UNIVERSITI TEKNOLOGI MARA

A SEQUENCE-BASED PHYLOGENETIC STUDY USING T-COFFEE ALIGNMENTS

NUR SYAMIMI DIANA BINTI HASNUL 'IZAD

Dissertation submitted in partial fulfilment of the requirements for the Bachelor of Pharmacy (Hons.)

Faculty of Pharmacy

July 2017

ACKNOWLEDGEMENT

First of all, I am grateful to Allah S.W.T for the good health and wellbeing that were necessary to complete this research. I wish to express my sincere gratitude to my supervisor, Dr Yuslina Zakaria for sharing expertise, and sincere and valuable guidance and encouragement extended to me. I would also wish to express my sincere thanks to the Bioinformatics laboratory and members for providing me with all the necessary facilities for the research. I also thank my parents for the unceasing encouragement, support and attention. I am also grateful to my partner who supported me through this venture.

TABLE OF CONTENT

ABSTRACT

T-Coffee is a multiple sequence alignments package used to align sequence. T-Coffee has two main features; first, it provides a simple and flexible ways of generating multiple sequence alignment using heterogeneous data sources .The data from these sources are provided to T-Coffee via a library of pairwise alignment. Second feature is the optimization method, which is used to find the multiple alignments that best fit the pairwise alignment. Sequence alignment plays an important role in producing an accurate phylogenetic tree. The best tool which capable of producing good alignment from sequence is needed to support the phylogenetic inference. The objectives of this study is mainly to build multiple sequence alignment using T-Coffee method, to build a phylogenetic tree using UPGMA method and to validate the reliability of the phylogenetic tree by comparing with the SCOP classification. The results showed that majority of the cluster resemble SCOP classification. It also showed that less difference in percentage identities for each protein pairs will give the best protein cluster.

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND OF STUDY

1.1.1 Protein Classification

There are two databases that are used in protein classification, which is SCOP and CATH. SCOP and CATH database arranged the 3-dimensional structure of protein into evolutionary classifications (Sillitoe, Dawson, Thornton, & Orengo, 2015). They enabled detailed molecular mechanism studies through which recent protein functions as well as structures emerge. The patterns of sequence and libraries of fold of SCOP and CATH have enabled structural relatives prediction, thus cater for structural interpretation of 50 million and more of domain sequences. SCOP and CATH also permit phylogenetic studies that recognize different chemistries within enzyme superfamilies.

1.1.2 Phylogenetics

Phylogenetics is the evolution relationships study. Phylogenetic analysis acts as a ways of estimating or inferring the relationships. The inferred evolutionary relationship is usually illustrated as branching, treelike diagrams, organism or both. Sometimes, phylogenetics is also known as cladistics. The word "clade" is a set of descendants which is a word that represent branch derived from the Greek that comes from a single ancestor. Particularly, cladistics is a process of hypothesizing the evolutionary relationships. Clade with similar past evolutionary and more relatable than members of another group is the basis behind cladistics.