ABSTRACT

Remote protein homology detection refers to the detection of structural homology in weak proteins. Remote protein homology is important to identify function for new proteins which could assist in curing genetic diseases, performing drug design, and identifying novel enzymes. To detect remote protein homology, several problems have been identified by researchers which are hard-to-align proteins homology detection and high dimensional feature vectors of proteins caused by redundant and noisy data. To address these problems, a new remote protein homology detection computational framework has been developed. The computational framework begins by extracting structural similarity of protein using highly sensitive structural similarity algorithm which consist of four steps: split protein sequences into substring, calculate similarity using pairwise protein substring alignment, build guide tree, and extract the high structural similarity using multiple protein sequence alignment. Then, Latent Semantic Analysis algorithm (LSA) is used to produce feature vectors. The LSA consist of three steps: generate protein pattern blocks using TEIRESIAS algorithm, remove redundant data using chi-square algorithm, and noisy data using Singular Value Decomposition (SVD) algorithm. Lastly, this computational framework uses SVM to classify all the proteins into homologue or non-homologue members. The proposed computational framework is analyzed using dataset from SCOP database version 1.53 and the performance has been compared with other methods such as PSI-BLAST and SVM-Pairwise sequence comparison models, SAM and HMMER generative models, and SVM-Fisher and SVM-I-Sites discriminative classifier models in terms of Receiver Operating Characteristic (ROC), Median Rate of False Positives (MRFP), and family by family comparison of ROC. The results show that the proposed computational framework successfully outperforms other remote protein homology detection methods.

ABSTRAK

Pengesanan homologi protein yang jauh merujuk kepada pengesanan homologi struktur dalam protein yang lemah. Homologi protein yang jauh adalah penting untuk mengenalpasti fungsi bagi protein yang baru dimana ia boleh membantu merawat penyakit genetik, mereka bentuk ubat dan enzim baru. Bagi mengesan homologi protein yang jauh beberapa masalah telah dikenalpasti oleh penyelidik-penyelidik iaitu protein-protein sukar diselarikan dan vektor-vektor ciri yang berdimensi tinggi disebabkan oleh data bertindan dan data hingar. Bagi menyelesaikan masalah tersebut, rangka kerja untuk homologi protein yang berasaskan komputer telah dibina. Rangka kerja tersebut bermula dengan mengekstrak persamaan struktur bagi protein menggunakan algoritma persamaan struktur yang bersensitif tinggi. Algoritma ini mengandungi empat langkah: membahagikan jujukan protein kepada berbilang subjujukan, mengira persamaan melalui penjajaran subjujukan secara berpasangan, membina pepohon pandu dan mengekstrak persamaan struktur yang tinggi melalui penjajaran subjujukan secara berganda. Kemudian, algoritma Analisis Semantik Pendam (LSA) digunakan untuk menghasilkan vektor-vektor ciri. LSA mengandungi tiga langkah: menghasilkan blok-blok corak protein menggunakan algoritma TEIRESIAS, membuang data bertindan menggunakan algoritma chi-square dan membuang data hingar menggunakan Penguraian Nilai Singular (SVD). Akhir sekali, rangka kerja yang dicadangkan ini menggunakan Mesin Sokongan Vektor (SVM) untuk mengelaskan semua protein ke dalam ahli-ahli homolog atau bukan homolog. Rangka kerja ini dianalisa menggunakan set data yang diperolehi dari pangkalan data SCOP versi 1.53 dan prestasinya telah dibandingkan dengan kaedah-kaedah lain seperti PSI-BLAST dan SVM-Pairwise daripada model perbandingan jujukan, HMMER dan SAM daripada model generatif dan SVM-Fisher dan SVM-I-Sites daripada model pengelasan perbezaan menggunakan ukuran Karakter Pengoperasian Penerima (ROC), Kadar Median Positif Palsu (MRFP) dan perbandingan keluarga dengan keluarga berasaskan ROC. Hasil keputusan menunjukkan rangka kerja yang dicadangkan ini berjaya mengatasi kaedah-kaedah lain dalam pengesanan homologi protein terpencil.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background

