level of prediction. There are bundles of methods that predict protein secondary structure and they reported prediction accuracies ranging from the 65% to the 75% level. Table 3.1 lists the names of several well established methods of protein secondary structure prediction with their reported accuracies and remarks about each method.

Table 3.1: Well established protein secondary structure prediction methods with their reported accuracies and remarks briefly describing each method.

Method Name	**Accuracy %**	**Remarks**

PROF	77.0	Cascade multiple classifier that uses quadratic and linear discrimination combiners (Ouali and King, 2000)
PSIPRED	76.6	Neural networks uses PSI-BLAST profiles (Jone, 1999)
SSpro	76.3	Based on an ensemble of 11 bidirectional recurrent neural networks (BRNNs).(Baldi, 1999)
JPred2	75.2	based on a consensus from several methods (Cuff and Barton, 1999)
PHD	71.9	Neural network systems of a sequence-to-structure level and structure-to-structure level (Rost and Sander, 1993)
PHDpsi	75.1	PSI-BLAST based predictor. Like NN-II (Rost and Sander, 1993)
PHDsec:	72.2	Multiple alignment-based neural network system focuses on hydrogen bond (Rost and Sander, 1993)
NSSP	71.0	Multiple alignment-based nearest-neighbour method.
GOR-IV	64.5	GOR IV uses all possible pair frequencies within the window of 17 amino acid residues. There is no defined decision constant. (Garnier et al., 1996)
GOR V	73.5	Uses different sizes of sliding windows and multiple sequence alignments.
SOPM	70.0	combining various other prediction programs. Based on the homologue method of Levin et al.
DSC	70.0	Based on residue conformation propensities (King and Sternberg, 1996)
SSPRED:	70.0	Multiple alignment-based program using statistics.
NNPREDICT	65.0	Single-sequence based neural network prediction. Like NN-I

3.4 Artificial Neural Networks

Artificial neural networks or neural networks are parallel, distributed information processing structures. The feed-forward net is the most widely used neural network architecture in solving problems and accomplishing pattern

classification and regression tasks. The feed-forward network is also known as multi-layer perceptron (MLP). One of the most important trends in neural computing over the past few years has been dealing with the neural networks as approach derived from statistical pattern recognition theory or probabilistic model (Baldi, 1995; Bishop, 1996; Devroye, *et al.*, 1996; Baldi and Brunak, 2002)

Neural networks have a fair chance to well suit the empirical approach to protein structure prediction. Like the process of protein folding, which is effectively finding the most stable structure given all the competing interactions within a polymer of amino acids, neural networks explore input information in parallel style.

3.4.1 Inside the Neural Networks

Inside the neural network as shown in Figure 3.1, many types of computational units exist; the most common type sums its inputs (x_i) and passes the result through a nonlinear approximation or activation function (a sigmoid function is used in this research) to yield an output (y_i). In artificial neural networks architecture generally, all the units in the same layer have the same transfer function and thus the total input is a weighted sum of incoming outputs from the previous layer. A transfer function may be a linear function like the function of the regression analysis, and hence the unit *i* is a linear unit. This is usually occurs in a network architecture that has no hidden units (Baldi and Brunak, 2002). However, in Artificial Neural Networks most of the time the transfer functions are non linear; examples of non linear transfer or activation functions are: hard limiters, sigmoid, and threshold logic elements.

Activation functions are often known as squashing functions. These functions simulate a dual state or binary decision. These threshold gates functions are discontinuous functions; this why sigmoid transfer functions are often used. The sigmoid transfer function of type logistic transfer function can estimate the probability of binary event.

An equivalent to logistic activation function is the softmax equation or normalized exponential unit which computes the probability of an event with *n* possible outcomes is also often used in classification tasks (Riis and Krogh,1996).

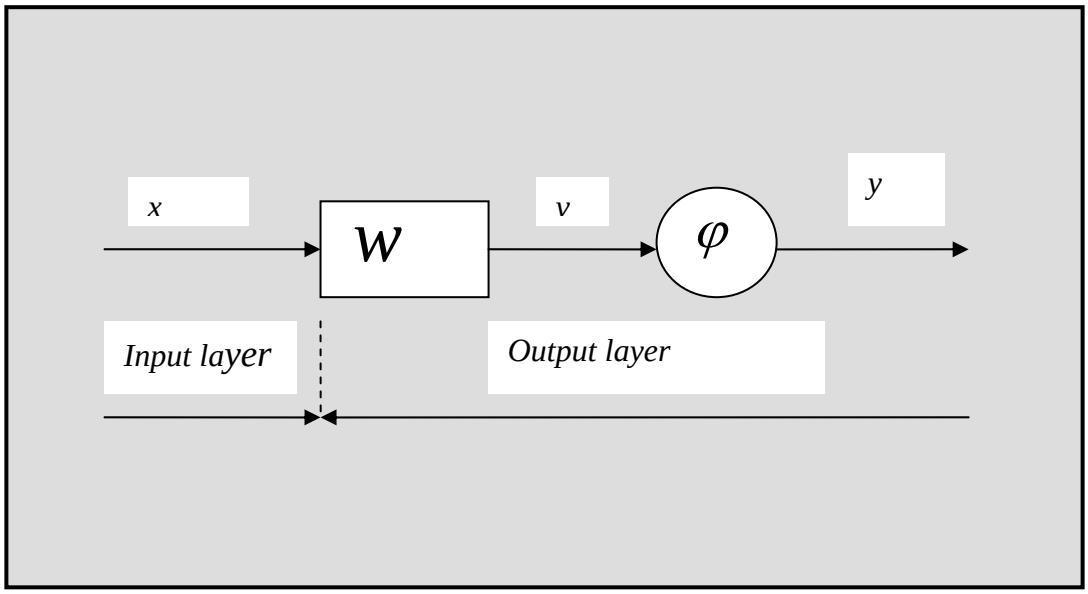


Figure 3.1: Basic graphical representations of a block diagram of a single neuron artificial neural networks.

One of the most important properties of artificial neural networks is that they can approximate any reasonable function to any degree of precision (Hornik *et al.*, 1990, Hornik *et al.*, 1994).

For neural networks models that classify an input into two classes (for example coil/not-coil), the target output can represented as 0 or 1. This model is a binomial model and can be estimated by a sigmoid transfer function. In consequence, in a binomial classification model, the output transfer function is logistic transfer function (Baldi, 1995).

If the classification task of the neural networks has *n* possible classes for a given input *x*, the target out put y is a vector with a single *1* and (*n-1)* zeros. The probabilistic model for this task is the multinomial or polynomial (multi-class classification) model (Farago and Lugosi, 1993).

One of the frequently used Artificial Neural Networks is the feedforward artificial neural networks trained with back-propagation for rule extraction purposes. It is termed feedforward because information is provided as input and propagated in a forward manner. The most well known artificial network is the feedforward neural networks will be reviewed in the following section.

3.4.2 Feedforward Networks

Feed-forward neural networks are the most widely used architecture of neural networks. The popularity of these networks originates from the fact that they have been applied successfully to a wide range of information processing tasks in many fields like financial prediction, speech recognition, image compression, medical diagnosis and of course protein structure prediction (Lisboa,1992).

In common with all neural networks, feed-forward networks are trained, rather than programmed, to carry out the chosen information processing tasks. Training a feed-forward network involves adjusting the network so that it is able to produce a specific output for each of a given set of input patterns. Since the desired inputs are known in advance, training a feed-forward net is an example of what is called supervised learning.

Feed-forward networks are characterized by “layers” architecture, with each layer comprising one or more simple processing units called neurons or nodes. Each node is connected to one or more other nodes by parameters values or weights to the nodes in other layers. All feed-forward networks are characterized by having at least ingle input layer and a single output layer. A network with only an input and an output layer is called a single layer network or single layer perceptron

Feedforward networks are often composed of visible and hidden units. The visible units are those in contact with the outside world such as input and output layers while invisible units are those called hidden layer or layers (Baldi and Brunak, 2002) Each network has connections between every node in one layer and every

other node in the layer above. Two layer networks, or perceptrons, are only capable of processing first order information and consequently obtain results comparable to those of multiple linear regression.

Hidden node networks, however, can extract from input information the higher order features that are ignored by linear models. Feedforward networks are trained to map a set of input patterns to a corresponding set of output patterns (Figure 3.2). In general, a network containing a large number of hidden nodes can always map an input pattern to its corresponding output pattern (Rumelhart and McClelland, 1986; Baldi, 1995).

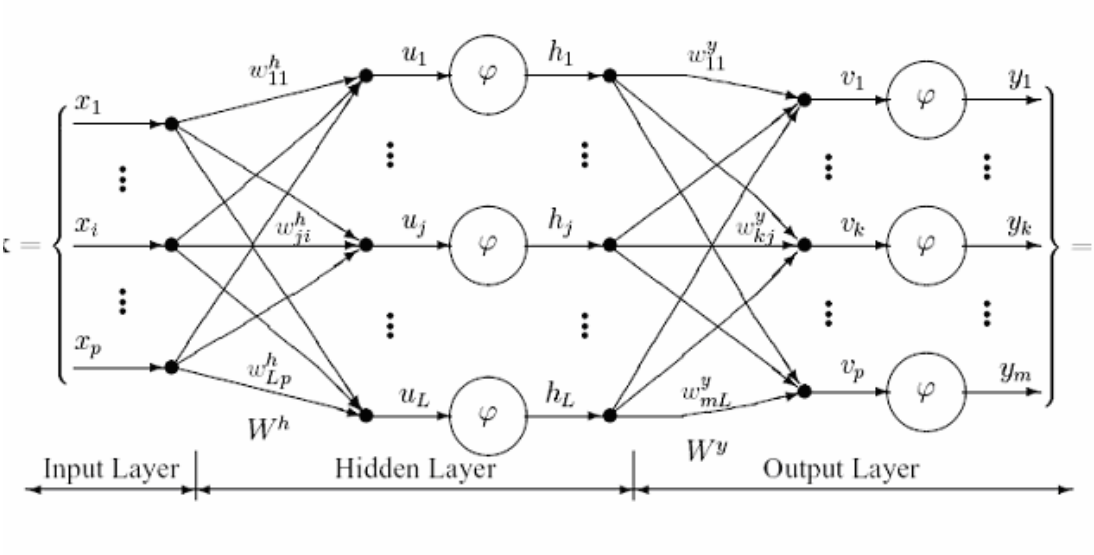


Figure 3.2: Representation of multilayer perceptron artificial neural networks.

3.4.3 Training the Networks

Feed-forward networks are trained using a set of patterns known as the "training set" for which the desired outputs are known in advance. This process is known in the neural network training as "supervised learning". In this type of learning, every pattern holds the same number of elements as the network input nodes, and every target pattern holds the same number of elements as the network output nodes (Rumelhart *et al*., 1986).

The network weights (w_{ij}) are initialised to small random values prior to training. A training algorithm is used to continuously reduce the total network error by iteratively adjusting the weights. There are two types of training; batch or offline training and stochastic or online training. With offline training, the whole set of patterns is repeatedly presented to the network, with the weights updated after each complete presentation. With online training, the weights are updated after the presentation of a subset of one or more training patterns. Online training is often more effective than offline training in practice since it performs fewer calculations if the training set contains redundant information, and is less likely to be trapped in the local minima (White, 1992; Swingler, 1996; Haykin, 1999,) which will be explained in the network optimization section.

While many algorithms exist for training, clearly the most frequently used technique is the method of back-propagation (Rumelhart, Hinton and Williams, 1986). Back-propagation involves two passes through the network, a forward pass and a backward pass. The forward pass generates the network output activities and is generally the least computation intensive. The more time consuming backward pass involves propagating the error initially found in the output nodes back through the network to assign errors to each node that contributed to the initial error (Qian and Sejnowski, 1988; Haykin, 1999). When all errors are assigned, the weights are changed so as to minimize these errors. Regardless of the training steps or equations, the main goal of the network is to minimize the total error of each output node over all training examples.(Haykin, 1999).

The time the neural networks learn this mapping for a set of training patterns, they are tested on examples that are usually different from those patterns used in training. While most feedforward networks are designed to maximize generalization from training examples to testing examples, some networks tend to memorize their training examples and hence over-fitting occurs in such networks

3.4.4 Optimization of Networks

Because the rules in most input-output mappings are complex and often unknown, a series of architecture optimizing simulations are required when testing each assumption. Examples of such optimizing experiments include varying input representation, numbers of hidden nodes, numbers of training examples, and others. In each case, some measure of network performance is evaluated and tabulated for each network architecture or training condition. The best performing network is chosen as that which performs the best on both the training and testing sets (Swingler, 1996).

With networks containing hidden nodes, training algorithms face the problem of multiple-minima when minimizing the output error across all training patterns. If the error space is uneven or rough, as is often the case in hidden node networks, the multiple-minima problem can be a serious one.

To solve the problem of local minima, researchers often permute their training and testing sets and train a number of times on each set (cross validation), while reporting the best performing network for each simulation. The variance between training and testing sets as well as between training sessions helps to describe the complexity of the weight space as well as the input-output mapping.

Usually smooth trends in performance levels point to optimal network architectures. Memorization or over-fitting is one of the main nuisances to the network where the network learns the training examples, rather than the general

mapping from inputs to outputs. Memorization reduces the accuracy of the network generalization to untrained examples. Clear signs of undesired memorization become apparent when the network performs much better on its training set than on its testing set; and typically, these results when the network contains far more weights than training examples. When undesired memorization results, the researcher is forced to increase the numbers of training examples, reduce node connectivity, or in more difficult situations, reduce the number of input, hidden, and/or output nodes. If it is not possible to increase the dataset of training examples, the next best choice is to reduce the network connectivity. Choice must be careful when deciding which connection to be removed. This process is known as network pruning, that often slows the already lengthy training process of the network.

Finally, reducing the number of network nodes is the least desirable of all approaches since it often results in losing important information from the network, especially if the number of input nodes is reduced. Similarly, reducing the number of hidden nodes often results in unacceptable input-output mappings; while reducing the number of output nodes, often results in mappings that are no longer useful. Anyhow, undesired memorization is one of the main drawbacks of Artificial Neural Networks solutions. Anyhow, the design and representations of Artificial Neural Networks should be smart and augmented.

Feedforward neural networks are powerful tools. They have the ability to learn from example, they are extremely robust, or fault tolerant, the process of training is the same regardless of the problem, thus few if any assumptions concerning the shapes of underlying statistical distributions are required.

All the above mentioned characteristics and advantages of the artificial neural networks made it a powerful and promising tool in the area of protein structure prediction. Many researchers applied the neural networks to solve the problem of prediction protein secondary structure successfully (Qian and Sejnowski, 1988; Rost, and Sander, 1994; Riis and Krogh, 1996; Chandonia and Karplus, 1999).

3.5 Information Theory

Information theory is a branch of the mathematical theory of probability and mathematical statistics that quantifies the concept of information. Shannon (1948) explained the information theory which considered communication as a strictly stated mathematical problem in statistics for the very first time. It concerns with information entropy, communication systems, data transmission and rate distortion theory, cryptography, data compression, error correction, and other related fields.

Possibly there is no review or explanation for the information theory without understanding quantum mechanics and physics, deliberate mathematical notations, and probabilities representation. In this review we present the information theory and entropy with minimal involvement in such diverged fields. The aim of this section is to give a general overview and understanding of the information that forms the basis of GOR algorithm. In the methodology chapter, more relevant details will be explained and mathematically represented.

The continuously increasing amount of protein structural information has urged researchers to develop several approaches that use this information for developing new ideas to predict protein structure and function. The most essential information applied here is to include statistical potentials to study and predict protein folding problem. Researchers in the past few years have used a variety of physical, chemical and biological measures of varying degrees complexities to understand the problem of protein folding. This concept was essentially applied by describing a protein representation by breaking up the amino acids atoms into functionally similar atom groups (Mintseris and Weng, 2004).

To adopt a quantitative measure for the information contained in an event, the proposed measure should have some perceptive properties; the following properties help forming such measure:

- Information contained in events has to be defined in terms of some measure of uncertainty of the events.
- Less certain events should contain more information than more certain events.

- The information of independent events taken as a single event should equal the sum of the information of the unrelated.

3.5.1 Mutual Information and Entropy

Researchers used the information theory approach to analyze the contributions of several traditional amino acid alphabets (residues) using mutual information (Cline *et al*., 2002).

Shannon(1948) arguments for entropy *H(X)* that it quantifies how much information is conveyed, on the average, by a letter drawn from the ensemble X; that is, it tells how many bits are required to encode such information.

The mutual information *I(X; Y)* quantifies how much correlated two bits are. How much do we know about an event drawn from X^n when we have read an event drawn from Y^n ? This can be explained by an example from signal communication field. Let a message sent from a transmitter to a receiver, given that the communication channel is noisy, so that the message received (y) might differ from the message sent (x). Then the noisy channel can be characterized by the conditional probabilities p($y|x$) which the probability that y is received when x is sent. Let us assume that the letter x is sent with a priori probability p(x). We would like to quantify how much we learn about x when we receive y? Or simply how much information or entropy we gain to describe x in the process of learning more about x. Bayesian statistics is usually used to update the probability distribution for x; that is:

$$p(x|y) = p(y|x).p(x)/\ p(y)$$

However, if for any reason x and y are absolutely not correlated, the information contained in x is zero, and the whole formula in this concept evaluates to nothing.

The following logarithmic definition of mutual information (MI) is similar and some time more convenient compared to the statistical definition:

$MI = Xi; j\ P(i; j)\ log\ P(i; j)\ P(i)P(j)$

Where $P(i; j)$ is the probability that an atom of type *i* forms a contact with an atom of type *j*, and
$P(i)$ and $P(j)$ are the marginal probabilities.

Interpretation of the reduced representation problem in information theory terms is straightforward. Mutual information between two variables *I* and *J* (representing a grouping of the protein atom types) is a measure of how much information one variable reveals about the other (Kullback *et al.*, 1987). If *i* and *j* are instances of *I* and *J*, where the number of such instances is governed by the size of the atom type alphabet, we want to define *i* and *j* such that the mutual information is maximized. Each instance *i* or *j* is a grouping of protein atoms of one type. It is easy to see from the equation that if *i* and *j* are chosen randomly, the probability of the joint distribution would be equal to the product of marginal distributions resulting in zero mutual information. On the other extreme, the maximum possible mutual information for a given alphabet size can be determined if we consider:

$P(i; j) = P(i) = P(j).$

This reduces to:

$MI = \sum_{i,j} P(i)\ log\ P(i)/P(i)P(j) = log(size)$

Another way to think about this is to realize that grouping atoms with similar biochemical properties; atoms that are commonly found in protein structures in similar environments, tends to increase mutual information by increasing the certainty that a specific atom type will occur in a given protein environment. Thus mutual information is a rigorous and intuitive measure suitable for optimization. It can be noticed that mutual information is also a measure of independence. If the variables *i* and *j* are randomly distributed, they reveal no information about each other, as shown above. Assuming under a null hypothesis (*H0*) that *i* and *j* are

independent and an alternative hypothesis (*H1*) that they are not, it can be shown that a log likelihood ratio test is exactly equivalent to the definition of mutual information (Shannon, 1948).

In the statistical context of the test of independence, the objective of finding the representation with maximum mutual information is equivalent to maximizing the significance of the test of independence between the atom types. The problem of finding such an optimal reduced protein representation for a given target alphabet size is essentially equivalent to maximum likelihood estimation.

3.5.2 Application of Information Theory to Protein Folding Problem

The application of the information theory to the problem of protein folding dates back to the 1970s of previous century (Chou and Fasman, 1974; Lim, 1974a; Lim, 1974b). The early versions of GOR method which was named after the first letters of its authors (Garnier *et al.*, 1978; Gibrat *et al.*, 1987) was based on single sequences and scored an accuracy of prediction below the 60% level.

Early works on the prediction of the secondary structure using information contained in residues based on the single residue statistics in various structural elements. The predictions were done by using a sliding window of a certain size and only single residue statistics for each residue within such a window were calculated for the prediction. A window of width of four residues, a characteristic length for helical contacts, was used in the Chou and Fasman method, and a width of 17 residues in the GOR-I method.

The pair-wise statistics for blocks of residues in secondary structure segments within the window was then used in GOR-III and GOR-IV which yielded in a significant improvement in protein secondary prediction and pushed it towards the 65% accuracy level. The implementation GOR algorithm is based on a window of a certain width, which is moved along the protein chain. Then the statistics of the

residues within the window are used to predict the conformational state of the residue at the centre of the window. The prediction process goes while the window moves along the chain, the secondary structure states of all residues from the N-terminal to the C-terminal along the chain (Garnier and Robson, 1989; Garnier *et al.*, 1996).

Significant progress has been made during the past few years in the accuracy of the prediction of secondary structure from sequence (Nishikawa and Nogughi, 1995). The improvement has been obtained by using multiple sequence alignments, instead of a single sequence. The multiple sequence alignments proved to contain distant and evolutionary information about protein structure.

Naderi *et al.* (2001) simple method based on information theory is introduced to predict the solvent accessibility of amino acid residues in various states defined by their different thresholds. Prediction is achieved by the application of information obtained from a single amino acid position or pair-information for a window of seventeen amino acids around the desired residue. Results obtained by pairwise information values are better than results from single amino acids. This reinforces the effect of the local environment on the accessibility of amino acid residues. The prediction accuracy of this method in a jack-knife test system for two and three states is better than 70 and 60 %, respectively. A comparison of the results with those reported by others involving the same data set also testifies to better prediction accuracy (Chen and Rost, 2002).

Rost (2003) and Przybylski and Rost (2002) argued that the main reason that information from the multiple sequence alignments improves the prediction accuracy is attributable to the fact that during evolution protein structure is more conserved than sequence, which consequently leads to the conservation of the long-range information. Many researchers suggest that some of this long-range information is exposed by multiple alignments (Kloczkowski *et al*., 2002).

Protein function is more fundamental for evolutionary information survival, than sequence conservation. Mutation on the other hand is important to the sequence that may destroy its function and usually cause the mutant sequence to be eliminated during evolution and hence change the conformation of the sequence (Branden and Tooze, 1991).

3.5.3 GOR Method for Protein Secondary Structure Prediction

The GOR method is one of the first major methods proposed for protein secondary structure prediction from sequence. GOR method fundamentally uses the information theory and naive Bayesian statistics. The method has been continuously modified and improved during the last two decades (Gibrat *et al.*, 1987; Garnier *et al.*, 1996). The first version of the method (GOR-I), used a rather small database of proteins which consisted few residues. GOR-II used database of 75 proteins containing about 13000 residues (Garnier and Robson, 1989).

GOR-I and GOR-II predicts four conformations rather than the three conventional states now predicted (helix (H), strand (E), and coil(C)), since turns (T) was used as the fourth confirmation. Both GOR-I and GOR-II algorithms use singlet frequency information within the window; this known in GOR literature as the directional information.

The advanced version GOR-III method utilized additional information about the frequencies of pairs (doublets) of residues within the window, based on the same database as the earlier GORs (Gibrat *et al.*, 1987). In GOR-III, the number of predicted conformations was brought to the now currently used (H, E, and C) three confirmations. The recently applied version of GOR methods is GOR-IV which uses 267 protein chains containing 63,566 residues (Garnier *et al.*, 1996) and available on the internet at http://abs.cit.nih.gov/gor/.

The GOR algorithm is a naive method based on the information theory combined with the Bayesian statistics. The information function *I(S,R)* which will be fully represented in mathematical notation together with other functions and formula in the methodology chapter, forms the basis of the information theory. The information function is described as the logarithm of the ratio of the conditional probability *P(S|R)* of observing conformation S, -where S is one of the three states:

helix (H), extended (E), or coil (C)- for residue R -where R is one of the 20 possible amino acids) and the probability P(S) of the occurrence of conformation S.

As mentioned above, GOR-IV uses a window of 17 residues, which means for a given residue, eight nearest neighbouring residues on each side are included in the calculations. The conformational state of a given residue in the sequence depends on the type of the amino acid R as well as the neighbouring residues along the sliding window. The information function of a complex event can be decomposed into information of simpler events and then summed up, according to the manipulation of the information theory. The GOR-IV method calculates the information function as a sum of information from single residues (*singlets*) and pairs of residues (*doublets*) within the width of the sliding window.

In GOR-IV, the first summation is over *doublets* and the second summation is over *singlets* within the window centred round the i^{th} residue. The pair frequencies of residues occurring in corresponding conformations are calculated from the database used for the GOR method. Using the above frequencies calculated from the databases, the GOR-IV algorithm can predict probabilities of conformational states for a new sequence.

A major advantage of the GOR method over other methods is that it obviously identifies all factors that are included in the analysis and calculates probabilities of all three conformational states. Another advantage of GOR algorithm over other algorithms is that it is computationally fast utilizing less CPU memory. It is possible to perform the full jack-knife procedure here where every single protein is removed from the database in turn and the frequencies is recalculated.

The GOR algorithm reads a protein sequence and predicts its secondary structure. For each residue *i* along the sequence, the program calculates the probabilities for each confirmation state, and the secondary structure prediction for such states (H, E, or C). Except in very few cases, the predicted conformational state usually corresponds to that with the highest probability.

GOR-V version applies the GOR-IV algorithm to the multiple sequence alignments. The gaps in the alignments are usually skipped by the GOR algorithm during the calculation of probabilities of conformation for each residue in the multiple alignments matrix but the information about position of gaps is kept (Kloczkowski *et al*., 2002). The main improvement made to GOR-IV was the systematic study of the GOR methods and the utilization of multiple sequence alignments to increase the accuracy of the secondary structure prediction. A full description of GOR-V will be presented in the methodology chapter.

3.6 Data Used In Protein Structure Prediction

The implementation of a practical approach to protein structure prediction is entirely dependent on the availability of experimental databases. The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of biological macromolecules (Berman *et al.,* 2002). The archives contain atomic coordinates, bibliographic citations, primary and secondary structure information, as well as crystallographic structure factors and NMR experimental data. It is produced and maintained at the Research Collaboratory for Structural Bioinformatics (RCSB). Other information included in the Protein Data Bank entries like protein name, relevant references, the resolution to which the structure was determined, the amino acid sequence, atomic connectivity, the researcher's judgement of secondary structure and disulfide bonding pattern, and other useful information

The PDB or Brookhaven Protein Data Bank database is updated continuously, at the year 2002 it contained 16,500 experimentally determined structures; and now (January 2005), the current holding of PDB is 28,992 structures. The PDB database is at http://helix.nih.gov/apps/bioinfo/pdb.html at the time of writing this report. The rate of adding structure to the current holdings is exponentially high.

Another database which is the SCOP (Murzin, 1995) classification of protein structure superfamilies are defined from careful analysis of structure, evolution and

function. The SCOP superfamilies contain protein domains that have the same fold and are likely to have evolved from a common ancestor. This database is also used by many researchers to generate their training sequences.

Rost and Sander (1994) defined non-redundancy to mean that no two proteins in the set share more than 25% sequence identity over a length of more than 80 residues. They presented 126 proteins set with which to train and test secondary structure prediction algorithms. Many well known algorithms and programs like PHD (Rost and Sander, 1994), NNSSP (Salamov and Solovyev, 1995), DSC (King and Sternberg, 1996), and PREDATOR (Frishman and Argos, 1997) have been trained on the Rost and Sander set of 126 proteins.

Cuff and Barton (1999) used sequences from the 3Dee (Siddiqui *et al*., 2001) database of structural domain definitions where a non-redundant sequence set was created by the use of a sensitive sequence comparison algorithm and cluster analysis, rather than a simple percentage identity cutoff. They then derived a set of 1233 domains where no pair shared obvious sequence similarity.

Using a more rigorous and stringent procedure which also included the 126 proteins of Rost and Sander, Cuff and Barton (1999) derived three non-redundant datasets suitable for cross-validation training and testing of secondary structure prediction methods. Finally they derived the sets CB396, CB497 and the then widely used in various experiments CB513 proteins data set which will be used in this research. These datasets, including secondary structure definitions and automatically generated multiple sequence alignments are available at http://barton.ebi.ac.uk.

3.7 Prediction Performance (Accuracy) Evaluation

With the advances of computer methods in bioinformatics and other related fields, researchers are always confronted with the problem of evaluating the accuracy of the prediction algorithms. It is of important to make sure that, for any type of prediction algorithm, the method or algorithm will be able to perform well on novel data that have not been used in the process of training the algorithm. Simply,

the method should be able to successfully generalize to new examples from the same data type.

Most secondary structure prediction methods include a set of parameters that must be estimated. Values for the parameters are obtained by statistical analysis or learning from a set of proteins data for which the 3D or tertiary structure. This set is known as the training set. Testing predictive accuracy on the training set leads to overestimated high accuracies. A practical test of a secondary structure prediction method should predict the structures of a test set of proteins that are not in the training set and show no sequence similarity with the training set.

An obvious problem facing methods of evaluating the performance of prediction methods is the redundancy of the data: if the sequence examples used for training and testing a particular algorithm are very similar the apparent predictive performance may be overestimated, reflecting the ability of the method to reproduce its own input rather than its ability to interpolate and extrapolate. Thus, the actual level of prediction accuracy is intimately related to the degree of similarity between the training and test sets, or in a cross-validated study, to the average degree of pair-wise similarity in a data set.

The most accurate method of predicting the secondary structure of a protein is to align the sequences by standard dynamic programming algorithms (Boscott *et al.*, 1993) when the protein sequence shows clear similarity to a protein of known three dimensional structure. On the other hand, when sequence similarity to a protein of known structure is not found, secondary structure prediction methods become the choice. It is therefore very important that there is no existence of sequence similarity between the training and testing sets of secondary structure prediction methods. (Sander and Schneider, 1991; Hobohm *et al*., 1992)

There are two empirical techniques to develop secondary structure prediction methods: cross-validation techniques, or full jack-knife or leave-one-out technique. Cross-validation techniques are less time consuming and use limited data. The full jack-knife test of *n* proteins, one protein is removed from the set, the parameters are

developed on the remaining *n-1* proteins, then the structure of the removed protein is predicted and its accuracy measured. This process is repeated *n* times by removing each protein in turn. However, care must be taken when using cross-validation since unrealistic high accuracies may be obtained for some methods if the set of proteins used in the cross-validation show sequence similarity to each other. (Nielsen *et al*., 1999).

In this chapter, more emphasis will be given to definition of relevant techniques and principles for the performance evaluation, and not on topics that relate to the selection of data The mathematical notation, graphical representation, and relevant illustrations for the following measures of performance are not presented here since they will be elaborated and explained in more details with mathematical notations in the methodology chapter.

3.7.1 Average Performance Accuracy (Q3)

The estimation of the global accuracy of a protein is usually conducted by a measure known as Q_3. The Q_3 is a measure of the overall percentage of predicted residues, to observed (Schulz and Schimer, 1979) and represented as: The summation of the number of residues identified in the (helix, strand, and coil) state, effectively observed in the state divided by the total number of residues

3.7.2 Segment Overlap Measure (SOV)

Segment overlap measure (Rost *et al.*, 1994) was performed for each data set. Segment overlap values attempt to capture segment prediction, and vary from an ignorance level of 37% (random protein pairs) to an average 90% level for homologous protein pairs.

The advanced version of SOV (Zemla *et al*., 1999) is a measure for the evaluation of secondary structure prediction methods that is based on secondary structure segments rather than individual residues. The algorithm is an extension of the segment overlap measure SOV, originally defined by Rost *et al.* (1994). The new definition of SOV corrects the normalization procedure and improves SOV ability to discriminate between similar and dissimilar segment distributions. SOV method has been comprehensively tested during the second Critical Assessment of Techniques for Protein Structure Prediction

SOV is a set of segment-based heuristic evaluation measures, where a correctly predicted segment position can give maximal score even though the prediction is not identical to the assigned segment. The score punishes broken predictions strongly, such as two predicted helices where only one is observed compared with one, too small, unbroken helix. In this manner the uncertainty of the assignment's exact borders is reflected in the evaluation measure (Baldi *et al.*, 2000).

3.7.3 Correlation

One of the standard measures used by statisticians is the correlation coefficient also called the Pearson correlation. In the framework of secondary structure prediction, this is also known as the Matthews correlation coefficient in the literature since it was first used by Matthews (1975). The correlation coefficient is always between -1 and + 1 and can be used with non-binary variables. It is a measure tends to have the same sign and magnitude. A value of -1 indicates total disagreement and + 1 total agreement. The correlation coefficient is 0 for completely random predictions. Therefore, it yields easy comparison with respect to a random baseline. If two variables are independent, then their correlation coefficient is 0. The converse in general is not true.

Baldi *et al.* (2000) argued that the correlation coefficient has a global form rather than being a sum of local terms. The correlation coefficient uses all four

numbers used to compare between predicted and observed classes which are: true positive (T P), true negative (TN), false positive (FP), and false negative (FN). These four numbers will be explained in details in the next section.

3.7.4 Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) curve is a method for visualizing, organizing and selecting classifiers based on their performance. ROC graphs have long been used in signal processing and detection theory to depict the tradeoff between hit rates and false alarm rates of classifiers (Egan, 1975; Swets *et al*., 2000). ROC analysis has been extended for use in visualizing and analyzing the behavior of diagnostic systems (Swets, 1988). The ROC techniques are then used extensively in biological sciences and specifically clinical medicine (Zweig and Campbell, 1993; Hand, 1997; Zou, 2002).

The ability of a test to discriminate abnormal cases from normal cases is evaluated using Receiver Operating Characteristic (ROC) curve analysis (Hand, 1997; Zweig and Campbell, 1993). ROC curves can also be used to compare the performance of two or more classifiers. ROC becomes popular in assessing a two-class or binary classifier and comparing many binary classifiers efficiently.

ROC can be explained when you consider the results of a particular test in two populations, one population with abnormal cases, the other population with normal cases. For every possible cut-off point or criterion value you select to discriminate between the two populations, there will be some cases with the abnormal cases correctly classified as positive (TP), but some cases with the abnormal cases will be classified negative (FN). On the other hand, some cases without the abnormal cases will be correctly classified as negative (TN), but some cases without the abnormal cases will be classified as positive (FP).

Sensitivity and Specificity are two important terms in the ROC literature which are defined as: Sensitivity is probability that a test result will be positive when the

abnormal cases is present (true positive rate) while Specificity is probability that a test result will be negative when the abnormal cases is not present (true negative rate).

To measure the performance accuracy of a binary classifier, a common method is to calculate the area under the ROC curve, which is known as AUC (Bradley, 1997). The AUC is a portion of the area of the unit square and hence its value will always be between 0 and 1.0 (Hand and Till, 2001).

Since the random guess produces the diagonal line between (0; 0) and (1; 1), which has an area of 0.5, no practical classifier have an AUC less than 0.5. The AUC has an important statistical property that the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Hand and Till, 2001).

3.7.5 Analysis of Variance Procedure (ANOVA)

The hypothesis that the means of two groups are equal can be fairly assessed by an appropriate *t*-test. Analysis of variance or ANOVA is the technique that is employed when there are more than two groups to compare. There are several versions of ANOVA. The corresponding version of the unpaired *t*-test is one-way ANOVA and this is the technique that is mostly used. The two-way ANOVA is the corresponding version of the paired *t*-test. In fact ANOVA is a very powerful technique to analyses variances and differentiate between means of random sample observations. As will be seen, mostly ANOVA assumes that the data or observations under analyses have a Normal or Gaussian distribution although it can be applied to other not Gaussian distributed data (Agresti, 2002).

The concept of combining models to improve the performance has long been established in statistical framework. The theoretical and background of this idea existed since (Bates and Granger, 1969; Dickinson, 1973; Jordan and Jacobs, 1994). Many well accurate methods are currently available to perform protein secondary

structure prediction. Since these methods are usually based on different principles, and different knowledge sources and approaches, significant benefits can be expected from combining them. However, the choice of an appropriate combiner may be a difficult task. However, up to the date of submitting this research work there is no work which combining involves the proposed GOR-V method with the neural network architecture.

3.8 Summary

To predict a protein 3D structure from its one dimensional amino acid sequence is one of the most major problems in the field of molecular biology. The conventional laboratory methods to solving this problem are extremely slow to rap the gap between the fast growing numbers of sequences and their predicted structures. Successful prediction of protein secondary structure is the right way to arrive at the 3D structure and possibly solve the protein folding problem. With the advances in computer methods and algorithms the possibility of designing and developing powerful methods and programs to predict protein secondary structure becomes practical.

In this chapter the theories, concepts, and implantation of protein secondary structure prediction algorithms and methods are presented. The methods that evaluate the prediction accuracies of the mentioned algorithms are briefly presented in this chapter without much details and elaboration of mathematical formula and notations. Full description of these methods will be found in the methodology chapter.

The chapter also presents a brief review of the artificial neural networks is presented. The biological inspiration of the neural networks and the theories and concepts underlying them are explained. The feedforward neural networks architecture is explained in a more details since they will be adopted in this research.

A thorough look inside the networks is presented with explanation of the networks training and optimization.

The information theory which delves to the quantum mechanics, physics, entropy, and mutual information is briefly explained in this chapter. The information function is explained in more details since it forms the basis for the GOR algorithms that uses this theory and Bayesian statistics. A brief presentation of the successive GOR algorithms which use information theory is shown due to its importance to this research.

The artificial neural network and the information theory are deliberately presented in this chapter without involving many mathematical representations since this will be explained in more empirical details in Chapter 5. Both the artificial neural network and the information theory form the basis for the new method that is developed and tested in this research.

The evaluation of the protein secondary structures prediction methods and algorithms is a vital task, since it shows how accurate a prediction method is. Most important in the evaluation procedure is to test the ability of a prediction algorithm to perform well on new test set of data. There are several measures that evaluate the performance, the quality, and the stability of a prediction algorithm which are discussed and explained in this chapter.

CHAPTER 4

METHODOLOGY

4.1 Introduction

Secondary structure prediction methods are of great use when a homologue to the sequence under consideration is not detected. If a complete homologue sequence is detected then the prediction accuracy is 100%. Several methods are proposed and implemented to predict protein secondary structure from the protein primary sequences.

This chapter describes the framework used in developing and implementing a method to achieve a better prediction method for the protein secondary structure from its primary sequence. The data set that is used in the experiments of this research is presented and discussed as well as the hardware and software utilized to implement the prediction method. This is an abstracted chapter that is presenting a brief description of the methodological framework followed in this research. Further details of the methodology are discussed and elaborated in the following Chapter 5.

4.2 General Research Framework

The general framework for predicting protein secondary structure from the amino acid sequences is presented in this section. Applying the conventional methods of machine learning approaches including neural networks without augmentation, to biological data bases does not achieve good performance. That is

true due to the nature of biological data which is dynamic rather than static data conventionally used in pattern recognition problem solving domain. The method used in this research combines the artificial neural network approach with the information theory to include more biological information to achieve a better and more accurate prediction method for protein secondary structure. Figure 4.1 elucidates the general framework for prediction of protein secondary structure from its amino acid sequences in this research. The following text of this section is an explanation of Figure 4.1 to elaborate the framework in more details.

The prediction framework is initiated by studying and investigating the protein secondary structure prediction problem by outlining a technique to solve this problem. The discussion of the literature in the previous chapters was the main motivation to adopt the approaches and developing the techniques to solve the problem of protein secondary prediction.

In order to understand the function of a protein and how it carries out this specific function, we need to understand its structure. Structural biology involves the study of the structure of biological molecules. Its 3D arrangement of atoms gives each protein a specific and unique structure. By understanding how atoms are arranged to produce an active binding site for a protein, we can understand how, why, and when a protein works (i.e. folds). Thus, the protein data that will be used in training and testing the method developed in this work is the amino acid sequences and their corresponding secondary structures that are determined by X-ray crystallography and NMR laboratory techniques.

The choice of the data set is discussed in more details in the next section. However, Cuff and Barton’s 513 protein data set (CB513) is chosen to train and test the prediction method developed in this research. CB513 is a benchmark data that is used by several researchers to develop prediction methods. The data is found in flat files with most secondary structure assignment schemes or methods included as well as some aligned sequences. PERL programming language is used to develop programs to extract and parse necessary data portion.

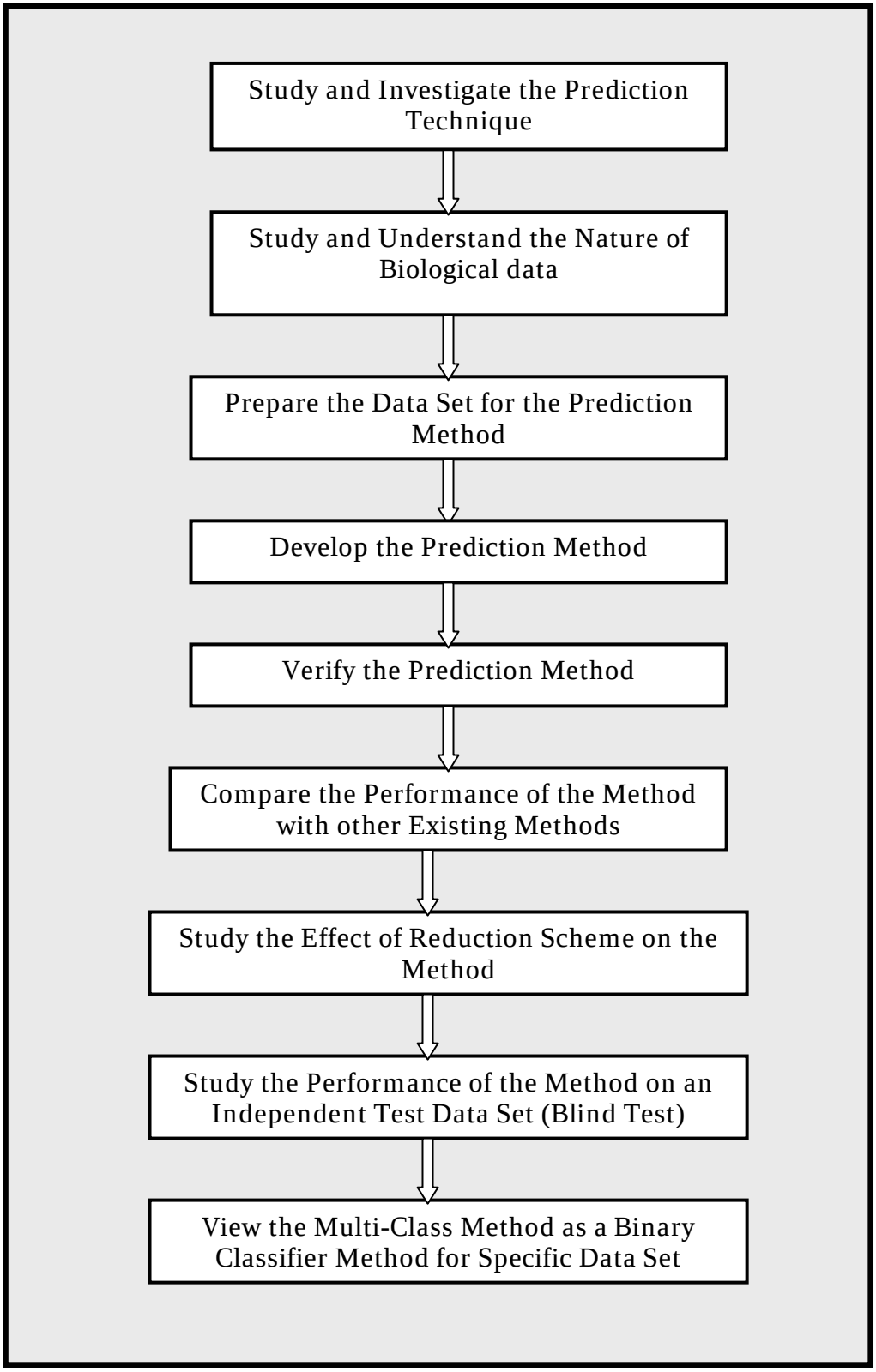


Figure 4.1: General framework for protein secondary structure prediction method

The prediction method is designed by combining neural networks model with a modified version of GOR-V information theory. This combination is based on strong statistical background which states that classification models which use different concepts and approaches may produce a better classification model. This assumption is only true when the errors of classification models are not correlated. The detail of the method is described in the next chapter.

The prediction method is verified during the training and testing stages. The seven fold cross validation is used where the CB513 data base is divided into almost equal seven sets. One set is used for verification and testing while the other six sets are used for the training procedure.

After developing and verifying the method, the performance of the method is compared with the other known methods investigated and implemented in this research to study the improvement achieved by the newly developed method. Comprehensive statistical analysis and test of significance is carried out.

The five well known DSSP eight-three reduction methods or schemes are obtained using PERL programming to study the effect of the different reduction methods on the performance and reliability of the newly developed prediction method.

The method is then tested using an independent data set that has not being used in training or testing. This is known as the blind test which is used to robustly test a newly developed classifier. The independent test set is the CASP3 which is found in the CASP and other Bioinformatics research groups' web sites.

Observing carefully the reduction methods, some methods assign almost half of the data set (48%) into the coil secondary structure states. This fact suggested the ROC curve to be used to assess the prediction methods and considering it as binary classifier instead of a multi-class classifier. This test can partially but accurately give another assessment procedure for the newly developed prediction method.

4.3 Experimental Data Set

Cuff and Barton's 513 non redundant proteins which contain 84,107 residues (Appendix B) are used for these series of experiments (Cuff and Barton, 1999). The CB513 data sets were selected by a stringent definition of sequence similarity or non redundancy, where no two proteins in the set share more than 25% sequence identity over a length of more than 80 residues. The sequences in the CB513 test set were developed from the 3Dee database of structural domain definitions where a non-redundant sequence set in this data base was created by the use of a sensitive sequence comparison algorithm and cluster analysis, rather than a simple percentage identity cutoff. This lead to a set of 1233 domains; the multi-segment domains were first removed to reduce the set size from 1233 to 988 sequences. The sequences were then filtered only to permit X-ray crystal structures with resolutions of less than or equal to 2.5 *Angstroms* which in turns reduced set of 554 domain sequences (CB554).

The Rost and Sander's (1993) 126 protein data set (RS126) was mostly used to develop early prediction methods and was used in the famous predictor PHD. The CB554 domain set and the RS126 set were combined and all pairs of both sequences were compared by BLOSUM62 matrix, and gap penalty of 10, in addition to alignments with SD score of >= 5 graded as similar sequence (Cuff et al 1999, 2000). With this stringent definition of sequence similarity, the 513 protein data set (CB513) was produced as shown in Figure 4.2. This data set was downloaded from web site http://barton.ebi.ac.uk/ and then extracted to LINUX RED HAT 9 platform.

During the preparation of data set for the experiments, among the CB513 proteins, few proteins did not generate valid PSIBLAST alignment profiles and others were not manipulated easy during the first stages of the experiment to translate them into codes readable by the neural networks or GOR-V C programs. However, the remaining proteins are 480 for training and testing of the seven prediction algorithms or methods. All the methods studied here are trained and tested on the same multiple sequence alignments data sets. This will allow a valid and reliable comparison of performance of the methods. In this experiment the data is split into seven more or less equal sets to perform seven folds cross validation. While the

seven fold cross validation is not as accurate as the full jackknife cross validation, it is not feasible to perform full Jackknife cross validation due the number of methods implemented, and the moderately huge size of the data set, and the very long CPU processing time.

RES:V,K,D,G,Y,I,V,D,D,V,N,C,T,Y,F,C,G,R,N,A,Y,C,N,E,E,C,T,K,L,K,G,E,S,G,Y,C,Q,W,A,
S,P,Y,G,N,A,C,Y,C,Y,K,L,P,D,H,V,R,T,K,G,P,G,R,C,H,
DSSP:_,E,E,E,E,B,B,_,T,T,S,_,B,_,_,_,S,_,H,H,H,H,H,H,H,H,H,H,T,T,_,S,E,E,E,E,E,E,E,E,T,T
,E,E,E,E,E,E,E,E,E,_,T,T,S,_,B,_,_,S,S,_,_,_,
DSSPACC:e,e,e,b,b,b,b,e,e,e,b,b,b,b,e,b,e,e,e,e,e,b,e,e,b,b,e,e,e,e,b,e,b,b,e,b,e,e,e,b,e,e,e,b,b,b,b,
b,e,e,b,b,e,e,b,e,b,b,e,e,e,e,e,e,
STRIDE:C,E,E,E,E,B,B,T,T,T,T,C,B,C,B,C,C,C,H,H,H,H,H,H,H,H,H,H,H,C,C,C,E,E,E,E,E,E,
E,E,T,T,E,E,E,E,E,E,E,E,E,T,T,T,T,C,B,C,C,C,C,C,C,C,C,
RsNo:1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,3
3,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,6
4,
DEFINE:_,_,_,E,E,E,E,E,_,_,_,_,E,E,E,E,E,H,H,H,H,H,H,H,H,H,H,H,H,H,H,_,_,_,_,_,E,E,E,E,_,_
,_,E,E,E,E,E,E,E,E,_,_,_,_,_,E,E,E,E,_,E,E,E,E,E,
align1:V,K,D,G,Y,I,V,D,D,V,N,C,T,Y,F,C,G,R,N,A,Y,C,N,E,E,C,T,K,L,K,G,E,S,G,Y,C,Q,W,
A,S,P,Y,G,N,A,C,Y,C,Y,K,L,P,D,H,V,R,T,K,G,P,G,R,C,H,
align2:K,R,D,G,Y,I,V,Y,P,N,N,C,V,Y,H,C,V,P,.,.,P,C,D,G,L,C,K,K,N,G,G,S,S,G,S,C,S,F,L,V,
P,S,G,L,A,C,W,C,K,D,L,P,D,N,V,P,I,K,D,R,K,.,C,T,
align3:A,R,D,A,Y,I,A,K,P,H,N,C,V,Y,E,C,Y,N,G,S,Y,C,N,D,L,C,T,E,N,G,A,E,S,G,Y,C,Q,I,L,
G,K,Y,G,N,A,C,W,C,I,Q,L,P,D,N,V,P,I,R,.,G,K,.,C,H,
.
.
.
align32:G,R,D,G,Y,I,A,Q,P,E,N,C,V,Y,H,C,F,P,S,S,G,C,D,T,L,C,K,E,K,G,A,T,S,G,H,C,G,F,L,
P,G,S,G,V,A,C,W,C,D,N,L,P,N,K,V,P,I,V,V,E,K,.,C,H,
align33:V,R,D,G,Y,I,A,Q,P,H,N,C,A,Y,H,C,L,K,S,S,G,C,D,T,L,C,K,E,N,G,A,T,S,G,H,C,G,H,
K,S,G,H,G,S,A,C,W,C,K,D,L,P,D,K,V,G,I,I,V,E,K,.,C,H,
align34:V,R,D,G,Y,I,A,Q,P,H,N,C,V,Y,H,C,F,P,S,G,G,C,D,T,L,C,K,E,N,G,A,T,Q,G,S,S,C,F,I,
L,G,R,G,T,A,C,W,C,K,D,L,P,D,R,V,G,V,I,V,E,K,.,C,H,

Figure 4.2: An example of a flat file of CB513 data base used in this research, 1ptx-1-AS.all file.

4.4 Hardware and Software Used

The data in bioinformatics field is usually diverse and huge and continuously increasing. In this research, ANSI C and PERL programming languages under Linux operating system are designed and developed to, implement, build, and run the prediction methods of this research. In addition to that, several hardware and dozens of systems and applications software are used to manipulate data and deploy the prediction methods or algorithms. To mention few, Cygwin, Linux Red Hat 9.0,

Fedora Core1, Windows 98, Windows XP, GNU gcc compiler, PERL interpreter, VIM editor, and a variety of FTP and Telnet utilities.

For statistical analysis several software are utilized including MS Excel application, SPSS, and SAS packages. To handle the matrices manipulations and the several curves, charts, and graphs representations the powerful Matlab package is exercised extensively in this work

4.5 Summary

This chapter explains a detailed framework of the methodology followed in this research in an attempt to solve the problem of protein secondary structure prediction. A benchmark data set is used in this work to allow a fair comparison with other published prediction methods.

The newly developed secondary structure prediction method framework is outlined and graphically represented. The chapter ends with a brief description of the environments and platforms used for the series of the experiments in this research with the bundles of software and hardware implemented in this research. This chapter presents a comprehensive framework methodology followed to solve the protein secondary structure prediction problem; however, a detailed step by step explanation for the whole method will be discussed in Chapter 5.

