



**Computer Science Department
Faculty of Computer and Information Sciences
Ain Shams University**

Intelligent Techniques for Protein Secondary Structure Prediction

Thesis submitted as a partial fulfillment of the requirements for the degree of
Master of Science in Computer and Information Sciences

By

Hanan Yousry Wahba Hendy

Teaching Assistant at Computer Science Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Under Supervision of

Prof. Dr. Abdel-Badeeh Mohamed Salem

Professor in Computer Science Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Prof. Dr. Mohamed Ismail Roushdy

Professor in Computer Science Department
and Dean of Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Wael Hamdy Khalifa

Assistant Professor in Computer Science Department,
Faculty of Computer and Information Sciences,
Ain Shams University

August – 2016
Cairo

Acknowledgment

First of all, I would like to thank God for his endless blessings, for giving me the power and strength to complete this work and for giving me people who kept on supporting me.

Second, I would like to express my sincere gratitude to my supervisors; Prof. Dr. Abdel-Badeeh Salem for his support, patience and guidance, Prof. Dr. Mohamed Roushdy for his encouragement and Dr. Wael Khalifa for the special supervision experience he gave me. I am deeply thankful.

Third, I would like to thank my family for being available all the time and for the love they gave me through the years. Thank you for accepting me through the tough times and for always believing in me.

My dear friends who have helped me through the past time and kept on encouraging me to get this work done; Kholoud Abdul Salam, Manal Mostafa, Yasmine Afify and Ghada Hamed, without you it would have been much harder.

Last but not least, I would like to thank all my professors, colleagues and students who kept on encouraging me. Thank you for being in my life.

Hanan

Abstract

Protein is considered the building block of any living organism. Protein performs various functions in the human body, these functions differ from one to another according to the way the protein bonds together. The protein is initially composed of a sequence of amino acids which are named as the primary structure. Then the protein forms its secondary, tertiary and quaternary structures by forming hydrogen bonds.

The primary structure can be extracted from raw protein using simple scientific experiments. There have been various amino acid sequences discovered through the years. However, the secondary structure sequences cannot be extracted in the same manner. Moreover, the diseases and protein disorders can be detected when examining the secondary structure not the primary one. That is why it is crucial to find a way to get the secondary structure of a given primary sequence. Prediction is considered a solution to this problem. Given only the knowledge of primary sequence, it is required to predict the corresponding secondary one.

Various machine learning techniques have been used through the last decade to try to predict the protein secondary structure. The most commonly used paradigm was the Artificial Neural Networks. Variations of ANN have been used to increase the protein secondary structure prediction accuracy. Then few used case based reasoning and mixed integer optimization.

This thesis presents a study on the different techniques used for protein secondary structure prediction. The techniques are divided into three generations starting with statistical generation and ending with Machine learning one.

Then the thesis discusses five different approaches that are used for predicting the protein secondary structure in detail along with their computation parameters. These approaches are: Case based reasoning, Artificial Neural Networks, Decision Tables, Decision Trees and Bayes Networks. Two different datasets are used with different sequence lengths and with proper distribution among different amino acids. In Case Based Reasoning, eight different experiments are conducted resulting in prediction accuracy of 88%. In ANN, one thousand twenty-four experiments are conducted using different computation parameters resulting in accuracy of 68%, 81% and 86% for predicting alpha, beta and alpha and beta together respectively.

Then for the statistical techniques, ZeroR is used to determine the baseline accuracy for the other three. Eight experiments are conducted for each of the Decision Tree, Decision Table and Bayes Network. The accuracies reach 70%, 71% and 75% respectively. Moreover, two ANN hybrid techniques are proposed to increase the prediction accuracy. The first predicts alpha and beta each alone using the same ANN and merges the result. The accuracy increased by 1-2%. The second, picks the best two ANNs and uses both for prediction and then merges the results. This increased the accuracy by about 2-3%. Finally, this thesis compares all the experiments and concludes the best among them. The discussed experiments reached a prediction accuracy of 88% for maximum and 75% on average.

