UNIVERSITI TEKNOLOGI MARA

CLASSIFICATION OF ZINC HYDROLASES USING A SEQUENCE-BASED PHYLOGENETIC METHOD

NUR SYAMIM ISMAIL

Written Thesis Submitted to
Universiti Teknologi MARA

Dissertation submitted in partial fulfillment of the requirement for the degree of

Bachelor of Pharmacy

Faculty of Pharmacy

JULY 2014

Acknowledgements

I would like to express my greatest gratitude for my supervisor; Madam Yuslina bt. Zakaria who has given me guidance and knowledge. I would like to thank Madam Zafirah Liyana Abdullah for her help and advice for this research.

My completion of this research could not been accomplished without the support of my family especially my beloved parents, Ismail b. Yusoff and Rogayah bt. Hashim for their support and encouragement.

Thank You.

TABLE OF CONTENT

Abstract

There are many methods in classifying proteins. Some classify proteins manually; some classify proteins using computational method whereas some classify proteins by combining both methods. In this research, the result of protein classifications using sequence-based phylogenetic method is compared with manual method and a previous work of classifying proteins. The objectives of this research are to assess the similarities of sequence-based phylogenetic tree base on the sequence alignment method; to construct a phylogenetic tree using similarity values from protein alignment; and to validate the phylogenetic tree against standard protein classifications. The dataset is obtained from Protein Data Bank (PDB) and the proteins sequences are aligned together using CLUSTALW. Then, a sequence-based phylogenetic tree is constructed and validated using PHYLIP. The final result is compared with the structure-based phylogenetic tree and a previous work of protein classifications. The results show that the sequence-based phylogenetic tree is more accurate in classifying proteins with high similarities. However, the sequence-based phylogenetic tree is less accurate in determining the proteins with low sequence similarities. As conclusion, sequence-based phylogenetic tree produced from this study is reliable in representing the proteins that have high sequence similarities.

Chapter 1: Introduction

This chapter introduces the research paper. It contains the introduction, problem statement and objectives. It answers the questions of why this research is done and the problems that arisen that lead to this research.

1.1 Introduction

Protein is one of the most important elements that can be found massively in our body. Proteins are made up of assembled amino acids and there are 20 different types of amino acids. The side chains of amino acids are varying from polar to non-polar. Different arrangement of amino acids will form different sequences.

Protein has four different levels of structure; that are primary, secondary, tertiary and quaternary structures (Peet, 2012). A primary structure of a protein is amino acids that are arranged to form linear sequences. A secondary structure of a protein consists of α-helix and β-sheets. Tertiary structures consist of three-dimensional form of the combination of α-helices, β-sheets and some globular structures. Quaternary structure is the three-dimensional multisubunit protein that is composed by polypeptide chains which joined by non-covalent bonds.

In organisms, there are many types of proteins such as enzymes, hormones, and collagen. These proteins can be further classified to its own family, superfamily, fold and