CHAPTER 5

A METHOD FOR PROTEIN SECONDARY STRUCTURE PREDICTION USING NEURAL NETWORKS AND GOR-V

5.1 Introduction

Secondary structure prediction methods are useful when we are unable to detect a homologue to the sequence under investigation. When a protein sequence shows clear similarity to a protein of known three dimensional structures, which is determined by laboratory methods, then the most accurate methods of predicting the secondary structure is sequence alignment methods. These methods of sequence alignments usually use dynamic programming algorithms in a process known as homology modeling. Sequence alignment methods are much more accurate than other secondary structure prediction methods (Cuff and Barton, 2000; Rost, 2003).

This chapter describes the techniques and methods used in developing and implementing algorithms and programs to achieve a better prediction method for the protein secondary structure from its primary sequence. The process of obtaining and generating the multiple sequence alignments that adds distant information to the prediction methods is explained in this chapter. The newly developed method combines neural networks with GOR-V (NN-GORV-I), and further improved by a filtering mechanism (NN-GORV-II). The framework of these experiments and all the methods studied and implemented in this research are discussed in this chapter.

5.2 Proposed Prediction Method – NN-GORV-I

Two newly developed protein secondary structure prediction methods (NN-GORV-I and NN-GORV-II) are described in details in this chapter. Other five well established prediction methods are studied in this research work and briefly described here. The seven methods are experimentally implemented in this research. This chapter mainly extends and explains the previous methodology chapter.

5.2.1 NN-I

Through the decades, the desire of people to produce artificial systems capable of sophisticated, intelligent computations similar to that of the human brain inspired the field of Artificial Neural Networks research (Wu and McLarty, 2000; Feraud and Clerot, 2002). However, most people agree on that Artificial Neural Networks is a network of many simplified unit or processors each have small amount of *local* memory. The units are connected by communication channels or connections which usually carry numeric and symbolic encoded data. These units operate only on their local data and input data they receive through the connections.

The neural network used for NN-I is the same as described by (Qian and Sejnowski, 1988). The NN-I uses no multiple alignments sequences in this experiment but it formed the basis for the NN-II when multiple sequence alignment is included. A detailed description of the network architecture, coding, training, and optimization is described in the NN-II section in this chapter.

5.2.2 GOR-IV

GOR method is based on information theory and is developed by (Garnier, *et al.,* 1978). GOR-IV uses all possible pair frequencies within a window of 17 amino acid residues Garnier *et al*. (1996) tested on a data base of 267 protein chains containing 63,566 residues. The GOR-IV algorithm is based on the information

theory combined with the Bayesian statistics. The theory behind GOR algorithm is described in details in the review chapters while the explanation and implementation of GOR method are described in details in the GOR-V section in this chapter. A fundamental difference between NN-I and GOR-IV against the other methods described in this chapter is that they use no multiple sequence alignment as explained earlier.

5.2.3 Multiple Sequence Alignments Generation

To automate the process of generating the multiple sequence alignment for large number of protein sequences in the experiments, a PSIBLAST search of the *nr* database (release 2004) which contains 198,742 entries is conducted. The method removes very long, very short and unrelated sequences. However it does allow sequences that are longer than the query, and are related, to be included after truncation. The sequence similar proteins selected by this method are then aligned by CLUSTALW (version 1.83) with default parameters.

The *nr* database is described by NCBI as "All non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF" protein. The *nr* contains essentially all the protein entries that there are. The same sequence may be present with different gi numbers as a GenBank entry, an EMBL entry, a SwissProt entry, etc. The "non-redundant" aspect of the organization is that the actual sequence for redundant entries is only represented once, hence only searched once. If there is a match in a BLAST search, links to all the entries corresponding to that sequence are then given (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein).

The multiple sequence alignments are modified so that they do not contain gaps in the first or *query* sequence, since with the current BLAST algorithms, gaps in the first sequence tend to reduce the accuracy of the prediction, or cause the program to fail to execute. This is slightly different method compared to the PHD (Profile network from HeiDelberg). This is why the gaps at the end of the target sequence are removed.

The reference secondary structures for the CB513 database is defined by DSSP (Kabsch and Sander, 1983), STRIDE, and DEFINE definitions. The DSSP (Dictionary of Secondary Structure Prediction) definition is reduced to 3 state models as will be shown later in this chapter. Cuff and Barton (1999) have shown that the exact mapping of DSSP output to three states secondary structure may have a significant effect on the estimated secondary structure prediction accuracy. Therefore, a consistent assignment or mapping of 3 states is used to test or to test and train all the methods used in this research.

Multiple sequence alignment is performed for all the training dataset sequences. The *nr* data base is formatted using the *formatdb* program form NCBI to generate sequences that could be searched by *blastpgp* program of PSI-BLAST (Altschul *et al*., 1997) to generate homologous sequences. Both *formatdb* and *blastpgp* are used with their default parameters.

CLUSTALW (version 1.83) (Thompson *et al*., 1994) is applied to generate multiple sequence alignments. CLUSTALW is implemented using *Gonnet* matrix and BLOSUM62 (Henikoff and Henikoff, 1994) matrix keeping other parameters as its default parameters.

The alignments are represented as profiles for input to the neural networks. The profiles are either presented to the networks as simple frequency counts for each amino acid through the column in the alignment and this is resemble a PHD like algorithm (Rost and Sander, 1994), or as each residue in an alignment column is scored by it corresponding BLOSUM62 matrix score and this is resemble a PSIPRED like algorithm (Jones *et al*., 1992; Jones, 1999a).

Since PSIBLAST is an iterative searching method, during iterations, it is possible for the searching profile to be populated with sequences of low similarity to the query sequence, or on the other hand sequences with high or significant similarity to the query sequence not to be included in the profile. This can be caused by matching sequences of biased composition. *pfilt* (Jones *et al.,* 1999a; Jones and Swindells, 2002) and *trimmer* (Saqi *et al.,* 1992) programs are used to filter the

searched database and to mask out regions of low complexity sequence and coiled coil regions and transmembrane helices.

The profiles generated from the multiple sequence alignment process are used in the prediction process for all the five remaining methods as will be explained in the following sections.

5.2.4 Neural Networks (NN-II)

NN-II represents the neural networks that have been described earlier in this report using the multiple sequence alignment. The mathematical representation, generation of the networks, optimization, training and testing the networks are described in this section.

5.2.4.1 Mathematical Representation of Neural Networks

A brief mathematical and logical description and representation of what is done in the NN experimental work is shown through this section. As shown in Figure 5.1, if the inputs (x_i) and the output (y_i); (y_i) which is a function of the inputs (x_i). That is (y_i) = f_i (x_i) is estimated as shown in equations (5.1) and (5.2).

$$x_i = \sum_{j \in N-(i)} w_{ij} y_j + w_i \tag{5.1}$$

$$y_i = f_i(x_i) = f_i\left(\sum_{j \in N-(i)} w_{ij} y_j + w_i \right) \tag{5.2}$$

Where, w_i is the bias or threshold of the unit i.

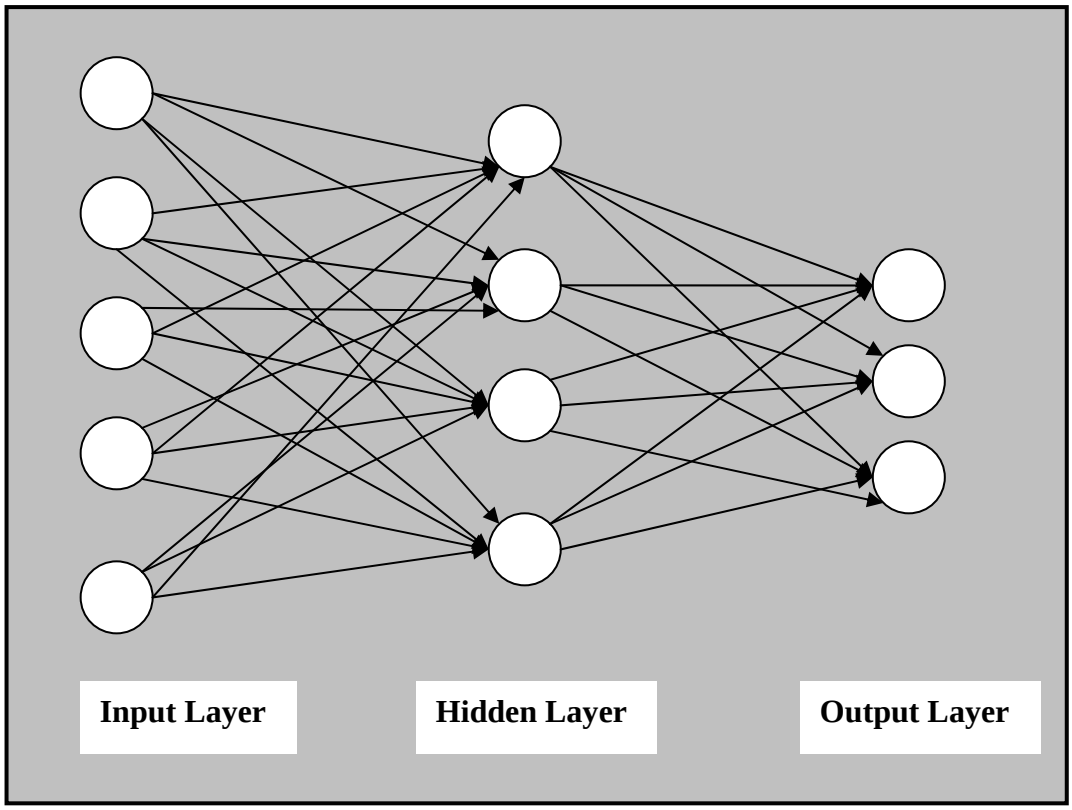


Figure 5.1: Basic representation of multilayer perceptron artificial neural network

Non linear transfer or activation functions (Figure 5.2) like sigmoid transfer function (Equation 5.4), logistic activation function (Equation 5.6), *tanh* or softmax functions (Equation 5.5) are used in the optimization process at the very beginning to observe which one contributes to better results. When f is a threshold or bias function, then

$$f(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

$$f(x) = \sigma(x) = \frac{1}{1+e^{-x}} \tag{5.4}$$

$$y_i = \frac{e^{-x_i}}{\sum_{k=1}^{n} e^{-x_k}} \tag{5.5}$$

$$y_1 = \frac{e^{-x_1}}{e^{-x_1} + e^{-x_2}} = \frac{1}{1+e^{-(x_2-x_1)}} \tag{5.6}$$

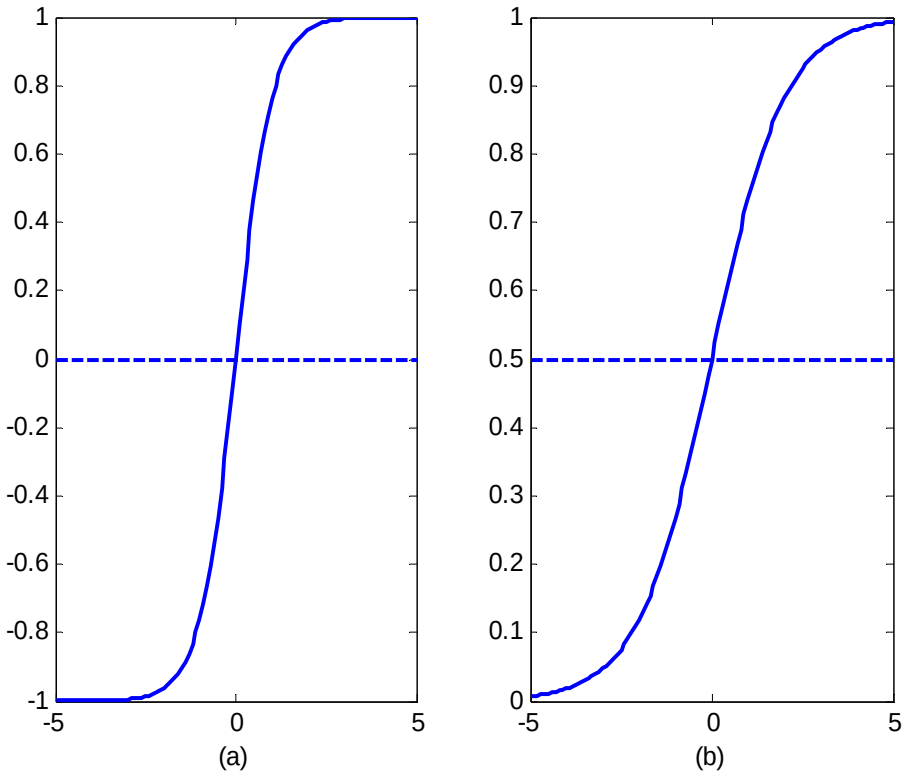


Figure 5.2: The sigmoidal functions usually used in the feedforward Artificial Network. (a) Hyperbolic tangent sigmoid transfer function or bipolar function (b) Log sigmoid transfer function or uniploar function

Considering equation (5.6) in a probabilistic model (Devroye *et al.*, 1996), any probability distribution (p_i) where ($1<=i<=n$) can be represented in normalized exponential equation from a set of variables x_j ($1<=j<=m$) as shown in equation (5.7).

$$p_i = \frac{e^{-x_i}}{\sum_{k=1}^{m} e^{-x_k}} \tag{5.7}$$

One of the most important properties of Artificial Neural Networks is that they can approximate any reasonable function to any degree of precision (Hornik *et al.*, 1990; Hornik *et al.*, 1994). If we have a continues function y = *f(x)* where both *y* and *x* are one dimensional units and if *x* changes in the interval [0,1], thus the value of x within a precision ε, where *f* is continuous over the compact interval [0,1], then there exists an integer *n* such that:

$$|x_2 - x_1| \le \frac{1}{n} \Rightarrow |f(x_2) - f(x_1)| \le \varepsilon \qquad (5.8)$$

Then *f* can be approximated with a the function g*(x)=f(k/n)* for any *x* in the interval *[(k-1/n, k/n)]* and any unit representing *k=1,…,n*.

If the data of our Artificial Neural Networks is assumed to be consisting of a set of independent input-output pairs $D_i = (d_i , t_i)$ where d_i is the input for unit *i* and t_i is the output for unit *i*. The Artificial Neural Networks operation is then a deterministic one as seen in equation (5.9).

$$\mathrm{P}\big((d_i, t_i)\,|\,w\big) = \mathrm{P}(d_i \mid w)\mathrm{P}(t_i \mid d_i, w) = \mathrm{P}(d_i)\mathrm{P}(t_i \mid d_i, w) \qquad (5.9)$$

Hence inputs *d* could be assumed as independent of the parameter *w*, using the Bayesian inference, equation (5.9) can be transformed into

$$-\log \mathrm{P}(w/D) = -\sum_{i=1}^{K} \log \mathrm{P}(t_i \mid d_i, w) - \sum_{i=1}^{K} \log \mathrm{P}(d_i) - \log \mathrm{P}(w) + \log \mathrm{P}(D) \qquad (5.10)$$

In the case of Gaussian regression, the probabilistic model assuming that the covariance matrix is diagonal and that there are *n* output units indexed by *j*, then

$$\mathrm{P}(t \mid d, w) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left(-\frac{(t_j - y_j)^2}{2\sigma_j{}^2}\right) \qquad (5.11)$$

With standard devotions as additional parameters assumed to be constant, then the negative log likelihood for this input is:

$$E = \sum_j \left(\frac{\left(t_j - y_j\right)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi - \log \sigma \right) \qquad (5.12)$$

The derivative of the log likelihood *E* with respect to an output y_i is shown in equation (5.13) which represents the regular least mean square (LMS) error function.

$$\frac{\partial E}{\partial y_j} = \frac{\partial E}{\partial x_j} = -\frac{t_j - y_j}{\sigma_j} = -\frac{t_j - y_j}{\sigma} \tag{5.13}$$

For Artificial Neural Networks that classify an input into two classes (a and $\bar{a}$) like Helix or not Helix, the target output can represented as 0 or 1. This model is a binomial model and can be estimated by a sigmoid transfer function as shown in equation (5.14).

$$y = y(d) = \mathrm{P}(d \in A) = \mathrm{P}(t \mid d, w) = y^t (1-y)^{(1-t)} \tag{5.14}$$

The relative entropy (the amount of information to describe a variable) between the output distribution and the observed distribution is expressed by:

$$E = -\log \mathrm{P}(t \mid d, w) = -t \log y - (1-t) \log (1-y) \tag{5.15}$$

Where d is data t is target.

If the output transfer function is the logistic function, then:

$$\frac{\partial E}{\partial y} = -\frac{t-y}{y(1-y)} \tag{5.16}$$

$$\frac{\partial E}{\partial x} = -(t-y) \tag{5.17}$$

Consequencely, in binomial classification, the output transfer function is logistic; the likelihood error function is the relative entropy between the predicted distribution and the target distribution.

If the classification task of our Artificial Neural Networks has *n* possible classes ($a_1,\ldots,a_n$) for a given input *d*, as it is in our case here three classes, the target output *t* is a vector with a single 1 and n-1 zeros. The probabilistic model for this task is a multinomial (polynomial) model. Thus the equations governing this classification task are shown in Equations (5.18-5.21).

$$\mathrm{P}(t \mid d, w) = \prod_{j=1}^{n} y_j^{\,t_j} \tag{5.18}$$

$$E = -\log \mathrm{P}(t \mid d, w) = -\sum_{j=1}^{n} t_j \log y_j \tag{5.19}$$

$$\frac{\partial E}{\partial y_j} = -\frac{t_j}{y_j} \tag{5.20}$$

$$\frac{\partial E}{\partial x_j} = -(t_j - y_j) \tag{5.21}$$

In general, a network containing a large enough number of hidden nodes can always map an input pattern to its corresponding output pattern (Rumelhart & McClelland, 1986). A number of different hidden layers are attempted in this study to reach an optimal mapping of the data set to the required classes.

5.2.4.2 Generating the Networks

The Stuttgart University SNNS neural network simulator program version 4.2 downloaded from the site: ftp://ftp.informatik.uni-stuttgart.de (Zell *et al.*, 1998) is used in this experimental work. SNNS for UNIX X Windows is used to generate many rapid prototypes of neural networks. SNNS's snns2c program is used to convert the simulated networks into ANSI C functions codes that are included in the main C program.

At the end of the experiments, several neural networks are generated using several coding and teaching methods:

i- The conventional method of Quian and Sejnowski (1988) where binary coding is adopted. The 20 amino acids and the three secondary structures are given binary codes to be fed to the neural network. The three target secondary structure outputs are coded as (1 0 0) for α helices, (0 1 0) for β strands and (0 0 1) for coils.

ii- The architecture and coding used in the PHD (Rost and Sander, 1994) is followed here with minor modification.

iii- The profile generated by PSI-BLAST is used in this experiment as explained earlier. This method uses the prior knowledge of amino-acid relationships embodied in the substitution BLOSUM matrix to generate residue pseudo-count frequencies, which are averaged with the observed frequencies to estimate the probability that a residue is at specific position in the query sequence (Henikoff and Henikoff, 1992). The different sequences are weighted accordingly to the amount of information they carry. (Altschul *et al*., 1997; Tatusov *et al*., 1994).

Sliding windows of 17 and13 for both the profiles and single sequences are used. This means that to predict a residue, eight (for windows of 17) and six (for windows of 13) previous residues and eight and six following residues are taken into consideration to predict the residue at the central position of the window. Then the window is shifted residue by residue through the protein (Qian and Sejnowski, 1988; Rost and Sander, 1993).

Many Artificial Neural Networks architectures with varying parameters are used in this work. The output of some neural networks is fed to other networks that classified this output into the three structures of protein (H, E, and C), and here the networks followed a polynomial model as explained in this chapter. Sigmoid transfer function and *tanh* function are attempted to optimize the networks. The artificial neural networks parameters are varied continuously in an attempt to arrive at a conclusion of a better if not a best optimized model(s).

5.2.4.3 Networks Optimization

Optimizing the Artificial Neural Networks that are designed includes varying the input representation, the numbers of hidden nodes, and the number of training examples, the biases, and the activation functions. In each case, the network performance is evaluated and tabulated for each network architecture or training condition. The best performing network or networks which performed best on both the training and testing sets is chosen for further network architectures or final evaluation (Siegelmann, 1998; Siegelmann and Sontag, 1999).

Cross validation which is the permutation of training and testing sets and train a number of times on each set, while reporting the best performing network for each simulation is used. This occurs when the error space is uneven or rough which leads to the local minima problem. Seven Cross validation is used in this experimental work.

Memorization or over-fitting is one of the main nuisances to the network where the network learns the training examples, rather than the general mapping from inputs to outputs. This problem is tackled by reducing node connectivity (network pruning), reducing the number of input nodes, and/or reducing the number of hidden nodes.

In addition, the training process of Artificial Neural Networks is not a one time event; it takes several rounds of training in order to arrive at a good parameter size and configuration. Several very powerful machines using LINUX platforms are used through a period of three years of this experimental work.

In this experiment, there are two levels of neural networks; a sequence to secondary structure network, with a window of 17 amino acids and a structure to structure network, with a window of 17 amino acids. The structure to structure network, improves prediction of the final length distributions of secondary structures. The training applied in this method is the unbalance training, where percentage of amino acid composition, sequence length, and insertions and deletions are not considered here.

The artificial neural network consisted of several networks. The first is a network with a sliding window of 17 residues over each amino acid in the alignment as input layer. The input layer is connected with nine nodes as hidden layer which in turn is connected to three nodes as output layer. The neural network used in this method is the standard three-layered fully connected feed-forward networks with the back-propagation having momentum learning rule in order to avoid oscillation problems. The width of the gradient steps is set to 0.05 and the momentum term is 0.2 (Rost and Sander, 1993). The initial weights of the neural networks are chosen randomly in the range of [-0.01, 0.01]. The learning process consists of altering the weights of the connections between units in response to a teaching signal which provides information about the correct classification in input terms. The difference between the actual output and the desired output is minimized.

All the neural networks have been trained on the 480 proteins set. The network outputs can be seen as estimated probabilities of correct prediction, and therefore they can indicate the confidence level of each predicted residue. (Riis and Krogh, 1996).

5.2.4.4 Training and Testing the Network

Seven-fold cross-validation is used on the 480 data sets to test the methods efficiencies. The whole data set is randomly divided into 7 subsets of equal size. In each validation, one subset is used for testing while the rest is used for training. Several parameters are regulated to optimize the training. Back-propagation with momentum networks which used the (0.05 -0.05) is implemented for this network.

The process of training the designed network involves presenting the network with an input pattern which is protein sequence data set, propagating the pattern through the architecture, comparing the network output to the desired output, and altering the weights in the direction so as to minimize the difference between the actual output and the desired output

Back-propagation algorithms which involve two passes through the network, a forward pass and a backward pass, are used in this training process. The online version of the back-propagation algorithm is simulated using gradient descent function. For each training pattern, if we have any weighted parameter w_{ij}, then

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_i}\frac{\partial y_i}{\partial w_{ij}} = \frac{\partial E}{\partial y_i} f_i^{'}(x_i) y_j \tag{5.22}$$

If n is the learning rate, y is the output of the unit from which the connection originated (*presynaptic activity*), and E is the back-propagation error (*postsynaptic activity*), the gradient descent learning equation is a product of these three terms (n, y, and E) as shown in equation (5.23).

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} = -\eta \in_i y_j \tag{5.23}$$

Then the back-propagation error is estimated by:

$$\in = (\partial E / \partial y_i) f_i^{'}(x_i) \tag{5.24}$$

A recursive implementation of this back-propagation error can be written as shown in equation (5.25).

$$\frac{\partial E}{\partial y_i} = \sum_{k \in N+(i)} \frac{\partial E}{\partial y_k} f_k^{'}(x_k) w_{ki} \tag{5.25}$$

Regardless of the training steps or equations, the main goal of the network is to minimize the total error of each output node over all training examples.

Pearson's correlation coefficient that measures the degree how much the input (X) is normalized with output (Y), is used as shown in equation (5.26):

$$Corr(X,Y) = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}} \qquad (5.26)$$

where $i \in [1, n]$

5.2.5 GOR-V

The idea of GOR-V was an experimental study to improve the existing GOR algorithms. It depends mainly on some important suggested modifications and improvements to the previous GOR algorithms to predict protein secondary structures from amino acid sequences (Kloczkowski *et al.*, 2002).

For understanding of GOR-V, it is better to introduce an accuracy matrix [Aij] of the size 3 x3 (where i and j stand for the three states H, E, C) to measure the quality of protein secondary structure prediction. The ij^{th} element A_{ij} of the accuracy matrix is then the number of residues predicted to be in state j, which according to the DSSP data are actually in state i. Then the sum over the columns of matrix A gives the number of residues n_j that are predicted to be in state j[3]:

$$n_j = \sum_{i=1}^{3} A_{ij} \qquad (5.27)$$

This equation can be written as:

$$N_j = \sum_{i=1}^{3} A_{ij} \qquad (5.28)$$

This can be viewed as that the diagonal elements of A count the correct predictions for each of three structural states, and the off-diagonal elements contain the information about wrong predictions (Kloczkowski *et al.*, 2002).

The Q_3 is the main parameter measuring the accuracy of the protein secondary structure prediction; it is calculated by the following equation which estimates the percentage of all correctly predicted residues within the three-state (H, E, C) classes.

$$Q_3 = \frac{\sum_{i=1}^{3} A_{ii}}{N} \times 100 \tag{5.29}$$

N is the total number of residues in the sequence which also can be written as:

$$N = \sum_{i=1}^{3} N_i = \sum_{i=1}^{3} n_i \tag{5.30}$$

The correctness of prediction for each of the structural classes (H,E,C) are measured by the following parameters:

$$q_i = \frac{A_{ii}}{N_i} \times 100 \tag{5.31}$$

where i=*H,E,C*

The first GOR work was based on the information theory and naive Bayesian statistics. The information function I(S,R), is one of the basic mathematical tools of the information theory, which is written as:

$$I(S;R) = \log[P(S \mid R) / P(S)] \tag{5.32}$$

The information function here is defined as the logarithm of the ratio of the conditional probability P(S|R) of observing conformation S, where S is one of the three states: helix (H), extended (E), or coil (C)] for residue R, where R is one of the 20 possible amino acids, and the probability *P(S)* of the occurrence of conformation *S*.

The data base used to calculate this information and naive probabilities is the 480 proteins of the CB513 proteins, where the secondary structure is known for each amino acid. The conformation state of a given residue (i.e. in which state (H, E, or C)

this residue will be) in the sequence depends not only on the type of the amino acid *R* but also on the neighboring residues along the chain within the sliding window.

The GOR algorithms used windows of 17 residues. This indicates that for a given residue *R*, eight immediate nearest neighboring residues on each side are analyzed. If R is considered as $R_{0,}$ then R_{+8} and R_{-8} are the immediate neighboring residues. GOR-IV method calculates the information function as the sum of the logarithmic information from single residues which is known as *singlets* and pairs of residues which is known as *doublets* within a window of width 2d +1, where d = 8, for the window of 17 residues.

Using the data base the first summation is calculated over doublets and the second summation is over singlets within the window centered around the j-th residue. The pair frequencies of residues Rj and Rj+m with Rj occurring in conformations Sj and n-Sj are calculated from the database. Thus, using the frequencies calculated from the databases, the algorithm can predict probabilities of conformational states for any new sequence. The prediction of secondary structure is performed by either predicting the secondary structure having the highest difference information functions, or computing the probability that the residue is in state S_j from the difference information.

GOR-V depends mainly on the fact that the objective study of the prediction of the secondary structure of the GOR method is by using multiple alignments. Several improvements have been applied to the GOR-IV algorithm to increase the accuracy of the secondary structure prediction from a single sequence and from the multiple alignments. In training this process, the seven fold cross-validation is used.

5.2.6 NN-GORV-I

The algorithms of the GOR-V and neural networks (NN) described above are combined in this method to attain a good performance predictor. NN-GORV-I depends on the assumption that combining information in prediction may increase the prediction accuracy. Up to date of writing this report there is no method implemented combining GORV with neural networks. NN-GORV-I is further implemented in slightly different way called NN-GORV-II which will be described in the next section. The general model for the newly developed method is shown in Figure 5.3. A general model for this method and its advanced version NN-GORV-II is shown in the next sections.

GOR-I to GOR-IV used windows of 17 residues. This indicates that for a given residue R, eight immediate nearest neighboring residues on each side are analyzed. If R is considered as $R_{0,}$ then R_{+8} and R_{-8} are the immediate neighboring residues. The information theory allows the information function of a complex event to be decomposed into the sum of information of simpler events, which can be written as:

$$I(\Delta S; R_1, R_2, .. R_n) = I(\Delta S; R_1) + (\Delta S; R_2 \mid R_1) + I(\Delta S; R_3 \mid R_1, R_2) + ... + I(\Delta S; R_n \mid R_1, R_2, .. R_{n-1}) \quad (5.33)$$

Where how much information difference is written as:

$$I(\Delta S; R_1, R_2, ...R_n) = I(S; R_1, R_2, ...R_n) - I(n - S; R_1, R_2, ...R_n) \quad (5.34)$$

Where n-S are the confirmations that are not S, i.e if S is happened to be E then n-S is the others two states H and C.

The previous GOR-IV method calculates the information function as the sum of the logarithmic information from single residues which is known as *singlets* and pairs of residues which is known as *doublets* within a window of width 2d +1, where d = 8, for the window of 17 residues:

$$\log\frac{P(S_j, LSeq)}{P(n-S_j, LSeq)} = \frac{2}{2d+1}\sum_{n,m=-d}^{d}\log\frac{P(S_j, R_{j+m}, R_{j+n})}{P(n-S_j, R_{j+m}, R_{j+n})} - \frac{2d-1}{2d+1}\sum_{m=-d}^{d}\log\frac{P(S_j, R_{j+m})}{P(n-S_j, R_{j+m})} \quad (5.35)$$

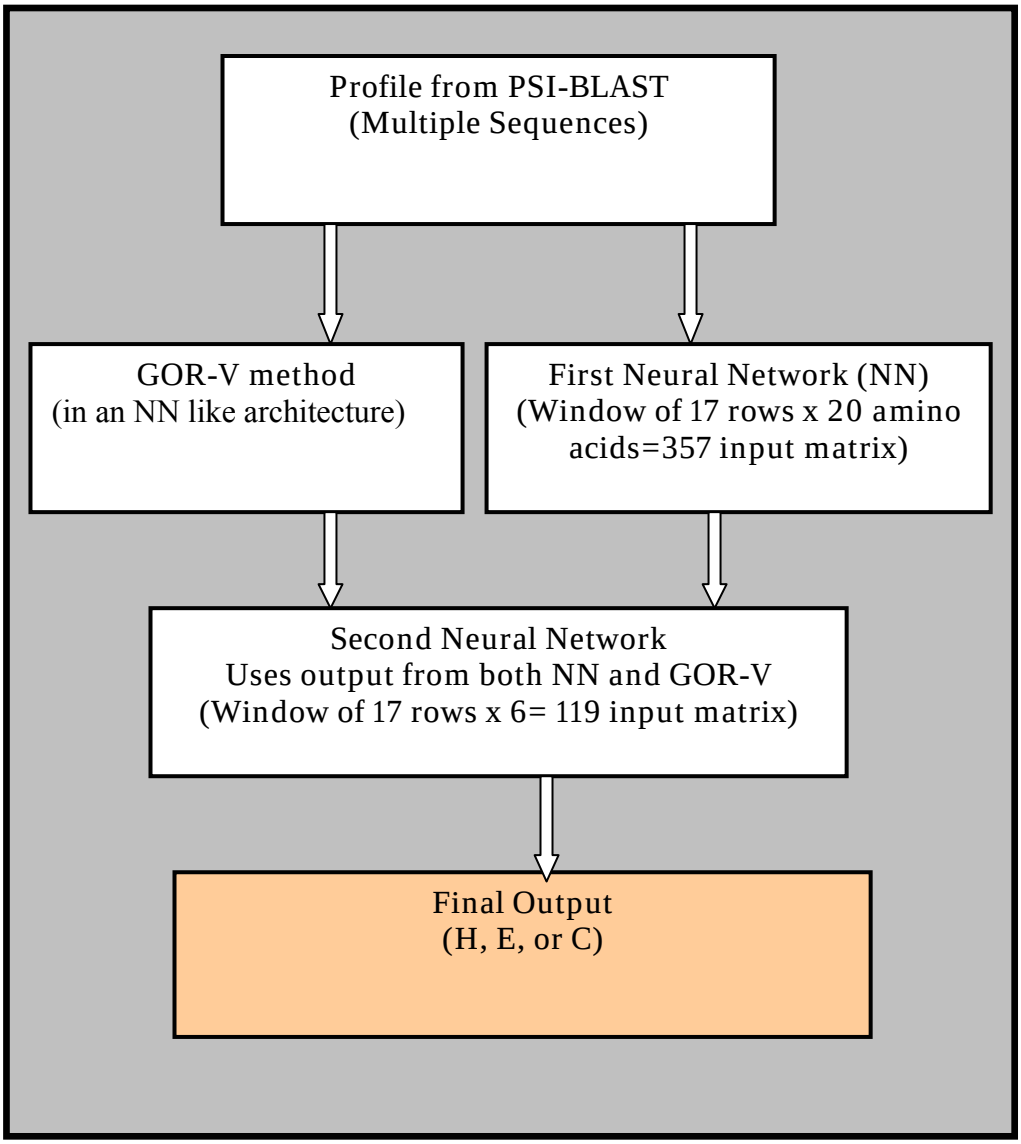


Figure 5.3: A general model for the newly developed protein secondary structure prediction method.

A detailed representation for the NN-GORV-I method is shown in figure 5.4. From the detailed figure it is elucidated that there is no filtering mechanism used in this version, unlike the advance version NN-GORV-II which uses the *pfilt* program to mask low complexity regions of the *nr* database sequences. The NN-GORV-II is explained in the next section.

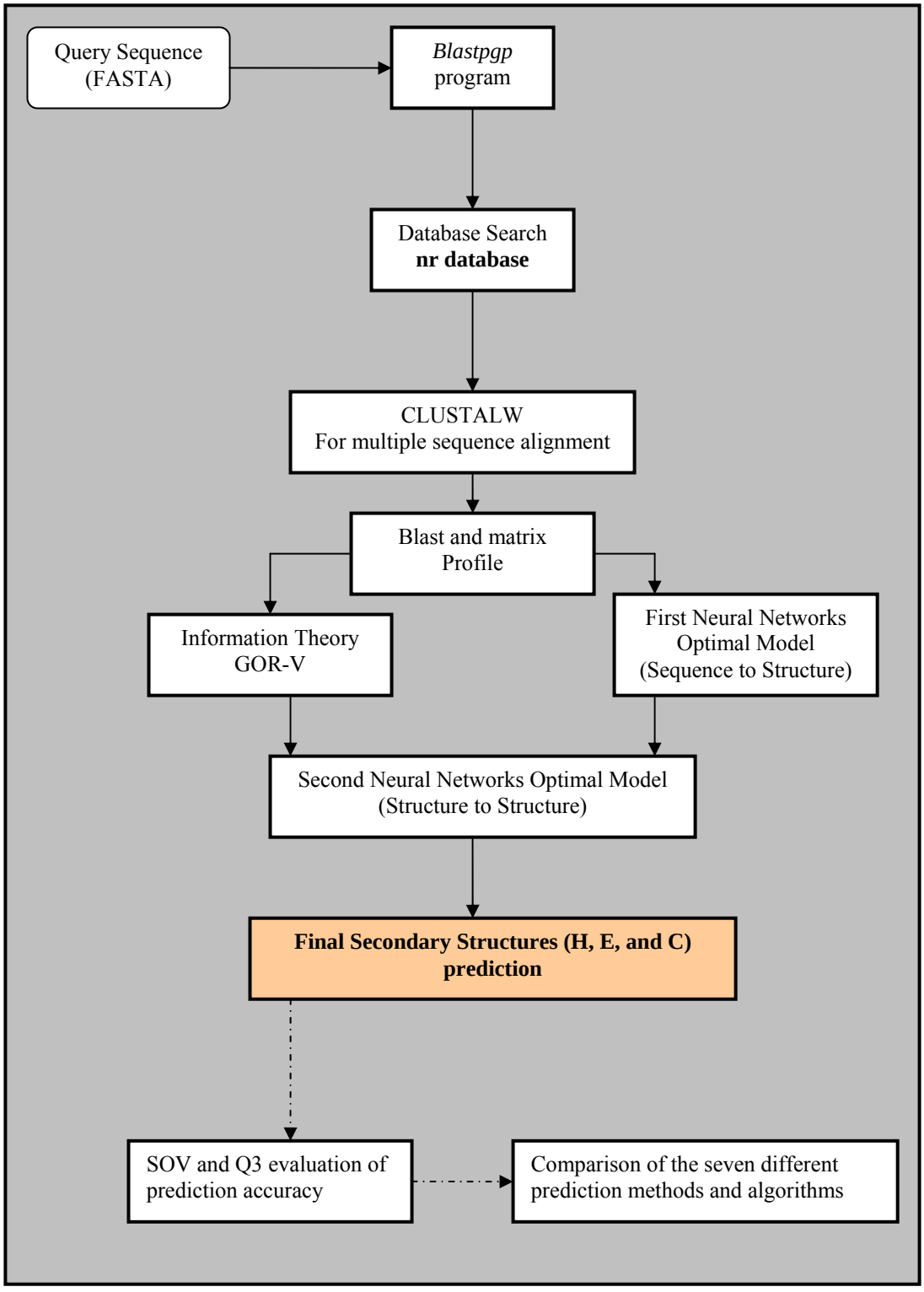


Figure 5.4: A detailed representation for the first version of the newly developed protein secondary structure prediction method (NN-GORV-I)

From the data base, the first summation is calculated over doublets and the second summation is over singlets within the window centered around the j-th residue. The pair frequencies of residues Rj and Rj+m with Rj occurring in conformations Sj and n-Sj are calculated from the database. Thus, using the frequencies calculated from the databases, the algorithm can predict probabilities of conformational states for any new sequence. The prediction then either to predict the secondary structure having the highest difference information function, or compute the probability that the residue is in state S_j from the difference information as follows:

$$p(S_j;R)=\frac{1}{1+\frac{p(\bar{S}_j)}{p(S_j)}\exp[-I(\Delta S_j;R)]} \tag{5.36}$$

The GOR algorithm reads a protein sequence in the FASTA format and then predicts its secondary structure. For each residue Ri along the sequence, the algorithm calculates the probabilities pH, pE, and pC of the secondary structure prediction (H, E, or C). The probabilities are then normalized to be in the range between 0 and 1 by the following formula:

$$p_H + p_E + p_C = 1 \tag{5.37}$$

GOR-V depends mainly on the fact that the objective study of the prediction of the secondary structure of the GOR method is by using multiple alignments. Several improvements have been applied to the GOR-IV algorithm to increase the accuracy of the secondary structure prediction from a single sequence and from the multiple alignments. Here the seven fold cross-validation is used.

The modifications and improvements to the original GOR algorithms are explained as follows:

1. The data base has been increased to 480 proteins, a manipulated set of CB513 proteins, compared to the previous GOR database of 267 sequences. The properties and source of this data base is explained previously in this chapter. The use of this database allows an objective and unbiased calculation of the

accuracy of the prediction, as well as easy comparison with results of other prediction algorithms that use similar non-redundant database sequences in their prediction methodologies.

2. The latest version of the GOR-IV algorithm used a window with a fixed width of 17 residues as explained earlier; with eight residues on both sides of the central residue or amino acid. A resizable window for the GOR-V algorithm is used here according to the length of the sequence. Studies showed that the accuracy of the prediction is slightly better for a smaller window of width of 13 residues. The number of triplets within a window of size N is:

$$N(N-1)(N-2)/6 \qquad (5.38)$$

According to this formula the conventional window of size 17 has 680 triplets, while and the window of size 13 has 286 triplets. The smaller sliding window of 13 will facilitate the computations compared to that of 17. More over the window of size 13 is expected to increase the accuracy of shorter sequences.

Different window sizes are used for different sequences lengths in the database as follows:

   i. Sequences 25 residues or shorter length, a sliding window size of seven residues is used.
   ii. Sequences greater than 25 and less than or equal to 50 residues length, a sliding window of nine residues is used.
   iii. Sequences greater than 50 residues long and less than 100 residues, a sliding window of 11 residues is used.
   iv. Sequences greater than 100 residues long and less than 200 residues, a sliding window of 13 residues is used.
   v. Sequences greater than 200 residues long, a window size of 17 residues is used.

3. The previous GOR algorithm had a tendency to over-predict the coil state (C) at the cost of the beta-strands conformation (E), and to a lesser extent at the

cost of alpha-Helical confirmation (H).These parameters haven optimized by introducing decision parameters. The idea of decision constant (adjustable weights) had been applied successfully in PSIPRED algorithm (Jones, 1999). The predicted probability of the coil (C) conformation was set to a value greater by some determined margins than the probability of either the (H) or (E) states to accept C as the predicted confirmation.

If the secondary structure of the j^{th} residue is assigned to the conformation with the largest (winning) probability value:

$$\max\left\{\left\langle P_H(j)\right\rangle, \left\langle P_E(j)\right\rangle, \left\langle P_C(j)\right\rangle\right\} \qquad (5.39)$$

The above assignment equation (Equation 5.39) is modified by introducing decision constant thresholds, such that:

$$\max\left\{\left\langle P_H(j)-0.075\right\rangle, \left\langle P_E(j)-0.15\right\rangle, \left\langle P_C(j)\right\rangle\right\} \qquad (5.40)$$

According to the above equation, the coil state will be selected as the predicted state only if the calculated probability of the coil conformation is greater than the probability of the other states by (0.15 for strands (E) and 0.075 for helices (H))

4. The previous versions of the GORs algorithms used only single residue statistics or combination of the single residue and pair residue statistics within the window. GOR-V algorithm estimates *singlets*, *doublets*, and *triplets* statistics of the secondary structure prediction. However, in this experiment, the triplet statistics complicated the optimization of the prediction and did not increase the prediction accuracy significantly. The triplet statistics within the sliding window had not been included in this experiment.

5. Unlike the previous GOR methods, PSIBLAST multiple sequence alignments for each protein sequence in the database had been used here for the secondary structure prediction. PSIBLAST program is executed as described earlier in this chapter using the nr database with default parameters. In cases where there is no convergence for the alignment process; that is the *blastpgp* program of PSIBLAST is unable to find any hits or alignments, the original

single sequence is used for the prediction algorithm. The alignments produced by PSIBLAST that are too similar to the query sequence are removed using *trimmer* program. Sequences in the alignment with sequence identity threshold greater than 97% to the query sequence are removed from the alignment.

5.2.7 Enhancement of Proposed Prediction Method - NN-GORV-II

The prediction of NN-GORV-I algorithm is further improved by implementing the *pfilt* program. The *pfilt* program is a filter that masks trans-membrane regions, coiled-coil and compositional bias in a query sequence (Jones and Swindells, 2002). The implementation of the *pfilt* program developed a different or advanced version of the prediction method called NN-GORV-II.

To portray a clear picture of the development and implementation of these algorithms and methods, the general framework of Figure 5.3 is extended and explained as shown in Figure 5.5. The process of predicting secondary structure of a protein (amino acid sequence) begins with the sequence in a FASTA format which is here is the query sequence. The prediction process is simply how to assign a given amino acid residue (there are 20 residues) one of the three secondary structure states or confirmation (helix, strand, or coil).

The query sequence will then be checked against a search database of very big number of non redundant sequences to find any homologue. If an exact homologue is found, the sequence is then predicted from the first step. The NCBI nr data is used as a searched non redundant database in these experiments.

The *pfilt* program is then implemented to mask the nr database sequences. CLUSTALW will then be implemented to perform multiple sequence alignment. The PSI-BLAST is the program used to find homologous sequences and to generate a profile for the query sequence. At the end of this step a profile and a matrix file that

contain information from other homologue sequences (evolutionary information) will be obtained.

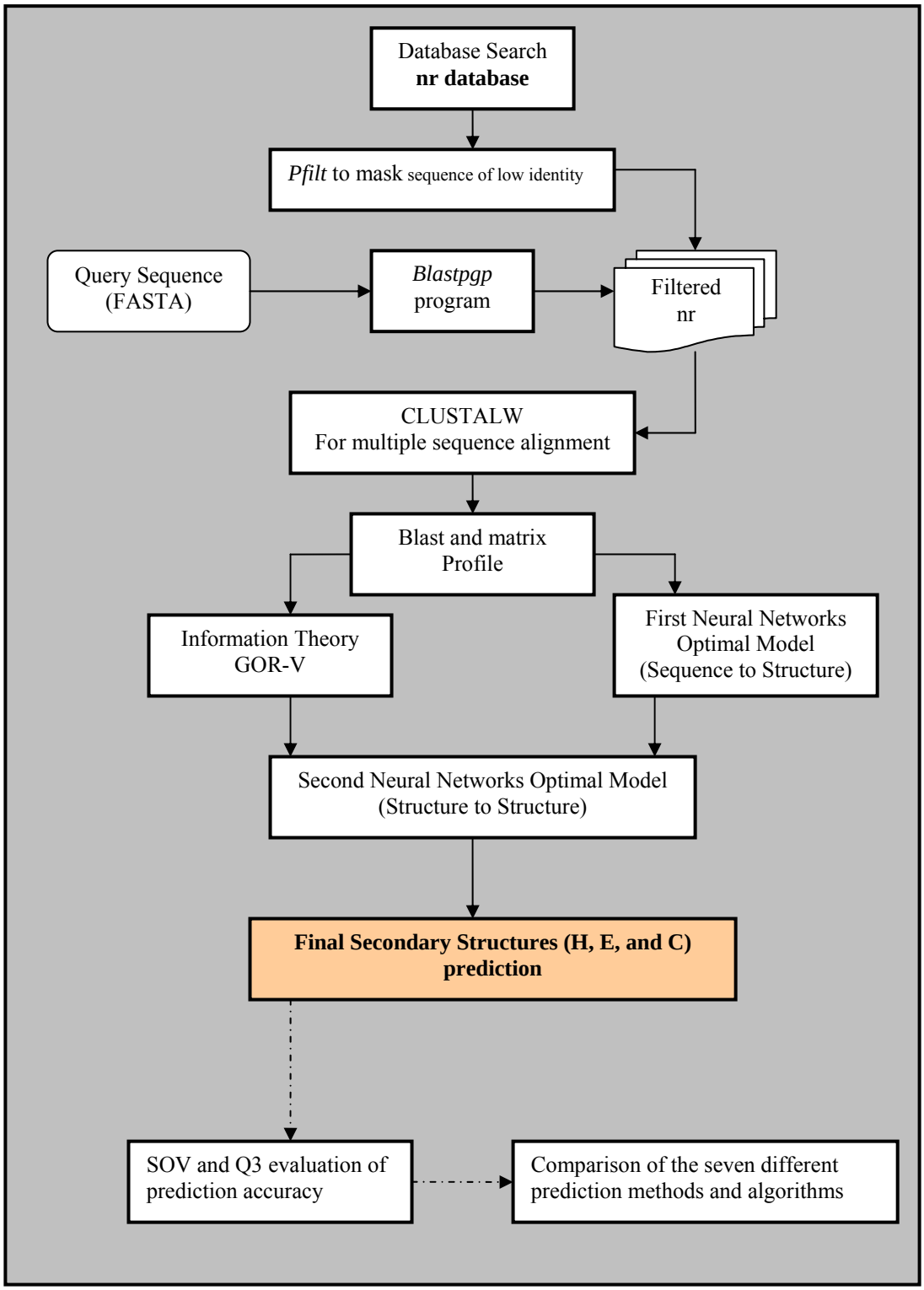


Figure 5.5: A detailed representation for the second version of the newly developed protein secondary structure prediction method (NN-GORV-II)

The GOR-V and the neural network uses the multiple sequence alignment to be fed to a second neural network and use all the information from the first network and GOR-V. The final prediction says this residue belongs to one of the three secondary states as shown in Figure 5.5.

The framework of the experiment continues to evaluate the resultant prediction of each of the seven methods. The Q performance and SOV measure is used to evaluate the prediction. The seven methods are then analyzed and compared using a variety of estimates and statistical measure to elucidate the power and weaknesses of each method.

5.2.8 PROF

PROF is a cascade multiple classifier combing many algorithms using quadratic and linear discrimination functions to group predictors in one classifier (Ouali and King, 2000). PROF is the advanced version of DSC (King and Sternberg, 1996) which applies GOR residue attributes in a quadratic model, with the addition of hydrophobicity and amino acid position, which are combined with information from the multiple sequence alignment. Optimal weights are deduced by linear discrimination, with filtering applied to remove erroneous predictions. This method is described as having an advantage that the prediction method is both implicit and effective.

PROF used a set of 496 non homologous domains that is part of the CB513 developed by Cuff and Barton (1999) described earlier. The original data base of PROF contains 82847 residues: 28678 in helix confirmation, 17741 in β strands and 36428 in coils. The secondary structure is assigned using the DSSP program (Kabsch and Sander, 1983) and assignment of DSSP eight states to three states is made using conservative mapping which corresponds to Method I in this work.

In this research experimental work PROF is tested using the 480 domains described earlier in this chapter. The 480 domains data set is almost the same as the

496 proteins that the original PROF is trained and tested on. The files of the 480 protein domains are renamed to other names since CLUSTALW may not read file names with dash '-' character as part of the file name. Separate RED HAT LINUX machines are dedicated for PROF and to the other methods implemented in this research to reduce the processing time of the experiments.

5.3 Reduction of DSSP Secondary Structure States

The predicted secondary structure is usually assigned to the experimentally determined tertiary or 3D structure by the DSSP (Kabsch and Sander, 1983), STRIDE (Frishman and Argos, 1995), or DEFINE (Richards and Kundrot, 1988) definitions. The 513 data set contains all these definitions. In this experiment, DSSP definition is used since it has been the most widely used secondary structure definition by researchers in this field. DSSP has eight secondary structure classes: H (α-helix), G(310-helix), I(π-helix), E(β-strand), B(isolated β-bridge), T(turn), S(bend), and - (coil).

The adopted reduction schemes of the mentioned eight classes to three states of helix (H), strands (E), and coil (C) is usually performed by using one of the following methods or schemes.

1. Method I: H,G and I to H ; E to E ; all other states to C
   (Riis and Krogh, 1996).

2. Method II: H,G to H ; E,B to E ; all other states to C
   Compared to the other reduction methods, this method is known as harder to predict. (Rost and Sander, 1994; Moult *et al.,* 1997; Moult *et al.,* 1999; Lesk *et al.*, 2001;Pollastri *et al*., 2002).

3. Method III: H,G to H ; E to E ; all other states to C
   (Kim and Park H., 2003).

4. Method IV: H to H ; E,B to E ; all other states to C
   (Kim and Park H., 2003).

5. Method V: H to H ; E to E ; all other states to C
   (Frishman and Argos, 1997; Salamov and Solovyev, 1995).

In this research, all the above mentioned schemes are attempted to study their effect on prediction performance and quality, while Method II is adopted for evaluating all the prediction algorithms. The 8-to 3-state reduction scheme can alter the prediction accuracy of an algorithm in a range of 1-3% (Cuff and Barton, 1999). In this experiment scheme 3 is adopted for the three states assignments because it is considered to be the stringent definition, which usually results in lower prediction accuracy than other definitions or reduction schemes. Scheme 5 is used to compare the affect of reduction schemes on prediction accuracy.

PERL (Practical Extraction and Reporting Language) is used to extract and parse the amino acids sequences or residues (RES) into corresponding files that contain standard FASTA format that in turn can be read for the seven methods to undergo predictions (Figure 5.6).

```
>
V K D G Y I V D D V N C T Y F C G R N A Y C N E E C T K L K G E S G Y C Q W A S P Y
G N A C Y C Y K L P D H V R T K G P G R C H
```

Figure 5.6: The 1ptx-1-AS.all file converted into a FASTA format (zptAS.fasta) readable by the computer programs.