The enormous growth of public sequence database and existing addition of fully sequenced genomes have become challenging problems in the field of bioinformatics. Based on the growth in GenBank (Benson *et al.*, 2008; http://www.ncbi. nlm.nih.gov/Genbank/), the number of sequence data has doubled approximately every 18 months. The unprecedented growth in biological information collection and the production of biological information have become much greater than its consumption. This causes problems on how to interpret and manage the huge amount of novel information so that the usage of protein sequence can be easier and more efficient. Remote protein homology detection is one of the methods used widely by researchers to manage protein sequences by classifying the protein sequences into their family (Kuang *et al.*, 2004).

Nowadays, various methods have been developed to solve the problem of a huge amount of novel sequence data in gene databases. However, in the effort to develop more accurate methods, a few problems have arisen such as high dimensional protein feature vectors (e.g. noisy and redundant data) and hard-to-align proteins. Examples of methods that have been developed to solve high dimensional protein feature vectors problem are Kearney *et al.* (2005) which apply data compression technique to reconstruct and compress all the protein feature vectors in order to remove redundant data and Liu *et al.* (2008) applied thresholding technique in order to remove noisy data. Both techniques are heading to reduce high dimensional protein feature vectors. Whereas the example of works which handle the hard-to-align proteins such are Pipenbacher *et al.* (2002) extend a graph-based clustering algorithm which uses an asymmetric distance measure. An asymmetric distance measure can be employed to distinguish between the two proteins being globally similar and one protein being similar to an individual domain of a multi-domain protein. While Mohseni-Zadeh *et al.* (2004) employs a simple and efficient clustering algorithm which avoids the problem of multi-domain, it nevertheless allows a partial chain effect with the result that all the elements within one cluster may not rigorously and absolutely share the same functional domain. In addition, one given protein may belong to several different clusters, which makes sense in the case of multi-domain proteins. Even though many tools have been developed, efforts on searching for the most accurate method should be continued. More researches need to be done in order to find the method that is capable to handle high dimensional protein feature vectors (e.g. noisy and redundant data) and hard-to-align proteins with more accurate result. Both problems are popular problems. It needs to be overcome because the existence of these problems will disturb result presentation and will present inaccurate results. An accurate result refers to the method that is able to classify all the protein sequences into homologue or non-homologue members precisely and generate more True Positive (TP) and True Negative (TN) protein sequences. When protein is classified into homologue and non-homologue members precisely, protein function could be easily identified and make all the protein sequences easy to interpreted, managed, found and used.

The following section in this chapter discusses the challenges involved in detecting remote protein homology. Then, current methods in remote protein homology detection are presented, followed by the problems to be solved in this study. The aim of this research as well as the research objectives, are also presented.

Further in this chapter, the scope and the significance of detecting remote protein homology are described before the overview of the organization of this thesis is presented.

1.2 Challenges of Detecting Remote Protein Homology

Remote protein homology detection refers to the detection of protein structural homology in protein sequences which contains little protein structural similarity. Remote protein homology detection is used to identify protein structural information in order to define functional properties of protein by means of homologies (similarity). High quality productions of protein structure similarity information are important to the classification process in order to define new functions precisely (Liu *et al.*, 2008). Detecting remote protein homology becomes increasingly difficult because of high dimensional protein feature vectors (e.g. noisy and redundant data) and hard-to-align proteins. The high dimensional protein feature vectors need corresponding dimension of feature space to map all the protein feature vectors and discriminate the protein feature vectors into homologue and non-homologue members accurately. The feature space mapping protein feature vectors depends on the vastness of feature spaces and if the feature space is full, the surplus protein feature vectors will be ignored (Cristianini and Shawe-Taylor, 2000). This causes the loss of similarity information between protein sequences (Aleksander *et al.*, 2002). The problems of hard-to-align proteins will give bad impact to the determination of protein sequence whether it is homologue or non-homologue. The detection of hard-to-align proteins is important because it assists to gain more protein structural similarity information and assists to identify the protein homologue or non-homologue accurately. If hard-to-align proteins problem is neglected, the proteins distribution into their clusters that refers to their structure similarity is inaccurate and it gives a bad impact to the determination of protein sequence whether it is homologue or non-homologue. Both problems will produce worse protein feature vectors presentation whereby the high dimensional feature space (e.g. noisy and redundant data) will present useless information in protein feature vectors, whereas the hard-to-align proteins have high possibility to miss the important part of protein structural similarity information from protein sequences. These problems will further affect in inaccurate classification of protein sequences into homologue or non-homologue members.