List of Publications

- 1- Hanan Hendy, Wael Khalifa, Mohamed Roushdy and Abdel Badeeh Salem. "A study of intelligent techniques for protein secondary structure prediction." *Information Models and Analyses Journal*, vol.4, no. 1, pp 3-12, 2015.
- 2- Hanan Hendy, Wael Khalifa, Mohamed Roushdy and Abdel Badeeh Salem. "The usage of Neural Networks Paradigm in the prediction of protein secondary structure." *Proceedings of the International Conference on Communications and Computers - Recent Advances in Electrical Engineering*, pp 14-17, October 2015.
- 3- Hanan Hendy, Wael Khalifa, Mohamed Roushdy and Abdel-Badeeh M. Salem. "The Usage of Machine Learning Paradigms on Protein Secondary Structure Prediction". *International Journal of Circuits and Electronics*, vol.1, no.1, pp 72-77, 2016.
- 4- Hanan Hendy, Wael Khalifa, Mohamed Roushdy and Abdel-Badeeh M. Salem. "The effect of using different Neural Networks architectures on the protein secondary structure prediction". *Egyptian Computer Science Journal*, vol.40, no. 3, pp 58-71, September 2016.

Submitted:

- 5- Hanan Hendy, Wael Khalifa, Mohamed Roushdy and Abdel-Badeeh M. Salem. "Protein Eight Secondary Structure Classes Prediction Using Artificial Neural Networks". *International Journal of Genomics, Proteomics, Metabolomics & Bioinformatics (IJGPMB)*, USA. 2016

Table of Contents

Acknowledgment	II
Abstract	III
List of Publications	V
Table of Contents	VI
List of Figures	VIII
List of Tables	XI
List of Abbreviations	XIV
Chapter 1. Introduction	2
1.1 Overview	2
1.2 Motivation	4
1.3 Objectives	5
1.4 Methodology	5
1.5 Contributions	6
1.6 Thesis Organization	7
Chapter 2. Biological Background	10
2.1 Protein Different Structures	11
2.1.1 Primary Structure	12
2.1.2 Secondary Structure	14
2.1.3 Tertiary Structure	15
2.1.4 Quaternary Structure	16
2.2 Protein Available Software	16
Chapter 3. Protein Secondary Structure Prediction Statistical and Intelligent Techniques	21
3.1 Secondary Structure Prediction Generations	22
3.1.1 Statistical Generation	24
3.1.2 Enhanced Statistical Generation	25
3.1.3 Machine Learning Generation	25
3.1.3.1 Artificial Neural Networks	25
3.1.3.2 Case Based Reasoning	29
3.1.3.3 Mixed Integer Linear Optimization	30
3.1.3.4 Swarm Intelligence	30
3.1.3.5 Hybrid Techniques	31
Chapter 4. Protein Data Preprocessing	35
4.1 Datasets	38
4.2 Data Preprocessing Stages	44
4.2.1 Raw Files Handling	46
4.2.2 Encode Data	47
4.2.3 Split Data	48

	4.2.4 Format Data	49
Chapter 5.	Case Based Reasoning Usage in Protein Secondary Structure	
Prediction	54
5.1	CBR Tool	56
5.2	Implementation	59
5.3	Results and Discussion.....	61
Chapter 6.	Artificial Neural Networks Usage in Protein Secondary	
Structure Prediction	64
6.1	Artificial Neural Network Tool.....	67
6.2	Implementation	71
6.3	Results and Discussion.....	73
	6.3.1 Experiment 1: Single ANN to Predict Three Main	
	Secondary Structure Classes	73
	6.3.2 Experiment 2: Hybrid ANN to Predict Three Main	
	Secondary Structure Classes	87
	6.3.3 Experiment 3: Single ANN to Predict Eight Secondary	
	Structure Classes	88
Chapter 7.	Statistical Techniques Usage in Protein Secondary Structure	
Prediction	94
7.1	Implementation and Results.....	98
	7.1.1 ZeroR	98
	7.1.2 Bayes Network.....	101
	7.1.3 Decision Table	104
	7.1.4 Decision Tree	106
Chapter 8.	Conclusion and Future Work	111
Appendix A:	<i>MATLAB</i> ANN Toolbox Walkthrough.....	115
Appendix B:	ANN Additional Experiment	121
References.....		129