The corresponding laboratory determined DSSP predictions of the residues are extracted and parsed into other files that contain the predicted sequences from the seven algorithms. The resulting final files are files that contain the amino acid sequence, the predicted secondary structure, and the observed secondary structure (DSSP) after being assigned into a three state scheme (Figure 5.7). PERL is used to make these files in format that is readable by SOV program (Zemla *et al.*, 1999). PERL is also used to convert the names of these files into format that is readable by CLUSTALW and PSIBLAST.

```
>OSEQ
C E E E E E E C C C C C E C C C C C H H H H H H H H H H C C C C E E E E E E E E C C E E E
E E E E E E C C C C C E C C C C C C C C
>PSEQ
C C C C E E E E C C C C C E E E E C C C C C C C C C C C H H H H C C C C E E E E E E C C C C C C E E E
E E E E C C C C C C C C E E E C C C C C C
>AA
V K D G Y I V D D V N C T Y F C G R N A Y C N E E C T K L K G E S G Y C Q W A S P Y G N
A C Y C Y K L P D H V R T K G P G R C H
```

Figure 5.7: The 1ptx-1-AS.all file parsed into a format readable by the Q_3 and SOV program

5.4 Assessment of Prediction Accuracies of the Methods

Several measures and methods are used in this work to estimate the prediction accuracy of the algorithms developed and studied in this research. The methods and measures used to assess the accuracy are further analyzed statistically to observe the significance of each. The methods implemented to assess the accuracy and significance of the predictions in this work are discussed in this section.

5.4.1 Measure of Performance (Q_H, Q_E, Q_C, *and* Q_3)

The Q_3 accuracy per residue which measures the expected accuracy of an unknown residue is computed as the number of residues correctly predicted divided by the total number of residues. The Q_3 per the whole protein is computed too using the same definition. The Q_H is defined as the total number of α helix correctly predicted divided by the total number of α helix. The same definitions are applied to Q_E (β strands) and Q_C (coils). The Q_3 is expressed as:

$$Q_3 = \sum_{(i=H,E,C)} \frac{predicted_i}{observed_i} x100 \qquad (5.41)$$

5.4.2 Segment Overlap (SOV) Measure

Segment overlap calculation (Rost *et al*., 1994; Zemla *et al*., 1999) is performed for each data set. Segment overlap values attempt to capture segment prediction, and vary from an ignorance level of 37% (random protein pairs) to an average 90% level for homologous protein pairs. In more details, the SOV aims to assess the quality of a prediction by taking into account the type and position of secondary structure segment, the natural variation of segment boundaries among families of homologous proteins, and the deviation at the end of each segment. Segment overlap is calculated by:

$$Sov = \frac{1}{N}\sum_{s}\frac{mnov(S_{obs};S_{pred}) + \delta}{mxov(S_{obs};S_{pred})} \times len(s_1) \qquad (5.42)$$

Where:

N : is the total number of residues,

$mnov$: is the actual overlap, with $mxov$ is the extent of the segment.

$len\,s_1$: is the number of residues in segment s_1.

δ is : the accepted variation which assures a ratio of 1.0 where there are only minor deviations at the ends of segments.

The Q_3 and SOV are implemented using the Q_3 and SOV ANSI C program downloaded from the web site: http://PredictionCenter.llnl.gov/

5.4.3 Matthews Correlation Coefficient (MCC)

As defined in the review chapter, the correlation is a measure of how strong two variables are related. Reconsidering the accuracy matrix $[A_{ij}]$ mentioned before, the general form of Matthews's correlation (Matthews, 1975) can be written as:

$$C_i = \frac{A_{ii}\left(\sum_{k\neq i}^{3}\sum_{j\neq i}^{3} A_{jk}\right) - \left(\sum_{j\neq i}^{3} A_{ij}\right)\left(\sum_{j\neq i}^{3} A_{ji}\right)}{\sqrt{\left(A_{ii}+\sum_{j\neq i}^{3} A_{ij}\right)\left(A_{ii}+\sum_{j\neq i}^{3} A_{ji}\right)\left(\sum_{k\neq i}^{3}\sum_{j\neq i}^{3} A_{jk}+\sum_{j\neq i}^{3} A_{ij}\right)\left(\sum_{k\neq i}^{3}\sum_{j\neq i}^{3} A_{jk}+\sum_{j\neq i}^{3} A_{ji}\right)}} \quad (5.43)$$

Matthews' correlation coefficient is performed for each of the three states. Calculating the four numbers (TP, FP, TN, and FN) discussed before, the formula of Matthews's correlation can be rewritten as:

$$C_i = \frac{p_i r_i - u_i o_i}{\sqrt{(p_i + u_i)(p_i + o_i)(r_i + u_i)(r_i + o_i)}} \quad (5.44)$$

Where:

p_i number of correctly predicted residues in conformation.

r_i number of those correctly rejected.

u_i number of the incorrectly rejected (false negatives).

o_i number incorrectly predicted to be in the class (false positive)

i = is one of the confirmation states H, E, or C.

5.4.4 Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) is typically used in a binary prediction or classification model like presence or absence, disease or normal. There are two possible prediction errors: false positives (FP) and false negatives (FN). The performance of a binary prediction model is normally summarized in a confusion or contingency matrix that cross-tabulates the observed and predicted patterns as shown in Table 5.1(Fielding and Bell,1997).

Table 5.1: The contingency table or confusion table for ROC curve

		Classified as	
Reference		**-**	**+**
	-	TN	FP
	+	FN	TP

The confusion matrix accuracy measures assume that data is real counts. The sensitivity of a test can be described as the proportion of true positives it detects of all the positives. All positives are the sum of (detected) true positives (TP) and (undetected) false negatives (FN). Sensitivity therefore can be rewritten as:

$$TP/(TP+FN) \tag{5.45}$$

While the specificity of a test can be described as the proportion of true negatives it detects all the negatives. It is thus a measure of how accurately it identifies negatives. All negatives are the sum of (detected) true negatives (TN) and (miss-predicted) false positives (FP). Specificity can therefore be rewritten as:

$$TN/(TN+FP) \tag{5.46}$$

As it can be seen from Table 5.1, the sensitivity and specificity do not use all information from the above confusion matrix. An ideal confusion matrix-based measure should meet four requirements and obey six additional constraints. In particular, it should measure agreement and not association. A classifier that yielded everything wrong would have a highly significant association but no agreement (Marzban, 2004).

Finally, sensitivity and specificity represent the measures of accuracy of a certain diagnostic test or classification. In fact, the measurements have to be sensitive in order to detect differences that are important to the research question, and specific enough to show only the feature of interest. Sensitivity describes how well a classification task classifies those observations in the right corresponding class (say coils). Similarly, specificity describes how well a classification task classifies those observations that are not coils. The definitions of sensitivity and specificity and can be depicted from the equations 5.45 and 5.46, respectively.

5.4.4.1 Threshold Value

Since a typical classifier generates a variable that has values within the range 0 -1, and all of the measures described in this section depend on the values in the confusion matrix, these values are obtained by application of a threshold criterion to a continuous variable generated by the classifier. A mid value between 0-1 which 0.5 is the threshold applied here. Thus, a continuous variable is converted into dichotomy variable in this case. If the threshold criterion is altered the values in the confusion matrix will change. Often, the raw scores are available so it is relatively easy to examine the effect of changing the threshold. FN errors are more serious than FP errors; the threshold can be adjusted to decrease the FN rate at the expense of an increased FP error rate.

The effect of the threshold on error rates can be explained by a cut-point of 0 where every case assigns as positive, while a cut-point of 1 assigns every case as negative. Therefore, as the cut-point is moved from 0 to 1 the false positive frequency falls while the false negative frequency increases. The point where these two curves cross is the point with the minimum overall error rate Thresholds can be amended to reflect different TP and FP rates according to different objectives.

5.4.4.2 Predictive Value

A certain test that may have high accuracy in terms of sensitivity and specificity values; it may yet perform poorly and have low positive predictive value. The predictive value of a test is an important index of actual test performance. The positive predictive value of a test indicates the probability that a (coil) is actually present when the test is positive, and can be calculated as:

$$\textit{Positive predictive value} = TP/\ TP + FP \tag{5.47}$$

The negative predictive value of a test indicates the probability that a (coil) is actually absent if the test is negative, and also can be calculated by the formula:

$$Negative\ predictive\ value = TN/\ TN + FN \tag{5.48}$$

5.4.4.3 Plotting ROC Curve

An efficient way to display the relationship between sensitivity and specificity and the cut-off point for positive and negative tests is with receiver operating characteristic (ROC) curves (Obuchowski, 2000; Gur *et al.*, 2003). The receiver operating characteristic (ROC) curve describes the performance of a test used to discriminate between normal and abnormal cases based on a variable measured on a continuous scale.

The ROC curve is a plot of the sensitivity and the 1-specificity. Each point on the curve represents a different cut-off value for the test indicated. Each cut-off value results in a true positive (y-axis) and false positive (x-axis) ratios. The test that yields the greatest number of true positives with the smallest number of false positives, representing a curve, which tends upwards and to the left, is good test. A perfect test has a curve of area equal to one. A poor diagnostic test has a low ROC curve approaching the diagonal with area of 0.5. Under the diagonal, true positives and false positives are equal at every cut-off points where the test is indifferent.

5.4.4.4 Area Under Curve (AUC)

The area under the ROC function (AUC) is usually taken to be an important index because it provides a single measure of overall accuracy that is not dependent upon a particular threshold (Hand, 1997; Hand and Till, 2001). With reference to Figure 5.8, the value of the AUC is between 0.5 and 1.0. If the value is 0.5, as in the diagonal line on the plot, the scores for two groups do not differ. A score of 1.0 indicates no overlap in the distributions of the group scores.

Typically, values of the AUC will not achieve these limits. A value of 0.8 for the AUC means that for 80% of the time a random selection from the positive group will have a score greater than a random selection from the negative class (Hand and Till, 2001). Usually the AUC for the training data is higher than that for the testing data. This is expected since most classification methods will perform best on the data used to generate the classification rule which the training data set, and less on the testing data set.

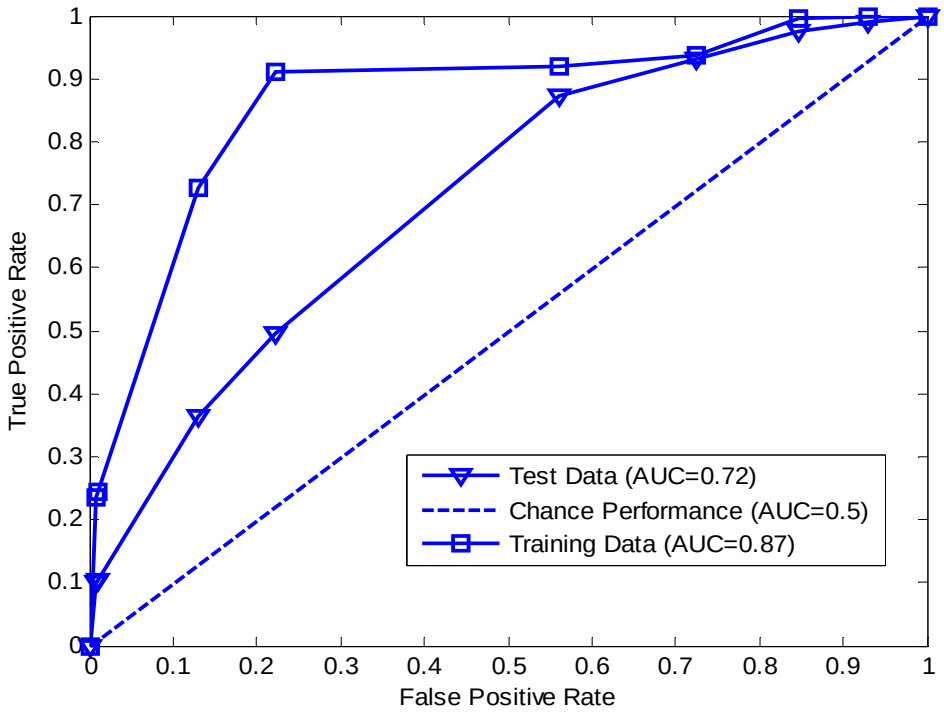


Figure 5.8: A typical example of area under curve (AUC) for training data, test data, and chance performance or random guess

Researchers argued that some caution is necessary when using ROC methods with biological data since biological cases may not be directly equivalent to the original definition. In particular, the original ROC model assumes that the group allocation is absolutely reliable and each signal is homogeneously presented and processed (Hanley and McNeil, 1983).

5.4.5 Reliability Index

The prediction reliability index (*RI*) offers an excellent tool for focusing on key regions having high prediction accuracy (Rost and Sander, 1993). It has been shown that prediction accuracy varied largely between different proteins. The RI is usually used to assess the effectiveness of the methods for the prediction of the secondary structure of a new sequence.

According to Rost and Sander, (1993), the value of RI can be normalized to be an integer between 0 and 9. The prediction accuracy of residues with higher RI values is much better than those with lower RI values. Therefore, the definition of RI reflects the prediction reliability and its index correlated with its accuracy.

In this research, the histograms and distribution of prediction analysis are considered as measures of reliability as well as accuracy. In fact, the representation, analysis, and discussion of the line graphs in the next chapter carry the same concept of the reliability index of Rost and Sander.

5.4.6 Test of Statistical Significance

It is commonplace that the probability principle is of utmost importance in statistics. A normal or *gausian* distribution of values is a bell-shaped curve with its x-axis representing the measurement of frequency of measurements and the y-axis representing the relative number of repetitions with the individual x values. The area under a portion of the curve is the probability that the true value is at or greater than the value of v at the line. In the normal distribution measurements, which occur with the greatest frequency occur at the center of the distribution and are known as the central tendency (Anderson, 2003).

Confidence intervals express the variation around the mean of a measurement, or a frequency. If a series of identical studies are performed on different samples from the same populations and a 95% confidence interval for the

difference between the sample means existed, then 95% of these confidence intervals would include the population difference between means. The researcher may select the degree of confidence, with 95% being the most common choice just as 5% level of statistical significance is widely used. The probability level of 0.01 is not uncommon too (Anderson, 2003).

5.4.6.1 The Confidence Level (P-Value)

One of the most commonly used statistical terms is the null hypothesis (Ho), which states that there is no difference between study groups except the one that is attributable to random phenomena. The alternate hypothesis (Ha) is the statement that there is a difference that cannot be explained by chance. The alternate hypothesis is proved by the exclusion of Ho. The p-value is the probability on the assumption that Ho is true of obtaining a measurement equal to or more extreme than that actually observed. In the graph of the normal distribution the p-value is represented by the area under the curve at and above the observed value marked by the line on the x-axis (Hand, 1977).

The level of statistical significance, also called type I error or false-positive result is the probability of rejecting Ho when Ho is actually true. It has been arbitrarily set at 0.05 as the threshold for statistical significance to distinguish whether an observed change in a set of measurements or frequencies may have arisen by chance or it represented something other than random variation. A type II error or false-negative result is the probability of accepting Ho as true when Ho is actually false, and as such missing a clinically significant difference. It is set at 0.1-0.2 as acceptable by most researchers. Practically, small p-values mean p-values of 0.05, which represent moderate evidence against to strong evidence; and those less than 0.001 represent strong to very strong evidence. (Lijmer *et al.,* 1999; Hand, 1997).

5.4.6.2 Analysis of Variance (ANOVA) Procedure

The analysis of variance or ANOVA tests the significant differences among the means of observations in an experiment (Anderson, 2003). The mathematical model for an observation X_{ij} in the experiment is can be written as in equation 5.49.

$$X_{ij} = u + v_i + e_{ij} \qquad (5.49)$$

Where:

u = the mean effect

v_i = effect of ijth entry

e_{ij} = experimental error effect

Then the ANOVA table will then look as illustrated in Table 5.2.

Table 5.2: ANOVA table based on individual observations (One way ANOVA)

Source of variation	**Degree of freedom**
observations	*d - 1*
Error	*d (n - 1)*
Total	*dn - 1*

5.5 Summary

This chapter explains in details the methodology and models followed in this research in an attempt to solve the problem of problem of protein folding. Collaborative programs in Bioinformatics like PSI-BLAST and CLUSTALW are utilized in this work to generate homologues sequences and conduct multiple sequence alignment. This provides standard procedure to incorporate evolutionary information in related sequences. Filtering programs are used to mask the data set and boost the prediction ability of NN-GORV-II method. Five DSSP eight-to-three

states secondary structure reduction methods or schemes are discussed and explained to be used in the series of the experiments of this research.

The newly developed secondary structure prediction algorithms, NN-GORV-I and NN-GORV-II together with other standard prediction methods are comprehensively explained. Emphasis is directed towards the explanation of the information theory and the neural networks since the newly developed methods (NN-GORV-I and NN-GORV-II) combine these two machines. To assess and evaluate the prediction accuracy and quality of the methods studied in this work, many algorithms and procedures are explained in this chapter, varying from Q_3, SOV measure, MCC, and ROC. Test of significance between means of the output of these algorithms using the ANOVA procedure is also explained.

CHAPTER 6

ASSESSMENT OF THE PREDICTION METHODS

6.1 Introduction

Different measures and methods are used to assess the accuracy of newly developed methods for protein secondary structure prediction. There are four assessment methods used in this study, namely: Q_3, SOV, MCC, ROC, and AUC, in addition to the ANOVA to test the significance of the prediction methods.

The Q_3 accuracy per residue and per the whole protein is used to calculate the percentage performance of the two methods developed in this research together with the other five methods investigated. Unlike the Q_3, the Segment overlap measure (SOV) is used as a measure of quality rather than performance. The third measure used in this research is the Matthews Correlation Coefficient (MCC) to measure the strength of the relation between the predicted and observed protein structure in a range between 0-1. After observing that the data set used in this research contains about 50% coils, the Receiver Operating Characteristic (ROC) and Area Under Curve (AUC) are used to partially assess the newly developed methods taking only the coil states in consideration. Finally, the test of the statistical significance of the prediction methods is conducted using the analysis of variance (ANOVA) procedure.

In the past few decades the prediction accuracy is oscillating slightly around or above 60% prediction accuracy. The reason for this low level of prediction is that all these algorithms used only local information to predict the secondary structure of

proteins. Researchers noticed and realized that information contained in multiple alignments can improve prediction accuracy (Kabsch, and Sander, 1983); However, the combination of large scale databases with more advanced algorithms and the use of distant or evolutionary information raised the level of prediction accuracy to the range of 70% (Rost, 2001; Chandonia and Karplus 1999).

Information from the position-specific evolutionary exchange is also recognized earlier that a profile of a particular protein family enhances discovering more distant members of that family (Kabsch and Sander, 1983). Automated database search methods successfully used position-specific profiles for searching (Frishman and Argos, 1996). The significance of high gain in prediction accuracy is achieved with the development of scoring matrices methods like PSIBLAST and probabilistic models like hidden Markov models (Krogh *et al.*, 1994). In particular, the gapped profile-based and iterated search tool PSI-BLAST continue to add to the field of protein sequence analysis due to its high speed and accuracy capabilities.

In this chapter the strength and weaknesses of the seven algorithms or methods of prediction are analysed and compared with respect to the newly developed two methods in this project experimental work. Several tests are used to assess the efficiency and accuracy of each method. Stringent statistical and procedures, tabular comparison, and graphical representation are used to enhance the discussion.

6.2 Data Set Composition

The set of 480 proteins that comprises a sub set of the CB513 proteins of Cuff and Barton (1999)is used in training and testing the seven algorithms. The set composed of 83392 residues as shown in Table 6.1. Alpha helices composed 35% of this set, beta sheets composed 21%, and coils constitute 44% of this data set. As discussed in the methodology chapter, the CB513 proteins of Cuff and Barton (1999) use the eight states DSSP secondary structure assignments beside others. Five

reduction methods are used in this research to assign the DSSP eight secondary structure states into three. The figures in Table 6.1 use reduction method I. These figures will differ when using another reduction method. The five reduction methods are used to study the effect of these reduction methods in the prediction process.

Table 6.1: Total number of secondary structures states in the data base

Structure	Total number	Percentage
Helix	28881	35
Sheet	17810	21
Coil	36701	44
ALL	83392	100

6.3 Assessment of GOR-IV Method

GOR-IV method is fully described in both the literature and methodology chapters of this report. It is further implemented in the series of experiments conducted in this work and also described in the methodology chapter. It is important to conduct GOR-IV method (and of course all the seven methods) on the same training and testing data, the same search database, and the same environments of the experiments.

Figure 6.1 shows the results of GOR-IV prediction. The figure shows a histogram that elucidates clearly the Q_3 which is a combination of the performance of helices, strands, and coils; less than 20 amino acids (proteins) scored the range of 20-30% and 30-40%. However, around 20 proteins of the 480 scored 80% accuracy. About 160 proteins scored between 60-70% and 140 proteins reached between 70-80% Q_3 accuracy. However, very few proteins (less than 5) scored a 100% Q_3 accuracy.

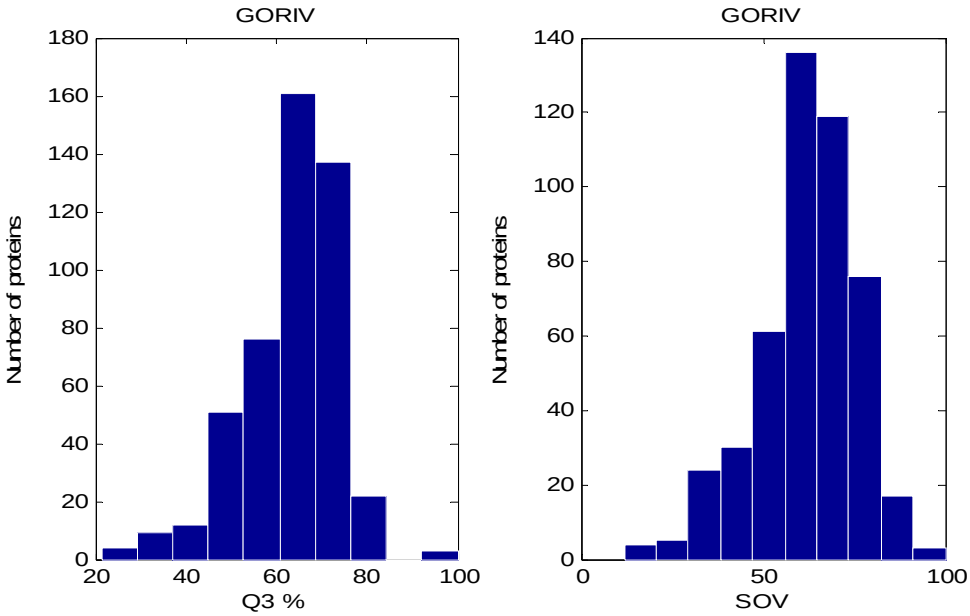


Figure 6.1: The performance of the GOR-IV prediction method with respect to Q_3 and SOV prediction measures

Table 6.2 shows the detailed results of GOR-IV predictions. The estimated accuracy for the alpha helices (Q_H) and beta strands (Q_E) are in the range of 57% and 51% with standard deviations as high as 29% and 27%, respectively. The coil states (Q_C) are estimated with higher accuracies that reached around 71% as expected. However, the standard deviation for coils is small (12.98) which indicate more even prediction estimates than the other two previous states. The overall Q_3 of GOR-IV is 63.19% with standard deviation of 10.61%. This results is slightly lower than which is reported in the original GOR-IV experiments (64.4%) and higher than that reported in the PROF experiments which is 61.3% (Ouali and King, 2000).

Table 6.2: The percentages of prediction accuracies with the standard deviations of the seven methods

Prediction Method	Q_3	Q_H	Q_E	Q_C
NN-I	64.05±12.68	57.29±30.64	57.39±28.49	74.10±13.36
GOR-IV	63.19±10.61	57.02±29.68	51.86±27.36	71.95±12.98
GOR-V	71.84±19.63	68.40±33.98	63.68±33.02	78.92±15.08
NN-II	73.58±17.82	70.77±31.62	68.72±30.01	78.33±15.18
PROF	75.03±14.74	70.65±31.39	68.29±28.09	79.38±13.68
NN-GORV-I	79.22±10.14	76.56±27.17	68.54±28.22	79.44±12.65
NN-GORV-II	80.49±10.21	77.40±26.53	77.12±24.19	79.99±11.75

Calculations are estimated from 480 amino acids (proteins)
Q_3 accuracy for amino acid
Q_H accuracy for α helices
Q_E accuracy for β strands
Q_C accuracy for coils

The GOR-IV segment overlap measure (SOV) showed that about 140 proteins scored between 55-65% and about 120 proteins scored between 65-75% SOV measure (Figure 6.1). The SOV measure is always considered as more reliable than Q_3 measure. Anyhow, both measures showed that the 480 proteins are distributed normally regarding GOR-IV method. However, this is not the case if each state (helices, strands, coils) is taken separately (histograms not shown for this part).

Table 6.3 shows the SOV prediction accuracies for the GOR-IV method. Prediction estimates are brought up to the level of 60% prediction accuracy for helices (SOV_H) and brought down to the level of 62% for coils (SOV_C). The SOV measure for strands (SOV_E) remained as low as 56%. The overall SOV measure for the three states is 62.07 % with standard deviation of 13.77. The SOV measure for GOR-IV method is higher than that reported by PROF method (56.9) which reflects good correlation between adjacent residues.

The SOV measure should be used to assess the quality of a prediction method rather than its performance since the SOV can be improved by applying a second

different structure network (Rost and Sander,1993) or sort of smoothing filters (King and Sternberg, 1996; Cuff and Barton, 2000).

Table 6.3: The SOV of prediction accuracies with the standard deviations of the seven methods

Prediction Method	SOV_3	SOV_H	SOV_E	SOV_C
NN-I	60.94±16.22	59.50±30.55	57.61±29.12	61.53±16.26
GOR-IV	62.07±13.77	60.81±29.47	56.01±29.36	62.34±14.89
GOR-V	69.33±22.96	70.87±32.51	64.00±33.94	66.63±21.16
NN-II	70.37±18.35	71.05±30.21	68.47±30.67	67.29±18.02
PROF	72.74±20.51	73.49±30.62	69.80±30.53	69.75±18.95
NN-GORV-I	76.55±14.39	76.93±27.82	70.76±29.33	72.90±14.47
NN-GORV-II	76.27±17.50	77.96±26.92	79.94±24.57	74.35±15.53

Calculations are estimated from 480 amino acids (proteins)
SOV_3 is the segment overlap measure per amino acid
SOV_H is the segment overlap measure for α helices
SOV_E is the segment overlap measure for β strands
SOV_C is the segment overlap measure for coils

Matthews's correlation coefficients (MCC) are shown in Table 6.4. The Matthews's correlation coefficient measures the predictive accuracy of an association between classes. A value that is near 0.1 indicates loose association between observed and predicted classes and hence less accurate prediction while a value that is near 0.9 indicates a tight association between observed and predicted classes and hence more accurate prediction. GOR-IV scored less than 0.5 MCC for β strands and coils which indicates less accurate prediction of these residues while it scored a value that greater than 0.5 for α helices which indicates that the prediction of the α helices is more accurate than the other two residues.

Table 6.4: The Mathew's correlation coefficients of predictions of the seven methods

Prediction Method	MCC_H	MCC_E	MCC_C
NN-I	0.4906	0.4124	0.4448
GOR-IV	0.5283	0.3756	0.4382
GOR-V	0.6859	0.5994	0.5675
NN-II	0.6503	0.5641	0.5304
PROF	0.7102	0.6291	0.5743
NN-GORV-I	**0.7736**	**0.6959**	**0.6494**
NN-GORV-II	**0.7744**	**0.6958**	**0.6501**

Calculations are estimated from 480 amino acids (proteins)
MCC_H is the Mathews correlation coefficient for α helices
MCC_E is the Mathews correlation coefficient for β strands
MCC_C is the Mathews correlation coefficient for coils

6.4 Assessment of NN-I Method

The NN-I prediction method is a neural network predictor that does not use the multiple sequence alignment. NN-I uses single sequences to predict novel proteins. PSI-BLAST or CLUSTALW are not utilized here which made this predictor looks like the early work of Quian and Sejnowski (1988). The network is a three layers network trained in an unbalanced way as mentioned in the methodology.

Figure 6.2 shows that the Q_3 and SOV for NN-I. Less than 20 proteins or amino acids scored a Q_3 of 10%, 20%, 30%, and 40% for each. Around 140 of the 480 protein scored Q_3 of 60% and more than 170 proteins scored 70%. This histogram revealed that NN-I performed almost similar or slightly better than GOR-IV.

SOV histogram for NN-I (Figure 6.2) shows that more proteins scored less than 20% accuracy unlike the case of Q_3. Less than 120 proteins scored 60% while

about 150 scored 70%. This revealed that NN-I method is better than GOR-IV method as far as SOV is concerned.

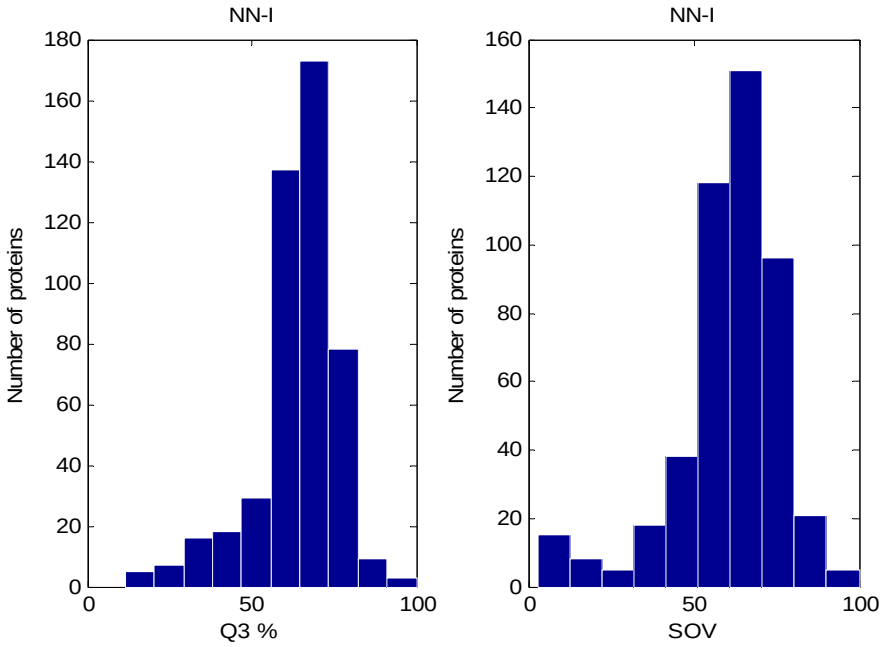


Figure 6.2: The performance of the NN-I prediction method with respect to Q_3 and SOV prediction measures

Table 6.2 shows that NN-I scored about 57% for helices and strands and 74% for coils. NN-I reached 64% as Q_3 accuracy which is better than that of GOR-IV for all the three states. However, in Table 6.3 the SOV of NN-I scored about 59%, 57%, and 61% for helices, strands and coils, respectively. The SOV for the three states is 60.94% indicting that the prediction of NN-I is of less quality than GOR-IV. This result is confirmed by the Matthews' correlation coefficients in Table 5.4 where NN-I correlation coefficients for the three states are less than 0.5.

6.5 Assessment of GOR-V Method

The GOR-V method is fully described and implemented using the same database that is used by their authors (Kloczkowski *et al.*, 2002). GOR-V uses multiple sequence alignment and resizable window size according to the length of the amino acid in an improvement that added triplet statistics to the previous GOR

methods. The original GOR-V combines the PSI-BLAST multiple sequence alignment with GOR methods with a full jack-knife training procedure.

The histogram of Figure 6.3 shows the performance of GOR-V in this experimental work with respect to Q_3 and SOV of the all three states. The figure clearly shows that there is a great shift of prediction accuracy towards the 100% compared to the previous methods GOR-IV and NN-I. Less than 30 proteins scored Q_3 of 10%, 20%, 30%, 40%, 50%, and 60% each, while the majority of the 480 proteins scored Q_3 of 70%, 80%, and 90%.

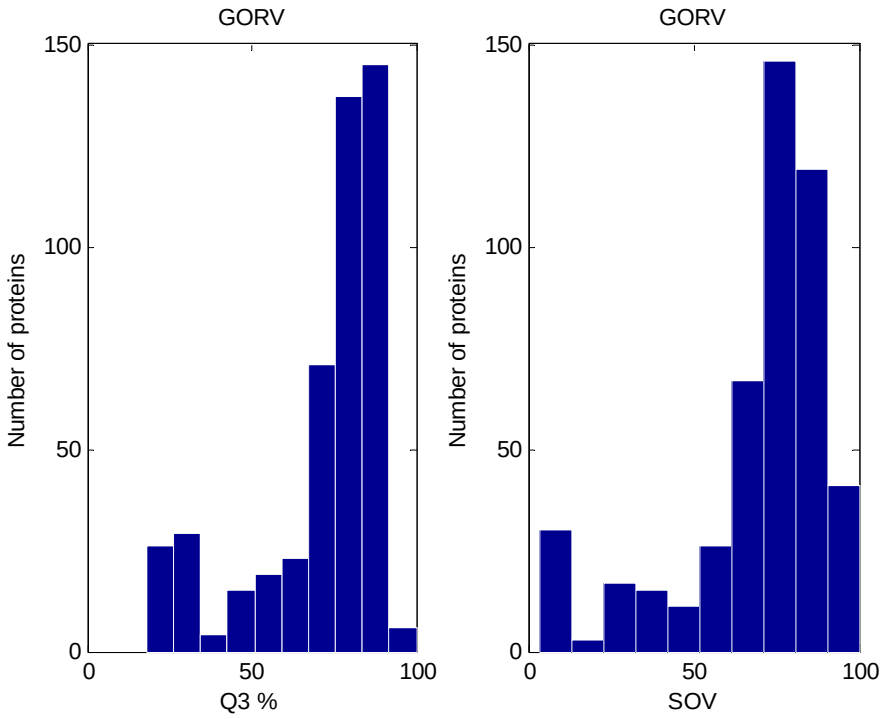


Figure 6.3: The performance of the GOR-V prediction method with respect to Q_3 and SOV prediction measures

GOR-V segment overlap measure (SOV) shows a similar trend to the Q_3 measure with about 30 proteins scored 10% and 40 proteins scored 100% SOV score. The majority of the proteins scored the range of 70%, 80%, and 90% as shown in Figure 6.3. Again the SOV is considered here as a measure of usefulness and quality of prediction rather than performance.

Table 6.2 shows the GOR-V performance regarding the percentage accuracies for helices, strands, coils, and all the states together (Q_H, Q_E, Q_C, and Q_3).

The GOR-V showed scores of 68.40% and 63.68% with relatively high standard deviations of around 33% for helices and strands, respectively. This indicates that prediction accuracies for these two states are oscillating from around 30% up to 90% or even more. The standard deviations reveal that helices and strands of some proteins within the data base are predicted with very high accuracies while others are predicted with very low accuracies. Coils in GOR-V are predicted with as high accuracy as 78.92% and low standard deviation. However, the overall Q_3 accuracy of GOR-V is 71.84% with a relatively reasonable standard deviation of 19.63%.

Kloczkowski *et al.,* (2002) reported an average accuracy of GOR-V prediction for the secondary structure with multiple sequence alignment and full jack-knife procedure as 73.5%. The accuracy of the prediction is further increased to 74.2% when limiting the prediction to 375 sequences of the CB513 database. However, the results of GOR-V which are presented in Table 5.2 showed a decrease of 2.36% (74.2-71.84) than that of Kloczkowski *et al.* (2002). This is in an agreement with Cuff and Barton (1999) who showed that a reduction of 3-4% of prediction accuracies when experiments are conducted in different environments.

Table 6.3 shows the SOV scores per residue and per protein for GOR-V. Significantly GOR-V has the highest score over all the three states compared to GOR-IV and NN-I. The SOV accuracy per protein is 69.33% with a moderate standard deviation of 22.96%. The score indicted that GOR-V method is superior in quality and usefulness compared to GOR-IV and NN-I methods.

The Matthews correlation coefficients (MCC) for GOR-V are shown in Table 6.4. The coefficients are 0.69, 0.60, and 0.57 for helices, strands, and coils, respectively. The figures show that helices are predicted with high accuracy and reliability since the correlation between predicted and observed residues is near 0.7. Strands and coils are predicted with better than average accuracy and reliability of around 0.6 which in turn less than the accuracy of helices states.

The results of the tables and figures of GOR-V showed that the method utilised the multiple alignment of PSI-BLAST in a way made it clearly superior

compared to GOR-IV and NN-I methods. The prediction accuracy jumped to a level above 70% from the previous level of 63% and 64% of GOR-IV and NN-I. These results elucidate and confirm the methodology that had been suggested and implemented by Rost and Sander (1993) boosting the secondary structure prediction level from 64% to above 70% level.

6.6 Assessment of NN-II Method

NN-II prediction method used in this experiment is basically similar to NN-I. It differs in the usage of the PSI-BLAST multiple sequence alignments to extract evolutionary information of similar proteins. PSI-BLAST profile is used to enable the network to slide over a window of 17 along the profile in contrast to the NN-I which slides over a window of 17 amino acids.

Figure 6.4 shows histograms of the Q_3 and the SOV for the NN-II prediction method. Most proteins are predicted at a level above 70%. About 50 proteins are predicted at the level of 70%, 180 proteins at the level of 80%, and more 140 of the 480 proteins are predicted at the level of 90% Q_3.

Less than 20 proteins of the 480 are predicted at the level of 10%, 30%, 40%, 50%, and 60% Q_3. About 20 proteins are predicted at the level of 20% and 100% Q_3. Figure 5.4 also shows that the SOV measure for NN-II scored less compared to the Q_3 measure. About 80 proteins of the 480 scored the level of 70%, 140 proteins scored 80% SOV value, and more than 100 proteins scored 90%. Since SOV is a measure of quality and usefulness of predictors, this value showed that NN-II method is of high quality and more useful than GOR-V method.

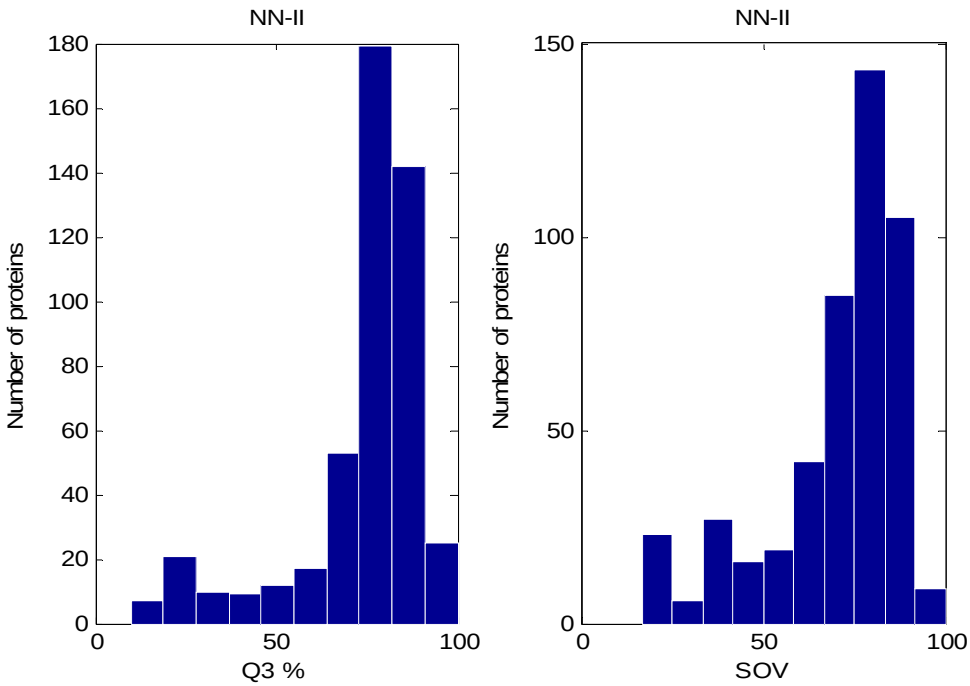


Figure 6.4: The performance of the NN-II prediction method with respect to Q_3 and SOV prediction measures

Table 6.2 shows the Q_3 predictions of the NN-II methods for all the three states of secondary structure separately and together. The prediction for helices (Q_H) is 70.77%, for strands (Q_E) is 68.72% with relatively high standard deviation of about 30% for each. The coils (Q_C) are predicted with a higher accuracy of 78.92% with a low standard deviation of 15.18%. The overall Q_3 for all states is 73.58% with standard deviation of 17.82%. The Q_3 of this neural network method (NN-II) is lower than that of the profile PHD method of where a similar architecture of the PHD is followed in the NN-II. The PHD scored 75.1% (Rost, 2001) Q_3 where NN-II method scored 73.6%. The different training procedure and different data set used for each method led to this drop in NN-II method prediction but the difference is very small and the two methods are still comparable.

Table 6.3 shows the results of the SOV measure regarding NN-II method. The SOV for strands and coils are below 70% which are 68.47 and 67.29, respectively, while for helices is 71.37%. This is better than the SOV of PHD method which is 70% (Rost, 2001).

The Matthews correlation coefficients (MCC) for the NN-II method are shown in Table 6.4. The MCC are 0.65, 0.56, and 0.53 for helices, strands, and coils, respectively. These correlation coefficients showed that the NN-II method could successfully relate unpredicted residues to their correspondent classes with relation that is better than that of NN-I and GOR-IV methods but not better than that of the GOR-V method. However, these coefficients are almost the same as for the PHD for the helices and coil states (0.64 and 0.53) and less as for the strand state which is 0 .62 for the PHD (Rost, 2001).

6.7 Assessment of PROF Method

The PROF Method is briefly described in the methodology chapter of this report and fully described in the work of Ouali and King (2000). PROF is cascading multiple protein secondary structure classifier or predictor that uses neural networks, GORI-IV, linear and quadratic discrimination, and voting methods. PROF uses a full jack-knife training method and reported reaching a Q_3 of 76.70% prediction accuracy.

The general performance of the PROF method is elucidated in Figure 6.5. Among the 480 proteins, proteins that scored a Q_3 accuracy of 10%, 20%, 30%, 40%, and 50% are less than 20 proteins for each. About 30 proteins scored a Q_3 of 60% and more than 50 proteins reached a Q_3 of 70%. More than 160 proteins scored an accuracy of 80% and 90% while less than 10 proteins reached the level of 10% Q_3 accuracy.

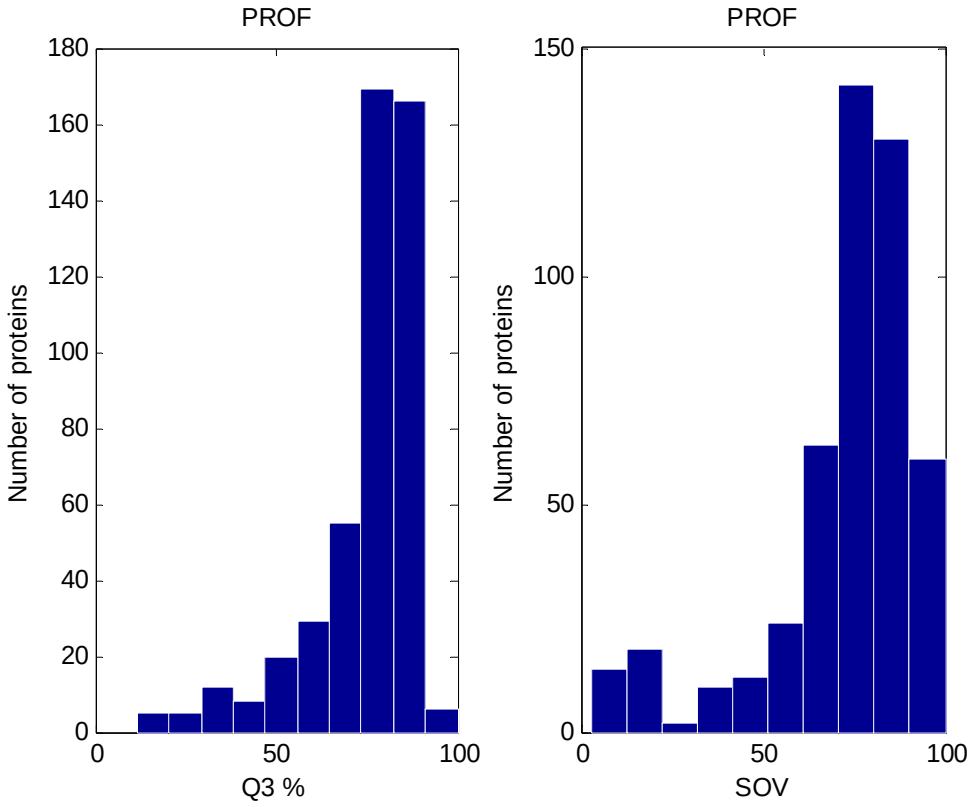


Figure 6.5: The performance of the PROF prediction method with respect to Q_3 and SOV prediction measures

The SOV measure for the PROF (Figure 6.5) achieved more or less similar scores to that of the Q_3 except that more than 60 proteins scored a SOV value that is equal to 100% and the 80% and 90% level is achieved by less than 140 proteins for each.

Table 6.2 shows that PROF has achieved accuracy of 70.65% and 68.29% for helices and strands, with standard deviations of 31.39% and 28.09%, respectively. These results are less than what had been reported by Ouali and King (2000) in their original work of PROF where their reported accuracy for helices and strands are 70.8% and 71.6% with standard deviations of 29.8% and 25.3%, respectively.

6.7.1 Three States Performance of PROF Method

The performance accuracies (Q) of the helices, strands, and coils states of PROF method in this work compared to other methods studied in this research are elucidated in Figure 6.6, Figure 6.7, and Figure 6.8, respectively.

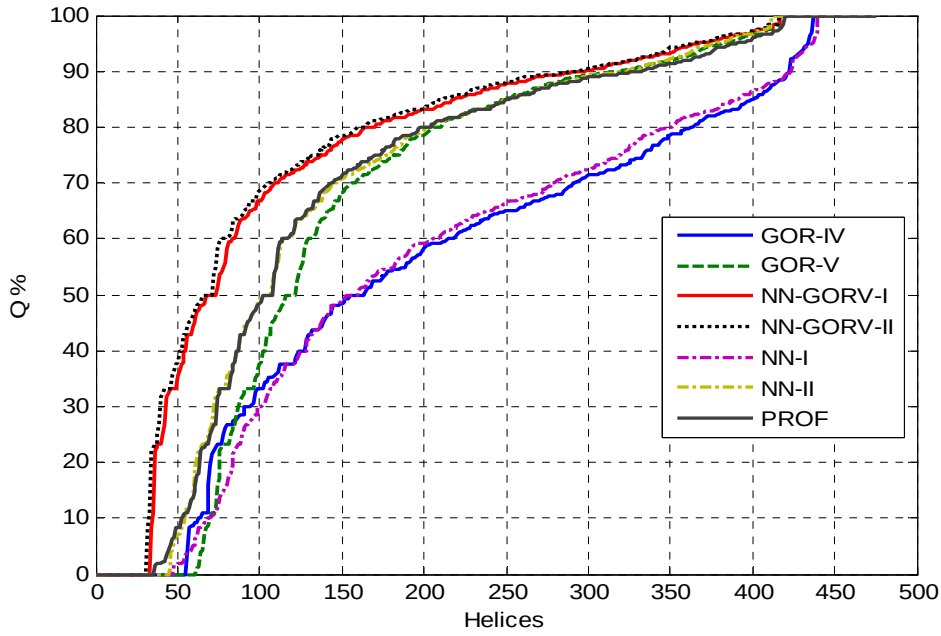


Figure 6.6: The α helices performance (Q_H) of the seven prediction methods

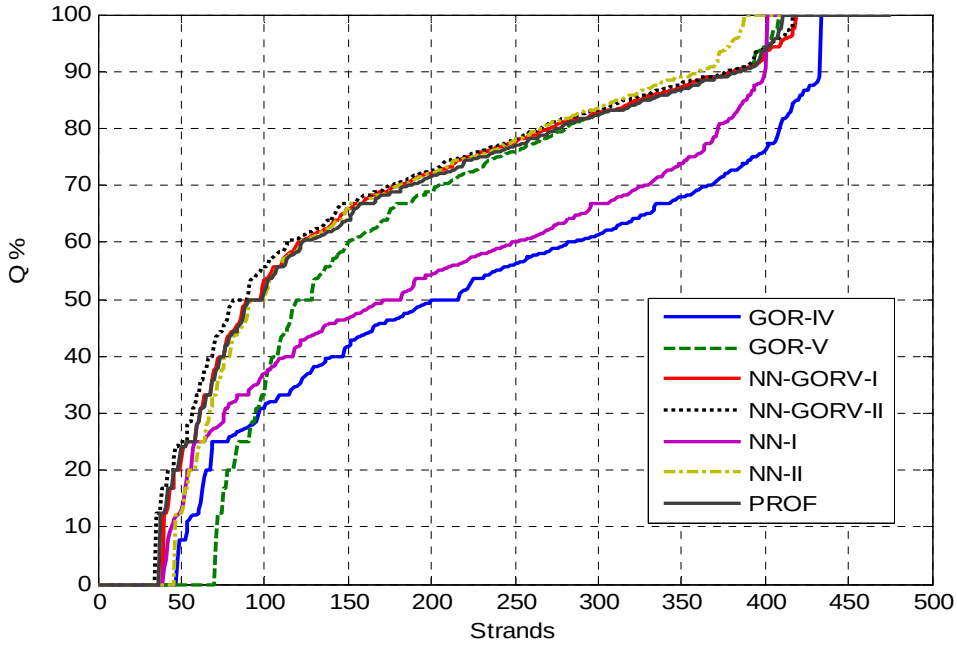


Figure 6.7: The β strands performance (Q_E) of the seven prediction methods

For coils, PROF in this work achieved an accuracy of 79.38% with standard deviation of 13.68%; both numbers showed an overestimation of coils and its standard deviation compared to the original work of PROF which scored 77.2% with standard deviation of 10.9%. Figure 6.8 explains the behaviour of coils prediction accuracies of the PROF in this wok with respect to the 480 proteins.

Figures 6.6, 6.7, and 6.8 elucidated how the three states of protein secondary structure (helices, strands, and coils) for the different proteins responded to the PROF classification in this work or how PROF predicted or classified these states of proteins to secondary structure from their original amino acids. In Figure 6.6 the curve of PROF helices creeps almost in pattern that is almost similar to NN-II prediction but far below that of which revealed that helices of PROF of this experiment and helices of NN-II are classified with almost the same accuracy and reliability. In Figure 6.7 the curve shows that the strands of PROF for the 480 proteins are predicted with a pattern that is relatively just better than that of NN-II strands a fact that is confirmed by the SOV measure shown in Table 6.3.

In Figure 6.8, the coils curves show a different pattern that of the helices and strands. There are no more than 30 proteins helices and coils are predicted at accuracy of zero for all the seven classification methods. The curves show that PROF in this experiment predicts the coils of the 480 proteins at higher accuracy than that of NN-II. However, a detailed comparison of the seven methods trends in prediction will be made in the next section of this chapter.