The presentation of high dimensional protein feature vectors is caused by noisy and redundant data. Noisy data is useless data. It can override genetic and environmental determinism (Ghim and Almaas, 2008). In fact, the presence of noisy data may significantly affect the fitness of an organism (Raser and O'Shea, 2004) whereas; redundant data is a duplicate data and containing in one dataset which exists from the origin dataset. In the detection of remote protein homology, noisy and redundant data contribute to the thousands matrix dimensions or high dimensional protein feature vectors that will disturb the presentation of protein structure similarity information with the presentation of useless protein feature vectors. The useless protein feature vectors will prevent classification algorithm to classify all the protein feature vectors into homologue or non-homologue members accurately.

Another problem is hard-to-align proteins. Hard-to-align proteins are a part of protein, which belongs to two or more cluster groups. If proteins cannot be determined to be hard-to-align proteins then two or more cluster groups will be linked together in one cluster (Hou *et al*., 2003). It will cause lost of similarity information between proteins. The given information is not precise. Whereas, if the hard-to-align proteins are defined to be a group member of two or more cluster groups, it means that the two or more cluster groups are not related. The detection of hard-to-align proteins is important because it assists to gain more protein structural similarity information and assists to identify the protein homologue or non-homologue accurately.

1.3 Current Methods for Detecting Remote Protein Homology

Basically, remote protein homology detection can be divided into three categories (Liao and Noble, 2003):

(i) Sequence comparison model is implemented by arranging all the protein sequences in order to identify regions of proteins similarity that have similar characteristics or in other words homologue. Protein homologue can be explained as proteins that have similarity in function, structure, and evolutionary relationships. Examples of works done are by Wittkop *et al.* (2007) and Chen *et al.* (2006).

(ii) Generative model is implemented by measuring the probabilistic value between protein sequences. The model involves building a model for an each protein family and then evaluating each protein sequence to see how well it fits the model. If the fit of the protein sequence is above some threshold value, then the protein is classified as homologue in the same family. Example algorithms are Hidden Markov Model (HMM) that has been used by HMMSTR (Hou *et al.*, 2004) and profile HMM (Bernardes *et al.*, 2007).

(iii) Discriminative classifier model is implemented by classifying protein into any protein similarity between characteristics that is due to their shared ancestry (homologue) or otherwise (non-homologue), referring to the model. Protein sequences will be tested whether it is included in homologue or non-homologue members by discriminating all the protein sequences using classification algorithm referring to the model. Example is Support Vector Machines (SVM) that is used by Rangwala and Karypis (2005) and Neural Network (NN) that is used by Stoffer and Volkert (2005).

1.4 Statement of the Problem

The remote protein homology detection problem to be solved in this study can be described as follows:

> "Given proteins, the challenge is to classify with higher accuracy all the proteins into homologue or non-homologue members by decreasing the high dimensional protein feature vectors while at the same time able to produce more protein structure similarity information by solving hard-to-align proteins. Both solution methods are capable to produce better results with higher Receiver Operating Characteristic (ROC), lower Median Rate of False Positives (MRFP), and higher family by family comparison of ROC."

