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 to 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

List of Figures

List of Tables

List of Abbreviations

Abbreviation		Stands for
ABC	:	**A**rtificial **B**ee **C**olony
ANN	:	**A**rtificial **N**eural **N**etworks
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 Bank
EMBOSS	:	The **E**uropean **M**olecular **B**iology **O**pen **S**oftware **S**uite
EVA	:	**EV**aluation 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**ibo**n**ucleic **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**orld**w**ide **P**rotein **D**ata **B**ank
ZeroR	:	**Z**eros of a **R**eal polynomial