UNIVERSITI TEKNOLOGI MARA

MODELLING THE CLASSIFICATION OF TWILIGHT ZONE PROTEINS USING STRUCTURE-BASED PHYLOGENETIC INFERENCES

SITI FATIMAH BINTI MOHD TAHA

Thesis submitted in fulfillment
of the requirements for the degree of
Master of Science
(Pharmacoinformatics)

Faculty of Pharmacy

October 2018

ABSTRACT

Structural studies of proteins have become a focus point for researchers as a result of vast growth of novel proteins and their huge contribution in drug discovery. Due to their highly conserved properties, modelling and predicting function of proteins commonly rely on their structural features with reference to protein classification. Proteins with close evolutionary relationship usually possess significant sequence similarity and are mostly studied using sequence-based approaches. However, evolutionary changes such as mutations can largely affect the sequences and thus, result in unreliable classification when dealing with highly dissimilar sequences of homologous proteins. As structures are highly conserved during evolution, the structure-based approach is the most suitable to infer homology between distantly related proteins. Previous studies have primarily focussed on finding protein homology rather than classifying proteins into families to represent evolutionary relationship. So far, there has been little discussion on the use of structural similarity for protein classification. Yet, no study has examined the accuracy of structural alignment tools to support an accurate phylogenetic classification of proteins. This thesis represents a study on structure-based methods in aligning twilight zone proteins to provide an accurate model for protein classification. A total of 716 proteins were chosen randomly from 4 major classes defined in the SCOPe database. All α proteins (A), all β proteins (B), α/β proteins (C) and α+β (D) proteins were represented in these classes. Structural alignment was conducted using six methods provided by five structural alignment tools namely CE, FATCAT, GANGSTA+, Matras and TM-Align. A sequence-based method was also conducted using T-COFFEE to provide a comparison with the accuracy and reliability of the structural methods. A distance-based phylogenetic approach, UPGMA, was then implemented using RMSD as inputs to produce classification trees. Evaluation of trees was performed by manually comparing the arrangement of clusters against the SCOPe v2.5 classification. External clustering metrics such as ARi were also used to validate the clusters. The results have shown that the structure-based approaches were more reliable than the sequence approach for classifying the twilight zone proteins. ARi scores obtained from structural trees outperformed the sequence approach for all folds at the superfamily level and 91.67% of folds at the family level. CE performed best for two major classes A, and C, whereas proteins from classes B, and D were best aligned using TM-Align. Based on the findings, a pipeline was developed to automate the classification analysis, and was tested in two case studies that involved Alzheimer's disease proteins, and substrate binding-proteins (SBP) respectively. Both case studies proved the feasibility of the proposed pipeline to provide a reliable classification of twilight zone proteins and serve as a guideline for future studies.

ACKNOWLEDGEMENT

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

In the Name of Allāh, the Most Gracious, the Most Merciful

Alhamdullilah, all praises to Allah SWT the All-Mighty, for blessing me with this opportunity to fulfill my dreams in achieving my Master's degree. I take this opportunity to express my profound gratitude and deepest regards to my resourceful yet humble supervisor, Dr. Yuslina binti Zakaria, for her exemplary guidance, patience, continuous encouragement and immense knowledge throughout my research. Despite her busy schedule, she has never failed to give support and guide me in my work during my years in UiTM. The blessings, help and guidance given by her shall carry me a long way in the journey of life on which I am about to embark.

I would also like to express a deep sense of gratitude to my co-supervisor Dr. Mashani binti Mohamad, Faculty of Pharmacy, UiTM Puncak Alam, and Puan Norfatimah binti Mohamed Yunus from Faculty of Applied Sciences, UiTM Shah Alam for their constant support and motivation for completing my studies. Not to forget, my sincere gratitude to the Malaysian Ministry of Higher Education for funding this research through the Fundamental Research Grant Scheme (FRGS).

Most of all, my deepest appreaciation to my family especially my beloved parents, Mohd Taha bin Othman and for their endless prayers and support they provided throughout my entire life. There are no greater gifts than their unconditional love and blessings. Not to forget, my dearest friends, and fellow colleagues in UiTM for their cordial support. I am grateful for their encouragement without which this thesis would not be possible.

TABLE OF CONTENTS

Page

CONFIRMATION BY PANEL OF EXAMINERS ii
AUTHOR'S DECLARATION iii
ABSTRACT iv
ACKNOWLEDGEMENT v
TABLE OF CONTENTS vi
LIST OF TABLES ix
LIST OF FIGURES x
LIST OF SYMBOLS xii
LIST OF ABBREVIATIONS xiii

CHAPTER ONE: INTRODUCTION 1
1.1 Background 1
1.2 Problem Statement 3
1.3 Hypothesis 3
1.4 Significance of Study 3
1.5 Objectives 4

CHAPTER TWO: LITERATURE REVIEW 5
2.1 Protein Structure and Classification 5
2.1.1 Levels of Protein Structure 5
2.1.2 Protein Family 6
2.1.3 Protein Evolution 7
2.1.4 Protein Database 8
2.2 Protein Alignment 10
2.2.1 Combinatorial Extension (CE) 11
2.2.2 FATCAT 15
2.2.3 GANGSTA+ 18
2.2.4 MATRAS 21
2.2.5 TM-Align 24

CHAPTER ONE
INTRODUCTION

1.1 Background

Proteins are key players in living organisms as they carry numerous important roles in the body, such as transporting/storing molecules, enzymes, messengers, and antibodies. Hence, many studies have been conducted to increase our understanding of the nature, properties, and functions of proteins in the body. Protein function prediction is one of the fundamental studies that relies on computational biology. To predict protein function, it is essential to understand the homology between proteins. The concept applied to this approach was explained by Sjölander (2004), where novel proteins were compared against a protein library, and significant similarities between the proteins may lead to its functional information. Therefore, the novel protein can be inferred based on the presumed functions of the homolog. Initially, protein prediction was performed by comparing sequences to a library, using bioinformatics tools such as FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990).

Today, proteins are classified into distinct groups based on their respective sequence or structural similarity. In structural classification, proteins are classified according to their domain, family, superfamily, fold, and class (Karim et al., 2016; Andreeva et al., 2008). These groups consist of well-characterised proteins of known functions, which enable researchers to study the functions of identified novel proteins based on prediction. Protein classification databases that serve this purpose are available freely on the web, such as Structural Classification of Proteins (SCOPe) (Fox et al., 2013) and CATH (Orengo et al., 1997). However, previous studies have shown the setbacks of using only homology-based methods for function prediction as they have the tendency to produce systematic errors (Nguyen et al., 2014; Gabaldón and Koonin, 2013; Delsuc et al., 2005; Koski and Golding, 2001). Therefore, studies have emphasised the need to include inferences from both structural and evolutionary analysis (Sjolander and Specht, 2007; Sjölander, 2004). For this purpose, the use of phylogenetic inference is the most suitable approach to provide a reliable classification of proteins and predict their functions by homology.

Phylogenetic analysis has been used to study the evolutionary relationship of