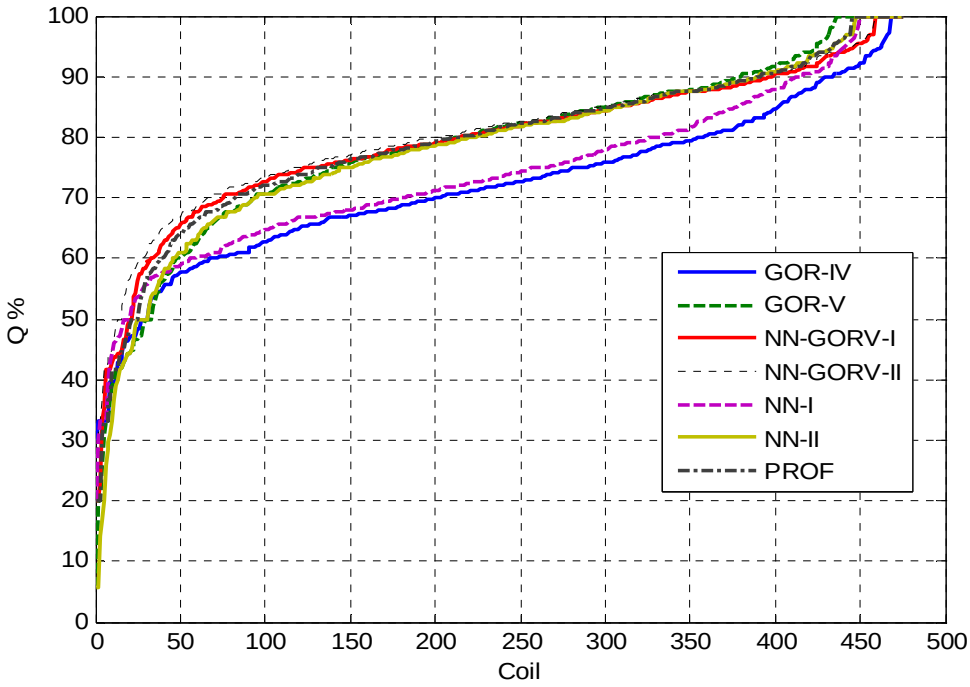


Figure 6.8: The coils performance (Q_C) of the seven prediction methods

6.7.2 Overall Performance and Quality of PROF Method

The overall Q_3 accuracy of PROF in this work is 75.03% with standard deviation of 14.74. This result shows that the Q_3 accuracy in this work is less than the previously reported 76.7% result for the PROF. Also the standard deviation in this PROF is greater than that of the original PROF which scored a standard deviation of 8.6%. The prediction of PROF of this work reveals that the proteins had been predicted in scattered and dispersed prediction rather than closely prediction compared to the original PROF. This result is supported by the histogram shown in Figure 6.5 as many proteins are predicted with very low accuracies and other with very high accuracies.

The SOV measures for PROF are shown in Table 6.3. The SOV for helices, strands, and coils are 73.49%, 69.80%, and 69.75% with standard deviations of 30.62%, 30.53%, and 18.95%, respectively. These results are almost similar to that reported for PROF with 71.1%, 75.6%, and 71.1% for helices, strands, and coils with

standard deviations of 29.9%, 26.0%, and 15.0%, respectively. The overall SOV for PROF in this experiment that combines all the three secondary structures states is 72.74 with standard deviation of 20.51 compared to 73.7 with standard deviation of 13.9 for the original PROF Ouali and King(2000). The above figures revealed that in general the original PROF experiment is of somewhat high quality and more useful than the PROF of this experiment. However, the margin of differences here is acceptable since each experimental work is conducted in a different environment (Rost, 2001; Cuff and Barton, 1999; Cuff and Barton, 2000).

Table 6.4 shows the Matthews correlation coefficients of helices, strands, and coils for the PROF experiment. The figures showed that the MCC are 0.71, 0.63, and 0.57 for helices, strands, and coils, respectively while the figures reported in the original PROF are 0.71, 0.63, and 0.57 for the same states, respectively. Surprisingly the figures are identical for each state in this experimental work and that of the original PROF. Matthews' correlation coefficients give an indication of how predicted states are in relation with observed states with a value near zero means that there is almost no relation between predicted states and observed states and a value near one means there is strong relation between predicted and observed states.

If we define the entropy as how much information a random variable carries or the amount of information needed to describe such a random variable (Baldi *et al*., 2000; Crooks and Brenner, 2004; Crooks *et al.*, 2004), we will recognize that Matthews' correlation coefficients carry a high entropy than the SOV measure since MCCs take into accounts the value of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). More discussion about correlations and entropy will be found in the next chapter of this report.

The PROF performance, quality, and reliability are far better than that of NN-II, GOR-V, GOR-IV, and NN-I ones. This concluding point could be clearly depicted from Table 6.2, Table 6.3 and Table 6.4 which is a true result because PROF combines several methods of predictions (Rost, 2001) as explained in the methodology chapter.

6.8 Assessment of NN-GORV-I Method

The NN-GORV-I method is the new method that has been developed in this research work. The method combines the new GOR-V method and the NN-II method which are explained and evaluated earlier in this chapter and the methodology chapter. At the beginning of this work GOR-V was just an idea and some theoretical points that had not yet being implemented. GOR-V is based on the information theory that founded the previous GOR methods while NN-II is based on the work of many researchers in the area of protein secondary structure that is sparked by the work of Quian and Sejnowski (1988) and refined by several recent workers (Rost and Sander, 1993; Cuff and Barton, 1999, Ouali and King 2000).

Figure 6.9 illustrates the performance of Q_3 and the SOV measure of NN-GORV-I method. The histogram Q_3 is significantly different of the other histograms of NN-II, GOR-V, and PROF. The figure shows that most proteins of the 480 proteins scored a Q_3 of above 50%. About 180 proteins scored a Q_3 of 80% while above 100 proteins scored a Q_3 accuracy of 70% and just below 100 proteins scored an accuracy of 90%. This sums up to 380 proteins of the 480 that achieved between 70%-90% Q_3 accuracies which is means that around 80% of the proteins achieved these high scores. However, few proteins which are less than 10 scored a Q_3 of 100% accuracy.

The SOV measure for NN-GORV-I (Figure 6.6) pushed up the 100% predictions to above 50 proteins and brought down the 80% predictions to about 120 proteins. The SOV scores for the 70% and 90% remained in the range of 100 proteins compared to Q_3 scores. The histogram of the SOV figure showed that there are more proteins predicted at high level of SOV accuracies than that of NN-II, GOR-V, and PROF; a result which revealed that the NN-GORV-I method is more useful and of high quality prediction than the previously discussed methods.

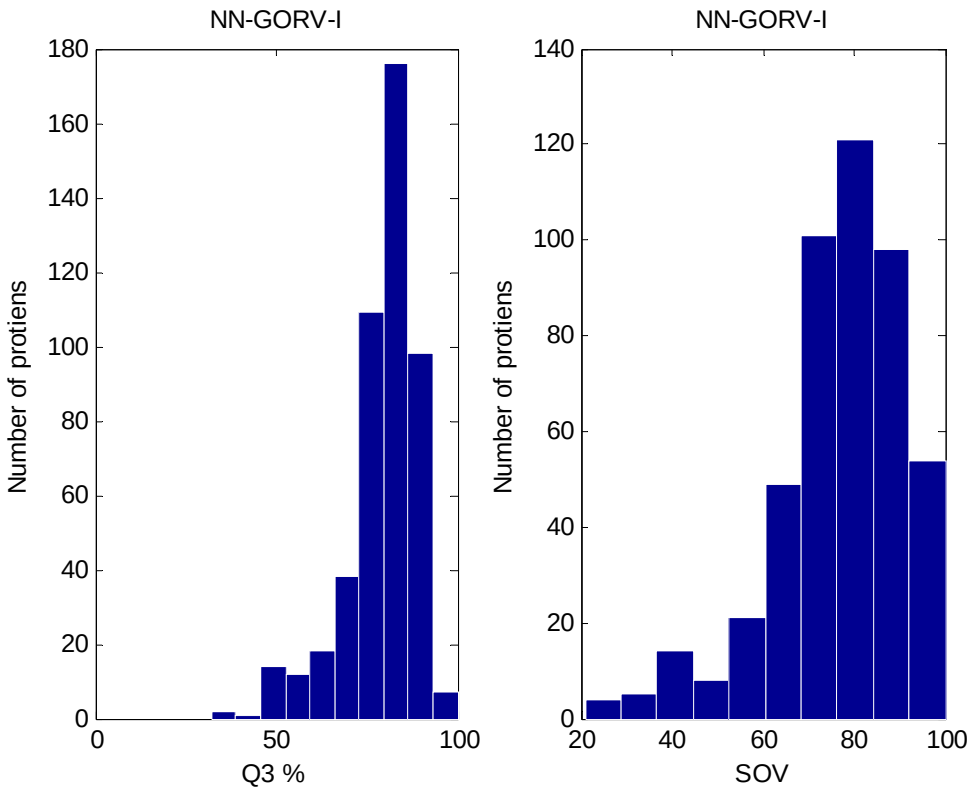


Figure 6.9: The performance of the NN-GORV-I prediction method with respect to Q_3 and SOV prediction measures

Table 6.2 shows the results of NN-GORV-I prediction accuracies for helices (Q_H), strands (Q_H), coils (Q_C), and all the three states together (Q_3). The results showed that NN-GORV-I gained an accuracy of 76.56 with standard deviation of 27.17 for alpha helices (Q_H), a result that is far better than the PROF prediction in this experiment for the same state which is 70.65 with standard deviation of 31.39. A gain of 6 points with lower standard deviation implied that the NN-GORV-I method is superior to PROF method in the performance of alpha helices with more closed or homogenous predictions towards the 100% accuracy side.

The same score of α alpha helices (Q_H) of the original PROF showed that the score for these states is 70.8% with standard deviation of 29.8% (Ouali and King, 2000) which almost behaved exactly like the PROF of this experiment and hence the same above conclusion which says the NN-GORV-I method is superior to PROF method in the performance of alpha helices with more closed or homogenous predictions towards better accuracy, applies in this case.

NN-GORV-I results for beta strands and coils are 68.54% and 79.44% with standard deviations of 28.22% and 12.65%, respectively. These results are almost the same as those are shown by PROF in this experiment (Table 6.2). However, the overall Q_3 accuracy of NN-GORV-I method is 79.22% with standard deviation of 10.14. The Q_3 accuracy result of this method is about 4% better and the standard deviation is also about 4% less (better) than that is scored by PROF in this experimental work (Table 6.2).

The Q_3 results of the original PROF for the beta strands and coils are71.6% and 77.2% with standard deviations of 25.3% and 10.9%, respectively. In comparing these results with Table 6.2, Figure 6.8, and 6.9, it suggests that the beta strands and coils of the original PROF performed better with slightly more homogenous predictions than that of the NN-GORV-I method.

6.8.1 Three States Quality (SOV) of NN-GORV-I Method

Table 6.3 shows the SOV measure for the NN-GORV-I method for the secondary structure separately as well as the overall SOV. The SOV measure is 76.93%, 70.76%, and 72.90% with standard deviations of 27.82%, 29.33%, and 14.47% for alpha helices, beta strands, and coils, respectively.

These results are further portrayed in the using the line graphs as shown in Figure 6.10 for helices, Figure 6.11 for strands, and Figure 6.12 for coils. Figure 6.10 curves depicted that alpha helices of the NN-GORV-I method for the 480 proteins are predicted with SOV measure pattern that exhibits a large margin above PROF of this work. This suggested that the dominant number of proteins helices is superior in their quality and usefulness to that of the PROF method.

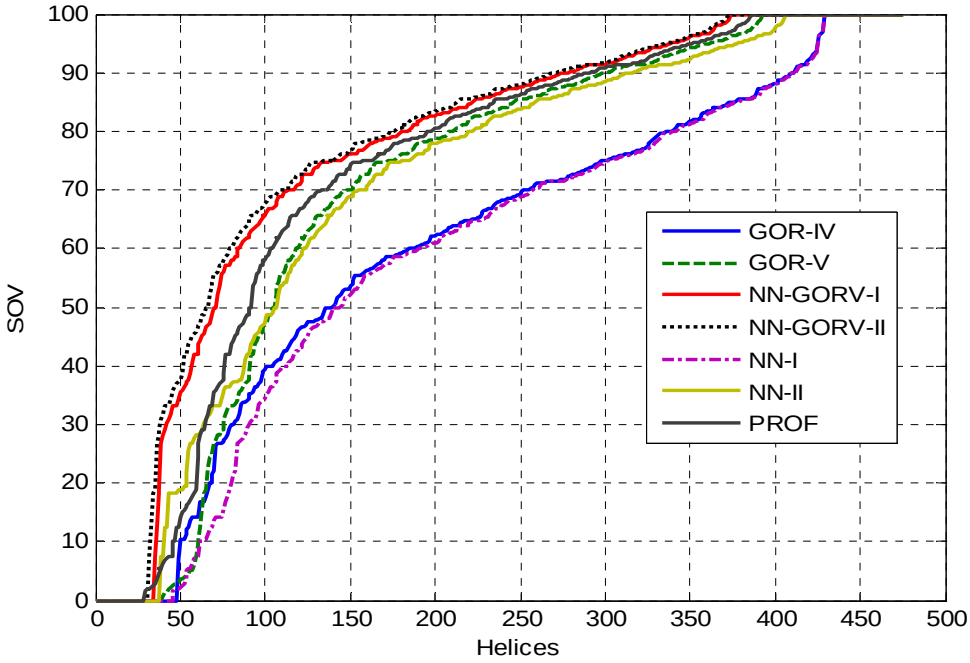


Figure 6.10: The helices segment overlap measure (SOV_H) of the seven prediction methods

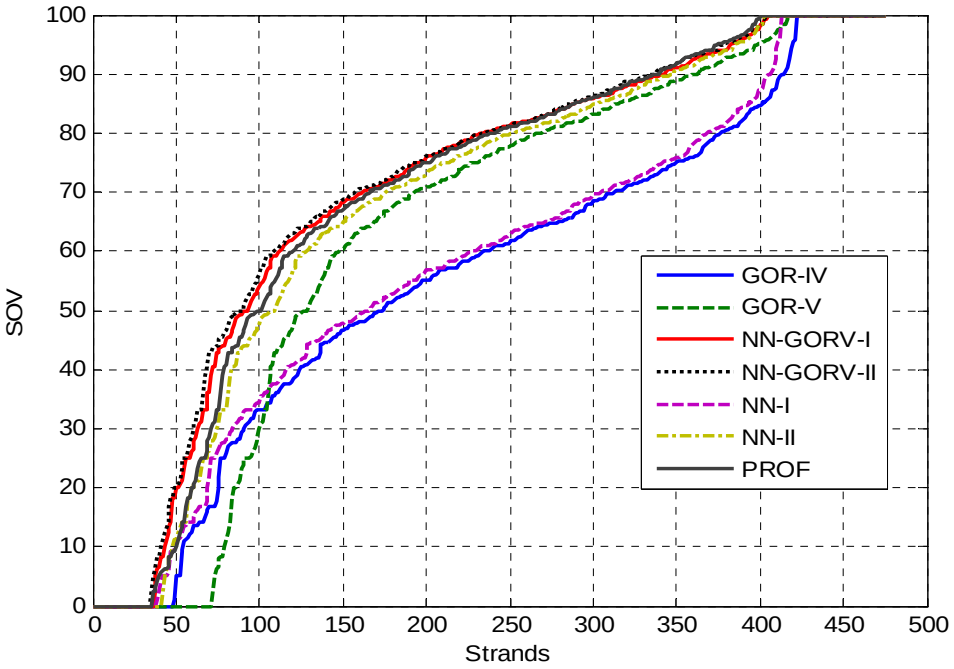


Figure 6.11: The strands segment overlap measure (SOV_E) of the seven prediction methods

In Figure 6.11, the curves illustrated that although the NN-GORV-I method SOV prediction of the strands states outperformed that of the PROF in this experiment, the margin is very small and the curves are running close to each others

through the 480 proteins. This is in agreement with which had been reported by Ouali and King (2000) that PROF predicts strands with high accuracy and reliability. This pattern, of course, shows that NN-GORV-I method prediction for strands has high quality and more useful.

As far as coils are concerned, Figure 6.12 presented the curves of the coils SOV measure for the seven classification or prediction methods. Unlike the SOV of helices and strands curves, the SOV of coils curves show that there are no proteins predicted the level of zero SOV measure. The figure also showed that the NN-GORV-I method curve pattern is always above that of PROF in this work.

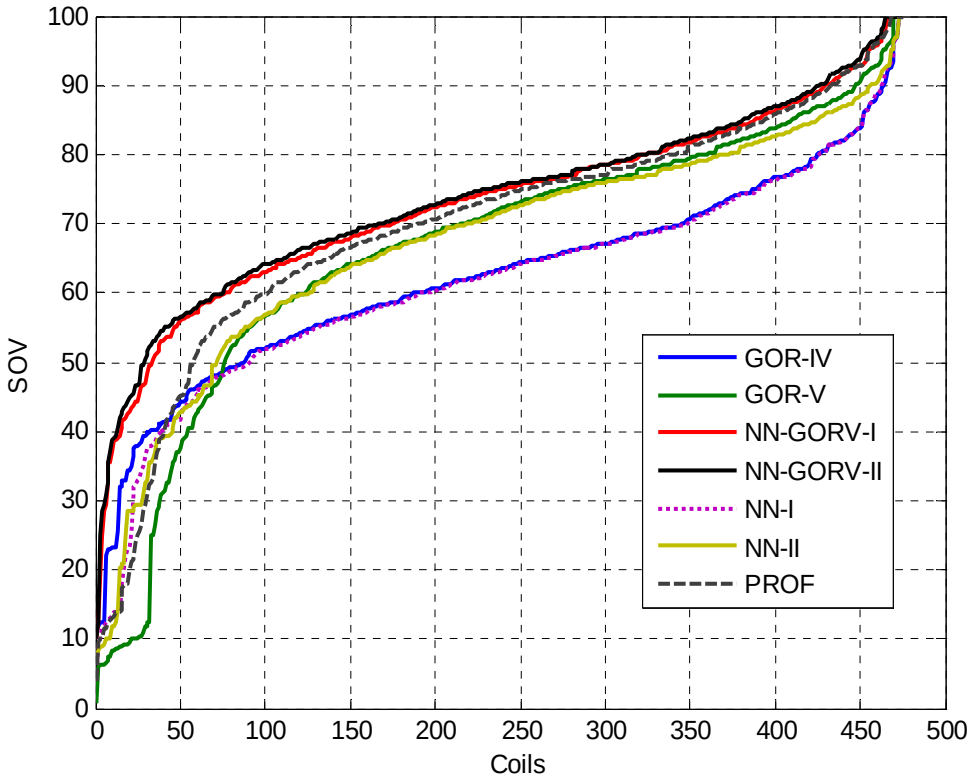


Figure 6.12: The coils segment overlap measure (SOV_C) of the seven prediction methods

6.8.2 Overall Performance and Quality of NN-GORV-I Method

The above analysis of the SOV measure for all the three states, helices, strands, and coils clarified that the predictions of the NN-GORV-I are far better than that scored by PROF method of this work. The results simply reveal that the NN-GORV-I method has high quality and more useful of than the PROF method.

Table 6.4 illustrate the results of the Matthews' correlation coefficients for the NN-GORV-I method. The coefficients are 0.77, 0.70, and 0.65 for alpha helices, beta strands, and coils, respectively. These figures are highly better than that of PROF and of course all the previously discussed methods (Table 6.4). These results revealed that NN-GORV-I method predicted states are more reliably related to the observed states. It is obvious that the figures and numbers of NN-GORV-I method carry more information about prediction than that of the PROF method.

Comparing the overall performance of the NN-GORV-I method with the original PROF needs a look at Table 6.2, Table 6.3, and Table 6.4 with all the corresponding figures. The prediction accuracies for helices (Q_H), strands (Q_E), and coils (Q_C) for the original PROF are 70.8%, 71.6%, and 77.2% with standard deviations of 29.8%, 25.3%, and 13.9%, respectively (Ouali and King, 2000). There is about 6% points in NN-GORV-I helices (Q_H) prediction higher than that of the original PROF while strands prediction of 3% have higher accuracy than of the NN-GORV-I strands (Table 6.2). This conclusion supports the findings reported by the authors of PROF that PROF predicts strands with relatively higher accuracy than other predictors. For coils, the NN-GORV-I prediction is more than the original PROF with about 3%. However, the overall performance of the NN-GORV-I Q_3 accuracy is about 2.5% better than original PROF. This result indicates that the NN-GORV-I outperform the original PROF in predicting protein secondary structure.

6.9 Assessment of NN-GORV-II Method

This section discusses and compares the findings of the seven prediction methods or algorithms examined in this research work. The NN-GORV-II is the method developed in this work to be an outstanding protein secondary structure classifier that predicts secondary structures from their amino acid sequences. As described in the previous section, the NN-GORV-I is developed by combining neural network method with GOR-V. The NN-GORV-I is further refined by using a filtering mechanism to the searched sequences database to mask low complexity regions. The *pfilt* program (Jones and Swindells, 2002) is used for this purpose. Although, there are limited changes to NN-GORV-I method, the use of the filtering mechanism to the searched database yields a different version of the NN-GORV-I which is called NN-GORV-II method.

6.9.1 Distributions and Statistical Description of NN-GORV-II Prediction

Figure 6.13 shows histograms of the performance of the Q_3 prediction accuracies and the segment overlap (SOV) measure of the 480 proteins. It shows that there is almost a negligible number of proteins that score a Q_3 below 50% and there are about 80 proteins score Q_3 predictions below 70% while other proteins scored above 70% with 180 proteins score 80% and about 140 proteins score 90%. This distribution of Q_3 scores have a tendency towards the 80% and 90% scores, making the average Q_3 score of NN-GORV-II method touches the 80% prediction accuracy.

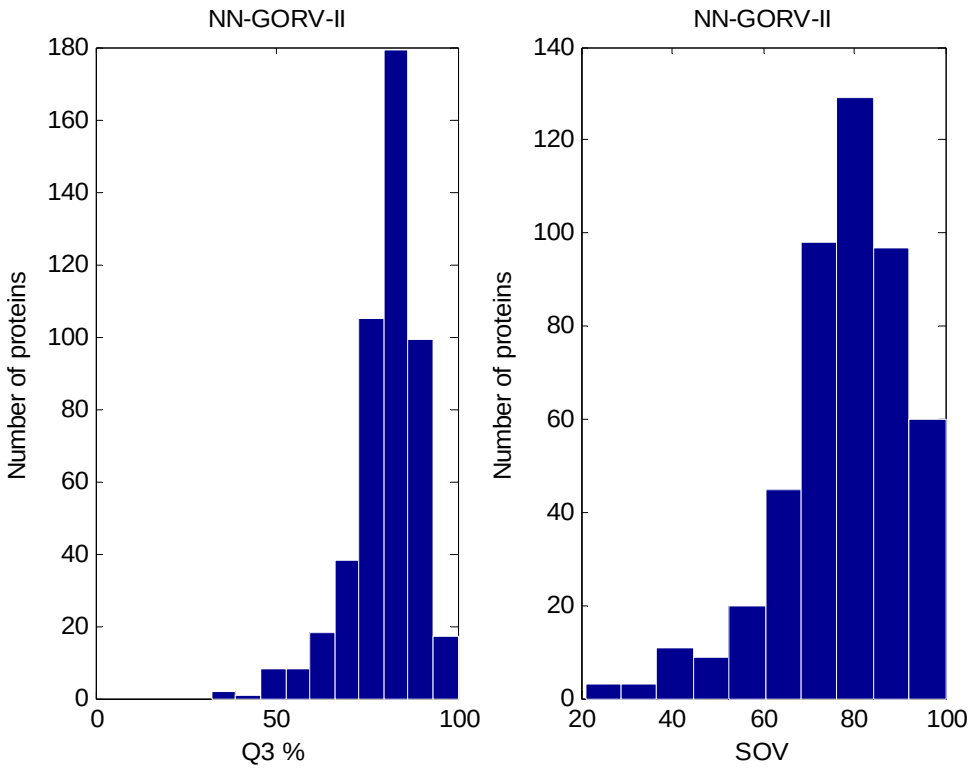


Figure 6.13: The performance of the NN-GORV-II prediction method with respect to Q_3 and SOV prediction measures

Table 6.5 elucidates these results in more details by rendering the Q_3 descriptive statistics of the secondary structure states. As for helices and strands of the NN-GORV-II method, the minimum predictions are 0.0% and the maximum are 100% and then the ranges are 100% for each state. The coils minimum prediction is 20% and maximum is 100% while the range is 80%.The minimum for the whole Q_3 prediction is 0.0% pushing the maximum to 97.4%. The mean standard deviation errors and variances are higher for the helices and strands states compared to the coil state and the whole Q_3 prediction.

The SOV measure in Figure 6.13 elucidates that NN-GORV-II method showed a different histogram than that of Q_3 performance. Among the 480 proteins there are about 60 proteins scored below 50% and about 60 proteins scored 100% SOV score. The rest of the proteins achieved score above 50% and below 100% with 120 proteins scored 80% SOV accuracy. This distribution of the SOV of NN-GORV-II method brought down the SOV score to the 76.27 level.

Table 6.5: Descriptive Statistics of the prediction accuracies of NN-GORV-II method

Structure	Min	Max	Range	Mean	Mean Std. Error	Std Dev	Variance
$\mathbf{Q_H}$	0.0	100.0	100.0	77.40	1.21	26.53	704.09
$\mathbf{Q_E}$	0.0	100.0	100.0	77.12	1.25	24.19	751.87
$\mathbf{Q_C}$	20.0	100.0	80.0	79.99	0.54	11.75	138.52
$\mathbf{Q_3}$	0.0	97.4	97.4	80.49	0.46	10.21	102.54

The above NN-GORV-II method SOV histogram of Figure 6.13 is further explained by the figures of Table 6.6. The minimum for helices and strands of the NN-GORV-II are 0.0% while the maximum are 100% with ranges of 100% each. The minimum for coils is 10% while the maximum is 100% with a range of 90%. The overall SOV minimum is 0.0% while the maximum and then the range is 98.8%. The high variances and standard deviations are shown by helices and strands while the low variances and standard deviations are shown by coils and the overall SOV. This indicates that the helices and strands scores are more dispersed than the scores of coils and overall predictions. The low mean SOV value of coils indicted that coils prediction for the NN-GORV-II method is of less quality and usefulness compared helices and strands that showed higher mean value and hence more useful and of high quality SOV scores.

Table 6.6: Descriptive Statistics of the prediction of SOV measure for NN-GORV-II method

Structure	Min	Max	Range	Mean	Mean Std. Error	Std Dev	Variance
$\mathbf{SOV_H}$	0.0	100.0	100.0	77.96	1.23	26.92	725.13
$\mathbf{SOV_E}$	0.0	100.0	100.0	79.94	1.32	24.57	840.46
$\mathbf{SOV_C}$	10.0	100.0	90.0	74.35	0.65	15.53	203.96
$\mathbf{SOV_3}$	0.0	98.8	98.8	76.27	0.75	17.50	267.58

Throughout the previous sections, results and discussion have been directed to explaining the performance, the quality, and the usefulness of the seven prediction methods. In the following section a detailed comparison of these methods will be explored.

6.9.2 Comparison of NN-GORV-II Performance with Other Methods

Figure 6.14 represents a histogram that elucidates the performance of the seven classification or prediction methods. It shows the seven classifiers Q_3 accuracy from the 50% level and above. Based on the nature of the composition of protein secondary structure, it is worth mentioning that prediction accuracy of 50% is worst than random guess. Baldi *et al.,* (2000) in their study about different protein data sets showed that the Q_3 accuracy for coil states is 48%. This number can be approximated to 0.5 probability of an event to occur; leading for detailed discussion about the dichotomous analysis in the next chapter.

Figure 6.14 shows the seven classifiers against their Q_3 accuracies. The NN-I method predicted about 30 proteins at the level between 50-55% and the PROF and NN-II methods predicted below 20 proteins for each respective level. This illustrates that these classifiers or predictors predict a considerable number of proteins at this low level of 50-55%. The NN-GORV-II predicts about 10 proteins at this level which suggested that the prediction ability of this method is negatively brought down by these proteins. However, the other three predictors which are GOR-IV, GOR-V, and NN-GORV-I methods are not shown at this low range of prediction accuracy.

NN-I and GOR-IV methods predict around 120 proteins each at the level of 55-65%. The rest of the prediction methods predicted less than 20 proteins each except the PROF which predicted about 30 proteins at the 55-65% level. This revealed that the NN-I and GOR-IV methods accuracies are much influenced by the 55-65% level of Q_3 prediction accuracy while the rest of the prediction methods are less influenced by this prediction level and PROF is somewhat influenced by this Q_3 level.

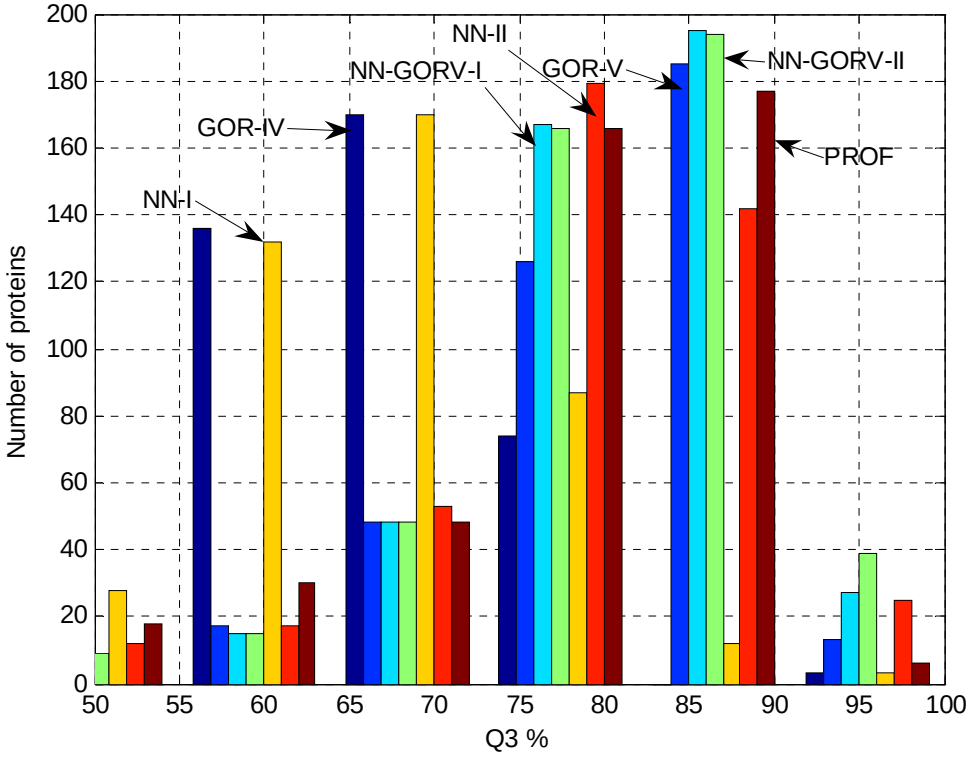


Figure 6.14: Histogram showing the Q_3 performance of the seven prediction methods

At the 65-72% Q_3 GOR-IV and NN-I predicted about 170 of the 480 proteins each while the rest of prediction methods predicted about 50 proteins at this Q_3 level. Again these results elucidated that GOR-IV and NN-I more predicted abundantly at this Q_3 level while the remaining prediction methods are predicted with less numbers of proteins at this Q_3 prediction level. This result explained that GOR-IV and NN-I methods predicted more proteins at this level and hence the final score for each will be affected by this Q_3 level and the level below it (55-65%) as

shown in Figure 6.14 while the rest of the methods predicted less proteins and hence these methods might be affected by other higher Q_3 prediction levels.

At the 75-80% Q_3 prediction level, NN-II method predicted about 180 proteins while NN-GORV-I, NN-GORV-II, and PROF methods predicted about 165 proteins each (Figure 6.14). GOR-V predicted above 120 proteins while NN-I and GOR-IV methods predicted around 80 proteins each. This revealed that NN-II, NN-GORV-I, NN-GORV-II, and PROF prediction methods predicted more proteins in the 75-80% level rather than lower levels of Q_3 prediction which will shift the prediction accuracies of these methods towards the high level of prediction accuracies. NN-I and GOR-IV methods predicted less protein at this level and more protein at lower levels as we discussed above and hence the predictive abilities of these two prediction methods are shifted towards lower prediction levels. GOR-V appears to have predictive accuracy between the two groups of prediction methods mentioned above.

At Q_3 prediction level of 85-90%, NN-GORV-I, NN-GORV-II, and GOR-V methods predicted above 180 proteins each, while PROF predicted below 180 proteins and the NN-II method predicted around 140 proteins. GOR-IV method did not predict any number of proteins at this level and NN-I predicted around 10 proteins. These results suggested that at this high level of prediction the NN-GORV-I, NN-GORV-II, GOR-V, and to a lesser extend PROF predicted many proteins at this level of Q_3 prediction (85-90%) which may push the level of accuracy of these predictors to a high level. The non appearance of GOR-IV at this Q_3 high level of prediction implied that GOR-IV is less accurate than the other predictors mentioned here.

Figure 6.14 shows the Q_3 prediction level of above 90-100% which is the highest level can be achieved to predict a protein. NN-GORV-II method predicted about 40 proteins while NN-GORV-I method and NN-II predicted about 25 proteins each at this level. GOR-V predicted about 15 proteins while the rest three prediction methods predicted less than 10 proteins each. These results supported the suggestion that NN-GORV-II predicts many proteins at Q_3 higher accuracy level compared to

the other prediction methods followed by NN-GORV-I method. NN-II predicted more proteins at this high level of prediction which suggested that this method will be pushed towards the high accuracy level while PROF predicted fewer proteins here which will drop its accuracy towards the previous levels.

However, Table 6.2 showed that NN-II scored a lower level of Q_3 accuracy than PROF; this can be explained by the fact that NN-II showed a higher standard deviation than PROF (Table 6.2) which made the prediction of NN-II scattered distribution prediction. GOR-IV and NN-II predicted very few proteins at this high level of accuracy (90-100%) while predicted many proteins at the level of 55-65% (Figure 6.14) a result suggested that these two methods among the low performance predictors of the seven prediction methods.

In conclusion, Figure 6.14 explains that the histograms distributions illustrate NN-GORV-II and NN-GORV-I outperform all other classifiers or prediction methods. However, NN-I and GOR-IV are the lowest performing classifiers and GOR-V, NN-II, and PROF are intermediate classifiers.

Figure 6.15 is a line graph designed to test the ability of the seven prediction methods, and how they behave in the prediction of the 480 proteins. An ideal line for an ultimate predictor is a line parallel to the x axis at a point of y axis equal to 100. When y equals to 50 for the same parallel line then the line represents a random guess for the coils states predictor. A line travels parallel to the x axis at y equals to 33.3 is as worst (poor) as random guess of a prediction. The figure resembles the reliability index (RI) for predicting proteins similar to that proposed by Rost (2003); that is to show the prediction methods did not only restrict their predictions to the most strongly predicted residues. It is also equivalent to the scale that discussed by Eyrich *et al.,* (2003) which plotted the accuracy versus coverage for subset of 205 proteins.

Figure 6.15 shows that NN-GORV-II line is travelling from Q_3 near 40% then steadily increasing accuracy to reach just below 100% assign the 480 proteins of the database. NN-GORV-II method line is above all the other six lines of other

prediction methods. The NN-GORV-I method line is just below the above line with small merging of dropping in accuracy. From the graph it can be concluded that the margin between NN-GORV-II line and NN-GORV-I line is the effect of *pfilt* program that mask low complexity regions of the data base as explained in the methodology. NN-GORV-II method is the second version of NN-GORV-I method that has been developed in this work, outperforming all the other methods as the figure shows.

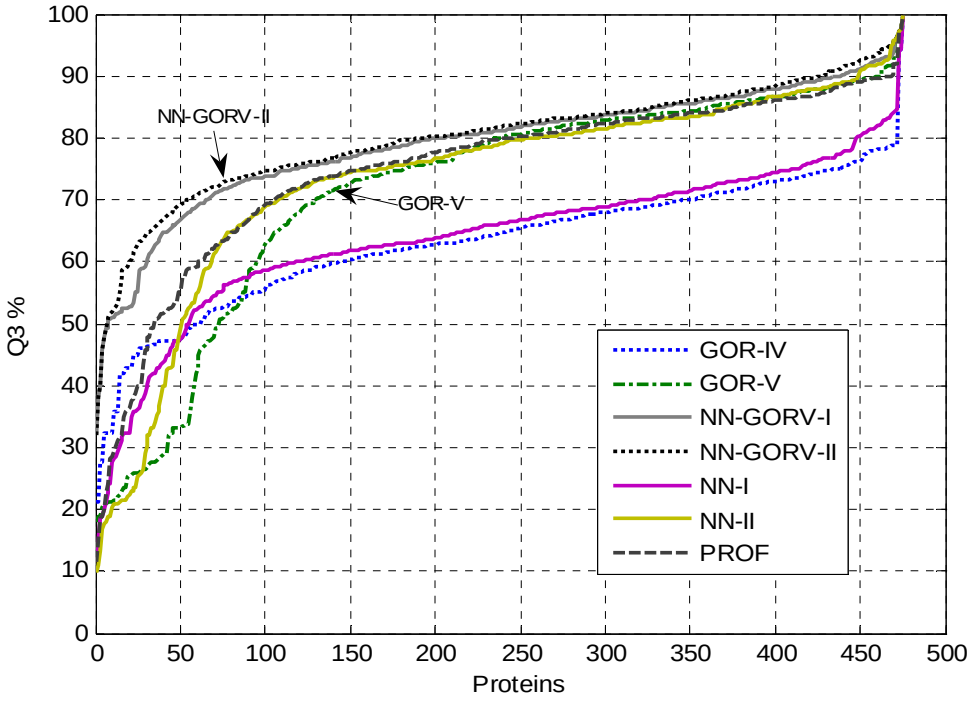


Figure 6.15: A graph line chart for the Q_3 performance of the seven prediction methods.

The same graph (Figure 6.15) shows that GOR-IV method travels from Q_3 prediction accuracy near 20% and then increases steadily until it reaches 85% spinning through the 480 proteins. GOR-IV line is under all the other six lines followed by NN-I method line just above it with very minor margin following a similar pattern indicting that GOR-IV method is the poorer performing prediction method followed by NN-I method. GOR-V method, NN-II method, and PROF method lines are in between the above mentioned four methods lines. GOR-V line is below the NN-II line while PROF line is above them and of course below the NN-GORV-I method and NN-GORV-II method lines. This graph elucidated that these

three methods are in between the NN-GORV-I and NN-GORV-II methods and GOR-IV and NN-I methods as far as Q_3 performance is concerned.

To conclude Figure 6.15, the newly developed method (NN-GORV-II) that combines GOR-V method and NN-II method is superior to all other methods studied in this work. Individual performance of GOR-IV and NN-I proved to be the poorest among other methods.

6.9.3 Comparison of NN-GORV-II Quality with Other Methods

Figure 6.16 shows a histogram of the SOV measure for the seven prediction methods. SOV has an ability to discriminate between similar and dissimilar segment distributions. This definition reflects the quality of prediction rather than a score or performance measure as discussed earlier.

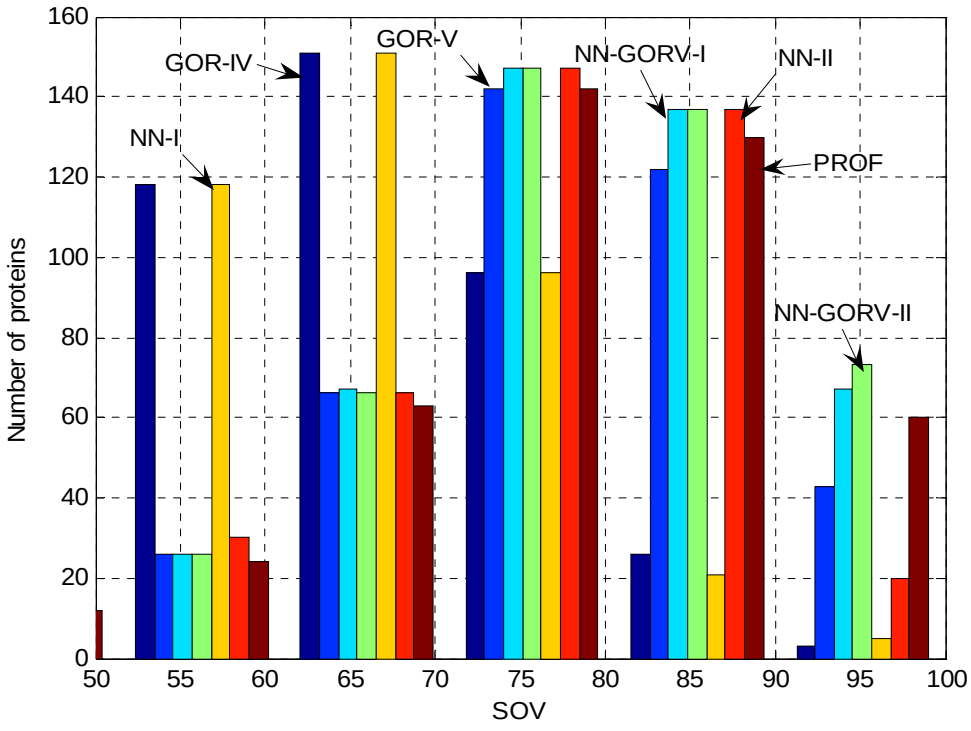


Figure 6.16: Histogram showing the SOV measure of the seven prediction methods

The distribution of the proteins according to each level of SOV followed almost the same pattern of Q_3 prediction accuracy. At the 50-60% SOV level, GOR-IV and NN-I methods predicted about 120 proteins each while the rest of the methods predicted 25 proteins each. For the 60-70% SOV level again GOR-IV and NN-I methods predicted about 150 proteins each while the rest of prediction methods predicted above 60 proteins each. At the 70-80% SOV level GOR-IV and NN-I methods predicted less than 100 proteins each while the rest of prediction methods predicted more that 140 proteins each.

Figure 6.16 also shows that when SOV level between 80-90% GOR-IV and NN-I methods predicted about 20 proteins each while the other five methods predicted about 125 proteins each. At the last SOV level which is 90-100%, GOR-IV and NN-I methods predicted less than five proteins each while NN-GORV-II and NN-GORV-I predicted about 65 proteins each. The PROF predicted 60 proteins; GOR-V predicted about 40 proteins, while NN-II predicted about 20 proteins at this high level of SOV.

These results elucidated that GOR-IV and NN-I methods predicted more proteins at lower levels of SOV while they predicted fewer proteins at higher levels of SOV. This is in contrast with the remaining five prediction methods which predicted more proteins at higher SOV levels. Among the five methods, NN-GORV-II, NN-GORV-I, and PROF predicted more proteins at the high level (90-100%) of SOV than the other two methods GOR-IV and NN-II. These results confirmed that NN-GORV-II and NN-GORV-I methods are of high quality prediction. The relatively many proteins predicted by GOR-V method at this high level of SOV compared to NN-II is confirmed by the Matthews correlation coefficients (Table 6.4) that is although NN-II outperformed GOR-V (Table 6.2), GOR-V prediction is of high quality and more useful than NN-II prediction.

It is clear that from the above results of Figure 6.14 and Figure 6.15, and Figure 6.16 NN-GORV-II method is superior and of high quality prediction method compared to other methods while NN-I and GOR-IV methods are the less accurate and of low quality methods. This concludes the discussion on the above two figures.

Figure 6.17 shows a line graph illustrates the same lines for the seven prediction methods but representing the SOV measure this time. Since the SOV measure is a measure of quality and reliability rather than performance, this figure shows the quality of each prediction method. NN-GORV-II and NN-GORV-I methods lines are above all the other five methods lines (Figure 6.17). The two lines are travelling through the proteins in the same pattern with a very small margin favouring NN-GORV-II method. This confirms the findings that NN-GORV-II and NN-GORV-I methods predictions are the most reliable and of high qualities predictions.

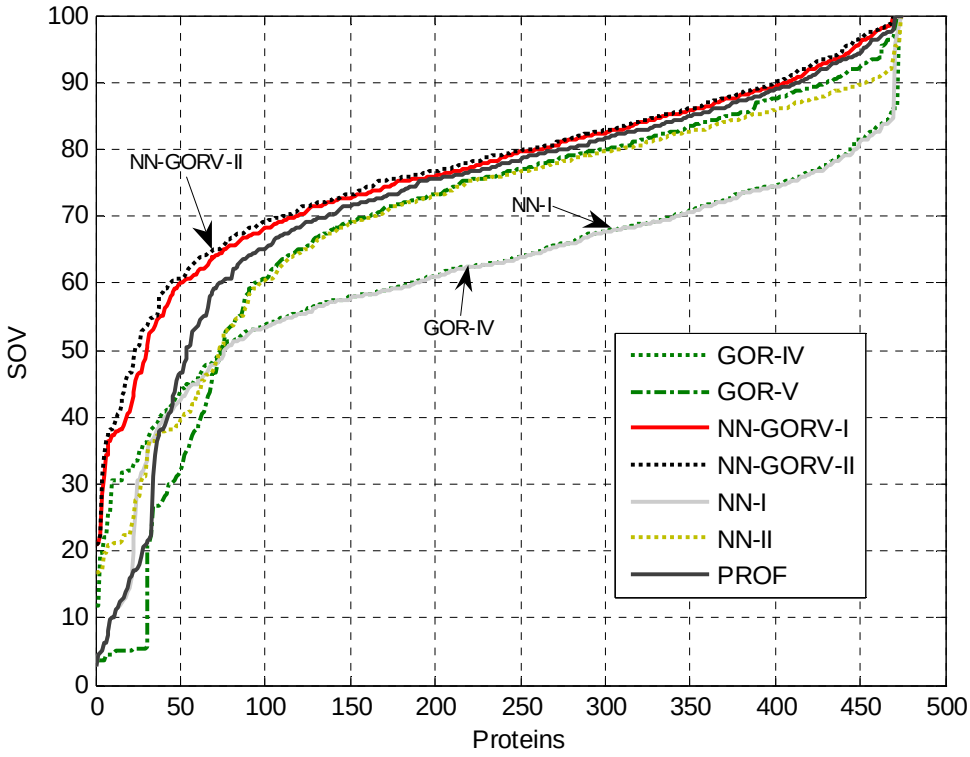


Figure 6.17: A graph line chart for the SOV measure of the seven prediction methods

The lines for NN-I and GOR-IV are almost identical but below all the other methods lines indicting that the prediction of these two methods are of low quality and less useful. NN-II and GOR-V methods lines are almost identical most of the time with a very little margin favouring GOR-V. This confirmed the fact that

although GOR-V performance is low compared to the NN-II performance, it exhibited a high quality prediction. Figure 6.17 shows also PROF line is travelling below NN-GORV-I and NN-GORV-II lines but above all the other four lines. This indicted that PROF is the third prediction method as far as quality is concerned.

The lines of Figure 6.17 confirm the facts revealed by Figure 6.16 that the newly developed method in this work (NN-GORV-II) method has the highest performance and the highest quality among all the seven methods studied in this work. This is followed by the NN-GORV-I, PROF, NN-II, GOR-V, NN-I, and GOR-IV methods, respectively.

6.9.4 Improvement of NN-GORV-II Performance over Other Methods

The following sections will discuss the gain and improvement of the prediction methods developed in this work. The NN-GORV-II is an advanced version of NN-GORV-I developed by combining two methods in this work, GOR-V and NN-II.

Table 6.7 shows the improvement of the prediction accuracy of helices, strands, coils, and all the three secondary structure sates together of NN-GORV-II over the other six methods. The improvement of NN-GORV-II method over NN-I and GOR-IV is very high which is above 19% improvement for the helices, and strands states but below 10% improvement for the coil states. However, the overall performance improvement (Q_3) of the NN-GORV-II method over NN-I and GOR-IV is above 16% which is a very big gain in secondary structure prediction accuracy. This result is not surprising since the two low performance predictors did not implement a multiple sequence alignment method to get use of the long range interactions of residues in the amino acid sequences.

Table 6.7: Percentage Improvement of NN-GORV-II method over the other six prediction methods

Prediction Method	Q_3	Q_H	Q_E	Q_C	Q_3 Improvement	Q_H Improvement	Q_E Improvement	Q_C Improvement
NN-I	64.05	57.29	57.39	74.1	**16.44**	20.11	19.73	5.89
GOR-IV	63.19	57.02	51.86	71.95	**17.3**	20.38	25.26	8.04
GOR-V	71.84	68.4	63.68	78.92	**8.65**	9.0	13.44	1.07
NN-II	73.58	70.77	68.72	78.33	**6.91**	6.63	8.40	1.66
PROF	75.03	70.65	68.29	79.38	**5.46**	6.75	8.83	0.61
NN-GORV-I	79.22	76.56	68.54	79.44	**1.27**	0.84	8.58	0.55
NN-GORV-II	80.49	77.4	77.12	79.99	**0**	0	0	0

GOR-V is one of the two methods that formed the NN-GORV-II method and hence the improvement over this method is of special importance. Table 6.7 showed that the improvements of the NN-GORV-II method over GOR-V are 9.0%, 13.44, and 1.07 for helices states (Q_H), strands (Q_E), and coils (Q_C), respectively. The improvements in helices and strands states are considerably high, especially for the strands since strands are known to be difficult to predict. The improvement in coil state is very low and this might be good sign that NN-GORV-II method is a high performance predictor since its gain is not from the coil states since most predictors over predict coil states.

When a prediction method gains an improvement in its helices and strands states, this means that this predictor is able to differentiate and discriminate between the three secondary structure states. That is because coils states are usually over predicted due to their high availability in the protein data set. As mentioned earlier a random guess of about 50% accuracy represent a right prediction for the coil states. The overall improvement (Q_3) of the NN-GORV-II method over the GOR-V method is 8.65%. The reported accuracy of GOR-V is 73.5% (Kloczkowski *et al*., 2002) which means an improvement of 6.99% is gained. Anyhow, whatever compared to the reported accuracy of GOR-V or the calculated accuracy in this experimental work, the improvement of the NN-GORV-II method performance over GOR-V is fairly high.

NN-II method is also one of the two methods that combined NN-GORV-II method. Table 6.7 shows the improvements of performance of NN-GORV-II method over the NN-II method are 6.63%, 8.4%, and 1.66% for helices (Q_H), strands (Q_E), and coils (Q_C) states, respectively. The improvement of Q_3 of NN-GORV-II over NN-II is 6.91%. The improvements in the helices and strand states are considerably high while the improvement in the coil states is low and as discussed before the gain in accuracies of beta strands is the most important among the three states of secondary structure. Most modern neural network methods of secondary structure prediction in the literature reported accuracies from 70.5% and below 76.4% (Riis and Krogh, 1996; Cuff and Barton 2000; Rost, 2003). However, an overall gain of accuracy of about 5- 7% in the NN-GORV-II method over NN-II in this experimental work and other works is an excitingly high gain.

Table 6.7 shows that the improvements of the NN-GORV-II method over the PROF method in this experimental work are 6.75%, 8.83%, and 0.61% for the helices (Q_H), strands (Q_E), and coils (Q_C), respectively. However, the improvements in the same states over the original PROF (Ouali and King, 2000), are 6.6%, 5.5%, and 2.8% for the helices (Q_H), strands (Q_E), and coils (Q_C), respectively. Unlike the original PROF, the gain of the NN-GORV-II method is very high for the helices and strands states over the PROF of this work while it is low for the coil states.

The improvement in the coil states over the original PROF is considerably high. However, the overall gain (Q_E) of the NN-GORV-II method over the PROF method is 5.46% for PROF this work and 3.8% over the reported Q_3 of the original PROF. The 3.8 -5.5% increment in the performance accuracy of the NN-GORV-II method over the PROF algorithm is considerably a significant gain in Q_3 accuracy if we compare this work with the work of Cuff and Barton (2000) where their Jnet algorithm achieved a 3.1% gain in Q_3 over the PHD (Rost and Sander, 1996) algorithm.

The improvement of the NN-GORV-II over NN-GORV-I method results are shown in Table 6.7. As explained earlier the NN-GORV-I method is the first version

of NN-GORV-II method and the increments in accuracies shown in the table is the affect of *pfilt* program. Except for strands states where the Q_3 accuracy improvement is 8.58%, the increments in accuracies for other states are very small and below 1% improvements. However, the overall increment in performance of Q_3 is 1.27% which is considered as significantly good gain since both experiments are conducted in identical environments except the invoking of *pfilt* program in the NN-GORV-II case.

Concluding the discussion about Table 6.7, the figures showed that the newly developed algorithm that combined the neural networks with information theory of GOR-V method is superior in performance to all methods tested here in this experimental work and most methods reported in the literature. The improvement in accuracies ranged from 5.5 % to 16.4% which is a significant gain in the domain of the protein secondary structure prediction. The *pfilt* program that masks low complexity regions in the searched database had even boosted the algorithm 1.27% further.

Table 6.8 shows the SOV measures improvements of the NN-GORV-II method over the other methods. The gain in the overall SOV_3 accuracies over the NN-I method and GOR-IV method are 15.33 and 14.20, respectively. The high gains in SOV over NN-I method and GOR-IV methods are expected since both methods did not use the multiple sequence alignment profile method to read more information from similar sequences (Cuff and Barton, 2000; Kaur and Raghava, 2003). Again the increments in SOV did reflect the fact that they are increments in prediction quality and usefulness rather that prediction performance.

