

UNIVERSITI TEKNOLOGI MARA

**A SEQUENCE – BASED PHYLOGENETIC
INFERENCE USING A BAYESIAN –BASED
MODEL**

SHAZANA BINTI MOHD NAWAWI

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

Faculty of Pharmacy

Jun 2017

ACKNOWLEDGMENT

First of all, I am grateful to Allah S.W.T for giving me opportunity and strength to complete this research. I would like to express my deepest appreciation, especially to my supervisor, Dr. Yuslina Zakaria for her non-stop guidance and support throughout my research. Not to forget to Faculty of Pharmacy for providing good facilities for me to make this study easier and successful. I am also thankful to my family and friends for their help throughout this study.

Thank you.

TABLE OF CONTENTS

Contents

ACKNOWLEDGMENT.....	iii
TABLE OF CONTENTS.....	iv
LIST OF FIGURES	vii
LIST OF TABLES	vii
ABSTRACT.....	1
CHAPTER ONE	2
INTRODUCTION	2
1.1 Background	2
1.2 Problem statement.....	4
1.3 Significance of study.....	4
1.4 Objectives	4
CHAPTER TWO	5
LITERATURE REVIEW	5
2.1 Level of Protein Structure.....	5
2.2 Protein classification.....	7
2.2.1 Protein Family.....	7
2.2.2 Families of Structurally Similar Proteins.....	8
2.2.3 Superfamily.....	9
2.2.4 Structural Classification of Protein.....	9

ABSTRACT

Phylogenetic methods play an important role in inferring protein classification from protein information i.e. sequence and structure. For this study, we construct a phylogenetic tree using MrBayes and validate the reliability of the phylogenetic tree with the “gold standard” protein classification (SCOP). There are five steps involved to produce a phylogenetic tree, which are retrieving the data from the protein data bank (PDB), doing multiple sequence alignments, identification of model parameters, building the phylogenetic tree using MrBayes and last but not least is validating the tree with the gold standard (SCOP) Classification. Protein sequence alignment is a step in molecular evolutionary analysis that is used for estimating divergence and commonly used for identifying protein functions. However, the resulting alignments produced from conventional methods become less accurate when the sequence identity between protein pairs decreases (Rosenberg, 2005). MrBayes consists of some advantages as compared to other techniques of phylogenetic inference. The result is easy to analyze and it has computational advantages like speed (J P Huelsenbeck & Ronquist, 2001). Previous studies (Grone & Maruska, 2016) (Müller, Rahmann, Dandekar, & Wolf, 2004) have shown that MrBayes produced accurate and robust phylogenetic classification for low similarity proteins (< 40% PID). To conclude MrBayes has been proposed as one of the best methods to classify the distantly related protein.

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND

Protein sequence alignment is a step in molecular evolutionary analysis that is used for estimation a divergence, and for identifying the functional protein. The alignments become less accurate when the evolutionary distance among the sequence is increase (Rosenberg, 2005) . Sequence that is similar to one another is aligned first and followed by another sequences.

Phylogenetic methods are used to make a protein sequence alignments that consist of multiple sequence alignments and also phylogenetic reconstruction. Multiple sequence alignment that used Clustal W methods is one of the modern bioinformatics that can make protein sequence alignments to know the similarity and dissimilarity of the protein. It is very important to have a research on the sequence patterns based on evolution and the relationships between different organisms. Sequence are aligned across their entire length or in certain region only. Clustal W was released in 1994 by J.D Thompson (Chenna et al., 2003) , it gives improvements to the alignments algorithm in term of position- specific gap penalties, sequence weighting and the automatic choice . Basically, individual weight are assigned to each sequence in a partial alignment. This is to down-weight-near duplicate sequence and up-weight the most divergent ones. For residue gap penalties and locally, it will