Universiti Teknologi MARA

Modelling Protein Prediction Using Data Mining Techniques

Nor Adila Mohd Zawawi

Thesis submitted in fulfillment of the requirements for
Bachelor of Science (Hons) Intelligent System
Faculty of Information Technology And Quantitative Science

MAY 2007

ACKNOWLEDGEMENTS

With the name of Allah the most Gracious, the most Merciful creator,
I seek His Blessing on His Prophet Muhammad S.A.W.

Alhamdulillah... All praise and glory be to Allah S.W.T whose infinite generosity has given me the strength to complete this final project in time.

My sincere gratitude, thanks and most appreciation goes to Puan Shuzlina Binti Abdul Rahman as my supervisor, thesis coordinator and also course tutor for her resourceful help, guidance, constant encouragement, comments and ultimately enthusiasm. She is such a great person who has paved the way for me throughout the overall research project. Without her great patience, this work would never have been finished.

I would also like to express my grateful thanks to all my colleagues for their opinion, suggestion and cooperation they gave. Also thanks for their morale support during the months I spent preparing this final project.

To my beloved parents and family, who are always there for me whenever I need them and million of thanks for all the supports, blessing, loves and financial support they give to me. Finally, to whom I failed to mention, who directly contributed to this project. Thank you very much.

ABSTRACT

Currently, many machine learning methods are employed in protein secondary structure prediction that includes neural network, support vector machine, bayesian segmentation, genetic algorithm and others. However, researchers have difficulty to determine and identify the right machine learning methods to be used because of the complex structure of protein which are primary, secondary, tertiary and quaternary structure. This is due to the different characteristics of protein which are structure, function, charge, acidity, hydrophilicity and molecular weight and the different type of shape of the protein which are alpha helices, beta strands and coils. Hence, researcher would like to propose a model that attempts to assist other researchers in order to choose a suitable classification technique to be used for protein structure prediction. The analysis is derived by excavating literature on protein secondary structure prediction starting from the year 1987 until 2006, focused on supervised learning algorithm. The model demonstrates general flow for protein secondary structure from sequence to structure (Q2T), structure to structure (T2T) and reliability index. As a result, a model of data mining techniques used in protein secondary structure prediction is built. Therefore, future work can be tackled by using more past work that had been applied in data mining techniques, other than protein structure but in other tasks like function prediction, location prediction, protein interaction and protein annotation, in order to get a better view of it.

TABLE OF CONTENTS **PAGE**

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

This chapter views the background of the study which is an introduction of the research. Then it is followed by the problem statement, objectives, scope and significances of the study. This chapter is an important part for the understanding of the project.

1.2 BACKGROUND OF THE RESEARCH

Data in the real world are often dynamic, incomplete which are lacking attribute values, lacking certain attributes of interest, or containing only aggregate data, containing errors or outliers that is known as noisy data (Han and Kamber, 2001). Data are very large and have a big scope. Sometimes it can be inconsistent because of the unknown data and the problem that occur from the data itself.

Bioinformatics approaches are often used for major initiatives that generate large data sets. Proteomics is one of the domains in bioinformatics. In the proteomic domain, the main application of computational methods is protein structure prediction (Larranaga *et al.*, 2005). Proteins are large biological molecules that are very hard to solve because of it complex structures which are primary, secondary, tertiary and quaternary structure, the different characteristics of protein which are structure, function, charge, acidity, hydrophilicity and molecular weight and the different type of shape of the proteins which are alpha helices, beta strands and coils.