List of Figures

Figure 1-1 Basic example for protein different structures.....	3
Figure 1-2 Real examples for protein different structures.....	4
Figure 2-1 Twenty common amino acids	10
Figure 2-2 Venn diagram of boundaries that symbolizes the universal set of 20 common amino acids.....	11
Figure 2-3 Four levels of protein structures	12
Figure 3-1 Feedforward neural network model.....	26
Figure 4-1 Distribution of all primary structures in the used datasets	42
Figure 4-2 Distribution of combined primary structures in the used datasets	42
Figure 4-3 Distribution of all secondary structures in the used datasets.....	43
Figure 4-4 Distribution of combined secondary structures in the used datasets	43
Figure 4-5 Data preprocessing stages	45
Figure 4-6 Dataset raw files samples.....	46
Figure 4-7 Data preprocessing stages example	50
Figure 4-8 Matrix file example.....	51
Figure 4-9 CSV file example.....	51
Figure 4-10 Data preprocessing stages example	52
Figure 5-1 CBR R-4 Cycle	55
Figure 5-2 myCBR workbench.....	58
Figure 5-3 CBR pseudocode.....	60
Figure 6-1 Biological vs Artificial neuron	64
Figure 6-2 Multilayer ANN	65
Figure 6-3 Sigmoid function.....	66

Figure 6-4 Master ANN script.....	68
Figure 6-5 Batch ANN script.....	69
Figure 6-6 Generic ANN script	70
Figure 6-7 Feedforward ANN prediction accuracy using numeric encoding and predicting alpha only	74
Figure 6-8 Feedforward ANN prediction accuracy using binary encoding and predicting alpha only	75
Figure 6-9 Feedforward ANN prediction accuracy using numeric encoding and predicting beta only	76
Figure 6-10 Feedforward ANN prediction accuracy using binary encoding and predicting beta only	77
Figure 6-11 Feedforward ANN prediction accuracy using numeric encoding and predicting alpha, beta and coil.....	78
Figure 6-12 Feedforward ANN prediction accuracy using binary encoding and predicting alpha, beta and coil.....	79
Figure 6-13 Feedforward ANN prediction accuracy variation between numeric and binary encoding using 75% training and no ambiguous amino acids	82
Figure 6-14 Feedforward ANN prediction accuracy variation between numeric and binary encoding using 80% training and no ambiguous amino acids	83
Figure 6-15 Feedforward ANN prediction accuracy variation among numeric and binary encoding using 70% training with ambiguous amino acids	84
Figure 6-16 Feedforward ANN prediction accuracy variation among numeric and binary encoding using 80% training with ambiguous amino acids	85

Figure 6-17 Combines ANN used in prediction	87
Figure 6-18 ANN compared prediction accuracies when predicting 8 classes without post processing.....	90
Figure 6-19 ANN compared prediction accuracies when predicting 8 classes with post processing	91
Figure 7-1 Weka explorer	95
Figure 7-2 Weka experiment sample	96
Figure 7-3 Wrong learning process methodology	97
Figure 7-4 Learning cyclic process.....	97
Figure 7-5 ZeroR pseudocode.....	99
Figure 7-6 Decision table structure.....	104
Figure 7-7 Greedy decision tree pseudocode	107
Figure A- 1 Import input matrix to ANN	115
Figure A- 2 Import output matrix to ANN	116
Figure A- 3 Divide samples to training, validation and testing.....	116
Figure A- 4 Choose architecture.....	117
Figure A- 5 Training process	118
Figure A- 6 Ready network model diagram	118
Figure A- 7 The model after training.....	119

