ABSTRACT

The goal of this thesis is to develop a computational method based on machine learning techniques for predicting disulfide-bonding states of Cysteine residues in proteins, which is a sub-problem of the bigger and yet unsolved problem of protein structure prediction. First, we preprocessed the datasets from Protein Data Bank (PDB) and filtered mutations and low resolution files out. A number of descriptors in two dimensional (2D) protein sequences are studied. These descriptors are based on local feature values of adjacent amino acid to Cystein residue, namely encoded, propensity value and averaged propensity value. We have used Artificial Neural Network (ANN) as a machine learning technique to develop our prediction method. We use 'trainlm', 'trainrp' and 'trainscg' training functions for training out network and also a 5-fold validation is implemented. Our results show that we can predict the state of Cystein disulphide bond formation. It shows that using propensity valued descriptor and 'trainscg' training function is better to be used for Cystein bond state prediction compared to the other training functions and descriptors in this study. The accuracy of prediction in this study is 80.85% on a propensity value descriptor dataset which had been trained by 'trainscg' with a dataset of over than 400 thousand protein patterns. Results of this work will have direct implications in site directed mutational studies of protein, protein engineering and the problem of protein folding.

ABSTRAK

Matlamat tesis ini adalah untuk membangunkan kaedah pengiraan yang berdasarkan teknik pembelajaran mesin untuk meramal ikatan-disulfida bagi residu Cysteine dalam protein di mana ini adalah sebahagian daripada masalah yang belum dapat diselesaikan dalam ramalan struktur protein. Pertama sekali, pra-proses set data dari Protein Data Bank (PDB) dilakukan dan fail mutasi serta fail resolusi rendah ditapis. Beberapa pemerihal di dalam susunan protein dua dimensi dikaji. Pemerihal adalah berdasarkan nilai ciri-ciri tempatan bagi asid amino bersebelahan dengan residu cystein, termasuk nilai yang diwakilkan, nilai kecenderungan dan purata nilai kecenderungan. Artificial Neural Network (ANN) digunakan sebagai teknik pembelajaran mesin untuk membangunkan kaedah ramalan. Fungsi pembelajaran 'trainlm', 'trainrp' dan 'trainscg' digunakan untuk menguji rangkaian dan juga melaksanakan pengesahan 5-lipatan. Keputusan menunjukkan pembentukkan ikatan cysteine disulfida dapat ramalkan. Ini menunjukkan dengan nilai kecenderungan dan fungsi pembelajaran 'trainscg' lebih sesuai digunakan untuk meramalkan keadaan ikatan cysteine berbanding fungsi pembelajaran dan pemerihal lain dalam kajian ini. Ketepatan ramalan di dalam kajian ini adalah 80.85% pada set data pemerihal nilai kecenderungan di mana ia telah diuji dengan 'trainscg' dengan nilai set dat lebih daripada 400 ribu corak protein. Keputusan ini memberi implikasi secara langsung kepada kajian mutasi terus setempat protein, kejuteraan protein dan masalah lipatan protein.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Protein three-dimensional (3D) structure prediction from its chain of amino acid residue (amino acid sequence) is a novel approach in structural bioinformatics domain which play role in pharmaceutical industry and drug design. One of the structure prediction components in the '*ab-initio*' 3D modeling is the determination of the disulphide bond in a protein structure. In this study, neural network technique in machine learning and computational methods are utilized to refine and optimize the descriptors would be applied to the prediction of a disulphide bond formation in 3D protein structure prediction.

1.2 Problem Background

The complexity of many problems in biochemistry prevents them to be immediately solved by methods that are based on first principles of theoretical calculations. This is exemplified by the protein structure prediction from amino acid sequence where the relationship between the effects of different features of amino acid sequences and the final structure of a protein is significant. In this case, we have to study, scrutinize and examine known structural data sourced from laboratory and experimental methods to create a desired model of relationships, such as between amino acid sequence and its biological activity and chemical interaction within the protein.

One of the challenges in computational biology is that the biological behavior of a component cannot yet be directly calculated from first principles by using theoretical methods in a process of deductive learning. Hence, we need to put the available data into context and see whether it allows us to make the prediction of new data. We should relate different features available to create useful information for better prediction. We have a many available biological and chemical databases which offer huge amounts of rapidly expanding data in various ways, but still in many cases, we lack the essential information. In this case, what comes to mind is if we can learn enough from the available data to obtain the knowledge for making our prediction when the necessary information is not directly offered? As an example, looking at the amino acid sequences inside the protein structures can barely help us to identify any specific rule of 3D structural formation. Considering different properties of each biological element inside the sequence and the combination of these elements with their associated information can boost us to generalize the relationships between these components and create a model that results to the knowledge of chemical effects inside the protein amino acid chains on its 3D structural formation. This process of deriving knowledge from data and observations is called inductive learning. Computational methods have now become

available for inductive learning, such as pattern recognition methods, artificial neural networks, or data mining methods. (Gasteiger, 2006)