6.9.5 Improvement of NN-GORV-II Quality over Other Methods

The overall SOV improvements of NN-GORV-II method over the GOR-V and NN-II methods that are the two methods which combined the NN-GORV-II

algorithm are 6.94% and 5.90%, respectively (Table 6.8). The most improvement in SOV is yielded from the strands states which recorded 15.94% and 11.47% for GOR-V and NN-II, respectively.

Table 6.8: SOV percentage improvement of NN-GORV-II method over the other prediction methods

Prediction Method	**SOV_3**	**SOV_H**	**SOV_E**	**SOV_C**	**SOV_3 Improvement**	**SOV_H Improvement**	**SOV_E Improvement**	**SOV_C Improvement**
NN-I	60.94	59.5	57.61	61.53	**15.33**	18.46	22.33	12.82
GOR-IV	62.07	60.81	56.01	62.34	**14.20**	17.15	23.93	12.01
GOR-V	69.33	70.87	64	66.63	**6.94**	7.09	15.94	7.72
NN-II	70.37	71.05	68.47	67.29	**5.9**	6.91	11.47	7.06
PROF	72.74	73.49	69.8	69.75	**3.53**	4.47	10.14	4.6
NN-GORV-I	76.55	76.93	70.76	72.9	**-0.28**	1.03	9.18	1.45
NN-GORV-II	76.27	77.96	79.94	74.35	**0**	0	0	0

A gain of about 6-7% in SOV over these two methods is significantly high gain and proved that combining two different methods of predictions that use different approaches might lead to an exciting improvement in protein secondary structure prediction usefulness and quality.

The improvement of NN-GORV-II algorithm over the PROF algorithm which is described as cascaded multiple classifier by its authors (Ouali and King, 2000) is shown in Table 6.8. SOV improvements of 4.47%, 10.14%, and 4.6% for helices, strands, and coils respectively are achieved. This is considerable improvement especially for the strands states which is very high and indicted that NN-GORV-II algorithm predicted the strands states in a high quality prediction compared to the PROF method in this work. In this work, the overall SOV accuracy of NN-GORV-II algorithm is increased by 3.53% compared to PROF which revealed that the new NN-GORV-II method is of high quality and useful in protein secondary structure prediction. However, the improvement in overall SOV (SOV_3) of the NN-GORV-II method over the published PROF SOV (Ouali and King, 2000) is 2.57%.

This fact leads to same conclusion as mentioned above that the method developed in this work is superior to the PROF method in predicting protein secondary structure.

The improvements in SOV of the NN-GORV-II method over the NN-GORV-I method are small in helices and coil states while is very high in strands sates and reached 9.18% (Table 6.8). However, the high improvement in the SOV of strands states did not reflect on the overall SOV where the NN-GORV-I proved to have slightly better SOV than the NN-GORV-II method. The negative value of 0.28 in Table 6.8 suggested that although there is an improvement in the overall performance accuracy of NN-GORV-II method over the NN-GORV-I method the quality of this prediction is not as good as the prediction of NN-GORV-I method.

6.9.6 Improvement of NN-GORV-II Correlation over Other Methods

Table 6.9 shows the improvements in the Matthews correlations coefficients (MCC) of NN-GORV-II method over the other methods. It is important to recall here that MCC is an index that shows how strong the relation between predicted and observed values. The nearest the coefficient to 1.0 the stronger the relation, while the nearest the coefficient to 0.0 the lesser the relation between observed and predicted values. There are significant improvements in the MCC of the NN-GORV-II method over the NN-I and GOR-V methods for all the secondary structure states ranging from 0.21-0.32 which indicated that the NN-GORV-II method is significantly containing high entropy or more information to describe the relation between predicted and observed values and its prediction is of more meaning than these two methods (Crooks *et al.,* 2004; Baldi, *et al*., 2000).

Table 6.9 also shows that the improvements in the MCC of the NN-GORV-II method over the GOR-V and NN-II are ranging from 0.08-0.13 for all the secondary structures sates; helices, strands, and coils. There are more improvements in the strand states compared to other states over both GOR-V and NN-II methods. This result revealed that the new developed algorithm by combining these two algorithms is superior in terms of describing more relations between predicted states and

observed ones with more emphasis to strands sates which are known to be difficult to predict.

Table 6.9: Matthews Correlation Coefficients improvement of NN-GORV-II method over the other six prediction methods

Prediction Method	**MCC_H**	**MCC_E**	**MCC_C**	**MCC_H Improvement**	**MCC_E Improvement**	**MCC_C Improvement**
NN-I	0.4906	0.4124	0.4448	0.2838	0.2834	0.2053
GOR-IV	0.5283	0.3756	0.4382	0.2461	0.3202	0.2119
GOR-V	0.6859	0.5994	0.5675	0.0885	0.0964	0.0826
NN-II	0.6503	0.5641	0.5304	0.1241	0.1317	0.1197
PROF	0.7102	0.6291	0.5743	0.0642	0.0667	0.0758
NN-GORV-I	0.7736	0.6959	0.6494	0.0008	-0.0001	0.0007
NN-GORV-II	0.7744	0.6958	0.6501	0	0	0

As far as the improvements of the MCC of the NN-GORV-II method over the PROF method are concerned, Table 6.9 shows that the increments in helices, strands, and coils are 0.06, 0.07, and 0.08, respectively. These are considerable improvements in the entropy of these states if we define the entropy as the information need to describe variables (Crooks and Brenner, 2004; Baldi, *et al.*, 2000). This result proved that the NN-GORV-II algorithm is not only superior in performance (Table 6.2) but also superior in describing the strength of the relations between observed and predicted states in its prediction.

The increments in the MCC achieved in the NN-GORV-II method over its previous version NN-GORV-I are shown in Table 6.9. The improvements in helices states and coils states are very small and counted to 0.001 each. Although this is very minor gain in MCC coefficients but it indicated that the improvement in the performance of the NN-GORV-II over NN-GORV-I method (Table 6.2) is accompanied by improvements in the strength of the predictions for the helices and coil states. However, Table 6.9 also shows a negative number (-0.0001) as the improvement in the MCC of the strand states of the NN-GORV-II method over NN-

GORV-I method. This elucidated that the amount of information described NN-GORV-I method prediction is a more than the information described NN-GORV-II method prediction.

This result also concluded that the gain in performance of the strands states (Table 6.2) of the NN-GORV-II method over NN-GORV-I method is not coupled by same gain or in the entropy or the information describing this prediction. This result is also confirmed by the results of the SOV values in Table 6.8 which suggested that the NN-GORV-I method prediction is of higher quality and more usefulness than the NN-GORV-II method; a fact that might questioned the improvement achieved in performance by using pfilt program.

6.10 Summary

In this chapter, the performance of the seven methods conducted in this work is described and assessed in detail. The results confirmed that methods or algorithms that did not use sequence alignment profiles like GOR-IV and NN-I are found to be of very low performance ranging between 63-64% compared to other methods. When the above two methods used multiple alignment profiles and hence named GOR-V and NN-II, a significant gain in the accuracy has been achieved and reached the range of 73-75%. The PROF method conducted in this work with almost the same database and environment of the original PROF and has achieved accuracy performance almost similar to that reported in the original PROF. This facilitates the statistical comparison with the method developed in this work.

The newly NN-GORV-II algorithm developed in this work which is an advanced version of NN-GORV-I algorithm developed in this work too, proved to be of superior performance that outperformed all algorithms implemented in the experimental work of this research. The NN-GORV-II algorithm outperformed the reported accuracy of the multiple cascaded classifier (PROF) method which is 76.7% (Ouali and King, 2000) and reached an accuracy of 80.84%.

The NN-GORV-II also proved that it is of high quality and more useful compared to the other methods. The method also proved that the entropy and the information used to describe its strength of prediction is more than the information used in the other prediction methods. However, the results proved that the NN-GORV-II method is superior to the NN-GORV-I method in performance of the prediction but might not in the quality of the prediction.

CHAPTER 7

THE EFFECT OF DIFFERENT REDUCTION METHODS

7.1 Introduction

The widely known and used DSSP (Dictionary of Protein Secondary Structure) algorithm to assign the secondary structure categories to the experimentally determined three-dimensional (3D) structure has been used in this experimental work. Among other algorithms to conduct the same task of assigning secondary structures are STRIDE and DEFINE. As described by the DSSP authors, the DSSP works by assigning potential backbone hydrogen bonds which based on the 3D coordinates of the backbone atoms and subsequently by identifying repetitive bonding patterns The DSSP database is a database of secondary structure assignments for all protein entries in the Protein Data Bank (PDB) and the DSSP program was designed by Kabsch and Sander to standardize these secondary structure assignments (Kabsch and Sander, 1983; Kabsch and Sander, 1984).

As mentioned in the methodology chapter, The DSSP algorithm classifies each residue into eight classes: H => α alpha helix; B =>residue in isolated β bridge; E =>extended strand, participates in β ladder; G => 3-helix [3/10 helix]; I => 5 helix [pi helix]; T => hydrogen bonded turn; S => bend; and "**.**". Since the methods developed and or implemented in this experimental work used the three states of protein secondary structure, these eight classes are collapsed or reduced into the three standard classes associated with helices (H), strands (E), and coils(C).

The adopted reduction schemes from the mentioned eight states or classes to three classes of helices, strands, and coils are usually performed by using one of the five assignment or reduction methods or schemes discussed previously in the methodology chapter.

7.2 Effect of Reduction Methods on Dataset and Prediction

The mentioned reduction methods are well established for a long time and some of them have been established for decades (Kabsch and Sander, 1983). It was argued that the eight-to-three state reduction scheme can alter the prediction accuracy of an algorithm in a range of 1-3% (Cuff and Barton, 1999). It is worth mentioning that the purpose of this chapter is to study the effect of the reduction methods on the newly developed algorithm NN-GORV-II and its affect on prediction accuracy and quality. The NN-GORV-II algorithm has been tested using the five reduction methods, which facilitates the comparison of this algorithm with other prediction algorithms adopting any of these five reduction methods.

In this experiment, Method II reduction has been adopted because it is considered to be among the stringent definitions of reduction. However, Method I usually results in lower prediction accuracy than other definitions or reduction methods. Method V is used to compare the effect of reduction schemes on prediction accuracy.

Table 7.1 shows the numbers of helices, strands, and coils according to each of the reduction methods from the eight states to the three states. A PERL program was developed to make these assignments and count the number of the total residues in the database and then the numbers and the ratio of each secondary structure state. From Method I to Method V the number of states assigned as coils increased gradually.

Table 7.1: Percentage of secondary structure state for the five reduction methods of DSSP definition (83392 residues)

Reduction Method	Helix		Strands		Coils	
	Number	%	Number	%	Number	%
Method I	28851	35	18951	23	35590	43
Method II	28881	35	17810	21	36701	44
Method III	28851	35	17810	21	36731	44
Method IV	25807	31	18951	23	38634	46
Method V	25807	31	17810	21	39775	48

The percentages of coils for Method I is 43% and then increased to 44% for Method II and III until it reached 48% for Method V. The helices are 35% for the first three methods and then decreased into 31% for methods IV and V. the Strands are 23% for method I and IV while they are 21% for the other reduction methods. The above table clearly explains that the least numbers of residues assigned to the coils states are for Method I while the best numbers are for Method V. Method V revealed that half of the residues are assigned to the coils states ($0.48 \approx 0.5$).

7.2.1 Distribution of Predictions

Table 7.2 shows the results of one way analysis of variance procedure (ANOVA) against the performance of prediction accuracy (Q_3) of the five reduction methods. The ANOVA procedure tests for the hypothesis that what ever all means of the five methods are similar or there are significant differences between them. In other words, the importance of this test is to accept or reject the fact that the means of the performance of the five reduction methods differ significantly at the 0.05 or 0.01 probability level or not. The same ANOVA test has been conducted for the SOV of the five reduction methods shown in Table 7.3.

Tables 7.2 and 7.3 show that the total degree of freedom of the test is 479 and that means 480 proteins (observations or entries) had been used in evaluating each method. Assignment for both between and within groups had been allocated at random; the total of sum of squares, is, however, the most important to determine the F-test. Method I is randomly chosen as a factor variable to compare methods with and among each others.

Table 7.2 presents the results of the five reduction methods. It shows that the means are significantly different from each others at the 0.001 probability level, as far as their performance accuracies are concerned. This probability level suggested that we are more than 99% sure that these methods differ from each others. The same conclusion applies for Table 7.3, that the five reduction methods are significantly different from each other as far as their SOVs are concerned. It elucidates that the five reduction methods are different in their quality and usefulness.

Table 7.2: The analysis of variance procedure (ANOVA) of the Q_3 for the five reduction methods*

Method		Sum of Squares	df	Mean Square	F-test	Significance
Method II	Between Groups	49578.977	252	196.742	122.356	.000
	Within Groups	365.003	227	1.608		
	Total	49943.980	479			
Method III	Between Groups	49633.031	252	196.956	132.267	.000
	Within Groups	338.023	227	1.489		
	Total	49971.053	479			
Method IV	Between Groups	44528.264	252	176.699	29.473	.000
	Within Groups	1360.915	227	5.995		
	Total	45889.180	479			
Method V	Between Groups	45300.225	252	179.763	24.194	.000
	Within Groups	1686.648	227	7.430		
	Total	46986.873	479			

* Method I is control

Table 7.3: The analysis of variance procedure (ANOVA) of SOV for the five reduction methods*

Method		Sum of Squares	df	Mean Square	F-test	Significance
Method II	Between Groups	134307.505	295	455.280	6.774	.000
	Within Groups	12367.493	184	67.215		
	Total	146674.998	479			
Method III	Between Groups	134764.938	295	456.830	6.833	.000
	Within Groups	12300.720	184	66.852		
	Total	147065.657	479			
Method IV	Between Groups	128010.211	295	433.933	15.716	.000
	Within Groups	5080.433	184	27.611		
	Total	133090.644	479			
Method V	Between Groups	144217.099	295	488.872	3.633	.000
	Within Groups	24761.180	184	134.572		
	Total	168978.279	479			

* Method I is control

Figure 7.1 shows how the 480 amino acids had been predicted and distributed through the different levels of Q_3 predictions by the five different reduction methods. As mentioned before the NN-GORV-II algorithm was screened using the five reduction methods to give a clear portray of this algorithm and study its response and stability towards each method. The descriptive statistics for the five reduction methods regarding Q_3 and SOV is shown in Appendix C.

Figure 7.1 elucidated that the performance accuracy Q3 for Method V predicted just below 250 of the 480 proteins tested at the level of 80-90%, just above 100 proteins for the level of 70-80, and below 100 proteins for the 90-100%. Method IV had a similar pattern of Method V, while other three reduction methods predicted just above 200 proteins at the 80-90% level. The five histograms for the five reduction methods illustrate that although they are entirely different reduction methods, the NN-GORV-II algorithm is stable in predicting the 480 proteins and each prediction took almost similar distribution.

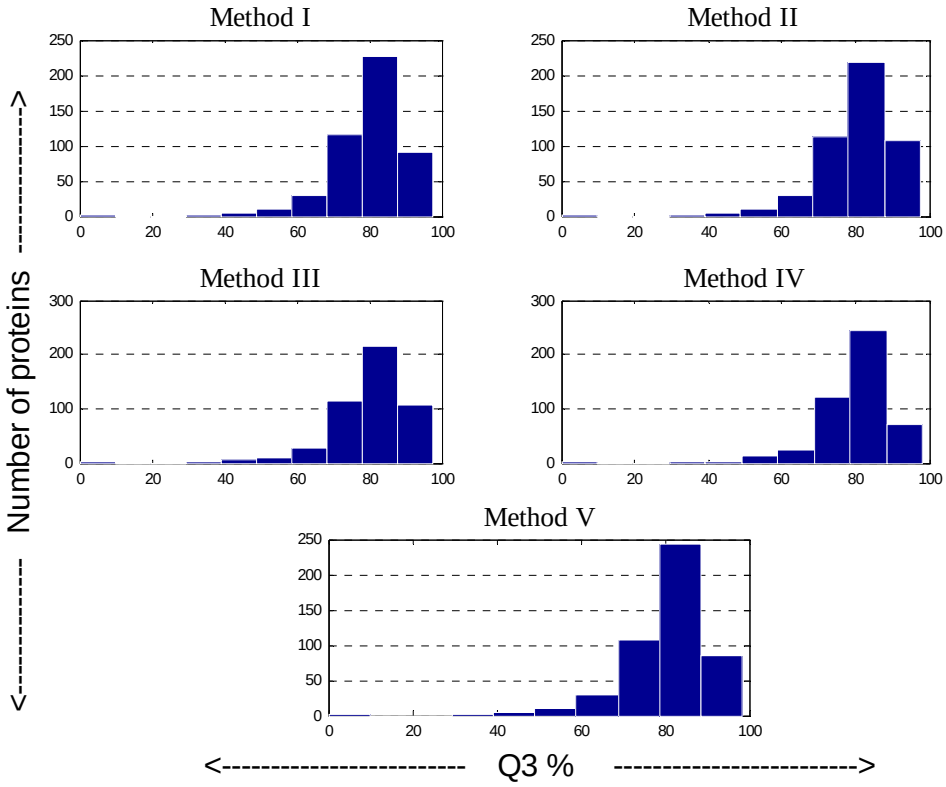


Figure 7.1: Five histograms showing the Q3 distribution of the test proteins with respect to the five reduction methods

The SOV measure distribution of the five reduction methods is shown in Figure 7.2. It is clearly elucidated in the histograms the variability of the SOV measures are more scattered than that of the Q3 variability (Figure 7.1). Method II and Method III predict more proteins at higher SOV range levels. This is followed by method I and Method IV, while Method V shows more proteins scoring SOV below 60%. This reveals that Method V was of low quality and less useful prediction followed by Method I and IV while Method II and III are of high quality and meaningful prediction. These results will be explained in more detail when studying the exact values of each reduction method and then arrive at solid conclusion.

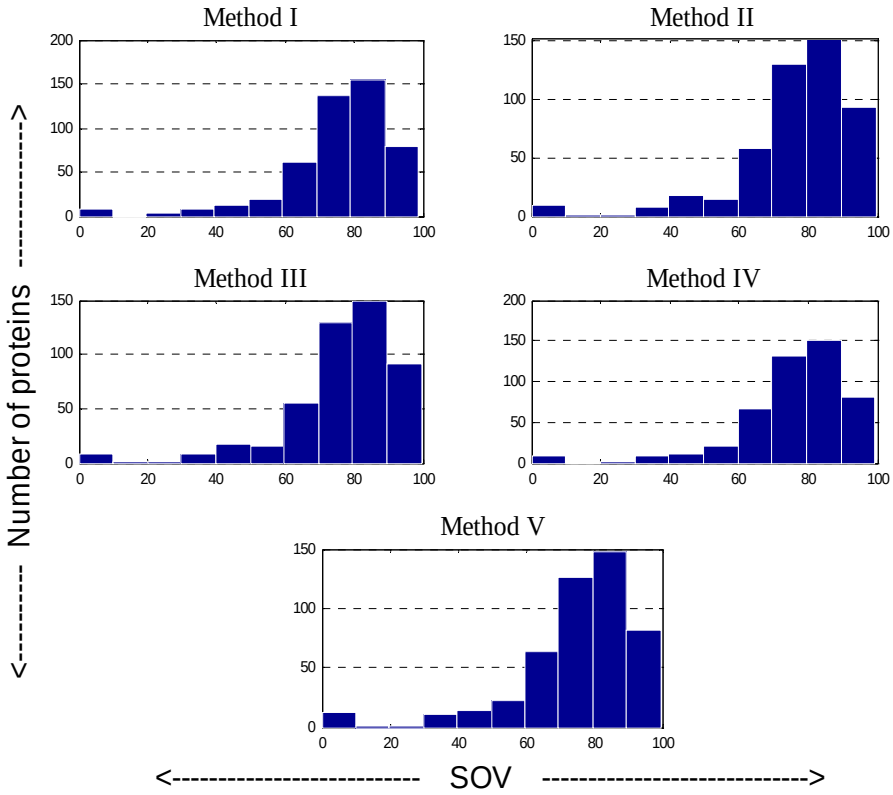


Figure 7.2: Five histograms showing the SOV distribution of the test proteins with respect to the five reduction methods

7.2.2 Effect of Reduction Methods on Performance

To explore the effect of the five reduction methods on the NN-GORV-II performance, Table 7.4 shows the scores of the helices (Q_H), strands(Q_E), coils(Q_C), and all the states together (Q_3) with respect to each reduction method. The performances of helices (Q_H) are almost the same and about 77.4% with standard deviations of 26.53% for all the first three methods, I, II, and III. The performances of the helices (Q_H) for Method IV and Method V are 87.03 with standard deviations 20.57 for each. There is about 10% Q_H increase in predicting helices for methods IV and V compared to methods I, II, and III. This increase in Q_H accuracy was accompanied by a 6% decrease in the standard deviations for methods IV and V. This result proves that methods IV and V predicted helices more accurately and the prediction is more homogenous compared to the other three methods I, II, and III.

The strands (Q_E) prediction accuracies are 77.12% with standard deviations of about 12% for methods II, III, and V while strands predictions are 69.49% with standard deviations of 27.42% for methods I and IV.

Table 7.4: The effect of the five reduction methods on the performance accuracy of prediction (Q_3) the of NN-GORV-II prediction method

Reduction Method	Q_3	Q_H	Q_E	Q_C
Method I	79.88±10.13	77.42±26.53	69.49±27.42	80.31±11.77
Method II	80.49±10.21	77.40±26.53	77.12±24.19	79.99±11.75
Method III	80.48±10.21	77.42±26.53	77.12±24.19	79.96±11.77
Method IV	80.38±9.79	87.03±20.57	69.49±27.42	78.34±11.78
Method V	80.98±9.90	87.03±20.57	77.12±24.19	78.07±11.76

Calculations are estimated from 480 amino acids
Q_3 is the accuracy per amino acid
Q_H is the accuracy for α helices
Q_E is the accuracy for β strands
Q_C is the accuracy for coils

This reveals that strands predictions have higher accuracies and more stable and homogenous for methods II, III, and V in comparison with other two methods. It had been reported in the literature that beta strands are difficult to predict compared to the other two states. Ouali and King, (2000) reveals that their algorithm (PROF) predicted strands with accuracy of 71.6% and that was the highest accuracy to be achieved by a protein secondary structure classifier or predictor.

As for the coils states prediction accuracy (Q_C), Table 7.4 shows that methods I, II and III scored about 80% prediction accuracies with standard deviations of 11% each while the prediction for the coil states scored about 78% with standard deviations of about 11%for methods IV, V each. This result proves that methods IV and V predicted the coil states with less accuracies but with the same stabilities and homogeneities compared to the other three methods.

Considering the overall prediction accuracies (Q_3) for the five reduction methods, Table 7.4 shows that Method I recorded the least accuracy of 79.88% while

Method V recorded the highest accuracy which is 80.98%. The other three methods recorded accuracies of 80.49%, 80.48%, and 80.38% for methods II, III, and IV, respectively. The standard deviations for all the five methods are almost the same and are around 10% which showed small standard deviations that reflected homogenous and stable predictions for all the five reduction methods. This observation is confirmed in Figure 7.3 which shows the trend of predicting the 480 proteins using the different five reduction methods. However, the graph portrays that the five reduction methods performed in more or less similar trend and the margin differences between the five methods are very small.

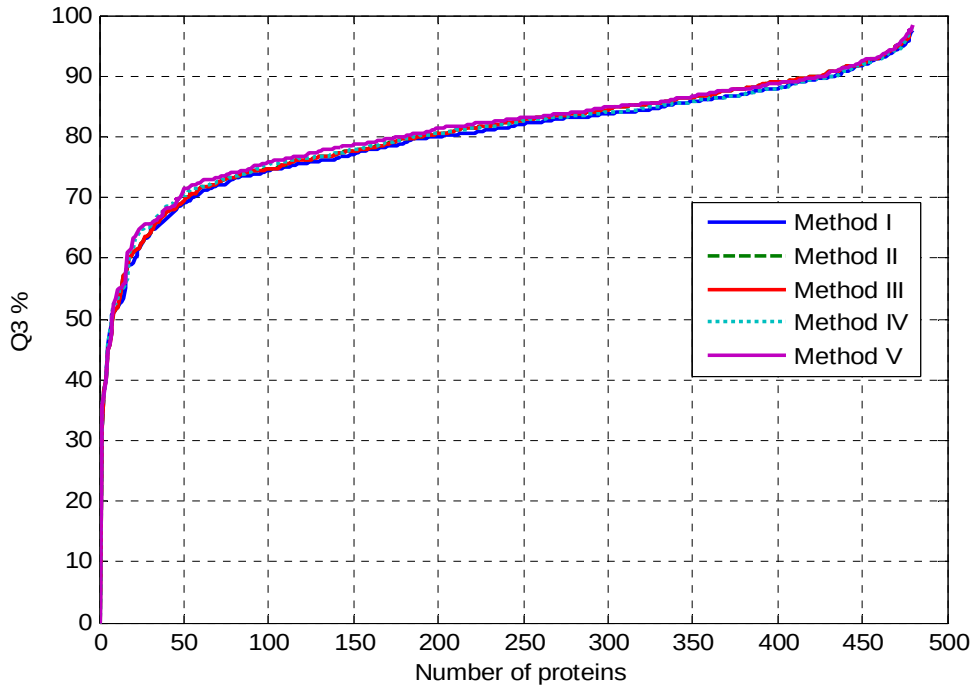


Figure 7.3: The performance accuracy (Q_3) of the five reduction methods on the test proteins

By further elaboration to Table 7.4, it is clear that Method I records the most rigorous and least accurate performance in assessing the NN-GORV-II algorithm. In contrast, Method V shows the highest accuracy demonstrating that it is the most optimistic method of assessing prediction algorithms. The difference in accuracy prediction (Q_3) between Method I and V is 1.1% which is a considerable and true difference in evaluating prediction algorithms since this difference has been resulted from experiments conducted in exactly the same environments. This result is

consistent with Cuff and Barton (1999) in leading the conclusion that different reduction methods can affect prediction accuracy of an algorithm with a range of 1-3%. Method II had a medium score between methods I and V, while having similar pattern score to methods III and IV. However, the difference in Q_3 score between Method II which is adopted in assessing the NN-GORV-II algorithm through out the experimental work in this research, and Method I is 0.61%. This is a very small difference that does not affect the reported accuracies of NN-GORV-II method.

7.2.3 Effect of Reduction Methods on SOV

The response of the five reduction methods to the SOV measures is shown in Table 7.5. The SOV_H of helices for methods I, II, and III are about 77% with standard deviations of about 26% each while the SOV_H for methods IV and V are 87.63% with standard deviations of 21.33% each. This indicates that methods IV and V predictions for the helices states are of higher qualities and stabilities compared to the other three methods.

Table 7.5: The effect of the five reduction methods on the segment overlap measure (SOV) of the NN-GORV-II prediction method*

Reduction Method	SOV_3	SOV_H	SOV_E	SOV_C
Method I	75.83±16.36	77.98±26.93	71.19±28.99	73.41±14.28
Method II	76.26±17.50	77.95±26.92	79.94±24.57	74.35±15.52
Method III	76.25±17.52	77.98±26.93	79.94±24.57	74.32±15.57
Method IV	75.84±16.67	87.63±21.33	71.19±28.99	72.69±14.84
Method V	74.93±18.78	87.63±21.33	79.94±24.57	72.50±16.33

*calculations are estimated from 480 amino acids
Q_3 is the accuracy for residue or amino acid
Q_H is the accuracy for α helices
Q_E is the accuracy for β strands
Q_C is the accuracy for coils

The SOV measures of strands SOV_E with respect to the five reduction methods is shown in Table 7.5. The SOV_E measures are 79.94 with standard deviations of 24.57 for methods II, III, and V though the SOV_E measures are 71.19%

with standard deviations of 28.99% for methods I and IV. These results indicate that method II, III, and V predict strands states with higher quality and more stability than methods I and IV.

The coils states SOV_C measures for the five reduction methods are shown in Table 7.5. Methods II and III scored about 74% SOV_C with standard deviations of about 16%. Methods IV and V achieved 72.69% and 72.5 SOV_C measurement for coils with standard deviations of 14.84% and 16.33, respectively while Method I achieved 73.41 SOV_C with standard deviation of 14.28. Referring to Table 7.4 which showed high performances for the coil states (Q_C) for the five reduction methods, the SOV_C results (Table 7.5) reflects that respective predictions of the coil states for the five methods are of low qualities, less usefulness, and less stabilities.

Table 7.5 shows the overall segment overlap (SOV_3) measures for the five reduction methods. Methods II and III achieve overall SOV_3 of 76.3% with standard deviations of 17.5 each. Method I and IV score SOV_3 of 75.8% with standard deviations of about 16% each. Method V achieves an overall SOV_3 for all the secondary structure states reached 74.93% with standard deviations of 18.78%. The figures of this table are rendered in Figure 7.4 which shows very small marginal differences between the five reduction methods.

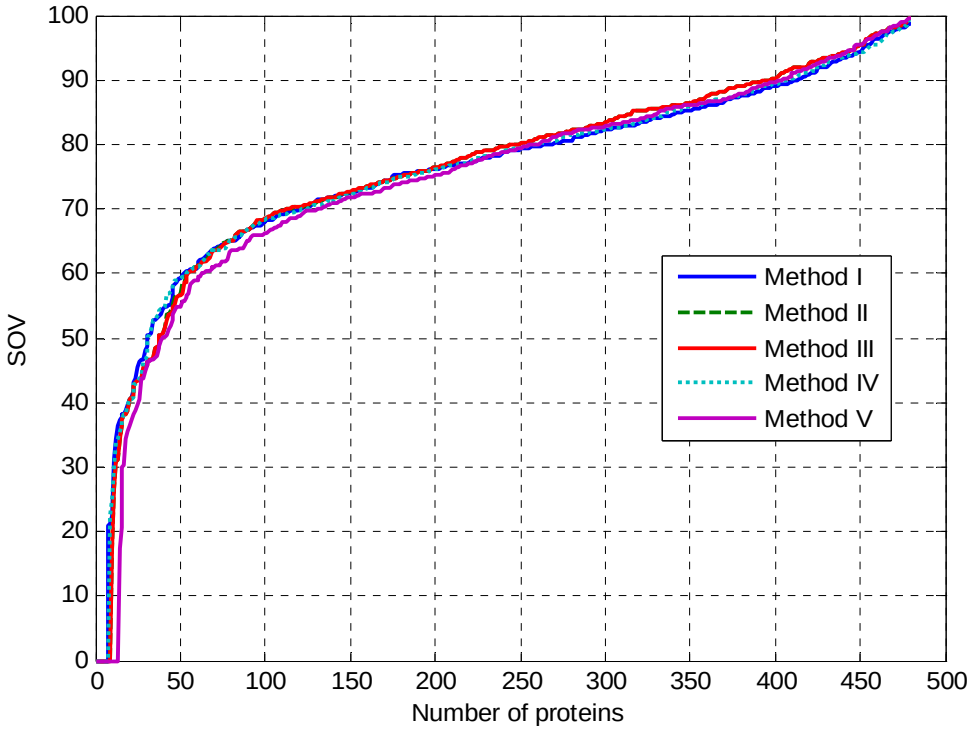


Figure 7.4: The SOV measure of the five reduction methods on the 480 proteins using NN-GORV-II prediction method

These results reveal that methods II and III predict the secondary structures of proteins with high quality and usefulness while methods I and IV predict proteins with comparatively less quality. However, Method V had achieved the highest apparent performance (Q_3) in prediction accuracy (Table 7.4). Method V as well had achieved the least SOV_3 and hence the least quality of prediction compared to the other five reduction methods. The above results also conclude that Method II which had been adopted in this work to evaluate the NN-GORV-II algorithm showed a higher quality and more usefulness than Method I.

7.2.4 Effect of Reduction Methods on Matthews's Correlation Coefficients

The effect of the five reduction methods on the Matthews's correlation coefficients (MCC) are shown in Table 7.6. The coefficients of the helices states (MCC_H) for methods IV and V are 0.79 each; they are 0.78 for methods I and III while the MCC_H for method II is 0.77. This indicates that although the correlation

coefficients for all the methods are almost similar, Method IV and V achieve the highest correlation coefficients which indicate that the relation between predicted and observed secondary helices structures is very strong for these two reduction methods.

Table 7.6: The effect of reduction methods on Matthews's correlation coefficients using NN-GORV-II prediction method

Reduction Method	**MCC_H**	**MCC_E**	**MCC_C**
Method I	0.779	0.700	0.654
Method II	0.774	0.696	0.650
Method III	0.779	0.714	0.666
Method IV	0.790	0.700	0.668
Method V	0.790	0.714	0.681

Calculations are estimated from 480 residues or amino acids
MCC_H is the Mathews correlation coefficient for α helices
MCC_E is the Mathews correlation coefficient for β strands
MCC_C is the Mathews correlation coefficient for coils

As for the strands states the Matthews's correlation coefficients (MCC_E) are 0.70 for methods I, II, and IV while they are 0.71 for methods III and V. The results reveal that the predicted strands states of methods I, II, and IV are less related to the observed ones compared to the other two methods but the differences are very minor.

The coils states Matthews's correlation coefficients (MCC_C) for the five reduction methods is 0.65 for methods I and II, 0.67 for methods III and IV, and 0.68 for Method V. Again these results reveal that methods I and II predictions for the coil states are less related to the observed coils while Method V coils predictions are more related to the observed coils states.

7.3 Summary

Five reductions methods that assign the DSSP eight protein secondary structural classes into the commonly used three structural classes are attempted in this work to test the ability of the newly developed NN-GORV-II algorithm performing under different assignment or reduction methods. The number of helices, strands, and coil states are affected by different reduction methods and the one way analysis of variance procedure showed that the five reduction methods varied significantly in their performance (Q_3) and quality (SOV_3) of predicting protein secondary structures.

Further analysis depicted that although there are differences between the five reduction methods in their performances, these are as half as had been estimated in other studies. Method I is the most pessimistic in its performance response while Method V is the most optimistic. Using method I will make a reliable comparison of the NN-GORV-II algorithm with other algorithms rather than using Method V. Method II which has been adopted in this work is in middle performance between method I and V and can let the NN-GORV-II algorithms to be fairly compared to other algorithms. However, for a reliable comparison of NN-GORV-II algorithm with other algorithms, 0.6% can be deducted from the NN-GORV-II algorithm performance. The evaluation of the five reduction method also proves and suggests that NN-GORV-II algorithm is stable and robust in performance and quality using different reduction methods.

CHAPTER 8

PERFORMANCE OF BLIND TEST

8.1 Introduction

As described by their founder, the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused. There have been several experiments in CASP every two years since 1994.The CASP3 competition gathered prediction groups from all around the world.

The goal of CASP experiments is to obtain an in depth and objective assessment of the current abilities and inabilities in the area of protein structure prediction. In the competition, participants will predict as much as possible about a set of soon to be known structures. This type of prediction was described by CASP initiators as true prediction and prediction made on already known proteins. Full details of these competition and results of predictions can be located at the CASP prediction center web site, http://PredictionCenter.llnl.gov/, and in the special issues of the Proteins journal (Moult *et al.*, 1997; Moult *et al.*, 1999).

CASP3 targets are used in this independent or blind test which represents sequences that have never been used in training the new NN-GORV-II algorithm developed in this work. The importance of these CASP3 proteins is that they are classified by the CASP organizers as proteins with no homologous sequences of known structure.

8.2 Distribution of CASP Targets Predictions

In this experiment, 42 CASP3 target proteins are extracted with their secondary structure predicted using the PHD (Rost and Sander, 199) program. It is not possible for this experiment to find predicted or observed CASP4 or CASP5 targets which are more recent and hence CASP3 was used to give an idea about an independent test set performance. According to Cuff and Barton (2000), the CASP3 data set was not included in the 480 proteins data set that had used in training and testing algorithms of this research work.

Figure 8.1 shows the distribution of the 42 CASP proteins predicted using the NN-GORV-II algorithm for all the secondary structure states. For the helices states, the histogram of Figure 8.1 shows that about 18 proteins (targets) are predicted at Q_H of above 95% and more than 5 proteins predicted at 85%, 75%, and 65% each.

Less than three proteins are predicted at 55% and about two proteins predicted at 45%, 35%, and 5%. The strands prediction accuracies (Q_E) are 8 proteins predicted at 95%, 6 proteins predicted at 85% and 5% each, and 7 proteins are predicted at 75%, and 65%. The rest of the proteins are predicted at 55% Q_E level and below. As for coils, Figure 8.1 shows that the about 15 proteins are predicted at level of 70-80% Q_C, about 13 proteins at level of 60-70%, and about 10 proteins at level of 80-90%. The rest three proteins are predicted at level 90-100%.

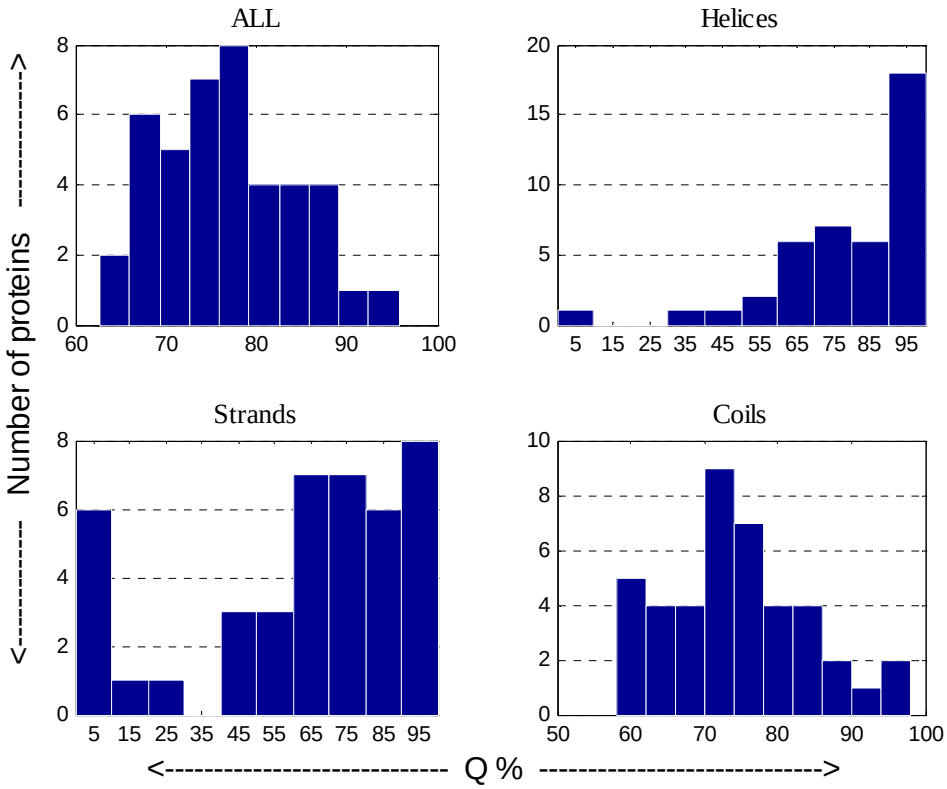


Figure 8.1: The distribution of prediction actuaries of the of the 42 CASP targets blind test for the secondary structure states.

The overall prediction accuracies Q_3 (ALL) for the 42 CASP targets are shown in Table 8.1. About 8 proteins are predicted at Q_3 accuracy between 60% and below 70%, about 20 proteins predicted at accuracy of 70-80%, about 12 proteins are predicted at Q_3 of 80-90%, and about two proteins predicted at accuracies above 90% and below 100%. It is clear that there is no protein predicted at accuracy below 60% of Q_3. These results are supported by the line graph of Figure 8.2 where each line indicates a secondary structure state travelling towards the 100% accuracy through the 42 CASP targets.

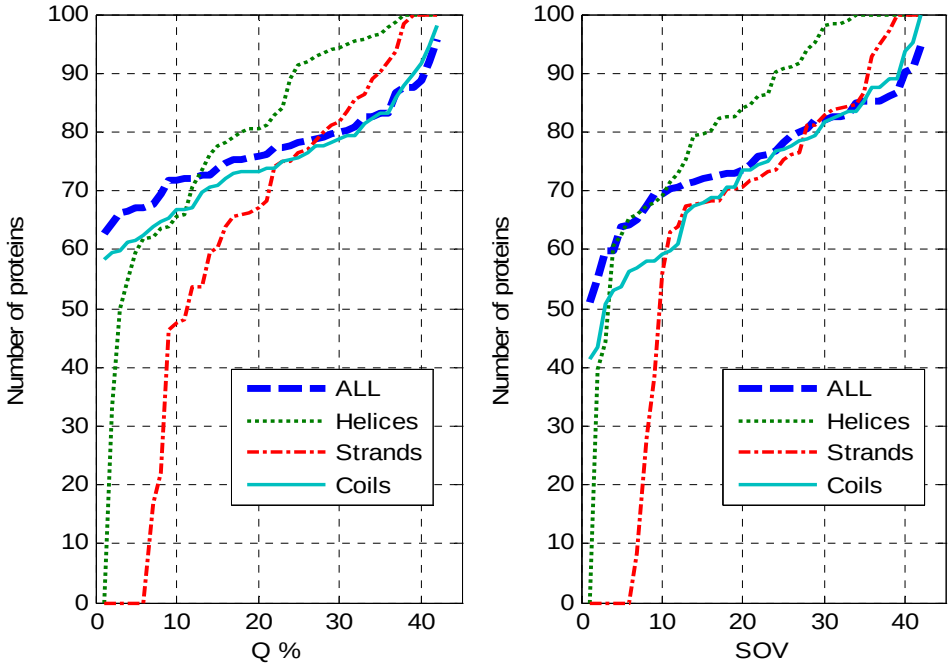


Figure 8.2: The performance of the 42 CASP targets with respect to Q_3 and SOV prediction measures

The figure elucidates that the helices (Q_H) and strands (Q_E) lines travelled from the zero prediction while coils (Q_C) and the overall performance (Q_3) travelled from below 60% and above 60%, respectively.

The histogram of Figure 8.1 and the line graph of Figure 8.2 show that the strands states are predicted by the NN-GORV-II in a more scattered distribution followed by the helices states while the overall prediction (ALL) was more homogenous and continuous followed by the coils states prediction. The results elucidated that the majority of protein are predicted at Q_3 accuracies between 70-80%.

The SOV measures for the helices, strands, coils, and all secondary structure states of the 42 CASP target proteins are shown in Figure 8.3. For the helices states NN-GORV-II method predicted about 3 proteins below the 65% SOV_H level, about 20 proteins are predicted between 60% and below 90%, and about 18 proteins are predicted above 90% SOV_H level.

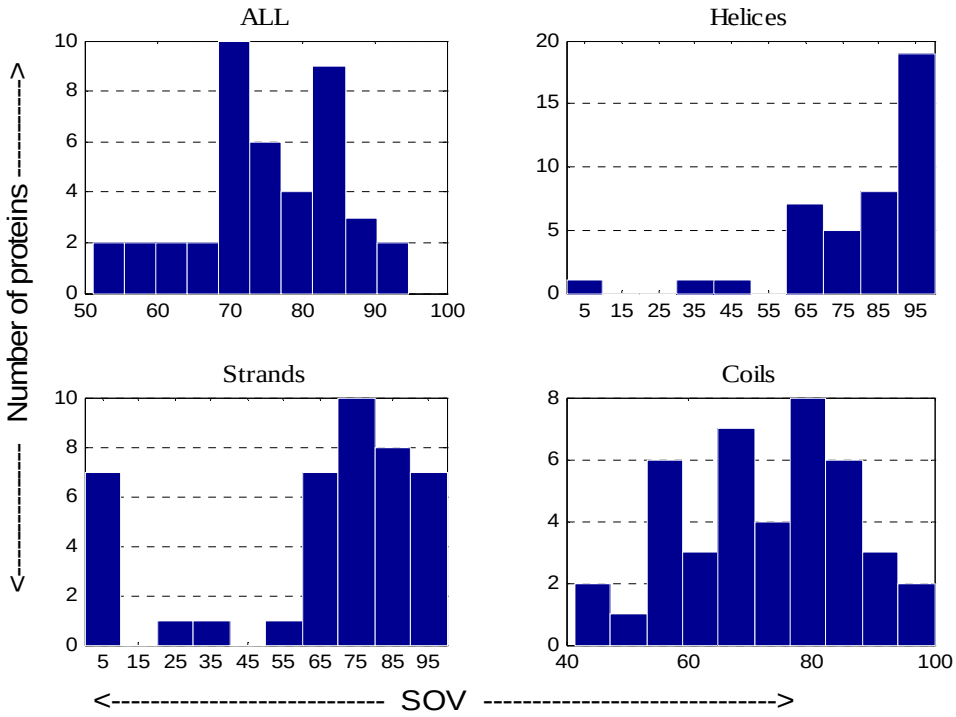


Figure 8.3: The distribution of SOV measure of the of the 42 CASP targets blind test for the secondary structure states.

The SOV_E measures of strands showed that 7 proteins are predicted at SOVE about 5%, in the range of above 5% and below 60% are only 3 proteins predicted while the rest of the 42 proteins predicted at level above 60% to 100% SOVE level (Figure 8.3).

As far as coils states are concerned, Figure 8.3 presents that the 42 proteins are distributed in a more homogenous manner. At SOV_C level of 80-90% about 13 proteins had been predicted, at level 60-80% about 19 proteins predicted while the remaining of the 42 proteins are predicted at SOV_C level of above 90% or below 60% but above 40%.

Figure 8.3 shows the estimations of overall SOV_3 (ALL). It reflects that about 8 proteins from the 42 are predicted at SOV_3 level of above 50% and below 70%.

About 24 proteins had been predicted at the level of above 70% and below 85% SOV_3 measure. The remaining 10 proteins had been predicted at level of 85-100%.

By reading the two figures (Figure 8.2 and Figure 8.3) together, it is clearly shown that the SOV prediction distribution of the 42 CASP proteins for the helices (SOV_H) and strands (SOV_E) states are more scattered than the distribution of the coils (SOV_C) and overall states (SOV_3). The line graph of SOV in Figure 8.2 illustrates that the lines of the three states travels through the 42 CASP target proteins towards the 100% SOV measure. It shows that helices and strands depart from 0.0% SOV prediction while the coils states SOV and overall SOV start above 40% and above 50%, respectively.

8.3 Performance and Quality of CASP Targets Predictions

Table 8.1 shows the performance of the NN-GORV-II method predicting the three secondary structures states: helices (Q_H), strands (Q_E), and coils (Q_C); and the overall accuracies (Q_3) of the 42 CASP targets. The observed secondary structure predictions of the 42 targets are referenced to the PHD predications of these target sequences. This independent test portrays a general view about the NN-GORV II algorithm predictions of data that has not been used in its training procedure.

Table 8.1: Percentages of prediction accuracies for the 42 CASP3 proteins targets

ID	Protein Name	Q_3	Q_H	Q_E	Q_C
T0042	NK-lysin from pig, 78a.a.	80.8	94.1	0.0	83.3
T0043	7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK)	66.5	62.0	81.0	66.7
T0044	RNA-3'terminal phosphate cyclase	72.0	94.9	66.1	64.9
T0045	HI1434	77.2	61.8	78.6	98.1
T0046	Gamma-Adaptin Ear Domain	79.0	33.3	92.0	75.0
T0047	Alpha(2u)-Globulin	87.7	100	98.5	75.3
T0048	Pterin-4-alpha-carbinolamine dehydratase, Pseudomonas aeruginosa	62.7	54.8	80.0	73.3
T0049	EstB, Pseudomonas marginata	71.7	79.2	46.2	73.9
T0050	Glutamate mutase component S - Clostridium cochlearium	69.3	95.6	47.6	64.0
T0051	Glutamate mutase component E - Clostridium cochlearium	74.7	84.1	53.8	70.7
T0052	Cyanovirin-N, Nostoc ellipsosporum	64.4	50.0	53.8	77.6
T0053	CbiK protein, S. typhimurium	72.7	80.5	65.6	61.4
T0054	VanX, Enterococcus faecium	75.7	76.3	48	82.2
T0055	lectin, Polyandrocarpa misakiensis	67.2	70.6	65.9	67.2
T0056	DnaB helicase N-terminal domain, E.coli	86.8	98.6	0.0	73.2
T0057	Glyceraldehyde 3-phosphate dehydrogenase, S. solfataricus	67.1	64.0	67.0	69.6
T0058	Uracil-DNA glycosylase, E.coli	79.9	95.7	59.6	78.8
T0059	Sm D3 protein (The N-terminal 75 residues)	82.7	100	85.4	79.4
T0060	D-dopachrome tautomerase, human	80.3	93.5	81.6	70.8
T0061	Protein HDEA, E. coli	66.3	78.3	16.7	59.5
T0062	Flavin reductase, E. coli	83.2	80.6	93.8	78.1
T0063	Translation initiation factor 5A, Pyrobaculum aerophilum	75.4	88.9	90.3	59.7
T0064	A SinR protein, Bacillus subtilis	77.5	92.7	0.0	79.5
T0065	B SinI protein, Bacillus subtilis	87.7	96.3	0.0	85.7
T0067	Phosphatidylethanolamine Binding Protein, Homo sapiens	75.9	100	68.4	77.5
T0068	Polygalacturonase, Erwinia carotovora subsp. carotovora	78.5	100	83.5	73.7
T0069	Recombinant conglutinin, bovine	78.8	91.3	77.1	72.0
T0070	Omp32 protein, Comamonas acidovorans	73.8	0.0	86.3	65.4
T0071	Alpha adaptin ear domain, rat	75.2	59.4	88.9	75.5
T0072	CD5 domain 1, human	78.2	63.6	60.5	91.8
T0074	The second EH domain of EPS15, human	88.8	97.7	100	81.5
T0075	Ets-1 protein (fragment), mouse	82.7	80.0	0.0	87.8
T0076	cdc4p, Schizosaccharomyces pombe	95.7	96.5	100	94.5
T0077	Ribosomal protein L30, Saccharomyces cerevisiae	76.2	94.3	74.2	61.5
T0078	Thioesterase, E. coli	67.7	82.8	76.4	58.2
T0079	MarA protein, E. coli	79.8	92.0	0.0	66.7
T0080	3-methyladenine DNA glycosylase, human	72.6	65.6	75.0	73.3
T0081	Methylglyoxal synthase, E. coli	71.7	73.4	64.0	73.0
T0082	Ribonuclease MC1, Momordica charantia (Bitter Gourd)	77.4	81.2	75.0	76.4
T0083	Cyanase, E.coli	83.3	77.5	100	90.0
T0084	RLZ, artificial construct	91.9	100	100	62.5
T0085	Cytochrome C554, Nitrosomonas europaea	72.0	65.8	22.2	83.3

Q_3 accuracy for amino acid
Q_H accuracy for α helices
Q_E accuracy for β strands
Q_C accuracy for coils

Table 8.2 also shows the SOV measures of the NN-GORV-II method predicting the three secondary structures states: helices (SOV_H), strands (SOV_E), and coils (SOV_C); and the overall accuracies (SOV_3) of the 42 CASP targets.