List of Tables

Table 2-1 Amino acids 1 and 3 letter codes	13
Table 2-2 Ambiguous amino acids	14
Table 2-3 Secondary structures classes	15
Table 2-4 Protein structures available software	17
Table 3-1 Three protein secondary structure prediction generations	23
Table 3-2 Comparison of main secondary structure prediction techniques .	32
Table 4-1 Protein databanks and datasets	36
Table 4-2 Datasets amino acid length distribution	39
Table 4-3 Occurrences of amino acids in the used datasets	39
Table 4-4 Occurrences of secondary structures in the used datasets.....	41
Table 4-5 Primary sequence encoding.....	47
Table 4-6 Secondary sequence encoding.....	48
Table 5-1 CBR prediction accuracy	61
Table 6-1 Feedforward ANN prediction accuracy using numeric encoding and predicting alpha only	74
Table 6-2 Feedforward ANN prediction accuracy using binary encoding and predicting alpha only	75
Table 6-3 Feedforward ANN prediction accuracy using numeric encoding and predicting beta only	76
Table 6-4 Feedforward ANN prediction accuracy using binary encoding and predicting beta only	77
Table 6-5 Feedforward ANN prediction accuracy using numeric encoding and predicting alpha, beta and coil.....	78
Table 6-6 Feedforward ANN prediction accuracy using binary encoding and predicting alpha, beta and coil.....	79

Table 6-7 Feedforward ANN prediction accuracy (no ambiguous amino acids) 75% training.....	81
Table 6-8 Feedforward ANN results using feedforward (no ambiguous amino acids) 80% training.....	82
Table 6-9 Feedforward ANN prediction accuracy using feedforward (with ambiguous amino acids) 75% training	83
Table 6-10 Feedforward ANN prediction accuracy using feedforward (with ambiguous amino acids) 80% training	85
Table 6-11 Hybrid ANN prediction accuracy	88
Table 6-12 ANN prediction accuracy when predicting 8 secondary structure classes	89
Table 7-1 ZeroR prediction accuracy	100
Table 7-2 Bayes network prediction accuracy	103
Table 7-3 Decision table prediction accuracy	105
Table 7-4 Decision tree prediction accuracy	108
Table B- 1 Patternet ANN results using numeric encoding and predicting alpha only	121
Table B- 2 Patternet ANN results using binary encoding and predicting alpha only	122
Table B- 3 Patternet ANN results using numeric encoding and predicting beta only	122
Table B- 4 Patternet ANN results using binary encoding and predicting beta only	123
Table B- 5 Patternet ANN results using numeric encoding and predicting alpha, beta and coil	123
Table B- 6 Patternet ANN results using binary encoding and predicting alpha, beta and coil	124

Table B- 7 Patternet ANN results (no ambiguous amino acids) 75% training	124
Table B- 8 Patternet ANN results (no ambiguous amino acids) 80% training	125
Table B- 9 Patternet ANN results (with ambiguous amino acids) 70% training	126
Table B- 10 Patternet ANN results (with ambiguous amino acids) 80% training	127

List of Abbreviations

<u>Abbreviation</u>	<u>Stands for</u>
ABC	: Artificial B ee Colony
ANN	: Artificial Neural Networks
BMRB	: B iological M agnetic R esonance Data B ank
CATH	: C lass, A rchitecture, T opology, H omologous superfamily database
CB513	: C uff and B arton data set
CBR	: C ase B ased R easoning
CSV	: C omma S eparated V alues (File Format)
DAG	: D irected A cyclic G raphs
DNA	: D eoxyribo N ucleic A cid
DSC	: D iscrimination of protein S econdary structure C lass
DSSP	: D atabase of S econdary S tructure assignments for entries in the P rotein Data B ank
EMBOSS	: The E uropean M olecular B iology O pen S oftware Suite
EVA	: E Valuation of A utomatic protein structure prediction
FSSP	: F amilies of S tructurally S imilar P roteins
GOR	: G arnier- O sguthorpe- R obson method
GUI	: G raphical U ser I nterface
HSSP	: H omology-derived S tructures of P roteins
IDE	: I ntegrated D evelopment E nvironment
ML	: M achine L earning
NMR	: N uclear M agnetic R esonance

PDB	:	P rotein D ata B ank
PDBe	:	P rotein D ata B ank E urope
PDBj	:	P rotein D ata B ank J apan
PHD	:	P rofile network from H ei D elberg
RCSB	:	R esearch C ollaboratory for S tructural B ioinformatics
RNA	:	R ibonucleic A cid sequence
SDK	:	S oftware D evelopment K it
SVM	:	S upport V ector M achine
WEKA	:	W aikato E nvironment for K nowledge A nalysis
WS	:	W indow S ize
wwPDB	:	W orldwide P rotein D ata B ank
ZeroR	:	Z eros of a R eal polynomial