The traditional methods of protein's 3D structure determination require various wet-lab experiments which should be carried out on the aqueous samples of proteins with high level of purification. The downsides are; all these methods are costly and very time-consuming, some techniques can be only applied to some sort of domains (e.g., X-Ray and NMR only on soluble proteins). On the other hand, the sequencing of proteins (determination of amino acid sequence inside protein) is comparatively simple, fast, and reasonably priced. In this study, neural network as a machine learning method is applied for the concept of protein 3D structure prediction from its primary structure.

Artificial Neural Networks (ANN) which are commonly used as Neural Networks (NN) are mathematical and computational models, which belong to class of general computational structures based on examination of biological nervous system. Neural Network is applied for classification or estimation. Neural network has been rapidly utilized in numerous applications in the field of biology and chemistry. Neural network with its good generalization features has been applied in many biological applications like drug likeliness classification and protein functionality prediction. Several advantages make neural network promising and active research area in the field of protein structure prediction. Theoretically, neural networks are highly competent at fitting functions and patterns recognition and there is proof that a neural network with a simple architecture can fit any practical function. Thus they are very good at inferring models in large or even infinite dimensions from a finite number of observations. Neural network performs man classification and estimation tasks for producing better outfits comparing the target data. The specifications of neural network caused an increasing popularity of its application in various fields like bioinformatics.

This project applies neural network methods to the prediction of disulphide bond formation in protein 3D structure prediction and compare its results for different input descriptors and training functions.

1.3 Problem statement

Currently, there is a lack of computational methods to predict the formation of the disulphide bond in protein structure prediction. In this project, an artificial neural network will be applied for disulphide bond prediction in protein 3D structure prediction using descriptors generated from its primary sequence information.

1.4 Objectives

- To apply neural network techniques as machine learning methods for predicting the disulphide bond connectivity in a 3D protein structure from its protein sequence.
- To determine which training functions of NN give the best performance on the selected descriptors in disulphide bond prediction.
- To determine which descriptor developed in this study gives better performance on disulphide bridge state prediction of Cystein residue.
- To search for suitable amino acid descriptors that can be fed into the artificial neural network to predict Cystein disulphide bonds.

1.5 Scope

- This study includes the pattern (descriptor) database development and preprocesses related to descriptor creation generated from protein structure derived from laboratory and experimental data archived in Protein Databank (PDB).
- This study focuses on developing a neural network model for prediction of Cystein disulphide bridge connectivity in 3D protein structure from its 2D structure information.
- The study compares the result of feeding different descriptors (Encoded, Propensity Value and Averaged Propensity value base) to the neural network.
- This study compares the result our neural network model for different training functions on our database of patterns.

1.6 Importance

Protein structure prediction is an important issue in the domain of medicine (e.g., design of drug) and in Biotechnology (e.g., designs of novel enzymes) which it comes to meaning when protein structure determines its biological function. Thus, study of protein structure with the intention of understanding the biological function of protein is essential. In this case, studying the disulphide bond connectivity formation of Cystein amino acid which plays a significant role in stabilizing the structure of protein is of high importance.

Previous studies on disulphide bond connectivity using machine learning approaches have met with relative success. Although prediction accuracy have improved, most used small homology datasets of several thousands and are not

comprehensive as compared to our study that utilized 67 thousand protein structures with more than 478 thousand sequence patterns. Another significance of this study is the use of amino acid biochemical characteristics and propensity values in the descriptor creation and using the different features of amino acid sequences on the disulphide bond connectivity prediction.

1.7 Outline of the Report

Following chapters will discuss literature review, methodology, results and conclusion. In Chapter 2 literature review will be discussed extensively. Chapter 3 discuses the methodology used in this project. It discusses the steps of preprocessing for descriptor development and how to achieve our objectives in the project. Results of the experiments conducted are recorded in Chapter 4. Finally, Chapter 5 concludes this report.

1.8 Summary

Application of machine learning methods is essential for increasing efficiency and accuracy in disulphide bond formation in protein structure prediction. Although previous work has tackled this challenge, the datasets were significantly smaller and the methods rely solely on the pattern formation and not the biochemical characteristics of the amino acids that form the primary structure. This study will compare the predicted results of the use of artificial neural network for Cystein

disulphide bond formation prediction using various descriptors (primary structure, biochemical characteristics and amino acid propensities) generated from laboratory and experimental data.