Table 8.2: Percentages of SOV measures for the 42 CASP3 proteins targets

ID	Protein Name	SOV_3	SOV_H	SOV_E	SOV_C
T0042	NK-lysin from pig, 78a.a.	71.6	73.0	0.0	100
T0043	7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK)	55.0	69.1	73.6	41.4
T0044	RNA-3'terminal phosphate cyclase	76.7	86.4	68.2	78.4
T0045	HI1434	82.9	80.3	81.0	87.5
T0046	Gamma-Adaptin Ear Domain	82.7	44.4	94.9	78.8
T0047	Alpha(2u)-Globulin	86.6	100	100	74.5
T0048	Pterin-4-alpha-carbinolamine dehydratase, Pseudomonas aeruginosa	70.6	65.4	73.3	81.7
T0049	EstB, Pseudomonas marginata	51.1	91.0	38.4	43.4
T0050	Glutamate mutase component S - Clostridium cochlearium	75.8	90.2	63.1	73.6
T0051	Glutamate mutase component E - Clostridium cochlearium	72.8	91.8	64.0	58.1
T0052	Cyanovirin-N, Nostoc ellipsosporum	69.4	71.4	68.3	68.9
T0053	CbiK protein, S. typhimurium	71.1	84.7	67.7	53.2
T0054	VanX, Enterococcus faecium	70.4	79.6	56.0	67.3
T0055	lectin, Polyandrocarpa misakiensis	59.8	82.4	75.4	50.7
T0056	DnaB helicase N-terminal domain, E.coli	79.5	98.4	0.0	61.0
T0057	Glyceraldehyde 3-phosphate dehydrogenase, S. solfataricus	72.2	66.0	70.6	79.4
T0058	Uracil-DNA glycosylase, E.coli	78.3	99.1	70.2	70.7
T0059	Sm D3 protein (The N-terminal 75 residues)	76.2	100	70.2	85.3
T0060	D-dopachrome tautomerase, human	90.9	100	92.8	83.6
T0061	Protein HDEA, E. coli	59.7	60.8	8.3	66.2
T0062	Flavin reductase, E. coli	90.3	86.2	97.1	89.0
T0063	Translation initiation factor 5A, Pyrobaculum aerophilum	72.5	100	84.6	59.2
T0064	A SinR protein, Bacillus subtilis	82.6	100	0.0	83.4
T0065	B SinI protein, Bacillus subtilis	85.2	100	0.0	77.1
T0067	Phosphatidylethanolamine Binding Protein, Homo sapiens	80.6	82.6	76.1	82.4
T0068	Polygalacturonase, Erwinia carotovora subsp. carotovora	74.4	39.6	81.0	70.7
T0069	Recombinant conglutinin, bovine	73.0	100	82.9	58.0
T0070	Omp32 protein, Comamonas acidovorans	64.1	0.0	84.0	53.8
T0071	Alpha adaptin ear domain, rat	82.0	67.4	84.3	89.1
T0072	CD5 domain 1, human	79.9	90.9	71.8	82.9
T0074	The second EH domain of EPS15, human	85.3	98.5	100	77.6
T0075	Ets-1 protein (fragment), mouse	85.0	79.4	0.0	95.1
T0076	cdc4p, Schizosaccharomyces pombe	94.7	100	100	87.5
T0077	Ribosomal protein L30, Saccharomyces cerevisiae	86.0	98.2	83.9	76.7
T0078	Thioesterase, E. coli	64.0	84.1	68.0	56.2
T0079	MarA protein, E. coli	85.3	100	0.0	68.8
T0080	3-methyladenine DNA glycosylase, human	65.0	94.1	67.5	59.8
T0081	Methylglyoxal synthase, E. coli	73.5	95.3	72.0	56.8
T0082	Ribonuclease MC1, Momordica charantia (Bitter Gourd)	67.4	62.3	76.4	68.1
T0083	Cyanase, E.coli	81.9	75.2	86.8	93.8
T0084	RLZ, artificial construct	69.5	68.0	100	75.0
T0085	Cytochrome C554, Nitrosomonas europaea	72.9	82.7	27.8	73.5

It is important to note that the SOV measure had been estimated by using the same observed and predicted data used in estimating performance accuracy (Q), and also the same program as discussed in the methodology chapter. Since the predicted secondary structures of the 42 targets of the PHD program are used here as observed structures, care should be taken when globally comparing the performances (Q_3) and qualities (SOV_3) of NN-GORV-II method with other prediction methods (Table 8.1 and Table 8.2).

Table 8.3 shows the mean performance (Q), the SOV measure, and the Mathew's Correlation Coefficients (MCC) of the NN-GORV-II method on the 42 CASP target sequences with the corresponding standard deviations. The values in the table confirmed what has been discussed previously in Chapter 6. Since they exhibit higher standard deviations, the strand states predictions have a higher variability and less homogeneity followed by the helices states. On the other hand the coils states exhibit less standard deviation and hence predicted in a continuous and homogenous pattern or distribution.

Table 8.3: The mean of Q_3 and SOV with and standard deviation, and Mathew's Correlation Coefficients (MCC) of CASP

Measure	ALL	H	E	C
Q	76.87 ± 7.52	79.69 ± 20.75	62.45 ± 31.10	74.58 ± 09.80
SOV	75.44 ± 9.75	81.87 ± 20.62	63.81 ± 31.03	72.33 ± 12.83
MCC	-	0.68	0.63	0.62

The performance of the NN-GORV-II method on the 42 CASP targets (Q_3) is 76.87% with a small standard deviation of 7.52% while the quality and usefulness (SOV_3) of the method reached 75.44% with relatively small standard deviation of 9.75%. The Mathew's Correlation Coefficients (MCC) is 0.68, 0.63, and 0.62 for helices, strands, and coils, respectively, indicating strong relationship between predicted and observed secondary structures states (Baldi *et al.*, 2000; Crooks *et al.*, 2004).

These results aim to give a general idea about the NN-GORV-II method performance on an independent test set and not accurate measures since the observed secondary structures are not produced with X-ray spectroscopy or NMR laboratory techniques.

8.4 Summary

This chapter assesses the performance and quality of the prediction of the NN-GORV-II algorithm by using an independent test set of protein data that has not been used in training the algorithm. CASP3 protein targets had been used for this purpose. The result of the test gives a good idea about the prediction performance and quality of the NN-GORV-II method despite the limitation of the data set.

The observed secondary structures states of these target sequences are determined by the PHD method and not laboratory methods; so a straightforward comparison with other methods might not be an accurate comparison. The NN-GORV-II method performance accuracy (Q_3) in predicting protein secondary structure is 76.9% and the quality of prediction (SOV_3) is 75.4%. These results are far better than what was reported by (Ouali and King, 2000) who used only 23 CASP3 targets instead of 42 CASP3 targets used in this test. The results are also in a comparative range with what reported by Kim and Park (2003) in their SVM predictor that used CASP5 targets.

CHAPTER 9

RECEIVER OPERATING CHARACTERISTIC (ROC) TEST

9.1 Introduction

Many researchers argue that dichotomous (binary) classification is convenient and powerful for decision making, while it may introduces distortions (Fielding and Bell 1997; Hand, 1997). In particular, the use of threshold-independent Receiver Operating Characteristic (ROC) curves has received considerable attention in recent years.

The Receiver Operating Characteristics (ROC) graphs are useful techniques for assessing the performance of classifiers. The ROC curves are well known in Biology and Medical decision making and they are well used in dichotomous classification. They have been increasingly adopted as a tool for analysing and visualizing many aspects of machine learning algorithms or methods. The ROC curve is a plot the true positive rate against the false positive rate for different possible cut points of a diagnostic test.

The ROC curve illustrates the trade-off between sensitivity and specificity in the sense that any increase in sensitivity will be accompanied by a decrease in specificity. It also shows that the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test while the closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. The area under the curve (AUC) is a measure of the algorithm accuracy.

Kloczkowski *et al.* (2002) argued that, regularly, proteins contain about 30% helical structure (H), about 20% strands (E), and about 50% coil (C) structure. This means that even the most trivial prediction algorithm which assigns all residues to the coil (C) state would give approximately 50% correct prediction. This chapter attempts to test the results of the prediction or classification task of the NN-GORV-II method discussed in this work while opening up a discussion about the reliability of ROC curve analysis in predicting coils only states in a multi-class classifier. The eight-to-three secondary structure reduction Method V discussed in the previous chapter showed that coils states composed 0.48 of the whole data set (Table 7.1). Several researchers in the protein secondary structure prediction reported similar ratio. Baldi *et al.* (2000) reported coil only random guess of 0.4765 while others argued that 50% accuracy of an algorithm is not better than a random guess in protein secondary structure prediction.

9.2 Binary Classes and Multiple Classes

For the problem of secondary structure prediction, if we have an amino acid sequence of length n , the secondary structures corresponding to these sequences are the three states helix, strand, and coils which can be considered as $d_i=d_1$, d_2, d_n.. The SOV measure mentioned before takes care of these assignment to give maximal score even though the prediction is not identical to the assigned segment.

In the case of the dichotomy problem of two alternative classes, that is if we would like to predict only one structural class, for instance: a coil versus non-coil, then, the d_i is in general equal to 0 or 1 which is a binomial model of 0.5 probability for a coil or non-coil state. In the case where d_i has a value between 0 and 1 revealing the uncertainty of our knowledge of the correct assignment at the corresponding position as our case in this work where we have three classes, the analysis for this multiple class case is very similar.

The problem of prediction accuracy is strongly related to the frequency of occurrence of each class. For instance, in protein secondary structure prediction the

non-helix class covers roughly 70% of the cases in natural proteins, while only 30% belong to the helix class. Thus a constant prediction of non-helix is bound to be correct 70% of the time, although it is highly non-informative and useless (Baldi *et al.*, 2000).

If we assume that the output of our prediction algorithm is $G=g_1, g_2, .. g_n$, of course g_i here is a probability between 0 and 1 showing the degree of confidence in the prediction. However, when both D and G are binary, their comparison can be entirely summarized by four numbers:

TP = the number of times d_i is coil, g_i is coil (true positive).
TN = the number of times d_i is non-coil, g_i is non coil (true negative).
FP = the number of times d_i is non-coil, g_i is coil (false positive).
FN = the number of times d_i is coil, g_i is non-coil (false negative).

Then

Sensitivity (True positive rate) = TP/(TP+FN)
Specificity (True negative rate) = FP/(FP+TN)

and *N* is the total sample size which defined as:
$N= TP + TN + FP + FN.$

When both D and G or one of them is not binary, then of course the situation is more complex and four numbers are not enough to summarize the situation. When G is not binary, binary predictions can still be obtained by using cut-off thresholds. The numbers TP, TN, FP, and FN will then vary with the threshold choice. These numbers are often arranged into a 2 x 2 contingency or confusion matrix as shown in Table 9.1.

Table 9.1: The contingency table or confusion matrix for coil states prediction

		Predicted	
Observed		C	$\overline{C}$
	C	TP	FN
	$\overline{C}$	FP	TN

C Coil

$\overline{C}$ Not Coil

The ROC curve does not provide a rule for the classification of cases. However, there are strategies that may be used to develop decision rules. Two elements are required to identify the appropriate threshold; the first is the relative cost of FP and FN errors while the second is the prevalence of positive cases. Assigning values to these costs are complex and subjective and dependent upon the context within which the classification rule will be used (Zweig and Campbell, 1993).

As discussed earlier, the numbers TP, TN, FP and FN depend on how the threshold is selected. In most cases, there is a trade-off between the amount of false positives and the amount of false negatives produced by the algorithm or the classifier. The Receiver operating characteristics (ROC) summarizes such results by displaying for threshold values within a certain range or *hit rate*; the sensitivity, against the false positive rate or *false alarm rate*. In a typical ROC curve the hit rate increases with the false alarm rate. It is also common to display the sensitivity versus the specificity in a similar curve or separately as a function of threshold in two different curves.

As illustrated in the methodology and shown in this chapter, the sensitivity can be defined as the probability of correctly predicting a positive example and the specificity is the probability that a positive prediction is correct. In biology and medical statistics, the word specificity is sometimes used in a different sense (Burset and Guigo, 1996) which is beyond our discussion in this research.

The sensitivity and specificity of a test depends also on what constitutes a not normal test. Figure 9.1 illustrates an idealized graph showing the number of normal and not normal observations arranged according to the value of a test. This distributions overlap does not distinguish normal from not normal with 100% accuracy. The area of overlap indicates where the test cannot distinguish normal from not normal. In practice, a cut-point (cut score) is chosen; above which the test will be considered as abnormal and below which the test will be considered as normal. The position of the cut point will determine the number of true positive, true negatives, false positives and false negatives. Different cut points may be chosen if we wish to minimize one of the errors types of the test results.

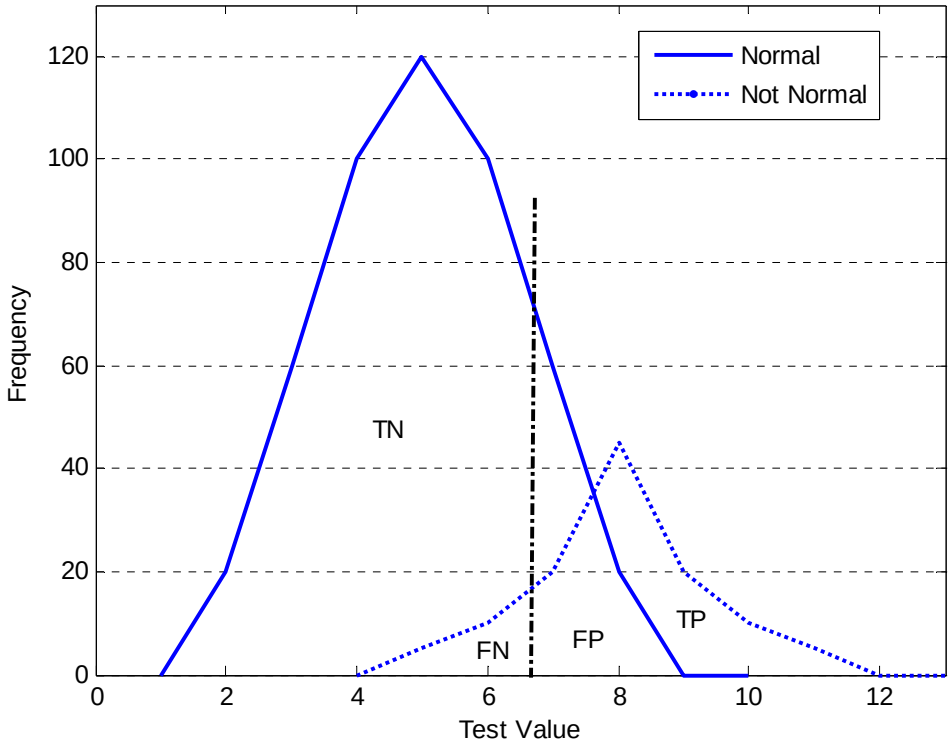


Figure 9.1: An idealized curve showing the (TP, TN, FP, and FN) numbers of a hypothetical normal and Not normal observations

Some researchers argued that even with four numbers alone, it is not immediately clear how a given prediction method fares. This is why a lot of the comparison methods aim at constructing a single number measuring the distance between D and G. But it must be clear from the outset, that information is always lost

in such a process, even in the binary case, i.e. when going from the four numbers above to a single one. In general, several different vectors (TP, TN, FP, and FN) will result in the same distance (Crooks and Brenner, 2004; Baldi *et al.*, 2000).

9.3 Assessment of NN-GORV-II

Table 9.2 shows nine cut scores of 10772 secondary structures outputs sample predicted by the NN-GORV-II algorithm with Method V reduction method. The true positive (TP) row represents the situation that coils states predicted by NN-GORV-II algorithm as coils (i.e. the number of times d_i is coil, g_i is coil) while the false positive (FP) represents the situation that not coils states predicted by NN-GORV-II algorithm as coils (i.e the number of times d_i is non-coil, g_i is coil).

Table 9.2: The cut scores for the NN-GORV-II algorithm considering coil only prediction

Cut Score	**C**	$\overline{C}$	**Sum**
1	544	33	577
2	625	45	670
3	929	139	1068
4	1244	185	1429
5	2588	1187	3775
6	710	415	1125
7	912	814	1726
8	18	14	32
9	56	314	370
10	0	0	0
Total	7626	3146	**10772**

As discussed in Chapter 6, the total number of residues in the database used in training and testing the algorithms and hence the number of predicted secondary structures is 83392 (Table 6.1). The test sample used in this experiment was chosen from 10772 secondary structure predicted states for its appropriate cut scores and convenience in calculations and representation (Table 9.2).

Figure 9.2 represents a curve resemble the idealized curve of Figure 9.1 where the cut scores were plot against the numbers of observations. The numbers of observation in this case represent the numbers of the true positives and the numbers of the false positives. Figure 9.2 there are nine cut scores plotting the two curves, but big number of selected cut scores will make the two curves look smoother. However, from this graph a very huge number of cut scores can be observed where the TP and FP change accordingly.

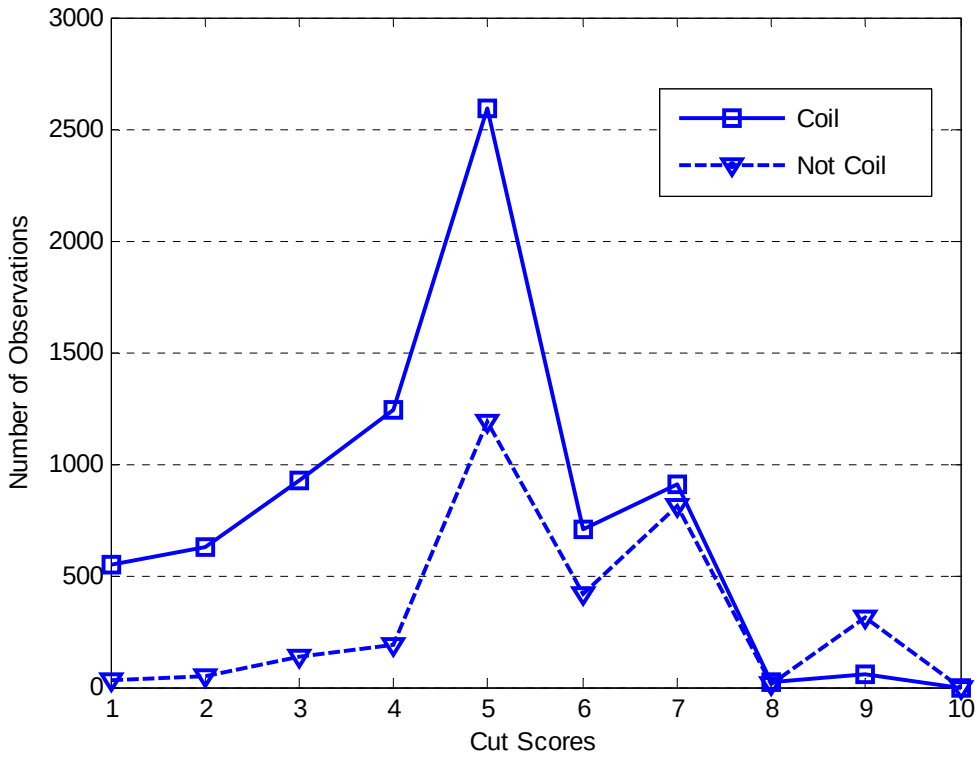


Figure 9.2: The cut scores of the coils and not coils secondary structure states predicted by the NN-GORV-II algorithm using Method V reduction scheme.

According to their respective cut scores, the true positive rate (TPR) which is the sensitivity of the test and the false positive rate which is (1- specificity) of the test are shown in Table 9.3. It shows the respective area for each cut score. The summation of the nine scores areas represents the area under the curve (AUC). This area under the curve measures the prediction accuracy. The AUC of this test as shown in the table is 0.7151 with standard error (SE) of 0.0057 as calculated from the nine cut scores.

Table 9.3: The cut scores, true positive rate (TPR), false positive rate (FPR), and area under ROC (AUC) for the NN-GORV-II prediction algorithm considering coil state only prediction

Cut Score	TPR	FPR	Area
1	1.0000	1.0000	0.0710
2	0.9895	0.9287	0.0805
3	0.9752	0.8467	0.1161
4	0.9310	0.7249	0.1471
5	0.8722	0.5618	0.2320
6	0.4949	0.2224	0.0399
7	0.3630	0.1293	0.0279
8	0.1043	0.0097	0.0002
9	0.0998	0.0073	0.0004
10	0.0000	0.0000	0.0000
AUC	-	-	**0.7151**
SE	-	-	**0.0057**

Figure 9.3 shows the ROC illustrates that the ROC curve travels above the diagonal line and below the top left corner of the graph indicting that the area of this curve is above null guess 0.5 and below the perfect prediction 1.0. The computed AUC as shown in the figure and described in Table 9.2 is 0.72 and the standard error

is 0.0057. This proves that the NN-GORV-II algorithm is able to discriminate the coils states from non coils with 72% prediction accuracy with a very minor experimental or standard error. Although there is a loss in the entropy in this procedure due to the 0.48 probability of the coils sates in the database instead of 0.5, this result is in-line with what has been reported by Kaur and Raghava (2003). This result also has a comparative agreement with the correlation coefficients of the NN-GORV-II method shown in Table 6.4.

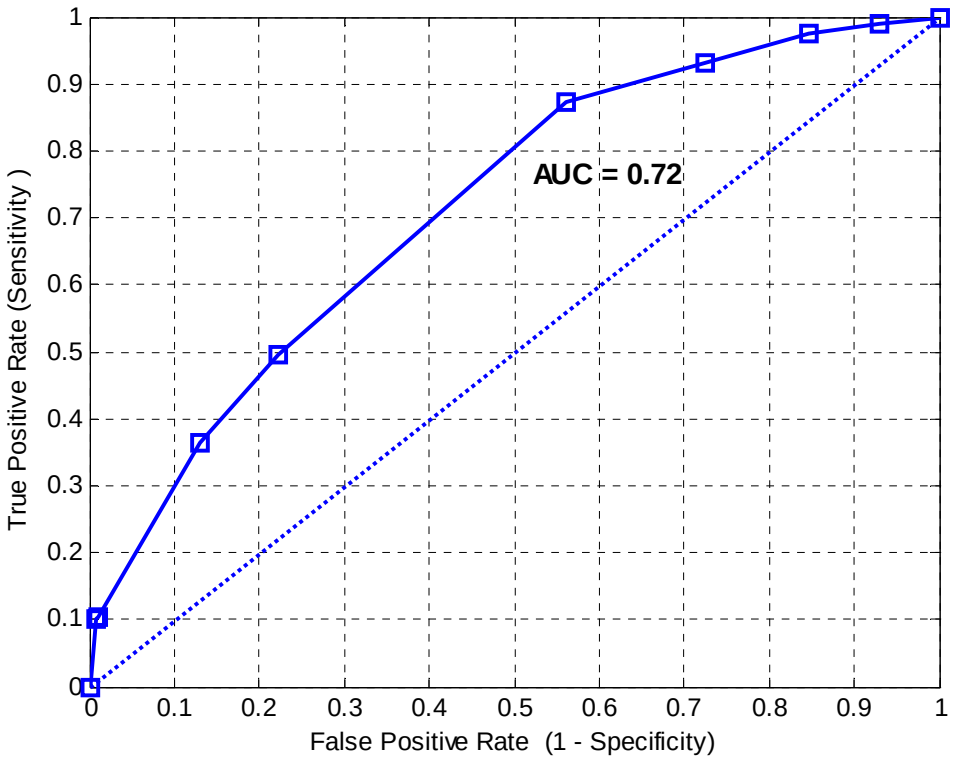


Figure 9.3: The area under ROC (AUC) for the NN-GORV-II prediction algorithm considering coil only prediction.

In this research, the adoption of the receiver operating characteristics (ROC) analysis aims to determine the discriminative ability of the NN-GORV-II algorithm to distinguish the coil states only since they constitute about 0.5 of the data. This test might be controversial since it is conducted on a three-class classifier and not a binary classifier. The nature of the data set that constitutes the three classes of secondary structure made the data set divided into two classes for the coil states that constitute half of the data set. The ROC analysis test arrived at a conclusion that the

NN-GOR-V-II algorithm was able to distinguish between two classes (coils/not coils) at 72% of the times.

9.4 Summary

The protein secondary structure coils states are further classified using the receiver operating characteristics ROC curve and analysis. The trade-off between the true positive rate (sensitivity) and the false positive rate was plotted in an ROC curve and the area under the curve (AUC) was estimated and found that the NN-GORV-II algorithm was able to correctly classify 72% of the coils states. Although this accuracy is less than the accuracy discussed in the previous chapter, this number can give an estimate for the NN-GORV-II algorithm.

The accuracy of ROC analysis should be less than the accuracy obtained by the SOV measure since there is loss in the entropy of the TP, FP, TN, and FN values as discussed. In addition, describing the data set as coils and not coils in its discrete binary meaning had not been accurately satisfied in this case.

CHAPTER 10

CONCLUSION

10.1 Introduction

Since the observations of the early researchers in the field of protein structure, it is concluded that the 3D structure of a protein is extremely related to its primary sequences of amino acids (Epstein *et al*., 1963; Anfisen, 1973). This observation made it possible to predict protein structure from sequences with considerably high accuracy. In the absence of a known 3D or a homologue of a certain protein, the secondary structure prediction of protein plays a great role in extracting the utmost possible information from the primary sequences. Large sequencing projects that generate an increasing number of amino acids sequences, made laboratory techniques like X-ray crystallography and NMR unfeasible to observe the secondary structures of such sequences. The demand for feasible and reliable structure prediction method becomes inevitable.

This chapter concludes the review of literature, methodology, experimental work, analysis, and the discussion of this research work. The output and results of the newly developed method of protein secondary structure prediction together with other methods studied in this research are concluded and summarised in this chapter. This chapter also presents the findings and the contributions of this research.

10.2 Summary of the Research

The research work of this project focuses on the protein folding dilemma that asks a vital question; how a protein folds from its primary sequence into its 3D structure? Predicting proteins 3D structures from amino acids directly is a very hard task. In molecular biology it is fairly easy to predict 3D structure of a protein from its secondary structure as explained in the text of this report. The problem of the protein secondary structure prediction from its amino acid sequences has been investigated in this work.

The research reviews the work done by other researchers and the literature cited in the area of amino acids sequences, proteins, and sequence homology and alignments. The types of protein structure as well as the laboratory methods of detecting and determining protein structures are reviewed.

The research also describes the artificial neural networks and the Information Theory which formed the basis of the new prediction method developed in this research work. Feed forward neural networks that are mainly used in the area of protein secondary structure prediction, the networks training and optimizations are fairly examined. The information theory that uses the statistics and the probabilities foundations with special reference to GOR theory is discussed.

The framework used in developing and implementing the new prediction method to achieve a better prediction accuracy protein secondary structure from its primary sequence is described and elucidated. The benchmark data set that is used in the experiments of this research is presented and discussed as well as the hardware and software utilized to implement the prediction methods.

The methods, algorithms, and modelling used to develop and implement the new prediction method, NN-GORV-I, and its advanced version NN-GORV-II are explained in detail. All the methods studied in this work are trained and tested on the same multiple sequence alignments data sets which allow a valid and reliable comparison of the performance of the seven methods studied in this research. The

multiple sequence alignment and the profile generation procedures to collect maximum possible biological information to be presented to the neural networks are clearly explained. Five reduction schemes that converted the DSSP eight classes to the conventional three secondary structure classes (helices, strands, and coils) are implemented in this research. The seven prediction methods developed or studied in this work are presented and discussed. The assessment of the performance and quality of the investigated methods is accomplished by several methods ranging from the accuracy per protein (Q_3), segment overlap measure (SOV_3), Matthews Correlation Coefficients (MCC), and the Receiver Operating Characteristic (ROC) procedure.

The results of the prediction methods together with the two newly developed methods are investigated and analysed in this research. The performances of GOR-IV and neural network (NN-I) method without utilizing multiple sequence alignment are shown to show the importance of including biological information in the prediction process. The newly developed methods NN-GORV-I and NN-GORV-II outperform all the investigated methods in terms of accuracy, quality, and reliability.

The effect of the five reduction methods on the NN-GORV-II performance and quality is discussed. The ANOVA procedure attests that the five reduction methods are significantly different in their predictions accuracies. The results show that it is advisable to use Method I or Method II rather than Method V in globally assessing the accuracy of a new prediction algorithm or method.

Chapter VIII explores the performance of a blind or an independent data set test on the NN-GORV-II method. CASP3 protein targets are predicted by the newly developed method. The output of NN-GORV-II method is then compared to the PHD algorithm prediction for the same targets. The performance of NN-GORV-II algorithm is found high and stable compared to other methods. The same conclusion applies for the SOV measure and the Mathew's Correlation Coefficients (MCC).

Observing the results of Chapter 9, Method V reduces the eight secondary structure states into almost 50% coils and 50% helices and strands structures. The

Receiver Operating Characteristics (ROC) is intelligently introduced to the multi-class classifier to assess it as a binary classifier. The ROC curve and the area under ROC curve (AUC) proved that the NN-GORV-II effectively and correctly classified 72% of the coil states.

10.3 Conclusions

The conclusions of this research may be listed and summarised in the followings remarks:

The accuracy of protein secondary structure has been significantly increased by the new methods NN-GORV-I and NN-GORV-II that are designed and developed in this research. NN-GORV-II method achieved 80.5 % prediction accuracy which is a very high accuracy in this domain.

The newly developed NN-GOR-V-II protein secondary structure prediction method achieves 5.46% additional accuracy over the one of the best prediction methods (PROF) in this domain. This is a significant improvement in the prediction accuracy.

The statistical bases of GOR-V information theory and the power of the neural networks are combined together to yield a new method of protein secondary structure prediction which is superior to both methods.

The effective and procedural implementation and generation of multiple sequence alignments enables the GOR-V and the neural network to fully utilize the evolutionary information of similar sequences in the searched repository sequence data bases which made the newly developed NN-GORV-II a high performing and high quality classifier.

The test of performance Q_3 and the test of quality and usefulness (SOV) conducted in this research proved that the NN-GORV-II method is of high accuracy

and good quality and more useful. The high values of Mathew's Correlation Coefficients (MCC) analysis conducted in this research provides strong evidence that the high accuracy and quality results obtained from NN-GORV-II method are reliable and consistent.

The newly developed method proved that it is highly stable and consistent when tested against the different DSSP secondary structure reduction methods conducted in this research. The output accuracies of NN-GORV-II according to each reduction methods also are also shown high accuracies compared to other existing methods.

The NN-GORV-II method proved additional high performance and high quality when the blind test is used for the method. An independent data from the CASP dataset is used for this test.

The NN-GORV-II method proved that it is capable of correctly and efficiently predict coils from non coils 72% of the times. The ROC curve has been intelligently introduced and implemented here to partially assess a multi-class prediction method (NN-GORV-II) by observing the composition of the secondary structure states in the data base.

The ROC curve has been intelligently introduced and implemented to partially assess a multi-class prediction method (NN-GORV-II) by observing the composition of the secondary structure states in the data base.

The new method for predicting protein secondary structure from the amino acid sequences developed by combing neural networks and GOR-V and hence named NN-GORV-I and further enhanced and improved to NN-GORV-II, provided evidence from the several tests conducted that the method is highly accurate, highly reliable, and robust.

10.4 Contributions of the Research

- This research proposes two new methods for predicting protein secondary structures from amino acid sequences. The proposed methods are then designed, developed, and implemented and proved highly accurate and robust.

- This research introduces and implemented several assessment or evaluation procedures to measure the success of the new methods. It has been proven that the newly developed methods (NN-GORV-I and NN-GORV-II) are highly accurate and reliable. The test also proved that the newly developed methods are highly consistent.

- The ROC test has been introduced as a novel procedure to test the ability of NN-GORV-II method to discriminate between two classes (coils/not-coils). This novel approach considers a multi-class classifier as a binary classifier or predictor. This new approach can be adopted to assess newly prediction methods developed in this domain in instances where the dataset consists 50% coils in its composition.

10.5 Recommendations for Further Work

Inspired from the work presented in this project, the recommendations of the author of this report for further work in the domain of protein secondary structure prediction are shown in the following points:

- A larger database for training and testing can be used instead of the 480 proteins used in this research. That is possible due to the collaborative sequencing projects in Bioinformatics where many proteins are added to the databases every time. This will allow the NN-GORV-II method to utilize more biological knowledge and evolutionary information in sequence data.

- Fine tuning the parameters of the neural network with better implementations of the different and optimized neural networks algorithms will enhance the prediction accuracy of NN-GORV-II method.

- NN-GORV-II exploits the biological information found in neighbouring residues and homologues sequences. A procedure for extracting biological information from the protein-protein interactions processes will add significantly extra reliable and biological information to the prediction process.

- The novel approach of using the ROC curve and the AUC to partially assess the multi-class prediction algorithm can further be validated and adopted to represent a powerful assessment tool when the data set consists 50% coils.

- The DSSP eight-to-three secondary structure states reduction methods together with other secondary structure assignments like DEFINE and STRIDE can be standardized and given unique names for each method. This will facilitate and standardize the comparison between prediction algorithms with more accuracy and minimum error.

- Similar methods of prediction and classification in domains rather than Bioinformatics can successfully utilize variety of techniques and tools used in this research.

- Since the research in Bioinformatics field in general and the protein secondary structure prediction domain in particular is increasing rapidly, the need for a “utility and statistical package for Bioinformatics” that successfully arranges data for input and helps in the analysis and assessment of the output becomes crucial. This will save considerable time for the research in Bioinformatics.

10.6 Summary

This chapter concludes and summarizes the research work discussed in this project. The chapter also presents and highlights the contributions and findings of this research. Recommendations for further work and future research directions in the domain of this work are also coined and proposed in this chapter.

REFERENCES

Abagyan, R., Frishman, D. and Argos, P. (1994). Recognition Of Distantly Related Proteins Through Energy Calculations. *Proteins: Structure, Function, and Genetics, Supplement.* 19: 132-140.

Agresti, A. (2002). *Categorical Data Analysis*. 2nd ed. New York, USA: Wiley and Sons.

Alexey, G. M. (1999). Structure Classification-Based Assessment Of CASP3 Predictions For The Fold Recognition Targets. *Proteins: Structure, Function, and Genetics, Supplement*. 3 (1): 88-103.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997). Gapped U BLAST and PSI-BLAST: A New Generation Of Protein Database Search Programs. *Nucleic Acids Research*. 25: 3899-3402.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*. 215:403-410.

Anderson, T. W. (2003). *An Introduction To Multivariate Statistical Analysis*. 3rd ed. N.Y., USA: Wiley and Sons.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. and Murzin, A. G. (2004). SCOP Database in: Refinements Integrate Structure and Sequence Family Data. *Nucleic Acids Research*. 32:226-229.

Anfinsen, C. B. (1973). Principles That Govern The Folding Of Protein Chains. *Science*. 181: 223-230.

Apostolico, A. and Giancarlo, R. (1998). Sequence Alignment in Molecular Biology. *Journal of Computational Biology*. 5: 173-196.

Attwood, T. K., Beck, M. E., Bleasby, A. J., Degtyarenko, K., Michie A. D. and Parrysmith, D. J. (1997). Novel Developments With The PRINTS Protein Fingerprint Database. *Nucleic Acids Research*. 25: 212-216.

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L, Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. and Zygouri, C. (2003). PRINTS and Its Automatic Supplement, Preprints. *Nucleic Acids Research*. 31: 400-402.

Aurora, R., Srinivasan, R. and Rose, G. D. (1994). Rules For Alpha-Helix Termination By Glycine. *Science.* 264:1126-1130.

Bairoch, A. and Apweller, R. (1997). The SWISS-PROT Protein Sequence Data Bank and Its Supplement TREMBL. *Nucleic Acids Research*. 25: 31-36.

Bairoch, A. and Boeckmann, B. (1991). The SWISS-PROT Protein-Sequence Data Bank. *Nucleic Acids Research*. 19: 2247-2248.

Bairoch, A., Bucher, P. and Hofmann, K. (1997). The PROSITE Database, Its Status in 1997. *Nucleic Acids Research.* 25: 217-221.

Baldi, P., Brunak, S., Frasconi, P., Pollastri, G. and Soda, G. (1999). Exploiting The Past and The Future in Protein Secondary Structure Prediction. *Bioinformatics*. 15(11): 937–946.

Baldi, P., Brunak, S., Frasconi, P., Pollastri, G. and Soda, G. (2001). *Bidirectional Dynamics For Protein Secondary Structure Prediction, Sequence Learning, Paradigms, Algorithms, and Applications.* 80-104. Springer-Verlag.

Baldi, P., Chauvin, Y., Hunkapillar, T. and Mcclure, M. (1994). Hidden Markov Models Of Biological Primary Sequence Information. *Proceedings of the National Academic of Science*. 91: 1059-1063.

Baldi, P. (1995). Gradient Descent Learning Algorithms Overview: A General Dynamical Systems Perspective. *IEEE Transactions On Neural Networks.* 6(1): 182-195.

Baldi, P. and Brunak, S. (2002). *Bioinformatics: The Machine Learning Approach*. MIT Press.

Baldi, P., Brunak, S., Chauvin, Y., andersen, C. A. F. and Nielsen, H. (2000). Assessing The Accuracy Of Prediction Algorithms For Classification: An Overview. *Bioinformatics*. 16: 412-424.

Barton, G. J. and Sternberg, M. J. E. (1987). A Strategy For The Rapid Multiple Alignment Of Protein Sequences: Confidence Levels From Tertiary Structure Comparisons. *Journal of Molecular Biology*. 198:327-337.

Barton, G. J. (1990). Protein Multiple Sequence Alignment and Flexible Pattern

Matching. *Method Enzymol*. 183: 403-428.

Barton, G. J. (1993). Alscript: A Tool To Format Multiple Sequence Alignments. *Protein Engineering*. 6:37-40.

Bates, P. A. and Sternberg M. J. E. (1999). Model Building By Comparison At CASP3: Using Expert Knowledge and Computer Automation. *Proteins: Structure, Function, and Genetic Supplement*. 3 (1): 47-54.

Benner, S. A. and Gerloff, D. (1991). Patterns Of Divergence in Homologous Proteins As Indicators Of Secondary and Tertiary Structure A Prediction Of The Structure Of The Catalytic Domain Of Protein-Kinases. *Advance in Enzyme Regulation*. 31: 121-181.

Benner, S. A., Badcoe, I., Cohen, M. A. and Gerloff, D. L. (1994). Bona-Fide Prediction Of Aspects Of Protein Conformation Assigning Interior and Surface Residues From Patterns Of Variation and Conservation in Homologous Protein Sequences. *Journal of Molecular Biology*. 235: 926-958.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Ravichandran, V., Schneider, B., Thanki, N., Padilla, D., Weissig, H., Westbrook, J. D. and Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallography*. 58 (6): 899–907.

Bishop, C. (1996). *Neural Networks For Pattern Recognition.* Oxford University Press.

Blundell, T., Sibanda, B. L. and Pearl, L. (1983). Three-Dimensional Structure, Specificity and Catalytic Mechanism Of Renin. *Nature*. 304: 273-275.

Boberg, J., Salakoski, T. and Vihinen, M. (1995). Accurate Prediction Of Protein Secondary Structural Class With Fuzzy Structural Vectors. *Protein Engineering*. 8: 505-512.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B. and Petersen, S. B. (1990). A Novel Approach To Prediction Of The 3-Dimensional Structures Of Protein Backbones By Neural Networks. *FEBS. Letters*. 261: 43-46.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M., Lautrup, B., Norskov, L., Olsen, O. H. and Petersen, S. B. (1988). Protein Secondary Structure and Homology By

Neural Networks. The Alpha-Helices in Rhodopsin. *FEBS Letters*. 241(1-2): 223-228.

Boscott, P. E., Barton, G. J. and Richards, W. G. (1993). Secondary Structure Prediction For Modelling By Homology. *Protein Engineering*. 6:261-266.

Bowie, J. U., Clarke, N. D., Pabo, C. O. and Sauer, R. T. (1990). Identification Of Protein Folds Matching Hydrophobicity Patterns Of Sequence Sets With Solvent Accessibility Patterns Of Known Structures. *Proteins: Structure, Function, and Genetics, Supplement*. 7: 257-264.

Bowie, J. U., Luthy, R. and Eisenberg, D. (1991). A Method To Identify Protein Sequences That Fold Into A Known 3-Dimensional Structure. *Science*. 253: 164-170.

Bradley, A. P. (1997). The Use Of The Area Under The ROC Curve in The Evaluation Of Machine Learning Algorithms. *Pattern Recognition*. 30 (7): 1145-1159.

Branden, Candtooze, J. (1991). *Introduction To Protein Structure*. Garland Publishing, Inc: New York.

Brenner, S. E. (1996). Molecular Propinquity: Evolutionary and Structural Relationships Of Proteins. University Of Cambridge: PhD Thesis.

Brian, H. (1998). Computing Science: The Invention Of The Genetic Code. *American Scientist.* 86 (1): 9-14.

Briffeuil, P., Baudoux, G., Lambert, C., De Bolle, X., Vinals, C., Feytmans, E. and Depiereux, E. (1998). Comparative Analysis Of Seven Multipl Protein Sequence Alignment Servers: Clues To Enhance Reliability Of Predictions. *Bioinformatics.* 14 (4): 357-66.

Brocchieri, L. and Karlin, S. (1998). A Symmetric-Iterated Multiple Alignment Of Protein Sequences. *Journal of Molecular Biology*. 276(1): 249-64.

Bryant, S. H. and Altschul, S. F. (1995). Statistics Of Sequence-Structure Threading. *Current Opinion in Structural Biology*. 5: 236-244.

Bult, C. J., White, O., Olsen, G. J., Zhou, L. X., Fleischmann, R. D., Sutton, G. G., Blake, J. A., Fitzgerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L.,

Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H. P., Fraser, C. M., Smith, H. O., Woese, C. R. and Venter, J. C. (1996). Complete Genome Sequence Of The Methanogenic Archaeon, Methanococcus Jannaschii. *Science*. 273: 1058-1073.

Burkhard, R. (1999). Twilight Zone Of Protein Sequence Alignments. *Protein Engineering*. 12(2): 85-94.

Burset, M. and Guigo, R. (1996). Evaluation Of Gene Structure Prediction Programs. *Genomics*. 34: 353-367.

Bystroff, C. and Baker D. (1997). Blind Predictions Of Local Protein Structure in Casp2 Targets Using The I-Sites Library. *Proteins: Structure, Function and Genetics Supplement*. 1: 167-171.

Carrington, M. and Boothroyd, J. (1996). Implications Of Conserved Structural Motifs in Disparate Trypanosome Surface Proteins. *Molecular and Biochemical Parasitology*. 81: 119-126.

Chandonia, J. M. and Karplus, M. (1999). New Methods For Accurate Prediction Of Protein Secondary Structure. *Proteins: Structure, Function and Genetics.* 35: 293-306.

Chen, C. P. and Rost, B. (2002). State-Of-The-Art in Membrane Protein Prediction. *Appl. Bioinformatics*. 1: 21-35.

Chothia, C. (1992). Proteins: One Thousand Families For The Molecular Biologist. *Nature*. 357: 543-544.

Chothia, C. and Janin, J. (1975). Principles Of Protein-Protein Recognition. *Nature*. 256: 705-708.

Chothia, C. and Lesk, A. M. (1986). The Relation Between The Divergence Of Sequence and Structure in Proteins. *EMBO Journal*. 5: 823-826.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D. R., Tulip, W. R., Colman, P. M., Alzri, P. M. and Poljak, R. J. (1989). Conformations Of Immunoglobulin Hypervariable Regions. *Nature*. 342: 877-883.

Chou, K. C. and Zhang, C. T. (1994). Predicting Protein-Folding Types By Distance Functions That Make Allowances For Amino-Acid Interactions. *Journal of Biological Chemistry*. 269: 22014-22020.

Chou, K. C. and Zhang, C. T. (1995). Prediction Of Protein Structural Classes. *Critical Reviews in Biochemistry and Molecular Biology*. 30: 275-349.

Chou, P. Y. and Fasman, G. D. (1974b). Prediction Of Protein Conformation. *Biochemistry.* 13: 222-245.

Chou, P. Y. and Fasman, G. D. (1974a). Conformational Parameters For Amino Acids in Helical, Sheet and Random Coil Regions From Proteins. *Biochemistry*. 13: 211.

Chou, P. Y. (1989). Prediction Of Protein Structural Classes From Amino Acid Composition: in: Fasman, G. D. ed. *Prediction Of Protein Structures and The Principles Of Protein Conformation*. Plenum Press. 549-586.

Cline, M. S., Karplus, K., Lathrop, R. H., Smith, T. F., Rogers, R. G., Jr. and Haussler, D. (2002). Information-Theoretic Dissection Of Pairwise Contact Potentials, *Proteins: Structure, Function, and Genetics, Supplement.* 49(1): 7-14.

Crick, F. (1989). The Recent Excitement About Neural Networks. *Nature*. 337: 129-132.

Crooks, G. E. and Brenner, S. E. (2004). Protein Secondary Structure: Entropy, Correlations and Prediction. *Bioinformatics*. 20:1603–1611.

Crooks, G. E., Jason, W. and Steven, E. B. (2004). Measurements Of Protein Sequence Structure Correlations. *Proteins: Structure, Function, and Bioinformatics*. 57:804–810.

Cuff, J. A. and Barton, G. J. (1999). Evaluation and Improvement Of Multiple Sequence Methods For Protein Secondary Structure Prediction. *Proteins: Structure, Function and Genetics*. 34: 508-519.

Cuff, J. A. and Barton G. J. (2000). Application Of Multiple Sequence Alignment Profiles To Improve Protein Secondary Structure Prediction. *Proteins: Structure, Function and Genetics*. 40: 502-511.

Daniel, F., Christian, B., Kevin, B., Arne, E., Adam, G., David, J., Kevin, K., Lawrence, A., Kelley, Robert, M., Krzysztof, P., Burkhard, R., Leszek, R. and Michael, S. (1999). CAFASP-1: Critical Assessment Of Fully Automated Structure Prediction Methods. *Proteins: Structure, Function, and Genetics, Supplement.* 3(1): 209-217.

Defay, T. R. and Cohen, F. E. (1996). Multiple Sequence Information For Threading

Algorithms. *Journal of Molecular Biology*. 262: 314-323.

Depiereux, E., Badoux, G., Briffeuil, P., Reginster, I., De Bolle, X., Vinals, C. and Feytmans, E. (1997). Match-Box-Server: A Multiple Sequence Alignment Tool Placing Emphasis On Reliability. *CABIOS*. 13(3): 249-256.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory Of Pattern Recognition*. NY: Springer-Verlag.

Dill, K. A. (1990). Dominanat Forces in Protein Folding. *Biochemistry*. 29: 7133–7155.

Dill, K. A., Bromberg, S., Yue, K. Z., Fiebig, K. M., Yee, D. P., Thomas, P. D. and Chan, H. S. (1995). Principles Of Protein-Folding A Perspective From Simple Exact Models. *Protein Science*. 4: 561-602.

Donnelly, D., Overington, J. P. and Blundell, T. L. (1994). The Prediction and Orientation Of Alpha-Helices From Sequence Alignments The Combined Use Of Environment-Dependent Substitution Tables, Fourier-Transform Methods and Helix Capping Rules. *Protein Eng*ineering. 7: 645-653.

Doolittle, R. F. (1981). Similar Amino-Acid Sequences Chance Or Common Ancestry. *Science*. 214: 149-159.

Dunbrack, R. L. (1999). Comparative Modelling Of CASP3 Targets Using PSI-BLAST Snd SCWRL. *Proteins: Structure, Function, and Genetics, Supplement*. 3(1): 81-7.

Dunbrack, R. L., Gerloff, D. L., Bower, M., Chen, X. W., Lichtarge, O. and Cohen, F. E. (1997). *Meeting Review: The Second Meeting On The Critical Assessment Of Techniques For Protein Structure Prediction (CASP2)*. Asilomar, California.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (2002). *Biological Sequence Analysis: Probabilistic Models Of Proteins and Nucleic Acids*. U.K.: Cambridge University Press.

Eddy, S. R. (1996). Hidden Markov Models. *Current Opinion in Structural Biology*. 6(3): 361-365.

Eddy, S. R. (1998). Profile Hidden Markov Models. *Bioinformatics*. 14(9): 755-63.

Eddy, S.R., Mitchison G. and Durbin R. (1995). Maximum Discrimination Hidden Markov Models Of Sequence Consensus. *Journal of Computational Biology*. 2: 9-23.

Egan, J. P. (1975). Signal Detection Theory and ROC Analysis. *Series in Cognitition and Perceptron*. New York: Academic Press.

Eisenhaber, F., Frommel, C. and Argos, P. (1996). Prediction Of Secondary Structural Content Of Proteins From Their Amino-Acid-Composition Alone: The Paradox With Secondary Structural Class. *Proteins: Structure, Function, and Genetics, Supplement.* 25: 169-179.

Elmasry, N. F. and Fersht, A. R. (1994). Mutational Analysis Of The N-Capping Box Of The Alpha-Helix Of Chymotrypsin Inhibitor-2. *Protein Engineering*. 7: 777-782.

Eyrich, V. A., Przybylski, D., Koh, I. Y. Y., Grana, O., Pazos, F., Valencia, A. and Rost, B. (2003). CAFASP3 in The Spotlight Of EVA. *Proteins: Structure, Function, and Genetics, Supplement*. 53 (6): 548–560.

Farago, A. and Lugosi, G. (1993). Strong Universal Consistency Of Neural Network Classifiers, *IEEE Transactions On Information Theory*. 39: 1146-1151.

Feng, D. F., Johnson, M. S, and Doolittle, R. F. (1985). Aligning Amino Acid Sequences: Comparison Of Commonly Used Methods. *Journal of Molecular*. 21: 112-125.

Feraud, R. and Clerot, R. (2002). A Methodology To Explain Neural Network Classification. *Neural Networks*. 15(2): 237-46.

Ferran, E. A., Pflugfelder, B. and Ferrara, P. (1994). Self-Organized Neural Maps Of Human Protein Sequences. *Protein Science*. 3: 507-521.

Fersht, A. R. (1984). Basis of Biological Specificity. *Trends in Biochemical Science*. 9: 145-147.

Fersht, A. R. (1987). The Hydrogen-Bond in Molecular Recognition. *Trends in Biochemical Science*. 12: 301-304.

Fielding, A. H. and Bell, J. F. (1997). A Review Of Methods For The Assessment Of Prediction Errors in Conservation Presence/Absence Models. *Environmental Conservation*. 24: 38–49.

Fischer, D. and Eisenberg, D. (1996). Protein Fold Recognition Using Sequence-Derived Predictions. *Protein Science*. 5: 947-955.

Fiser, A., Simon, I. and Barton, G. J. (1996). Conservation Of Amino-Acids in Multiple Alignments: Aspartic Acid Has Unexpected Conservation. *FEBS Letters*. 397: 225-229.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., Mckenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., Mcdonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. and Venter, J. C. (1995). Whole-Genome Random Sequencing and Assembly Of Haemophilus Influenzae Rd. *Science*. 269: 496-512.

Flockner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M. and Sippl, M. J. (1995). Progress in Fold Recognition. *Proteins: Structure, Function, and Genetics, Supplement.* 23: 376-386.

Francesco, V. D., Garnier, J. and Munson, P. J. (1997). Protein Topology Recognition From Secondary Structure Sequences: Application Of The Hidden Markov Models To The Alpha Class Proteins. *Journal of Molecular Biology*. 267(2): 446-463.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J. F., Dougherty, B. A., Bott, K. F., Hu, P. C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A. and Venter, J. C. (1995). The Minimal Gene Complement Of Mycoplasma Genitalium. *Science*. 270: 397-403.

Freedman, R. B. (1995). The Formation Of Protein Disulfide Bonds. *Current Opinion in Structural Biology*. 5: 85-91.

Frishman, D. and Argos, P. (1995). Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Structure, Function, and Genetics, Supplement*. 23:566-579.

Frishman, D. and Argos, P. (1997). Seventy-Five Percent Accuracy in Protein Secondary Structure Prediction. *Proteins: Structure, Function, and Genetics, Supplement.* 27: 329-335.

Frishman, D. and Argos, P. (1996). Incorporation Of Non-Local Interactions in Protein Secondary Structure Prediction From The Amino-Acid Sequence. *Protein Engineering.* 9: 133-142.

Garnier, J. and Robson, B. (1989). The GOR Method For Predicting Secondary Structures in Proteins. in: Fasman GD, ed. *Prediction Of Protein Structure and The Principles Of Protein Conformation.* New York: Plenum Press. 417-465.

Garnier, J. Gibrat, J. and Robson, B. (1996). GOR Method For Predicting Protein Secondary Structure From Amino Acid Sequence. *Method Enzyme*. 266: 540-553.

Garnier, J., Osguthorpe, D. J. and Robson, B. (1978). Analysis Of The Accuracy and Implications Of Simple Methods For Predicting The Secondary Structure Of Globular Proteins. *Journal of Molecular Biology*. 120: 97-120.

Gibrat, J. F., Garnier, J. and Robson, B. (1987). Further Developments Of Protein Secondary Structure Prediction Using Information-Theory - New Parameters and Consideration Of Residue Pairs. *Journal of Molecular Biology*. 198: 425-443.

Gilbrat, J., Madej, T. and Bryant, S. (1996). Surprising Similarities in Structure Comparison. *Current Opinion in Structural Biology.* 6: 377-85.

Gobel, U., Sander, C., Schneider, R. and Valencia, A. (1994). Correlated Mutations and Residue Contacts in Proteins. *Proteins: Structure, Function, and Genetics, Supplement.* 18: 309-317.

Gotoh, O. (1996). Significant Improvement in Accuracy Of Multiple Protein Sequence Alignments By Iterative Refinement As Assessed By Reference To Structural Alignments. *Journal of Molecular Biology*. 264(4): 823-38.

Gotoh, O. (1999). Multiple Sequence Alignment: Algorithms and Applications. *Advances in Biophysics*. 36(1): 159-206.

Greer, J. (1981). Comparative Model-Building Of The Mamlian Serine Proteases. *Journal of Molecular Biology*. 153: 1027-1042.

Gribskov, M., Luthy, R. and Eisenberg, D. (1990). Profile Analysis. *Method Enzymol*. 183: 146-159.

Gribskov, M., Mclachlan, A. D. and Eisenberg, D. (1987). Profile Analysis Detection Of Distantly Related Proteins. *Proceedings of the National Academic of*

Science. USA. 84:4355-4358.

Grundy, W. N., Bailey, W., Elkan, T. and Baker, C. (1997). Meta-MEME: Motif-Based Hidden Markov Models Of Protein Families. *CABIOS*. 13(4): 397-406.

Gur, D., Rockette, H., and Armfield, D. (2003) Prevalence Effect in A Laboratory Environment. *Radiology*. 228: 10-14.

Han, K. F. and Baker, D. (1995). Recurring Local Sequence Motifs in Proteins. *Journal of Molecular Biology*. 251: 176-187.

Han, K. F. and Baker, D. (1996). Global Properties Of The Mapping Between Local Amino-Acid Sequence and Local-Structure in Proteins. *Proceedings of the National Academic of Science*. USA. 93: 5814-5818.

Hand, D. J. (1997). *Construction and Assignment Of Classification Rules*. NY: John Wiley and Sons.

Hand, D. J. and Till, R. J. (2001). A Simple Generalisation Of The Area Under The ROC Curve For Multiple Class Classification Problems. *Machine Learning*. 45: 171-186.

Hanke, J., Beckmann, G., Bork, P. and Reich, J. G. (1996). Self-Organizing Hierarchical Networks For Pattern-Recognition in Protein-Sequence. *Protein Science*. 5: 72-82.

Hanley, J. A. and Mcneil, B. J. (1983). The Meaning and Use of The Area Under The Receiver Operating Characteristic (ROC) Curve. *Radiology*. 148: 839-43.

Hartl, F. U. (1996). Molecular Chaperones in Cellular Protein-Folding. *Nature***.** 381: 571-580.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Upper Saddle River, NJ.

Heilig, R., Eckenberg, R., Petit J. L., Fonknechten ,N, Da Silva C., Cattolico L., Levy M., Barbe, V., De Berardinis, V., Ureta-Vidal, A., Pelletie,R E., Vico, V., Anthouard, V., Rowen, L., Madan, A., Qin, S., Sun, H., Du, H., Pepin, K., Artiguenave, F, Robert, C, Cruaud, C, Bruls, T., Jaillon, O., Friedlander, L., Samson, G., Brottier, P., Cure, S., Segurens, B., Aniere, F., Samain, S., Crespeau, H., Abbasi, N., Aiach, N., Boscus, D., Dickhoff, R., Dors, M., Dubois, I., Friedman, C., Gouyvenoux, M., James, R., Madan, A., Mairey-Estrada, B., Mangenot, S., Martins, N., Menard, M., Oztas, S., Ratcliffe, A.,

Shaffer, T., Trask, B., Vacherie, B., Bellemere, C., Belser, C., Besnard-Gonnet, M., Bartol-Mavel, D., Boutard, M., Briez-Silla, S., Combette, S., Dufosse-Laurent, V., Ferron, C., Lechaplais, C., Louesse, C., Muselet, D., Magdelenat, G., Pateau, E., Petit, E., Sirvain-Trukniewicz, P., Trybou, A., Vega-Czarny, N., Bataille, E., Bluet, E., Bordelais, I., Dubois, M., Dumont, C., Guerin, T., Haffray, S., Hammadi, R., Muanga, J., Pellouin, V., Robert, D., Wunderle, E., Gauguet, G., Roy, A., Sainte-Marthe, L., Verdier, J., Verdier-Discala, C., Hillier, L., Fulton, L., Mcpherson, J., Matsuda, F., Wilson, R., Scarpelli, C., Gyapay, G., Wincker, P., Saurin, W., Quetier, F., Waterston, R., Hood, L. and Weissenbach, J. (2003). The DNA Sequence and Analysis Of Human Chromosome 14. *Nature*. 421(6923): 601-607.

Henikoff, S and Henikoff, J.G. (1992). Amino Acid Substitution Matrices From Protein Blocks. *Proceedings of the National Academic of Science*. USA 89: 10915-10919.

Henikoff, S., Henikoff, J. G., Alford, W. J. and Pietrokovski, S. (1995). Automated Construction and Graphical Presentation Of Protein Blocks From Unaligned Sequences. *Gene.* 163(2): 17-26.

Henikoff, S. and Henikoff, J. G. (1994). Protein Family Classification Based On Searching A Database Of Blocks. *Genomics*. 19: 97-107.

Henikoff, S. and Henikoff, J. G. (1997). Embedding Strategies For Effective Use Of Information From Multiple Sequence Alignments. *Protein Science*. 6: 698-705.

Henikoff, S. (1996). Scores For Sequence Searches and Alignments. *Current Opinion in Structural Biology*. 6: 353-360.

Higgins, D. G., Thompson, J. D. and Gibson, T. J. (1996). Using CLUSTAL For Multiple Sequence Alignments. *Methods Enzymol.* 266: 383-402.

Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992). Selection Of A Represetative Set Of Structures From The Brookhaven Protein Data Bank. *Protein Science* 1: 409-417.

Holley, L. H. and Karplus, M. (1989). Protein Secondary Structure Prediction With A Neural Network. *Proceedings of the National Academic of Science*. USA. 86(1): 152-6.

Holm, L. and Sander, C. (1993). Protein-Structure Comparison By Alignment Of

Distance Matrices. *Journal of Molecular Biology*. 233**:** 123-138.

Hornik, K., Stinchocombe, M., White, H. and Auer, P. (1994). Degree Of Approximation Results For Feedforward Networks Approximating Unknown Mappings and Their Derivatives. *Neural Computation*. 6(6): 1262-1275.

Hornik, K., Stinchocombe, M. and White, H. (1990). Universal Approximation Of Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks. *Neural Networks*. 3: 535-549.

Huang, X. (1994). On Global Sequence Alignment. *CABIOS*. 10(3): 227-35.

Hubbard, T., Murzin, A., Brenner, S. and Chothia, C. (1997). SCOP: A Structural Classification of Proteins Database. *Nucleic Acids Research*. 25(1): 236-9.

Hubbard, T. J. and Park, J. (1995). Fold Recognition and Abinitio Structure Predictions Using Hidden Markov-Models and Beta-Strand Pair Potentials. *Structure, Function, and Genetics, Supplement.* 23: 398-402.

Hubbard, T. J. P. (1997). New Horizons in Sequence Analysis. *Current Opinion in Structural Biology*. 7: 190-193.

Hudak J. and Mcclure M. A. (1999). A Comparative Analysis Of Computational Motif-Detection Methods. *In Pacific Symposium On Biocomputing*. 138-49.

Hutchinson, E. G. and Thornton, J. M. (1994). A Revised Set Of Potentials For Beta-Turn Formation in Proteins. *Protein Science*. 3: 2207-2216.

Islam, S. A., Luo, J. C. and Sternberg, M. J. E. (1995). Identification and Analysis Of Domains in Proteins. *Protein Engineering*. 8: 513-525.

Jacob, F. (1977). Evolution and Tinkering. *Science*. 196: 1161-1166.

Jimenez, M. A., Munoz, V., Rico, M. and Serrano, L. (1994). Helix Stop and Start Signals in Peptides and Proteins The Capping Box Does Not Necessarily Prevent Helix Elongation. *Journal of Molecular Biology*. 242: 487-496.

Johnson, M. S., Sali, A. and Blundell, T. L. (1990). Phylogenetic-Relationships From 3-Dimensional Protein Structures. *Method Enzymol*. 183: 670-690.

Jones, D. T. (1999b). Genthreader: An Efficient and Reliable Protein Fold Recognition Method For Genomic Sequences. *Journal of Molecular Biology*. 287(4): 797-815.

Jones, D. T. and Thornton, J. M. (1996). Potential-Energy Functions For Threading. *Current Opinion in Structural Biology*. 6: 210-216.

Jones, D. T. (1999a). Protein Secondary Structure Prediction Based On Position-

Specific Scoring Matrices. *Journal of Molecular Biology*. 292: 195-202.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). A New Approach To Protein Fold Recognition. *Nature*. 358: 86-89.

Jones, D.T. and Swindells, M. B. (2002). Getting The Most From PSI-BLAST. *Trends in Biochemistry Science*. 27: 161-164.

Julie, D. T., Frederick, P. and Oliver, P. (1999). Balibase: A Benchmark Alignment Database For The Evaluation of Multiple Alignment Programs. *Bioinformatics*. 15 (1): 87-88.

Julie, D., Thompson, Desmond, G., Higgins, T. and Gibson, J. (1994). CLUSTAL W: Improving The Sensitivity Of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties, and Weight Matrix Choice. *Nucleic Acids Research*. 2(22): 4673-4680.

Julie, D., Thompson. F. P. and Oliver, P. (1999). A Comprehensive Comparison Of Multiple Sequence Alignment Programs. *Nucleic Acids Research*. 27(13): 2682-90.

Kabsch, W. and Sander, C. (1983). A Dictionary Of Protein Secondary Structure: Pattern Recognition Of Hydrogen-Bonded and Geometrical Features. *Biopolymers*. 22: 2577-2637.

Kabsch, W. and Sander, C. (1984). On The Use Of Sequence Homologies To Predict Protein-Structure: Identical Pentapeptides Can Have Completely Different Conformations. *Proceedings of the National Academic of Science. USA*. 81: 1075-1078.

Kaur, H. and Raghava, G. (2003). A Neural-Network Based Method For Prediction Of Beta-Turns in Proteins From Multiple Sequence Alignment. *Protein Science*. 12: 923-929.

Kendrew, J. C. Dickerson RE, Strandberg BE, Hart RG, and Davies D.R. (1960). Structure Of Myoglobin. *Nature*. 185: 422-427.

Kevin, K., Christian, B. and Richard, H. (1998). Hidden Markov Models For Detecting Remote Protein Homologies. *Bioinformatics*. 14(10): 846-856.

Kevin, K., Christian, B., Melissa, C., Mark, D., Leslie, G. and Richard, H. (1999). Predicting Protein Structure Using Only Sequence Information. *Proteins: Structure, Function, and Genetics, Supplement*. 3(1): 121-125.

Kevin, K., Kimmen, S., Christian, B., Melissa, C., David, H., Richard, H., Liisa, H.

and Chris, S. (1997). Predicting Protein Structure Using Hidden Markov Models. *Proteins: Structure, Function, and Genetics, Supplement.* 1:134-139.

Kim, H. and Park, H. (2003). Protein Secondary Structure Prediction Based On An Improved Support Vector Machines Approach. *Protein Engineering.* 16(8): 553-60.

King, R. D. and. Sternberg, M. J. E. (1996). Identification and Application Of The Concepts Important For Accurate and Reliable Protein Secondary Structure Prediction. *Protein Science*. 5: 2298-2310.

Klein, P. and Delisi, C. (1986). Prediction of Protein Structural Classes From Amino Acids Sequence. *Biopolymers.* 25: 1659–1672

Kloczkowski, A., Ting, K. L., Jernigan, R. L. and Garnier, J. (2002). Combining The GOR V Algorithm With Evolutionary Information For Protein Secondary Structure Prediction From Amino Aid Sequence. *Proteins: Structure, Function, and Genetics, Supplement*. 49: 154-166

Koretke, K. K., Russell, R. B., Copley, R. R. and Lupas, A. N. (1999). Fold Recognition Using Sequence and Secondary Structure Information. *Proteins: Structure, Function, and Genetics, Supplement*. 3(1): 141-8.

Krigbaum, W. R. and Knutton, S. P. (1973). Prediction of The Amount Of Secondary Structure in A Globular Protein From Its Amino acid Composition. *Proceedings of the National Academic of Science. USA. 70*(10): 2809-2813.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994). Hidden Markov-Models in Computational Biology Applications To Protein Modelling. *Journal of Molecular Biology*. 235: 1501-1531.

Kullback, S., Keegel, J. C. and Kullback, J. H. (1987). *Topics in Statistical Information Theory*. Berlin; New York: Springer-Verlag.

Kulp, D., Haussler, D., Reese, M. G. and Eeckman, F. (1996). A Generalized Hidden Markov Model For The Recognition Of Human Genes in DNA. *Proceedings of the 4th Intelligent Systems for Molecular Biology*.134-142.

Ladunga, I. and Smith, R. F. (1997). Amino Acid Substitutions Preserve Protein Folding By Conserving Steric and Hydrophobicity Properties. *Protein Engineering*. 10: 187-196.

Lathrop, R. H. and Smith, T. F. (1996). Global Optimum Protein Threading With

Gapped Alignment and Empirical Pair Score Functions. *Journal of Molecular Biology*. 255: 641-665.

Lathrop, R. H. (1994). The Protein Threading Problem With Sequence Amino-Acid Interaction Preferences Is NP-Complete. *Protein Engineering*. 7: 1059-1068.

Lattman, E. E. (1995). Protein-Structure Prediction: A Special Issue. *Protein: Structure, Function, and Genetics, Supplement.* 23: 1.

Lawrence, M. C. and Colman.,P.M. (1993). Shape Complementarity at Protein/Protein Interfaces. *Journal. Molecular Biology*. 234: 946–950.

Lesk, A. M., Lo Cont., L. and Hubbard, T. J. P. (2001). Assessment Of Novel Folds Targets in CASP4: Predictions Of Three-Dimensional Structures. Secondary Structures and Inter-Residue Contacts. *Structure, Function, and Genetics, Supplement.* 45(S5): 98–118.

Levitt, M. and Chothia, C. (1976). Structural Patterns in Globular Proteins. *Nature.* 261: 552-557.

Lichtarge, O., Bourne, H. R. and Cohen, F .E. (1996). An Evolutionary Trace Method Defines Binding V Surfaces Common to Protein Families. *Journal of Molecular Biology*. 257: 342-358.

Liisa, H. and Chris, S. (1996). Mapping The Protein Universe. *Science*. 273(5275): 595-603.

Lijmer, J., Mol, B., Heisterkamp, S. (1999). Empirical Evidence Of Design-Related Bias in Studies Of Diagnostic Tests. *Journal of the American Medical Association*. 282: 1061-1066.

Lim, V. I. (1974a). Structural Principles Of The Globular Organisation Of Protein Chains. A Stereochemical Theory Of Globular Protein Secondary Structure. *Journal of Molecular Biology*. 88: 857-872.

Lim, V. I. (1974b). Algorithms For The Prediction Of Alpha-Helical and Beta-Structural Regions in Globular Proteins. *Journal of Molecular Biology*. 88: 873-894.

Lipman, D. J., Altschul, S. F. and Kececioglu, J. D. (1989). A Tool For Multiple Sequence Alignment. *Proceedings of the National Academic of Science.* April USA. 86: 4412-4415.

Lisboa, P. G. J.(Ed) (1992). *Neural Networks: Current Applications*. London: Chapman Hall.

Lise, S. and Jones, D. T. (2005). Sequence Patterns Associated With Disordered Regions in Proteins. *Proteins: Structure, Function, and Bioinformatics*. 58: 144-150.

Maclin, R. and Shavlik, J. W. (1994). Incorporating Advice Into Agents That Learn From Reinforcements. *In Proceedings Of The 12th National Conference On Artificial Intelligence.*

Madej, T., Gibrat, J. F. and Bryant, S. H. (1995). Threading A Database Of Protein Cores. *Structure, Function, and Genetics, Supplement.* 23: 356-369.

Marcella, A., Mcclure, Tatha, K., Vasi and Walter, M. (1994). Comparative Analysis Of Multiple Protein Sequence Alignment Methods. *Molecular Biology and Evolution*. 11(4): 571-592.

Marcella, M., Chris, S. and Pete, E. (1996). Parameterization Studies For The SAM and HMMER Methods Of Hidden Markov Model Generation. *Proceedings of 4th International Conference on Intelligent Systems for Molecular Biology*. 155-164.

Marchler-Bauer, A. and Bryant, S. H. (1997). A Measure Of Success in Fold Recognition. *Trends in Biochemistry Science.* 22: 236-240.

Mark, G. and Michael, L. (1998). Comprehensive Assessment Of Automatic Structural Alignment Against A Manual Standard. The SCOP Classification Of Proteins. *Protein Science*. 7: 445-456.

Marzban, C (2004). A Comment On The ROC Curve and The Area Under It As Performance Measures. *Http://Www.Nhn.Ou.Edu/˜Marzban*.

Matsuo, Y. and Nishikawa, K. (1995). Assessment Of A Protein Fold Recognition Method That Takes Into Account 4 Physicochemical Properties Side-Chain Packing, Solvation, Hydrogen-Bonding, and Local Conformation. *Structure, Function, and Genetics, Supplement.* 23: 370-375.

Matthews, B. B. (1975). Comparison Of The Predicted and Observed Secondary Structure Of T4 Phage Lysozyme. *Biochima et Biophysica Acta.* 405(2): 442-451.

May, A. C. W. and Johnson, M. S. (1994). Protein-Structure Comparisons Using A Combination Of A Genetic Algorithm, Dynamic-Programming and Least-Squares Minimization. *Protein Engineering.* 7: 475-485.

May, A. C. W. and Johnson, M. S. (1995). Improved Genetic Algorithm-Based

Protein-Structure Comparisons Pairwise and Multiple Superpositions. *Protein Engineering.* 8: 873-882.

May, A. C. W. (1996). Pairwise Iterative Superposition Of Distantly Related Proteins and Assessment Of The Significance Of 3-D Structural Similarity. *Protein Engineering.* 9: 1093-1101.

Mcgregor, M. J., Flores, T. P. and Sternberg, M. J. (1989). Prediction Of Beta-Turns in Proteins Using Neural Networks. *Protein Engineering.* 2(7): 521-6.

Metfessel, B. A., Saurugger, P. N., Connelly, D. P. and Rich, S. S. (1993). Cross-Validation Of Protein Structural Class Prediction Using Statistical Clustering and Neural Networks. *Protein Science.* 2: 1171-1182.

Michael, L. (1997). Competitive Assessment Of Protein Fold Recognition and Alignment Accuracy. *Proteins: Structure, Function, and Genetics, Supplement.* 1(1): 92-104.

Michie, A. D., Orengo, C. A. and Thornton, J. M. (1996). Analysis Of Domain Structural Class Using An Automated Class Assignment Protocol. *Journal of Molecular Biology*. 262: 168-185.

Michie, D., Spiegelhalter, D. J. and Taylo, C. C. (1994). *Machine Learning, Neural and Statistical Classification.* Ellis: Horwood.

Mintseris, J. and Weng, Z. (2004). Optimizing Protein Representations With Information Theory. *Genome Informatics.* 15(1): 160-169.

Morgenstern, B., K. Frech, A. Dress and Werner, T. (1998). DIALIGN: Finding Local Similarities By Multiple Sequence Alignment. *Bioinformatics*. 14: 290-294.

Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K., Pedersen, J. T. (1997). Critical Assessment Of Methods Of Protein Structure Prediction (CASP): Round II. *Proteins: Structure, Function, and Genetics, Supplement.* 1(29): 2-6 and113-136.

Moult, J., Hubbard, T., Fidelis, K. and Pedersen, J. (1999). Critical Assessment Of Methods Of Protein Structure Prediction (CASP): Round II. *Proteins: Structure, Function, and Genetics, Supplement.* 3(1): 2-6.

Murzin, A., Brenner S. E., Hubbard, T. and Chothia, C. (1995). SCOP: A Structural Classification Of Proteins Database and The Investigation Of Sequences and Structures. *Journal of Molecular Biology.* 247: 536-540.

Muskal, S. M. and Kim, S. H. (1992). Predicting Protein Secondary Structure-Content A Tandem Neural Network Approach. *Journal of Molecular Biology.* 225: 713-727.

Muskal, S. M., Holbrook, S. R. and Kim, S. H. (1990). Prediction Of The Disulfide-Bonding State Of Cysteine in Proteins. *Protein Engineering. 3*(8): 667-72.

Naderi-Manesh, H., Sadeghi, M., Sharhriar, A., Moosavi and Movahedi, A. A. (2001). Prediction Of Protein Surface Accessibility With Information Theory. *Proteins: Structure, Function, and Genetics, Supplement.* 42: 452–459.

Nagano, K. (1973). Logical Analysis Of The Mechanism Of Protein Folding. *Journal of Molecular Biology*. 75: 401-420.

Nakai, K., Kidera, A. and Kanehisa, M. (1988). Cluster-Analysis Of Amino-Acid Indexes For Prediction Of Protein-Structure and Function. *Protein Engineering*. 2: 93-100.

Nakashima, H., Nishikawa, K. and Ooi, T. (1986). The Folding Type Of A Protein Is Relevant To The Amino-Acid Composition. *Journal of Biochemistry.* 99: 153-162.

Needleman, S. B. and Wunsch, C. D. (1970). A General Method Applicable To The Search For Similarities in The Amino Acid Sequence Of Two Proteins. *Journal of Molecular Biology*. 48: 443-453.

Neuwald, A., Liu, J., Lipman, D. and Lawrenc,E C. E. (1997). Extracting Protein Alignment Models From The Sequence Database. *Nucleic Acids Research*. 25: 1665-1677.

Nielsen, H., Brunak, S. and Von Heijne, G. (1999). Machine Learning Approaches For The Prediction Of Signal Peptides and Other Protein Sorting Signals. *Protein Engineering.*12: 3-9.

Nishikawa, K. and Nogughi, T. (1995). Predicting Protein Secondary Structure Based On Amino Acid Sequence. *Method Enzymol.* 202: 31-44.

Nishikawa, K. and Ooi, T. (1982). Correlation Of The Amino-Acid Composition Of A Protein To Its Structural and Biological Characters. *Journal of Biochemistry*. 91: 1821-1824.

Nishikawa, K., Kubota, Y. and Ooi, T. (1983). Classification Of Proteins Into Groups Based On Amino-Acid Composition and Other Characters (2) Grouping Into Types. *Journal of Biochemistry.* 94: 997-1007.

Norel, R., Lin, S. L., Wolfson, H. J. and Nussinov, R. (1994). Shape Complementarity At Protein-Protein Interfaces. *Biopolymers*. 34: 933-940.

Notredame, C., Holm, L. and Higgins, D. (1998). COFFEE: An Objective Function For Multiple Sequence Alignments. *Bioinformatics.* 14(5): 407-422.

Obuchowski, N. (2000). Sample Size Tables For Receiver Operating Characteristic Studies. *American Journal of Roentgenology*. 175: 603-608.

Olmea, O. and Valencia, A. (1997). Improving Contact Predictions By The Combination Of Correlated Mutations and Other Sources Of Sequence Information. *Folding and Design*. 2: 25-32.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindelis, M. B. and Thornton, J. M. (1997). CATH - A Hierarchic Classification Of Protein Domain Structures. *Structure*. 5(8): 1093-108.

Ouali, M. and King, R. D. (2000). Cascaded Multiple Classifiers For Secondary Structure Prediction. *Protein Science*. 9: 1162-1176.

Pace, C. N., Shirley, B. A., Mcnutt, M. and Gajiwala, K. (1996). Forces Contributing To The Confor- Mational Stability Of Proteins. *Journal of the American Societies for Experimental Biology*. 10(1): 75-83.

Pauling, L. and Corey, R. B. (1951). Configurations Of Polypeptide Chains With Favoured Orientations Around Single Bonds: Two New Pleated Sheets. *Proceedings of the National Academic of Science.* USA. 37**:** 729-740.

Pauling, L. and Corey, R. B. (1951). Configurations Of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proceedings of the National Academic of Science*. USA. 37**:** 729–740.

Pearson, W. and Lipman, D. (1988). Improved Tools For Biological Sequence Comparison. *Proceedings of the National Academic of Science*. USA 85: 2444-2448.

Pearson, W. R. (1990). Rapid and Sensitive Sequence Comparison With FASTP and FASTA. *Method Enzymol*. 183: 63-98.

Periti, P. F., Quagliarotti, G. and Liquori, A. M. (1967). Recognition Of Alpha Helical Segments in Proteins Of Known Primary Structure. *Journal of Molecular Biology*. 24: 313-322.

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: Structure, Function, and Genetics, Supplement.* 47: 228-235.

Ponder, J. W. and Richards, F. M. (1987). Tertiary Templates For Proteins Use Of Packing Criteria in The Enumeration Of Allowed Sequences For Different Structural Classes. *Journal of Molecular Biology*. 193: 775-791.

Przybylski, D. and Rost, B. (2002). Alignments Grow Secondary Structure Prediction Improves. *Proteins: Structure, Function, and Genetics, Supplement*. 46: 197-205.

Ptitsyn, O. B. (1969). Statistical Analysis Of The Distribution Of Amino Acid Residues Among Helical and Non-Helical Regions in Globular Proteins. *Journal of Molecular Biology*. 42: 501-510.

Qian, N. and Sejnowski, T. J. (1988). Predicting The Secondary Structure Of Globular Proteins Using Neural Network Models. *Journal of Molecular Biology*. 202(4): 865-84.

Rao, S. T. and Rossmann, M. G. (1973). Comparison Of Super-Secondary Structures in Proteins. *Journal of Molecular Biology*. 76: 241-256.

Rice, D. W. and Eisenberg, D. (1997). A 3D-1D Substitution Matrix For Protein Fold Recognition That Includes Predicted Secondary Structure Of The Sequence. *Journal of Molecular Biology*. 267: 1026-1038.

Richard, H. and anders, K. (1996). Hidden Markov Models For Sequence Analysis: Extension and Analysis Of The Basic Method. *Computational Application for* BioScience. 12(2): 95-107.

Richards, F. M. and Kundrot, C. E. (1988). Identification Of Structural Motifs From Protein Coordinate Data: Secondary Structure and First-Level Supersecondary Structure. *Proteins: Structure, Function, and Genetics, Supplement*. 3: 71-84.

Richardson, J. S. (1981). The Anatomy and Taxonomy Of Protein Structure. *Advances in Protein Chemistry*. 34: 168-339.

Richardson, J. S. (1986). The Greek Key Topology As A Favoured Form in Folding and Structure. *Federation Proceedings*. 45: 1829.

Riis, S. K. and Krogh, A. (1996). Improving Prediction Of Protein Secondary Structure Using Structured Neural Networks and Multiple Sequence Alignments. *Journal of Computational Biology*. 3: 163–183.

Rooman, M. J. and Wodak, S. J. (1991). Weak Correlation Between Predictive

Power Of Individual Sequence Patterns and Overall Prediction Accuracy in Proteins. *Proteins: Structure, Function, and Genetics, Supplement.* 9: 69-78.

Rost, B. and Sander, C (1996). Bridging The Protein Sequence-Structure Gap By Structure Predictions. *Annual Review Of Biophysics and Biomolecular Structure*. 25: 113-136.

Rost, B. and Sander, C. (1993). Prediction Of Protein Secondary Structure At Better Than 70% Accuracy. *Journal of Molecular Biology*. 232. 584-599.

Rost, B. (1995). TOPITS: Threading One-Dimensional Predictions Into Three-Dimensional Structures. *Proceedings of the Intelligent System in Molecular Biology*. 314-21.

Rost, B. and Sander, C. (1994). Combining Evolutionary Information and Neural Networks To Predict Protein Secondary Structure. *Proteins: Structure, Function, and Genetics, Supplement.*19: 55-72.

Rost, B. (2001). Review: Protein Secondary Structure Prediction Continues To Rise. *Journal of Structural Biology*. 134: 204–218.

Rost, B. (2003). Neural Networks Predict Protein Structure: Hype Or Hit? Paolo Frasconi ed. *in: Artificial Intelligence and Heuristic Models For Bioinformatics*. CITY:ISO Press. Page

Rost, B. R., Sander, C. and Schneider, R. (1994). Redefining The Goals Of Protein Secondary Structure Prediction. *Journal of Molecular Biology*. 235: 13-26.

Rost, B. and Sander, C. (1993). Prediction Of Protein Secondary Structure At Better Than 70% Accuracy. *Journal of Molecular Biology*. 232: 584–599.

Rost, B., Schneider, R. and Sander, C. (1997). Protein Fold Recognition By Prediction-Based Threading. *Journal of Molecular Biology*. 270: 471-480.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning Representations By Back-Propagating Errors. *Nature.* 323: 533-536.

Rumelhart, D. E. and Mcclelland, J. L. (1986). *Parallel Distributed Processing: Explorations in The Microstructure Of Cognition.* Cambridge. MA: MIT Press.

Russell, R. B. and Barton, G. J. (1992). Multiple Protein-Sequence Alignment From Tertiary Structure Comparison -- Assignment Of Global and Residue Confidence Levels. *Proteins: Structure, Function, and Genetics, Supplement.* 14: 309-323.

Russell, R. B. and Barton, G. J. (1993). The Limits Of Protein Secondary Structure Prediction Accuracy From Multiple Sequence Alignment. *Journal of Molecular Biology*. 234: 951-957.

Russell, R. B. and Barton, G. J. (1994). Structural Features Can Be Unconserved in Proteins With Similar Folds An Analysis Of Side-Chain To Side-Chain Contacts Secondary Structure and Accessibility. *Journal of Molecular Biology*. 244: 332-350.

Russell, R. B., Copley, R. R. and Barton, G. J. (1996). Protein Fold Recognition By Mapping Predicted Secondary Structures. *Journal of Molecular Biology*. 259: 349-365.

Russell, R. B., Saqi, M. A. S., Sayle, R. A., Bates, P. A. and Sternberg, M. J. E. (1997). Recognition Of Analogous and Homologous Protein Folds: Analysis Of Sequence and Structure Conservation. *Journal of Molecular Biology*. 269: 423-439.

Salamov, A. A. and Solovyev, V. V. (1995). Prediction Of Protein Secondary Structure By Combining Nearest-Neighbour Algorithms and Multiple Sequence Alignments. *Journal of Molecular Biology*. 247: 11-15.

Salamov, A. A. and Solovyev, V. V. (1997). Protein Secondary Structure Prediction Using Local Alignments. *Journal of Molecular Biology*. 268: 31-36.

Sanchez, R. and Sali, A. (1997). Advances in Comparative Protein-Structure Modelling. *Current Opinion in Structural Biology*. 7: 206-214.

Sander, C. and Schneider, R. (1991). Database Of Homology-Derived Protein Structures and The Structural Meaning Of Sequence Alignment. *Proteins: Structure, Function, and Genetics, Supplement*. 9(1): 56-68.

Saqi, M.A., Bates, P. A. and Sternberg, M. J. (1992). Towards An Automatic Method Of Predicting Protein Structure By Homology: An Evaluation Of Suboptimal Sequence Alignments. *Protein Engineering*. 5: 305–311.

Sauder, S. M., Arthur J. W. and Dunbrack, R. L. (2000). Large-Scale Comparison Of Protein Sequence Alignment Algorithms With Structural Alignments. *Proteins: Structure, Function, and Genetics, Supplement.* 40: 6-22.

Schulz, G. E. and Schirmer, R. H. (1979). *Principles Of Proteins Structure*. Springer-Verlag, New York.

Schulz, G. E. (1977). Recognition Of Phylogenetic Relationships From Polypeptide

Chain Fold Similarities. *Journal of Molecular*. 9: 339-342.

Sean, E. (1995). Multiple Alignment Using Hidden Markov Models. in Christopher Railings. *Proceedings in the International Conference of Intelligent Systems for Molecular* Biology. 114-120.

Shannon, C. E. (1948). The Mathematical Theory Of Communications. *Bell System Technical Journal.*

Shindyalov, I. N. and Bourne, P. E. (1998). Protein Structure Alignment By Incremental Combinatorial Extension (CE) Of The Optimal Path. *Protein Engineering*. 11(9): 739-47.

Siddiqui, A. S. and Barton, G. J. (1995). Continuous and Discontinuous Domains An Algorithm For The Automatic-Generation Of Reliable Protein Domain Definitions. *Protein Science*. 4: 872-884.

Siddiqui, A. S., Dengler, U. and Barton, G. J. (2001). 3Dee: A Database Of Protein Structural Domains. *Bioinformatics.* 17: 200-201.

Siegelmann, H. T. (1998). *Neural Networks and Analog Computation: Beyond The Turing Limit*, Boston, Birkhauser.

Siegelmann, H. T. and Sontag, E. D. (1999). During Computability With Neural Networks. *Applied Mathematics Letters.* 4: 77-80.

Sippl, M. J. (1990). Calculation Of Conformational Ensembles From Potentials Of Mean Force An Approach To The Knowledge-Based Prediction Of Local Structures in Globular-Proteins. *Journal of Molecular Biology*. 213: 859-883.

Sippl, M. J., Lackner, P., Domingues, F. S., Prlic, A., Malik, R., andreeva, A. and Wiederstein, M. (2001). Assessment Of The CASP4 Fold Recognition Category. *Proteins: Structure, Function, and Genetics, Supplement.* 5: 55–67.

Sjolander, K., Karplu,S K., Brown, M. P., Hugheym, R., Krogh, A., Mian ,I. S. and Haussler, D. (1996). Dirichlet Mixtures: A Method For Improving Detection Of Weak But Significant Protein Sequence Homology. Computer Application in the Biosciences. 12 (4): 327-345.

Smith, R. F. and. Smith, T. F. (1992). Pattern-Induced Multi-Sequence Alignment (PIMA) Algorithm Employing Secondary Structure-Dependent Gap Penalties For Use in Comparative Protein Modelling. *Protein Engineering*. 5(1): 35-41.

Smith, T. F. (1999). The Art of Matchmaking: Sequence Alignment Methods and

Their Structural Implications. *Structure With Folding and Design.* 7(1): 7-12.

Smith, T. F and Waterman, M. S. (1981). Identification Of Common Molecular Subsequences. *Journal of Molecular Biology*. 147: 195-197.

Sonnhammer., E. L. L. and Kahn., D. (1994). Modular Arrangement Of Proteins As Inferred From Analysis Of Homology. *Protein Science*. 3: 482-492.

Srinivasan, N., Guruprasad, K. and Blundell, T. (1996). Comparative Modelling Of Proteins. in: M. J.Sternberg., ed. *Protein Structure Prediction.* IRL Press. 1-30.

Stephen, R. Holbrook, Steven, M., Muskal and Sung-Hou Kim. (1990). Predicting Protein Structural Features With Artificial Neural Networks. in: Lawrence Hunter ed. *Artificial Intelligence and Molecular Biology*. UK.

Sternberg, M. J. E. and Thornton, J. M. (1976). On The Conformation Of Proteins: The Handedness Of The Beta-Strand - Alpha-Helix - Beta-Strand Unit. *Journal of Molecular Biology*. 105: 367-382.

Swets, J. A., Dawes, R. M and Monahan, J. (2000). Better Decisions Through Science. *Scientific American.* 283: 82-87.

Swets, J. (1988). Measuring The Accuracy Of Diagnostic Systems. *Science*. 240: 1285-1293.

Swindells, M. B. (1995b). A Procedure For The Automatic-Determination Of Hydrophobic Cores in Protein Structures. *Protein Science*. 4: 93-102.

Swingler, K. (1996). *Applying Neural Networks: A Practical Guide.* London: Academic Press.

Tatusov, R., Altschul, S. and Koonin, E. (1994). *Proceedings Of The National Academy Of Sciences Of The United States Of America*. 91(25): 12091-12095.

Taylor, W. R. (1998). Dynamic Sequence Databank Searching With Templates and Multiple Alignments. *Journal of Molecular Biology*. 280(3): 375-406.

Taylor, W. R. and Orengo, C. A. (1989). Protein-Structure Alignment. *Journal of Molecular Biology*. 208: 1-22.

Taylor, W. R. and Thornton, J. M. (1984). Recognition Of Super-Secondary Structure in Proteins. *Journal of Molecular Biology*. 173. 487-514.

Taylor, W. R. (1997). Multiple Sequence Threading: An Analysis Of Alignment

Quality and Stability. *Journal of Molecular Biology*. 269: 902-943.

Thomas, D. J., Casari, G. and Sander, C. (1996). The Prediction Of Protein Contacts From Multiple Sequence Alignments. *Protein Engineering.* 9: 941-948.

Thomas, P. D. and Dill, K. A. (1996). Statistical Potentials Extracted From Protein Structures How Accurate Are They? *Journal of Molecular Biology*. 257: 457-469.

Thompson, J. D., Higgins, D. G. and. Gibson, T. J. (1994). CLUSTAL W: Improving The Sesitivity Of Progressive Multiple Sequence Alignment Through Sequence Weighting, Positions-Specific Gap Penalties and Weigh Matrix Choice. *Nucleic Acids Research*. 22: 4673-4680.

Timothy, L., Bailey and Charles, E. (1994). *Fitting A Mixture Model By Expectation Maximization To Discover Motifs in Biopolymers.* in /SMB-94. 28-36. Menlo Park. CA: AAI/MIT Press.

Tomii, K. and Kanehisa, M. (1996). Analysis Of Amino-Acid Indexes and Mutation Matrices For Sequence Comparison and Structure Prediction Of Proteins. *Protein Engineering.* 9: 27-36.

Valiant, L. (1988). Functionality in Neural Nets, Learning and Knowledge Acquisition. *Proceeding of the American Association for Artificial Intelligent*. 629-634.

Van-Heel, M. (1991). A New Family Of Powerful Multivariate Statistical Sequence-Analysis Techniques. *Journal of Molecular Biology*. 220: 877-887.

Warme, P. K., Momany, F. A., Rumball, S. V., Tuttle, R. W. and Scheraga, H. A. (1974). Computation Of Structures Of Homologous Proteins. Alpha-Lactalbumin From Lysozyme. *Biochemistry. 13*: 768-782.

Weiner, P. K. and Kollman, P. A. (1981). AMBER: Assisted Model Building With Energy Refinement. A General Program For Modeling Molecules and Their Interactions. *Journal of Computational Chemistry.* 2: 287-303.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P. (1984). A New Force Field For Molecular Mechanical Simulation Of Nucleic Acids and Proteins. *Journal of American Chemical Societies.* 106: 765-784.

Weiss, S. M. and Kulikowski, C. A. (1991). *Computer Systems That Learn*. Morgan Kaufmann Publishers, Inc, San Mateo. CA.

White, H. (1992). *Artificial Neural Networks: Approximation and Learning Theory*. Blackwell. Oxford.

Woody, R. W. (1995). Circular-Dichroism. *Method Enzymol*. 246. 34-71.

Wu, C. H. and Mclarty, J. W. (2000). Neural Networks and Genome Informatics. *Elsevier Science*.

Wu, C., Whitson, G., Mclarty, J., Ermongkoncha, A. and Chang, T. C. (1992). Protein Classification Artificial Neural System. *Protein Science*. 1: 667-677.

Yang, A. S, Hitz, B. and Honig, B. (1996). Free-Energy Determinants Of Secondary Structure Formation (3) Beta-Turns and Their Role in Protein-Folding. *Journal of Molecular* Biology. 259: 873-882.

Yi, T. M. and Lander, E. S. (1993). Protein Secondary Structure Prediction Using Nearest-Neighbor Methods. *Journal of Molecular* Biology. 232: 1117-1129.

Zachariah, M. A., Crooks, G. E., Holbrook, S. R. and Brenner, S. E. (2005). A Generalized Affine Gap Model Significantly Improves Protein Sequence Alignment Accuracy. *Proteins: Structure, Function, and Bioinformatics*. 58: 329–338.

Zell, A., Mamier, G., Vogt, M., Mache, N., Hubner, R., Doring, S., Herrmann, K. U., Soyez, T., Schmalzl, T., Sommer, T., Hatzigeorgiou, A., Posselt, D., Schreiner, T., Ket., B., Clemente, G. and Wieland. (1998). *The SNNS Users Manual Versi*on 4.1.http://Www.Informatik.Uni-Tuttgart.De/ Ipvr/Bv/Projekte / Snns/Usermanual/User Manual.Html

Zemla, A., Venclovas, C., Fidelis, K. and Rost, B. (1999). A Modified Definition Of SOV: A Segment Based Measure For Protein Secondary Structure Prediction Assessment. *Proteins: Structure, Function, and Genetics, Supplement.* 34: 220-223.

Zhou, G. F., Xu, X. H. and Zhang, C. T. (1992). A Weighting Method For Predicting Protein Structural Class From Amino-Acid-Composition. *European Journal of Neuroscience*. 210: 747-749.

Zou, K. H. (2002). Receiver Operating Characteristic (ROC) Literature Research. *Http://Splweb.Bwh.Harvard.Edu:8000/Pages/ Ppl/Zou/Roc.Html*.

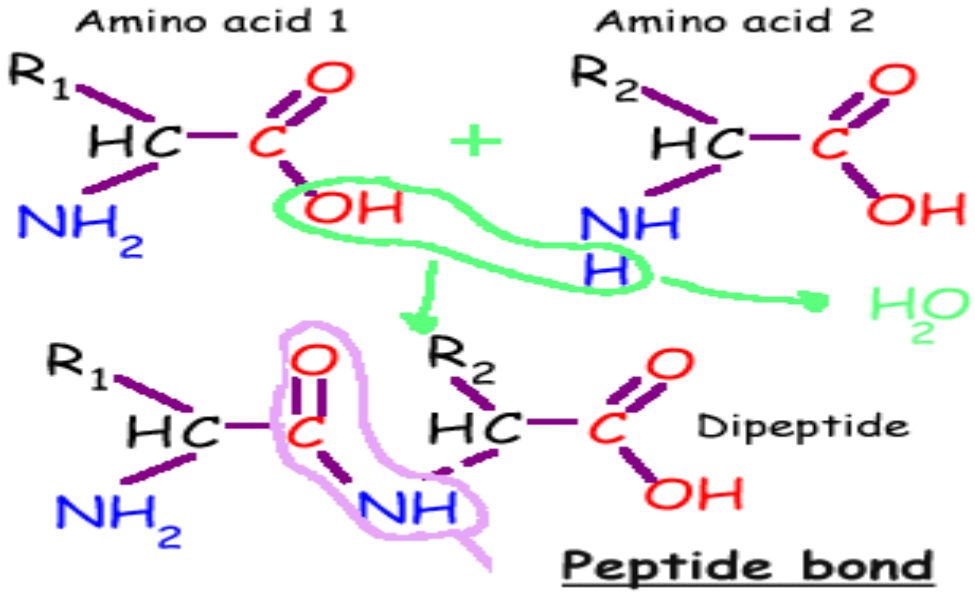
Zvelebil, M. J., Barton, G. J., Taylor, W. R. and Sternberg, M. J. E. (1987). Prediction Of Protein Secondary Structure and Active-Sites Using The Alignment Of Homologous Sequences. *Journal of Molecular* Biology. 195:

957-961.

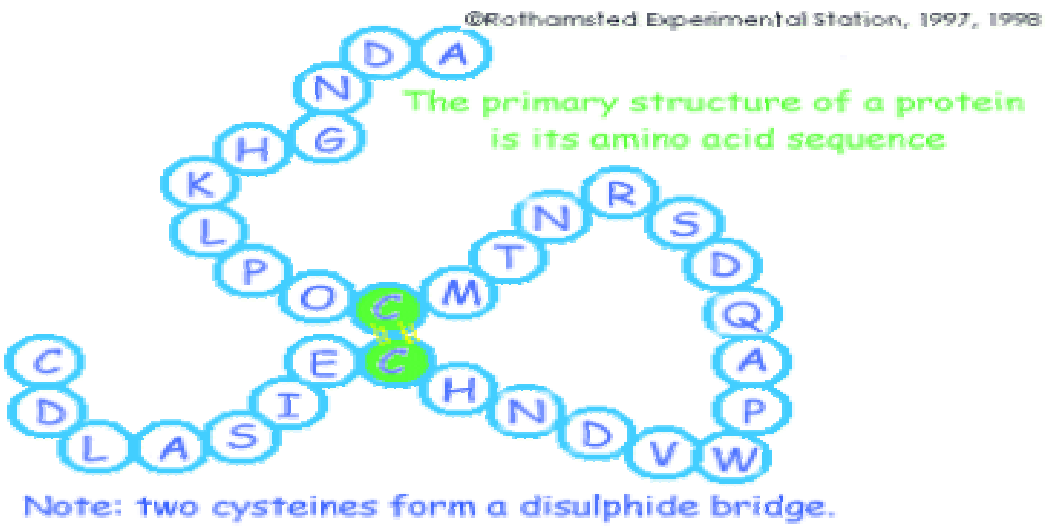
Zweig, G. and Campbell. C. C. (1993). Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chemistry*. 39(4): 561-77.

Appendix A

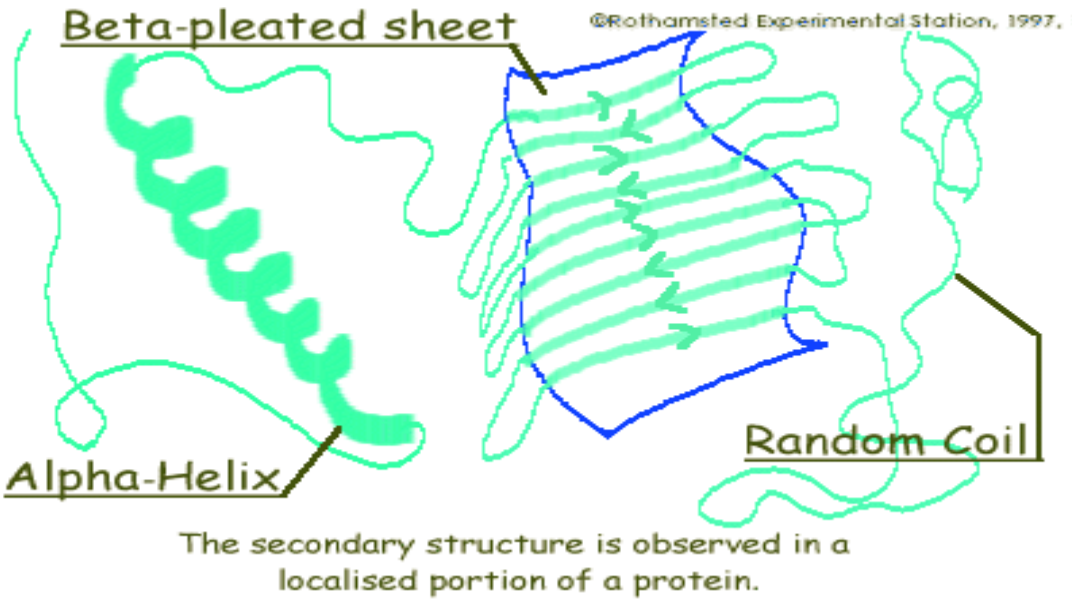
PROTEIN STRUCTURES



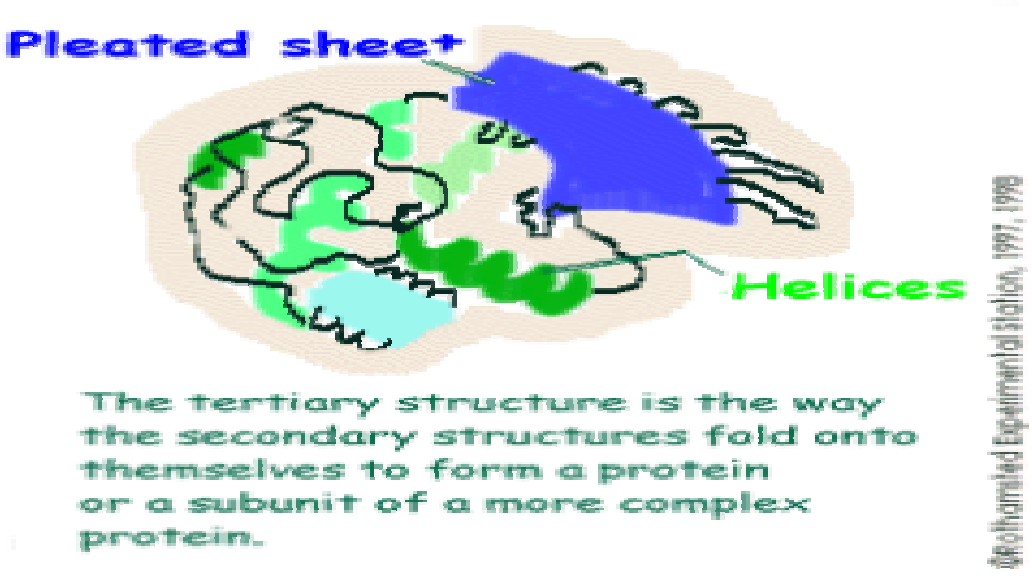
a) Amino acid sequences and peptide bond linking



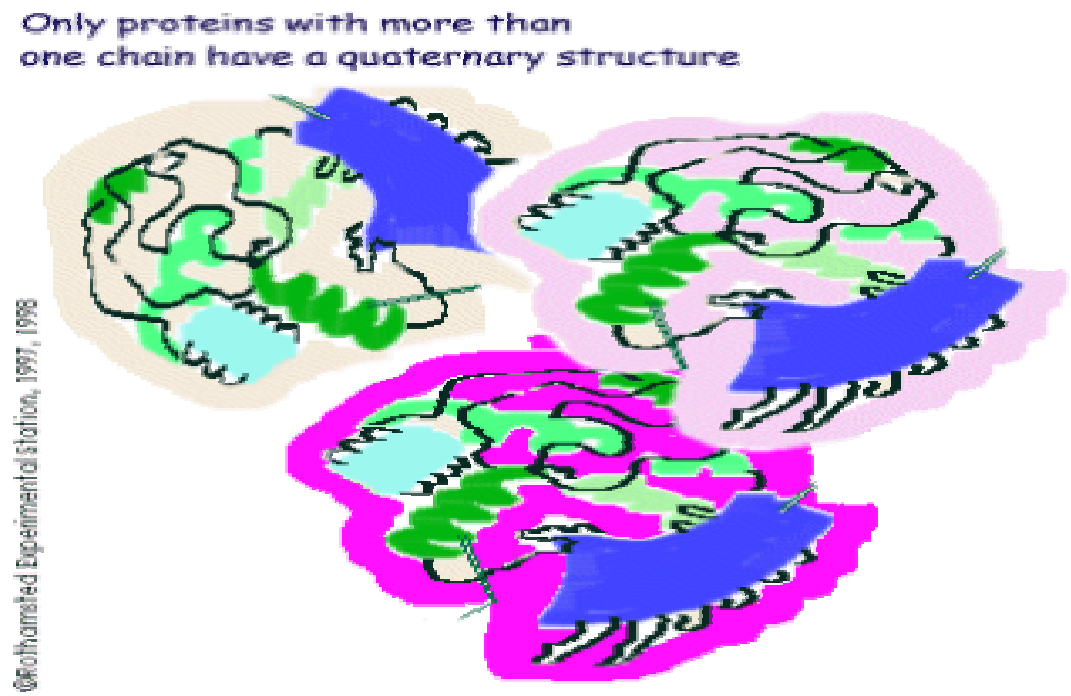
b) Primary structure of a protein (Amino acid sequences)



c) Secondary structure of a protein



d) Tertiary or 3D structure of a protein



e) Quaternary structure of a protein

Source: http://www.rothamsted.bbsrc.ac.uk/notebook/courses/guide/prot.htm

Appendix B

CUFF AND BARTON'S 513 PROTEIN DATA SET

Name	PHD	Length	Class	Fold
1aozb-1-AS	82.3	130.0	All beta	Cupredoxins
1atpi-1-DOMAK	85.0	20.0	Peptides	Protein kinases (PK) Inhibitor
1ayab-1-GJB	83.1	101.0	Alpha and beta (a+b)	SH2-like
1bsdb-1-DOMAK	74.7	107.0	Alpha and beta (a+b)	Microbal ribonucleases
1coi-1-AS	96.5	29.0	Peptides	Antifreeze polypeptide HPLC-6
1cthb-1-DOMAK	59.4	79.0	Small proteins	Cytochrome c3
1ctm-2-DOMAK	81.6	60.0	All beta	Barrel-sandwich hybrid
1ctn-1-AS.1	80.7	109.0	All beta	Immunoglobulin-like beta-sandwich
1edmc-1-AUTO.1	97.4	39.0	Small proteins	EGF-like module
1fc2c	65.1	43.0	All alpha	Immunoglobulin-binding protein A, fragment B
1gln-3-AS	75.0	48.0	All alpha	Anticodon-binding (C-terminal) domain of glutamyl-tRNA Domain I
1gp2a-1-AUTO.1	89.2	28.0	Peptides	Mellitin
1grj-2-AS	71.4	77.0	Alpha and beta (a+b)	FKBP-like
1hcgb-1-AS	80.3	51.0	Small proteins	EGF-like module
1htrp-1-AS	67.4	43.0	Small proteins	Acid protease presegment
1hup-1-AS.1	100.0	24.0	All alpha	Oligomers of long helices
1ilk-2-AS	95.5	45.0	All alpha	4-helical cytokines fragment
1isub-1-DOMAK	66.1	62.0	Small proteins	HIPIP (high potential iron protein)
1lpe-1-DOMAK	84.0	144.0	All alpha	Four-helical up-and-down bundle
1mcti-1-AUTO.1	53.5	28.0	Small proteins	Small inhibitors, toxins, lectins
1mdta-1-AS	74.3	187.0	Alpha and beta (a+b)	ADP-ribosylation toxins
1mrt	100.0	31.0	Small proteins	Metallothionein
1ndh-2-AS	69.3	147.0	Alpha and beta (a/b)	Ferredoxin reductase-like, C-terminal NADP-linked domain
1ovoa	69.6	56.0	Small proteins	Ovomucoid/PCI-like inhibitors
1pga-1-DOMAK	75.0	56.0	Alpha and beta (a+b)	beta-Grasp
1powb-4-DOMAK	77.2	44.0	All alpha	Pyruvate oxidase and decarboxylase, C terminal domain
1ppt	100.0	36.0	Peptides	Pancreatic polypeptide
1reqc-1-AS	69.8	53.0	Unknown	Unknown
1rpo-1-AUTO.1	96.7	61.0	All alpha	ROP protein
1svb-2-AS	69.7	96.0	Unknown	Unknown
1tabi-1-DOMAK	86.1	36.0	Small proteins	Small inhibitors, toxins, lectins

1ubdc-1-AS	70.3	27.0	Small proteins	Classic zinc finger
1ubq	80.2	76.0	Alpha and beta (a+b)	beta-Grasp
1wapv-1-AUTO.1	73.1	67.0	All beta	Double-stranded beta-helix, jelly-roll domain
1wfbb-1-AUTO.1	97.3	37.0	Peptides	Antifreeze polypeptide HPLC-6
2aaib-2-DOMAK	69.3	124.0	All beta	beta-Trefoil
2erl-1-AUTO.1	45.0	40.0	Unknown	Unknown
2mhu	90.0	30.0	Small proteins	Metallothionein
2mltb-1-GJB	80.7	26.0	Peptides	Mellitin
2or1l	84.1	63.0	All alpha	lambda repressor-like DNA-binding domains
2tgpi	87.9	58.0	Small proteins	BPTI-like
3b5c	62.3	85.0	Alpha and beta (a+b)	Cytochrome b5
3pmgb-2-AS	76.3	114.0	Alpha and beta (a/b)	Phosphoglucomutase, first domains
6rlxd-1-DOMAK	64.0	25.0	Small proteins	Insulin-like
9wgaa	59.6	171.0	Small proteins	Small inhibitors, toxins, lectins
1nga-2-AS.1	78.4	190.0	Alpha and beta (a/b)	Ribonuclease H-like motif
1gpmd-5-AS	78.0	178.0	Alpha and beta (a/b)	ATP pyrophosphatases
1asw-1-AUTO.1	82.4	148.0	Alpha and beta (a/b)	Ribonuclease H-like motif
1eca	80.8	136.0	All alpha	Globin-like
1fuqb-1-AUTO.1	75.0	136.0	Unknown	Unknown
1zymb-2-AUTO.1	87.5	128.0	Unknown	Unknown
2cab	75.7	256.0	All beta	Carbonic anhydrase
5lyz	65.1	129.0	Alpha and beta (a+b)	Lysozyme-like Domain I
1cnsb-1-AUTO.1	68.7	243.0	Alpha and beta (a+b)	Lysozyme-like
1mspb-1-AS	79.5	122.0	All beta	Immunoglobulin-like beta-sandwich
1mai-1-JAC	70.5	119.0	Unknown	Unknown
1dlc-1-AS.1	83.8	229.0	Membrane and cell surface proteins and peptides	Toxins' membrane translocation domains
1dynb-1-AUTO.1	63.7	113.0	All beta	PH domain-like
2hmza	80.7	114.0	All alpha	Four-helical up-and-down bundle
3mddb-2-AS	72.0	111.0	All beta	Acyl-CoA dehydrogenase (flavoprotein), middle domain, barrel like
1vcab-2-AUTO.1	68.1	110.0	All beta	Immunoglobulin-like beta-sandwich
1acx	81.4	108.0	All beta	Immunoglobulin-like beta-sandwich
1cewi-1-DOMAK	69.4	108.0	Alpha and beta (a+b)	Cystatin-like
1ilk-1-AS	77.3	106.0	All alpha	4-helical cytokines Short chain
1sesa-2-AS	64.9	317.0	Alpha and beta (a+b)	Class II aaRS and biotin synthetases
1irk-2-AS	76.4	204.0	Alpha and beta (a+b)	Protein kinases (PK), catalytic core C terminal Domain
1cfb-1-AS	79.2	101.0	All beta	Immunoglobulin-like beta-sandwich
2alp	67.6	198.0	All beta	Trypsin-like serine proteases Domain I
1stfi-1-DOMAK	77.5	98.0	Alpha and beta (a+b)	Cystatin-like
1thtb-1-AUTO.1	67.5	293.0	Alpha and beta (a/b)	alpha/beta-Hydrolases
1nal4-1-AUTO.1	84.1	291.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
1ris-1-DOMAK	67.0	97.0	Alpha and beta (a+b)	Ferredoxin-like
1tml-1-AS	84.2	286.0	Alpha and beta (a/b)	Cellulases
2ebn-1-AS	81.4	285.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
1gep-2-AS	81.0	179.0	Unknown	Unknown
1dpgb-1-AUTO.1	87.5	177.0	Alpha and beta (a/b)	NAD(P)-binding Rossmann-fold domains
1tig-1-AUTO.1	78.4	88.0	Alpha and beta (a+b)	IF3-like
1celb-1-AUTO.1	65.1	433.0	Unknown	Unknown
2hpr-1-DOMAK	72.4	87.0	Alpha and beta (a+b)	Histidine-containing phosphocarrier proteins (HPr)
1cc5	72.2	83.0	All alpha	Cytochrome c
1fuqb-2-AUTO.1	75.6	250.0	Unknown	Unknown
1pht-1-AUTO.1	48.1	83.0	All beta	SH3-like barrel

2spt-2-DOMAK	81.7	82.0	Small proteins	Kringle modules
1mdta-3-AS	77.3	159.0	All beta	Common fold of diphtheria toxin/transcription factors/cytochrome f
1onrb-1-AUTO.1	77.2	316.0	Unknown	Unknown
1mns-2-AS	71.4	228.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
1nfp-1-AS	77.6	228.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
3icb	85.3	75.0	All alpha	EF-hand
1latb-1-AUTO.1	74.3	74.0	Small proteins	Glucocorticoid receptor-like (DNA-binding domain)
4fisb-1-DOMAK	84.9	73.0	All alpha	FIS protein
1fdlh	73.3	218.0	All beta	Immunoglobulin-like beta-sandwich
3cln	89.5	143.0	All alpha	EF-hand
1il8a	78.8	71.0	Alpha and beta (a+b)	Interleukin 8-like chemokines
1oacb-4-AS.1	69.9	426.0	All beta	Supersandwich
2utga	84.2	70.0	All alpha	Uteroglobin-like
1ctf-1-DOMAK	76.4	68.0	Alpha and beta (a+b)	Ribosomal protein L7/12, C-terminal fragment
1rsy-1-AS	71.8	135.0	All beta	Immunoglobulin-like beta-sandwich
1fuqb-3-AUTO.1	86.3	66.0	Unknown	Unknown
1dik-2-AS.1	62.3	130.0	Alpha and beta (a+b)	ATP-grasp sub-domain II
1dsbb-2-AUTO.1	79.6	64.0	All alpha	Disulphide-bond formation facilitator (DSBA), insertion domain
2pgd-2-AUTO.1	79.8	253.0	All alpha	6-phosphogluconate & Acyl-CoA dehydrogenases, C-terminal domain
1csei	71.4	63.0	Alpha and beta (a+b)	CI-family of serine protease inhibitors
7rsa	68.5	124.0	Alpha and beta (a+b)	Ribonuclease A-like
2nadb-2-AS.1	74.5	185.0	Alpha and beta (a/b)	NAD(P)-binding Rossmann-fold domains
1qbb-2-AUTO.1	77.0	122.0	Unknown	Unknown
3inkd-1-DOMAK	59.5	121.0	All alpha	4-helical cytokines
2pgd-1-AUTO.1	70.7	181.0	Alpha and beta (a/b)	NAD(P)-binding Rossmann-fold domains
1dnpb-2-AUTO.1	68.8	180.0	Unknown	Unknown
1esl-1-GJB	74.1	120.0	Alpha and beta (a+b)	C-type lectin
1gp2g-2-AS	83.2	298.0	All beta	7-bladed beta-propeller
1bncb-4-AS	76.2	118.0	All beta	Barrel-sandwich hybrid
6cpp	75.8	405.0	All alpha	Cytochrome P450
1sftb-2-AS	70.0	230.0	Unknown	Unknown
1seib-2-AUTO.1	68.4	57.0	Unknown	Unknown
9apia	71.6	339.0	Multi-domain (alpha and beta)	Serpins
2bat-1-GJB	70.8	388.0	All beta	6-bladed beta-propeller
2gsq-2-AS	85.5	111.0	All alpha	Glutathione S-transferases, C-terminal domain
821p-1-DOMAK	80.7	166.0	Alpha and beta (a/b)	P-loop containing nucleotide triphosphate hydrolases
1isab-2-GJB	80.7	109.0	Alpha and beta (a+b)	Fe,Mn superoxide dismutase (SOD), C-terminal domain
1fkf	72.9	107.0	Alpha and beta (a+b)	FKBP-like
1tcba-1-AS	57.7	317.0	Alpha and beta (a/b)	alpha/beta-Hydrolases
1hxn-1-AS	76.1	210.0	All beta	4-bladed beta-propeller
1pnt-1-AS	77.7	157.0	Alpha and beta (a/b)	Phosphotyrosine protein phosphatases I
1chbe-1-DOMAK	73.7	103.0	All beta	OB-fold
1hiws-1-AS	64.0	103.0	All alpha	Retroviral matrix proteins
1dpgb-2-AUTO.1	73.0	308.0	Alpha and beta (a+b)	Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain
1kinb-1-AUTO.1	71.7	308.0	Unknown	Unknown
3mddb-3-AS	84.4	154.0	All alpha	Four-helical up-and-down bundle
1bncb-3-AS	64.7	51.0	Alpha and beta (a+b)	ATP-grasp sub-domain II
1gdj	87.5	153.0	All alpha	Globin-like

2hft-1-AS	65.6	102.0	All beta	Immunoglobulin-like beta-sandwich
1gky-2-AS	60.0	50.0	Alpha and beta (a+b)	P-loop containing nucleotide triphosphate hydrolases, inserted domain in Guanylate Kinase
1krca-1-AUTO.1	78.0	100.0	Alpha and beta (a+b)	Urease, gamma-subunit
1smpi-1-AS	69.0	100.0	All beta	Streptavidin-like
7cata	72.0	498.0	All alpha	Heme-linked catalases N-terminal fragment
1ncg-1-AUTO.2	76.7	99.0	All beta	Immunoglobulin-like beta-sandwich
1gln-4-AS	81.6	98.0	All alpha	Anticodon-binding (C-terminal) domain of glutamyl-tRNA Domain II
1hmy-2-AS	56.1	98.0	Alpha and beta (a/b)	S-adenosyl-L-methionine-dependent methyltransferases Domain II
1dnpb-1-AUTO.1	85.4	289.0	Unknown	Unknown
1lap	74.8	481.0	Alpha and beta (a/b)	Leucine aminopeptidase, N-terminal domain
1sh1	62.5	48.0	Small proteins	Defensin-like
1wsyb	73.7	385.0	Alpha and beta (a/b)	Tryptophan synthase, beta-subunit Domain I
1clc-2-AS.1	80.3	239.0	All alpha	Glycosyltransferases of the superhelical fold Domain I
2ltnb	80.8	47.0	All beta	ConA-like lectins/glucanases
2sns	75.1	141.0	All beta	OB-fold
3pmgb-1-AS	75.5	188.0	Alpha and beta (a/b)	Phosphoglucomutase, first domains
1cpcl-1-DOMAK	82.8	140.0	All alpha	Globin-like
1bcx-1-DOMAK	82.7	185.0	All beta	ConA-like lectins/glucanases
1s01	71.2	275.0	Alpha and beta (a/b)	Subtilases
1powb-1-DOMAK	76.3	182.0	Alpha and beta (a/b)	Thiamin-binding
4rhv1	73.6	273.0	All beta	Viral coat and capsid proteins
1vcab-1-AUTO.1	78.6	89.0	All beta	Immunoglobulin-like beta-sandwich
1mdta-2-AS	72.3	177.0	Membrane and cell surface proteins and peptides	Toxins' membrane translocation domains
1han-1-AUTO.1	78.7	132.0	Alpha and beta (a+b)	2,3-Dihydroxybiphenyl dioxygenase (DHDB, BPHC enzyme)
1kuh-1-AS	67.4	132.0	Alpha and beta (a+b)	Metzincins, catalytic (N-terminal) domain
1aazb-1-DOMAK	78.1	87.0	Alpha and beta (a/b)	Thioredoxin-like
1pda-3-AS	79.3	87.0	Alpha and beta (a+b)	dsRBD & PDA domains
1dkza-1-JAC	80.9	215.0	Unknown	Unknown
1pdo-1-GJB	86.0	129.0	Unknown	Unknown
1svb-1-AS	66.5	299.0	Unknown	Unknown
1trb-2-AS	67.1	128.0	Alpha and beta (a/b)	FAD (also NAD)-binding motif
1cei-1-GJB	82.3	85.0	Unknown	Unknown
1r092	62.7	255.0	All beta	Viral coat and capsid proteins
1vid-1-JAC	78.8	213.0	Unknown	Unknown
1rie-1-GJB	77.1	127.0	Unknown	Unknown
2sil-1-AS	72.7	381.0	All beta	6-bladed beta-propeller
1masb-1-AUTO.1	78.3	295.0	Unknown	Unknown
1powb-2-DOMAK	73.3	169.0	Alpha and beta (a/b)	Pyruvate oxidase and decarboxylase, middle domain
1cgu-3-GJB	75.0	84.0	All beta	Immunoglobulin-like beta-sandwich
1isab-1-GJB	67.4	83.0	All alpha	Long alpha-hairpin
1vpt-1-JAC	74.9	291.0	Unknown	Unknown
1epbb-1-DOMAK	81.1	164.0	All beta	Lipocalins
2npx-3-AS.1	65.8	123.0	Alpha and beta (a+b)	FAD/NAD-linked reductases, dimerisation (C-terminal) domain
2polb-1-AS	75.6	123.0	Alpha and beta (a+b)	DNA clamp
2fxb	77.7	81.0	Alpha and beta (a+b)	Ferredoxin-like
1scud-1-AS	76.8	121.0	Alpha and beta (a/b)	NAD(P)-binding Rossmann-fold domains
1chd-1-AS	80.8	198.0	Alpha and beta (a/b)	CheB methylesterase domain (C-terminal residues

				152-349)
1hjrd-1-AUTO.1	79.7	158.0	Alpha and beta (a/b)	Ribonuclease H-like motif
1srja-1-DOMAK	78.8	118.0	All beta	Streptavidin-like
1hvq-1-AUTO.1	67.7	273.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
3pmgb-3-AS	70.9	117.0	Alpha and beta (a/b)	Phosphoglucomutase, first domains
1din-1-AS	81.1	233.0	Unknown	Unknown
1gln-2-AS	75.8	116.0	Alpha and beta (a/b)	ATP pyrophosphatases inserted Domain I
1ghsb-1-GJB	70.9	306.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
1gog-1-AS.1	75.1	153.0	All beta	Galactose-binding domain-like
1ktq-1-AUTO.1	73.2	153.0	Alpha and beta (a/b)	Ribonuclease H-like motif
2rsla-1-GJB	72.1	115.0	Alpha and beta (a/b)	gamma,delta Resolvase, large fragment
6cpa	80.4	307.0	Alpha and beta (a/b)	Zn-dependent exopeptidases
1lehb-3-AS	77.2	229.0	Unknown	Unknown
1pnmb-2-AS	70.6	191.0	All alpha	N-terminal nucleophile aminohydrolases (Ntn hydrolases) B chain Domain
1tnfa	75.0	152.0	All beta	Tumor necrosis factor
2paba	74.5	114.0	All beta	Prealbumin-like
2tsca	70.8	264.0	Alpha and beta (a+b)	Thymidylate synthase
1hyp-1-DOMAK	70.6	75.0	All alpha	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin
2afnc-1-AUTO.1	76.5	149.0	All beta	Cupredoxins
2tgi-1-DOMAK	51.7	112.0	Small proteins	Cystine-knot cytokines
154l-1-AUTO.1	56.2	185.0	Alpha and beta (a+b)	Lysozyme-like
1dih-2-AS	76.3	110.0	Alpha and beta (a+b)	Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain
2dln-3-AS	61.6	73.0	Alpha and beta (a+b)	ATP-grasp sub-domain II
1cem-1-GJB	71.9	363.0	Unknown	Unknown
1nol-1-AUTO.2	70.0	107.0	All alpha	Hemocyanin, N-terminal domain
4xiaa	77.8	393.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
5sici-1-DOMAK	80.3	107.0	Alpha and beta (a+b)	Subtilisin inhibitor
3cd4	69.1	178.0	All beta	Immunoglobulin-like beta-sandwich
1wsya	86.2	248.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
1aorb-1-AS	75.3	211.0	Alpha and beta (a+b)	Aldehyde ferredoxin oxidoreductase, N-terminal domains
1kptb-1-AUTO.1	52.3	105.0	Alpha and beta (a+b)	Virally encoded KP toxin
1mla-2-AS.1	68.5	70.0	Alpha and beta (a+b)	Ferredoxin-like
1rbp	72.9	174.0	All beta	Lipocalins
1cpn-1-DOMAK	67.7	208.0	All beta	ConA-like lectins/glucanases
1ecl-1-AS	64.0	139.0	Alpha and beta (a/b)	Type I DNA topoisomerase Rossmann-fold like domain
3rnt	76.9	104.0	Alpha and beta (a+b)	Microbal ribonucleases
1bovb-1-DOMAK	69.5	69.0	All beta	OB-fold
5cytr	66.0	103.0	All alpha	Cytochrome c
1clc-1-AS.1	70.5	102.0	All beta	Immunoglobulin-like beta-sandwich
1find-1-AUTO.1	78.6	136.0	Unknown	Unknown
1pkyc-2-AUTO.1	66.1	68.0	All beta	Pyruvate kinase beta-barrel domain
1ecpf-1-AUTO.1	76.3	237.0	Alpha and beta (a/b)	Purine and uridine phosphorylases
1vhrb-2-AUTO.1	78.2	101.0	Unknown	Unknown
1xvab-1-GJB	65.4	269.0	Unknown	Unknown
1euu-2-JAC	80.0	100.0	All beta	Immunoglobulin-like beta-sandwich
1oyc-1-AS	74.1	399.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
2cpo-1-AUTO.1	68.7	298.0	Unknown	Unknown
1gcmc-1-AUTO.1	87.8	33.0	All alpha	Oligomers of long helices
2aat	76.0	396.0	Alpha and beta (a/b)	PLP-dependent transferases Domain I

2trt-1-AUTO.1	86.8	198.0	Unknown	Unknown
1fnd	70.2	296.0	All beta	Reductase/elongation factor common domain
1rlr-2-JAC	67.6	526.0	Unknown	Unknown
1l58	73.1	164.0	Alpha and beta (a+b)	Lysozyme-like Domain I
1lib-1-DOMAK	83.2	131.0	All beta	Lipocalins
1ctu-2-AUTO.1	56.1	130.0	Alpha and beta (a/b)	Cytidine deaminase
1tupc-1-AUTO.1	68.2	195.0	All beta	Common fold of diphtheria toxin/transcription factors/cytochrome f
1gnd-2-JAC	67.0	97.0	Unknown	Unknown
1tplb-3-AS	79.8	129.0	Alpha and beta (a/b)	PLP-dependent transferases Domain III
2ak3a	80.9	226.0	Alpha and beta (a/b)	P-loop containing nucleotide triphosphate hydrolases
3blm	77.4	257.0	Alpha and beta (a+b)	beta-Lactamase/D-ala carboxypeptidase Domain I
1cgu-2-GJB	70.8	96.0	All beta	alpha-Amylases, beta-sheet domain
1fxia	77.0	96.0	Alpha and beta (a+b)	beta-Grasp
1ptx-1-AS	62.5	64.0	Small proteins	Small inhibitors, toxins, lectins
1vnc-1-JAC	68.5	576.0	Unknown	Unknown
2ccya	82.6	127.0	All alpha	Four-helical up-and-down bundle
1chkb-2-AUTO.1	76.8	95.0	Unknown	Unknown
1cyx-1-AUTO.1	76.5	158.0	All beta	Cupredoxins
1cfr-1-GJB	65.3	283.0	Unknown	Unknown
1dts-1-AUTO.1	79.0	220.0	Alpha and beta (a/b)	P-loop containing nucleotide triphosphate hydrolases
3bcl-1-DOMAK	58.1	344.0	All beta	Bacteriochlorophyl A protein
1gpc-1-AS	59.6	218.0	All beta	OB-fold
1gal-3-AS	59.1	186.0	Alpha and beta (a+b)	FAD-linked reductases, C-terminal domain
1knb-1-AS	76.8	186.0	All beta	Adenovirus type fiber protein, knob domain
6dfr	75.9	154.0	Alpha and beta (a/b)	Dihydrofolate reductases
1tcra-2-GJB	78.0	91.0	Unknown	Unknown
1sra-1-AS	67.5	151.0	All alpha	EF-hand
1regy-1-AUTO.1	64.1	120.0	Alpha and beta (a+b)	Ferredoxin-like
3mddb-1-AS	70.8	120.0	All alpha	Acyl-CoA dehydrogenase (flavoprotein), N-terminal domain
9insb	83.3	30.0	Small proteins	Insulin-like
1trkb-1-AS	79.0	329.0	Alpha and beta (a/b)	Thiamin-binding
1gog-2-AS.1	63.9	388.0	All beta	7-bladed beta-propeller
1comc-1-DOMAK	79.8	119.0	Alpha and beta (a+b)	Chorismate mutase
1vjs-3-GJB	80.9	89.0	All beta	alpha-Amylases, beta-sheet domain
2reb-2-DOMAK	64.4	59.0	Alpha and beta (a+b)	Anti-LPS factor/recA domain
1ecl-4-AS	80.3	117.0	All alpha	Winged DNA binding like
1lmb3	75.8	87.0	All alpha	lambda repressor-like DNA-binding domains
1rhgc-1-DOMAK	90.3	145.0	All alpha	4-helical cytokines
1ubdc-2-AS	62.0	29.0	Small proteins	Classic zinc finger
2gn5	64.3	87.0	All beta	OB-fold
2gcr	73.4	173.0	All beta	Crystallins/protein S
1oacb-3-AS.1	75.6	115.0	Alpha and beta (a+b)	Cystatin-like
1amg-2-AS	77.1	57.0	All beta	alpha-Amylases, beta-sheet domain
1bncb-1-AS	75.4	114.0	Alpha and beta (a/b)	Biotin carboxylase N-terminal domain-like
2asr-1-DOMAK	86.6	142.0	All alpha	Four-helical up-and-down bundle
2hhmb-1-DOMAK	60.5	142.0	Alpha and beta (a+b)	Sugar phosphatases alpha+beta N terminal domain
1fbab-1-DOMAK	80.8	360.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
5er2e	70.0	330.0	All beta	Acid proteases Domain I
1ctu-1-AUTO.1	59.7	164.0	Alpha and beta (a/b)	Cytidine deaminase
1lbu-1-AS	82.9	82.0	Unknown	Unknown

1pii-2-DOMAK	73.3	191.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
1cbg-1-AS	72.8	490.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
1powb-3-DOMAK	75.2	190.0	Alpha and beta (a/b)	Thiamin-binding
1fdx	70.3	54.0	Alpha and beta (a+b)	Ferredoxin-like
1horb-1-AUTO.1	76.3	266.0	Alpha and beta (a/b)	Glucosamine 6-phosphate deaminase
2spt-1-DOMAK	75.4	53.0	Small proteins	Kringle modules
3ecab-1-AS	72.6	212.0	Alpha and beta (a/b)	Glutaminase/Asparaginase Domain I
1aorb-3-AS	60.0	185.0	All alpha	Aldehyde ferredoxin oxidoreductase, C-terminal domain
1cxsa-4-AUTO.1	74.6	158.0	Unknown	Unknown
3pgk-2-AS	71.4	210.0	Alpha and beta (a/b)	Phosphoglycerate kinase Domain II
1lbu-2-AS	71.7	131.0	Unknown	Unknown
1hcra-1-DOMAK	82.6	52.0	All alpha	DNA-binding 3-helical bundle
1sfe-1-AS	74.3	78.0	Unknown	Unknown
1umub-1-AS	69.2	104.0	Unknown	Unknown
3gapa	71.6	208.0	All beta	Double-stranded beta-helix, jelly-roll domain
1rvvz-1-AUTO.1	79.2	154.0	Unknown	Unknown
1znbb-1-AS	74.7	230.0	Unknown	Unknown
1pda-2-AS	73.5	102.0	Alpha and beta (a/b)	Periplasmic binding protein-like II Domain II
4sgbi	80.3	51.0	Small proteins	Ovomucoid/PCI-like inhibitors
1oxy-3-AS	75.4	228.0	All alpha	Hemocyanin, middle domain II
1hnf-1-AS	45.5	101.0	All beta	Immunoglobulin-like beta-sandwich
1ese-1-AUTO.1	67.5	302.0	Alpha and beta (a/b)	Flavodoxin-like
1otgc-1-AS	58.4	125.0	Alpha and beta (a+b)	Tautomerase/MIF
1ptr-1-AUTO.1	64.0	50.0	Small proteins	Protein kinase cystein-rich domain (cys2)
8adh	72.4	374.0	All beta	GroES-like
1qrdb-1-AUTO.1	66.3	273.0	Unknown	Unknown
1oacb-2-AS.1	69.7	99.0	Alpha and beta (a+b)	Cystatin-like
1gep-3-AS	60.1	148.0	Unknown	Unknown
1grj-1-AS	77.0	74.0	All alpha	Long alpha-hairpin
1gym-1-AUTO.1	72.6	296.0	Unknown	Unknown
1dlc-3-AS.1	74.1	197.0	All beta	beta-Prism I
6hir	83.6	49.0	Small proteins	Thrombin inhibitors
1jud-1-GJB	79.5	220.0	Unknown	Unknown
1find-2-AUTO.1	77.0	122.0	Unknown	Unknown
1pbwb-1-AS	70.7	195.0	Unknown	Unknown
1rhd	76.1	293.0	Alpha and beta (a/b)	Rhodanese
1lba-1-DOMAK	73.9	146.0	Alpha and beta (a+b)	Bacteriophage T lysozyme (Zn amidase)
1seib-1-AUTO.1	75.3	73.0	Unknown	Unknown
1hplb-1-AS	60.6	338.0	Alpha and beta (a/b)	alpha/beta-Hydrolases
1qbb-3-AUTO.1	69.3	483.0	Unknown	Unknown
1nar-1-DOMAK	64.3	289.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
1reqc-2-AS	75.8	506.0	Unknown	Unknown
1smnb-1-AUTO.1	64.3	241.0	Alpha and beta (a+b)	Endonuclease
1dik-3-AS.1	59.7	144.0	Alpha and beta (a/b)	The "swivelling" beta/beta/alpha domain
1gmpb-1-DOMAK	76.0	96.0	Alpha and beta (a+b)	Microbal ribonucleases
2olba-3-AS	81.4	216.0	Alpha and beta (a/b)	Periplasmic binding protein-like II Domain II
1edd-1-DOMAK	67.1	310.0	Alpha and beta (a/b)	alpha/beta-Hydrolases
1gd1o	72.1	334.0	Alpha and beta (a/b)	NAD(P)-binding Rossmann-fold domains
1daab-1-AS	73.9	119.0	Alpha and beta (a+b)	D-amino acid aminotransferase Domain I
5ldh	67.5	333.0	Alpha and beta (a/b)	NAD(P)-binding Rossmann-fold domains
1tie-1-DOMAK	78.3	166.0	All beta	beta-Trefoil
1spbp-1-AS	64.7	71.0	Alpha and beta (a+b)	Ferredoxin-like

1pyta-1-AS	77.6	94.0	Alpha and beta (a+b)	Ferredoxin-like
2glsa	74.5	468.0	Alpha and beta (a+b)	Glutamine synthetase smaller domain
1ppi-2-AS	67.7	93.0	All beta	alpha-Amylases, beta-sheet domain
1gal-2-AS	69.8	116.0	All alpha	Inserted domain into FAD (also NAD)-binding motif for Glucose oxidase
1trh-1-AS	61.8	534.0	Alpha and beta (a/b)	alpha/beta-Hydrolases
1crn	41.3	46.0	Small proteins	Crambin-like
1gflb-1-AS	59.5	230.0	Unknown	Unknown
1gtqb-1-AUTO.1	71.7	138.0	Alpha and beta (a+b)	Tetrahydrobiopterin biosynthesis enzymes
1ignb-2-GJB	71.7	92.0	Unknown	Unknown
1mjc-1-DOMAK	81.1	69.0	All beta	OB-fold
3pgm	73.0	230.0	Alpha and beta (a/b)	Phosphoglycerate mutase-like
1udh-1-AUTO.1	75.0	228.0	Alpha and beta (a/b)	Uracil-DNA glycosylase
4pfk	79.3	319.0	Alpha and beta (a/b)	Phosphofructokinase Domain I
1gcb-2-AS	79.9	204.0	Alpha and beta (a+b)	Cysteine proteinases Domain II
1inp-1-AS.1	66.4	247.0	Alpha and beta (a+b)	Sugar phosphatases alpha+beta N terminal domain
1eceb-1-AUTO.1	72.6	358.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
1efud-2-AUTO.1	79.7	89.0	Unknown	Unknown
2gbp	78.6	309.0	Alpha and beta (a/b)	Periplasmic binding protein-like I Domain I
1qbb-1-AUTO.1	72.0	154.0	Unknown	Unknown
2dkb-2-AS	75.7	264.0	Alpha and beta (a/b)	PLP-dependent transferases Domain II
2reb-1-DOMAK	70.4	220.0	Alpha and beta (a+b)	P-loop containing nucleotide triphosphate hydrolases, small a+b insert
1inp-2-AS.1	52.9	153.0	Alpha and beta (a/b)	Sugar phosphatases alpha/beta C terminal domain
1tfr-1-GJB	53.3	283.0	Unknown	Unknown
1bbpa	74.5	173.0	All beta	Lipocalins
1scue-3-AS	84.5	149.0	Alpha and beta (a/b)	Flavodoxin-like
1lpba-1-DOMAK	48.2	85.0	Small proteins	Small inhibitors, toxins, lectins
1azu	73.8	126.0	All beta	Cupredoxins
1kte-1-AS	75.2	105.0	Alpha and beta (a/b)	Thioredoxin-like
2mtac-1-AS	70.7	147.0	All alpha	Cytochrome c
3cox-1-AS.1	69.7	314.0	Alpha and beta (a/b)	FAD (also NAD)-binding motif
2phy-1-GJB	54.4	125.0	Alpha and beta (a+b)	Profilin-like
4sdha	82.0	145.0	All alpha	Globin-like
7icd	74.8	414.0	Alpha and beta (a/b)	Isocitrate & isopropylmalate dehydrogenases
3cox-2-AS.1	67.2	186.0	Alpha and beta (a+b)	FAD-linked reductases, C-terminal domain
1fua-1-AUTO.1	77.6	206.0	Unknown	Unknown
1rec-2-DOMAK	68.6	102.0	All alpha	EF-hand
1scue-2-AS	80.2	81.0	Alpha and beta (a+b)	ATP-grasp sub-domain II
1stme-1-AUTO.1	70.2	141.0	Unknown	Unknown
1mdaj-1-GJB	61.4	342.0	All beta	7-bladed beta-propeller
2ltna	81.2	181.0	All beta	ConA-like lectins/glucanases
1bdo-1-AS	70.0	80.0	Unknown	Unknown
1nox-1-GJB	77.0	200.0	Unknown	Unknown
1ovb-1-GJB	66.0	159.0	Alpha and beta (a/b)	Periplasmic binding protein-like II Domain II
1irk-1-AS	72.7	99.0	Alpha and beta (a+b)	Protein kinases (PK), catalytic core N terminal Domain
6tmne	58.5	316.0	Alpha and beta (a+b)	Metzincins, catalytic (N-terminal) domain
2fox	78.9	138.0	Alpha and beta (a/b)	Flavodoxin-like
2admb-1-AUTO.1	65.2	216.0	Unknown	Unknown
1gog-3-AS.1	79.5	98.0	All beta	Immunoglobulin-like beta-sandwich
1hnf-2-AS	73.0	78.0	All beta	Immunoglobulin-like beta-sandwich
2dnja-1-AS	77.0	253.0	Alpha and beta (a+b)	DNase I-like
1dupa-1-AS	68.3	136.0	All beta	beta-Clip

2olba-2-AS	58.0	136.0	Alpha and beta (a+b)	Phosphate binding protein-like inserted domain
1csmb-1-AUTO.1	76.1	252.0	All alpha	Chorismate mutase II
3tima	77.5	249.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
2i1b	69.2	153.0	All beta	beta-Trefoil
1hmpb-1-AUTO.1	76.0	209.0	Alpha and beta (a/b)	Phosphoribosyltransferases (PRTases)
1tif-1-AS	77.6	76.0	Alpha and beta (a+b)	beta-Grasp
2tmdb-3-AS	75.6	152.0	Alpha and beta (a/b)	FAD (also NAD)-binding motif
1vokb-1-AS	73.9	188.0	Unknown	Unknown
1bmv2	64.4	374.0	All beta	Viral coat and capsid proteins
3hmga	59.4	328.0	All beta	Segmented RNA-genome viruses' proteins
4rhv3	77.9	236.0	All beta	Viral coat and capsid proteins
1mmoh-1-AS	76.5	162.0	All alpha	Methane monooxygenase hydrolase, gamma subunit
1nlkl-1-DOMAK	74.1	143.0	Alpha and beta (a+b)	Ferredoxin-like
1mof-1-AS	69.8	53.0	Unknown	Unknown
1ndh-1-AS	76.4	123.0	All beta	Reductase/elongation factor common domain
1tsp-1-AS	55.1	544.0	All beta	Single-stranded right-handed beta-helix
1dar-3-AS	42.8	35.0	Unknown	Unknown
1sfe-2-AS	77.0	87.0	Unknown	Unknown
2wrpr	79.8	104.0	All alpha	Trp repressor
1taq-2-AS	44.9	69.0	Unknown	Unknown
1brse-1-DOMAK	70.9	86.0	Alpha and beta (a/b)	Barstar (barnase inhibitor)
1krcb-1-AS	76.7	86.0	All beta	beta-Clip
2hft-2-AS	68.9	103.0	All beta	Immunoglobulin-like beta-sandwich
6cts	77.6	429.0	All alpha	Citrate synthase Domain I
4gr1	70.0	461.0	Alpha and beta (a/b)	FAD (also NAD)-binding motif
1delb-2-AUTO.1	62.1	119.0	Unknown	Unknown
1hslb-2-DOMAK	65.6	102.0	Alpha and beta (a/b)	Periplasmic binding protein-like II Domain II
2bopa-1-DOMAK	60.0	85.0	Alpha and beta (a+b)	Ferredoxin-like
2phh	62.9	391.0	Multi-domain (alpha and beta)	p-Hydroxybenzoate hydroxylase as a single domain
2sodb	78.1	151.0	All beta	Immunoglobulin-like beta-sandwich
1qbb-4-AUTO.1	76.1	67.0	Unknown	Unknown
1alkb-1-AS	63.4	449.0	Alpha and beta (a/b)	Alkaline phosphatase
1aozb-3-AS	64.3	216.0	All beta	Cupredoxins
2cmd-2-GJB	75.9	166.0	Alpha and beta (a+b)	Lactate & malate dehydrogenases, C-terminal domain
2afnc-2-AUTO.1	64.8	182.0	Unknown	Unknown
1nbac-1-AS	74.3	214.0	Alpha and beta (a/b)	N-carbamoylsarcosine amidohydrolase
2rspa	68.7	115.0	All beta	Acid proteases Domain I
1oacb-1-AS.1	52.4	82.0	Alpha and beta (a+b)	Copper amino oxidase, domain 1
1vmob-1-AS	73.0	163.0	All beta	beta-Prism I
1pmi-2-GJB	79.8	114.0	Unknown	Unknown
3ecab-2-AS	79.8	114.0	Alpha and beta (a/b)	Glutaminase/Asparaginase Domain II
1amp-1-AS	74.5	291.0	Alpha and beta (a/b)	Zn-dependent exopeptidases
2yhx-3-DOMAK	48.8	129.0	Alpha and beta (a/b)	Ribonuclease H-like motif
6acn	71.2	753.0	Alpha and beta (a/b)	Aconitase, Domain I
1mdam-1-DOMAK	64.2	112.0	Small proteins	Methylamine dehydrogenase, L-chain
3chy-1-DOMAK	83.5	128.0	Alpha and beta (a/b)	Flavodoxin-like
1hplb-2-AS	72.0	111.0	All beta	Colipase binding domain-like
3pmgb-4-AS	71.8	142.0	Alpha and beta (a+b)	TBP-like
1bfg-1-DOMAK	65.0	126.0	All beta	beta-Trefoil
1lki-1-AS	68.0	172.0	All alpha	4-helical cytokines
1vcc-1-AS	83.1	77.0	Alpha and beta (a+b)	A DNA topoisomerase I domain

2stv	67.9	184.0	All beta	Viral coat and capsid proteins
1gp1a	67.7	183.0	Alpha and beta (a/b)	Thioredoxin-like
2end-1-DOMAK	75.1	137.0	All alpha	T endonuclease V
9pap	65.0	212.0	Alpha and beta (a+b)	Cysteine proteinases Domain I
1dik-1-AS.1	70.5	241.0	Alpha and beta (a+b)	ATP-grasp sub-domain I
1dfnb-1-DOMAK	83.3	30.0	Small proteins	Defensin-like
1fdt-1-AS	72.6	285.0	Alpha and beta (a/b)	NAD(P)-binding Rossmann-fold domains
1nozb-2-AUTO.1	71.5	225.0	Unknown	Unknown
1paz	76.6	120.0	All beta	Cupredoxins
3ait	70.2	74.0	All beta	alpha-Amylase inhibitor
1dik-4-AS.1	59.3	354.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
1pyp	74.6	280.0	All beta	OB-fold
1lis-1-DOMAK	68.7	131.0	All alpha	Lysin
1tndb-2-DOMAK	80.1	116.0	All alpha	Transducin (alpha subunit), insertion domain
1daab-2-AS	77.2	158.0	Alpha and beta (a+b)	D-amino acid aminotransferase Domain II
1vhh-1-AS	76.4	157.0	Alpha and beta (a+b)	Hedgehog/DD-peptidase
1rlds-1-DOMAK	84.2	114.0	Alpha and beta (a+b)	RuBisCO, small subunit
1fjmb-2-AS	81.9	111.0	Unknown	Unknown
1rec-1-DOMAK	78.3	83.0	All alpha	EF-hand
1cqa-1-AUTO.1	82.9	123.0	Alpha and beta (a+b)	Profilin-like
1eft-3-DOMAK	72.6	95.0	All beta	Elongation factor Tu (EF-Tu), the C-terminal domain
1thx-1-AUTO.1	71.3	108.0	Alpha and beta (a/b)	Thioredoxin-like
3hmgb	67.4	175.0	Membrane and cell surface proteins and peptides	Influenza hemagglutinin (stalk)
1bet-1-DOMAK	63.5	107.0	Small proteins	Cystine-knot cytokines
2cyp	61.4	293.0	All alpha	Heme-dependent peroxidases Domain I
1ceo-2-AUTO.1	33.9	53.0	All alpha	small domain attached to TIM barrel
1bmv1	74.5	185.0	All beta	Viral coat and capsid proteins
1cksc-1-AUTO.1	64.1	78.0	Alpha and beta (a+b)	Cell cycle regulatory proteins
4bp2	64.9	117.0	All alpha	Phospholipase A2
1tul-1-JAC	57.8	102.0	Unknown	Unknown
1dfji-1-AUTO.1	55.0	456.0	Unknown	Unknown
1yrna-2-AS	84.1	63.0	All alpha	DNA-binding 3-helical bundle
1bam-1-AS	60.0	200.0	Alpha and beta (a/b)	Restriction endonucleases
1trkb-3-AS	78.1	137.0	Alpha and beta (a/b)	Transketolase, C-terminal domain
4ts1a	69.0	317.0	Alpha and beta (a/b)	ATP pyrophosphatases
1gtmc-2-AUTO.1	61.9	134.0	Unknown	Unknown
1tssb-2-DOMAK	64.3	73.0	All beta	OB-fold
1hip	58.8	85.0	Small proteins	HIPIP (high potential iron protein)
1mrrb-1-DOMAK	76.4	340.0	All alpha	Ferritin like
1aozb-2-AS	68.9	206.0	All beta	Cupredoxins
2admb-2-AUTO.1	53.8	169.0	Unknown	Unknown
1cdta	75.0	60.0	Small proteins	Snake toxin-like
1tiic-1-GJB	66.6	36.0	Peptides	Antifreeze polypeptide HPLC-6
9apib	86.1	36.0	Multi-domain (alpha and beta)	Serpins
2mev4	46.5	58.0	All beta	Viral coat and capsid proteins
1gpmd-4-AS	66.5	206.0	Alpha and beta (a/b)	Class I glutamine amidotransferases
1han-2-AUTO.1	69.0	155.0	Alpha and beta (a+b)	2,3-Dihydroxybiphenyl dioxygenase (DHDB, BPHC enzyme)
1pkyc-3-AUTO.1	76.6	120.0	Alpha and beta (a/b)	Pyruvate kinase, C-terminal domain
4rxn	64.8	54.0	Small proteins	Rubredoxin-like
3cla	70.8	213.0	Alpha and beta (a/b)	CoA-dependent acetyltransferases
1edn-1-AS	57.1	21.0	Small proteins	Endothelin-like

2dln-1-AS	67.8	84.0	Alpha and beta (a/b)	Biotin carboxylase N-terminal domain-like
1cgu-4-GJB	77.8	104.0	All beta	Prealbumin-like
1chmb-1-DOMAK	61.9	155.0	Alpha and beta (a/b)	Ribonuclease H-like motif
1poc-1-DOMAK	50.7	134.0	All alpha	Phospholipase A2
2hipb-1-DOMAK	45.0	71.0	Small proteins	HIPIP (high potential iron protein)
1adeb-2-AUTO.1	78.0	100.0	All alpha	P-loop containing nucleotide triphosphate hydrolases all helical domain
4rhv4	50.0	40.0	All beta	Viral coat and capsid proteins
1add-1-AS	69.3	349.0	Alpha and beta (a/b)	beta/alpha (TIM)-barrel
1etu	73.4	177.0	Alpha and beta (a/b)	P-loop containing nucleotide triphosphate hydrolases
1pbp-2-DOMAK	55.1	176.0	Alpha and beta (a/b)	Periplasmic binding protein-like II Domain II
2scpb-1-DOMAK	66.6	174.0	All alpha	EF-hand
256ba	75.4	106.0	All alpha	Four-helical up-and-down bundle
1pdnc-2-AS	76.3	55.0	All alpha	DNA-binding 3-helical bundle
1colb-1-DOMAK	73.6	197.0	Membrane and cell surface proteins and peptides	Toxins' membrane translocation domains
1fbl-1-AS	69.7	175.0	Alpha and beta (a+b)	Metzincins, catalytic (N-terminal) domain
1bds	69.7	43.0	Small proteins	Defensin-like
2abk-2-AS	74.5	110.0	All alpha	Endonuclease III
1ahb-2-GJB	55.2	67.0	Alpha and beta (a+b)	Ribosome inactivating proteins (RIP) Domain II
1avhb-4-AS	67.5	74.0	All alpha	Annexin Domain
2bltb-2-AUTO.1	64.3	73.0	All alpha	beta-Lactamase/D-ala carboxypeptidasea inserted domain
1avhb-3-AS	76.7	86.0	All alpha	Annexin Domain
1clc-3-AS.1	68.0	200.0	All alpha	Glycosyltransferases of the superhelical fold Domain II
4cpai	75.6	37.0	Small proteins	Small inhibitors, toxins, lectins
1yptb-1-AUTO.1	57.1	280.0	Alpha and beta (a/b)	Phosphotyrosine protein phosphatases II
1bpha-1-DOMAK	57.1	21.0	Small proteins	Insulin-like
2hhmb-2-DOMAK	60.7	130.0	Alpha and beta (a/b)	Sugar phosphatases alpha/beta C terminal domain
1rlr-1-JAC	59.7	211.0	Unknown	Unknown
1whi-1-AS	66.3	122.0	Unknown	Unknown
1cdlg-1-DOMAK	75.0	20.0	Peptides	Simple helix
2tmvp	60.3	154.0	All alpha	Four-helical up-and-down bundle
1ctn-3-AS.1	60.2	73.0	Alpha and beta (a+b)	FKBP-like
1cbh	72.2	36.0	Small proteins	Small inhibitors, toxins, lectins
6rlxc-1-DOMAK	37.5	24.0	Small proteins	Insulin-like

Q_3 Values for each sequence with DSSP as the definition model (PHD only is rendered here)
Source: http://barton.ebi.ac.uk/

Appendix C

DESCRIPTIVE STATISTICS

Table C-1: Descriptive Statistics of the Q_3 for the five reduction methods

Method	Num of AA	Range	Min	Max	Mean	Mean Std. Error	Standard Deviation	Variance
Method I								
ALL	480	97.4	.0	97.4	79.876	0.462	10.1263	102.542
H	480	100.0	.0	100.0	77.418	1.211	26.5348	704.094
E	480	100.0	.0	100.0	69.494	1.252	27.4202	751.867
C	480	80.0	20.0	100.0	80.306	0.537	11.7696	138.523
Method II								
ALL	480	97.6	.0	97.6	80.491	0.466	10.2111	104.267
H	480	100.0	.0	100.0	77.403	1.211	26.5316	703.926
E	480	100	0	100	77.120	1.10	24.193	585.283
C	480	72.7	27.3	100.0	79.989	0.536	11.7515	138.098
Method III								
ALL	480	97.6	.0	97.6	80.484	0.466	10.2139	104.324
H	480	100.0	.0	100.0	77.418	1.211	26.5348	704.094
E	480	100	0	100	77.120	1.10	24.193	585.283
C	480	72.7	27.3	100.0	79.965	0.537	11.7748	138.646
Method IV								
ALL	480	98	0	98	80.38	0.45	9.788	95.802
H	480	100.0	.0	100.0	87.031	0.939	20.5739	423.285
E	480	100.0	.0	100.0	69.494	1.252	27.4202	751.867
C	480	80.0	20.0	100.0	78.339	0.538	11.7773	138.705
Method V								
ALL	480	98.4	.0	98.4	80.984	0.452	9.9042	98.094
H	480	100.0	.0	100.0	87.031	0.939	20.5739	423.285
E	480	100	0	100	77.12	1.10	24.193	585.283
C	480	72.7	27.3	100.0	78.067	0.537	11.7615	138.332

Table C-2: Descriptive Statistics of SOV measure for the five reduction methods

Method	Num of AA	Range	Min	Max	Mean	Mean Std. Error	Standard Deviation	Variance
Method I								
ALL	480	98.8	.0	98.8	75.830	0.747	16.3579	267.582
H	480	100.0	.0	100.0	77.982	1.229	26.9282	725.130
E	480	100	0	100	71.19	1.32	28.991	840.459
C	480	90.0	10.0	100.0	73.414	0.652	14.2813	203.956
Method II								
ALL	480	99.5	.0	99.5	76.265	0.799	17.4989	306.211
H	480	100.0	.0	100.0	77.955	1.229	26.9177	724.565
E	480	100.0	.0	100.0	79.938	1.122	24.5743	603.895
C	480	87.5	12.5	100.0	74.349	0.709	15.5282	241.125
Method III								
ALL	480	99.5	.0	99.5	76.248	0.800	17.5222	307.026
H	480	100.0	.0	100.0	77.982	1.229	26.9282	725.130
E	480	100.0	.0	100.0	79.938	1.122	24.5743	603.895
C	480	87.5	12.5	100.0	74.323	0.711	15.5726	242.507
Method IV								
ALL	480	99.3	.0	99.3	75.844	0.761	16.6689	277.851
H	480	100.0	.0	100.0	87.633	0.974	21.3347	455.168
E	480	100	0	100	71.19	1.32	28.991	840.459
C	480	90.0	10.0	100.0	72.693	0.677	14.8422	220.291
Method V								
ALL	480	99.5	.0	99.5	74.932	.857	18.7823	352.773
H	480	100.0	.0	100.0	87.633	.974	21.3347	455.168
E	480	100.0	.0	100.0	79.938	1.122	24.5743	603.895
C	480	82.6	17.4	100.0	72.503	.745	16.3328	266.761

Appendix D

SELECTED PUBLICATIONS

Abdalla, S. O. and Deris, S. (2005). Combining Artificial Neural Networks and GOR V Information Theory to Predict Protein Secondary Structure from Amino Acid Sequences. *International Journal of Intelligent Information Technologies, USA. (accepted, now waiting for copy right signature).*

Abdalla, S. O. and Deris, S. (2005). Predicting Protein Secondary Structure Using Artificial Neural Networks: Current Status and Future Directions *Information Technology Journal 4(2): 189-196,2005.*

Abdalla, S. O. and Deris, S. (2005). An Improved Method for Protein Secondary Structure Prediction by Combining Neural Networks and GOR V Theory. *Second Middle East Conference on Healthcare Informatics (MECHCI 2005).* Dubai Knowledge Village, Dubai 9-10 April 2005. UAE.

Abdalla, S. O, Deris, S. and Mohamad, M. S. (2005). A hybrid Classifier for Protein Secondary Structure Prediction. *Information Technology Journal 4 (3): 000-000,2005. (accepted, now under press)*

Abdalla, S. O. and Deris, S. (2005). Protein Secondary Structure Reduction Schemes Significantly Affect Prediction Accuracy. *2005 IEEE Computational Systems Bioinformatics Conference (CSB2005).* Stanford University, California 8-11 August 2005. USA (submitted)

Abdalla, S. O. and Deris, S. (2005). Blind Test is a Pragmatic Test for a New Protein Secondary Structure Classifier. *The 7th International Conference on BIOINFORMATICS (BIO 2005).* June 9-12, 2005 , Tartu, Estonia . (submitted)

Mohd Saberi Mohamad ,Safaai Deris, Saad Osman Abdalla, and Rosli Md Illias(2005). An Improved Hybrid Of Genetic Algorithm And Support Vector Machine For Gene Selection And Classification Of Gene Expression Data.. *Journal of Bioinformatics and Computational Biology (Submitted).*