In this study, in order to develop a computational framework, which will produce more accurate results, two problems are considered. The first problem is reducing the high dimensional protein feature vectors. This problem is caused by noisy and redundant data in protein feature vectors. The noisy data identified exist in dataset SCOP 1.53 (Dong *et al.*, 2006; Liu *et al.*, 2008). To reduce the noisy data, protein feature vector is seen as factor to be taken into consideration to remove all the noisy data. The protein feature vectors keep value score of similarity between proteins. In this study, all protein feature vectors will be identified which is free from noisy data and containing noisy data using thresholding technique. The thresholding technique had been proven able to refine noisy data (Liu *et al.*, 2008). It removes noisy data based on E-value equal to 0. The E-value is selected based on the corresponding limit of value. The protein feature vector, which is less than E-value, is containing noisy data. The second problem is the redundant data. The redundant data is also identified to exist in dataset SCOP 1.53 (Dong *et al.*, 2006; Liu *et al.*, 2008). To reduce the redundant data, protein patterns are seen as factor to be taken into consideration to remove all the redundant data. In this study, all the redundant data will be separated into independent protein patterns and non- independent protein patterns. The independent protein pattern is a protein pattern that is not redundant to any other protein patterns. Whereas, the non-independent protein pattern is a protein pattern that is redundant to any other protein patterns.

The hard-to-align proteins will reduce production of protein structural similarity information between protein sequences. While searching protein structural similarity process, the possibility to find part of proteins that have structural similarity is neglected. The neglected part of proteins is considered as hard-to-align proteins that occur in multi-domain protein. The hard-to-align proteins are the important parts, which possess a high percentage containing protein structural similarity. Recently, many computational methods are not concerned with this problem. But the computational methods, which try to handle this problem, are faced with other problems related to their algorithm. It is the computational methods, which cannot identify a part of hard-to-align proteins more precisely and require the best computational methods, which can identify the hard-to-align proteins more sensitively. In this study, to overcome this problem, searching protein structural similarity process is a factor to identify the hard-to-align proteins more sensitively. In this study, protein structural similarity technique to identify hard-to-align proteins is applied. The technique will identify the hard-to-align proteins first, and then extract protein structural similarity information for all protein sequences. This technique is effective compared with the existed techniques.

Considerations of both factors are expected to generate high protein structural similarity information and low dimensional protein feature vectors with less noise and redundancy. At the same time this will lead to the more accurate classification.

1.5 Objectives of the Study

The goal of this study is to develop a computational framework to reliably assign protein sequences to its predefined family. In order to realize this goal, several objectives need to be achieved:

(i) To study and investigate current remote protein homology methods in order to understand the behaviour, data, design and flow of method.

(ii) To develop Support Vector Machine-Latent Semantic Analysis-Structural Similarity 1 algorithm (SVM-LSA-SS1) using combination of SVM with String Kernel, LSA, and Smith-Waterman algorithm to reduce high dimensional protein feature vectors caused by having noisy and redundant data in protein dataset in order to assist in getting higher ROC, lower MRFP, and higher family by family comparison of ROC.

(iii) To develop Support Vector Machine-Latent Semantic Analysis-Structural Similarity 2 algorithm (SVM-LSA-SS2) by enhancing sequence comparison model from SVM-LSA-SS1 with Substitution Matching Similarity (SMS), guide tree, and multiple alignment algorithm in order to produce more protein structural similarity information by handling hard-to-align proteins that can assist to get higher ROC, lower MRFP, and higher family by family comparison of ROC.

(iv) To test the algorithms mentioned above using ROC and MRFP in order to measure its performance.

1.6 Significance and Scope of the Study

Remote protein homology detection is the task of classifying protein sequence into homologue or non-homologue. Homologue refers to the same ancestor and non-homologue is otherwise. In order to infer the function of an unknown remote protein, identification the homologue or non-homologue members of proteins too helpful, which the accurate homologue protein leads to define the new protein's function precisely according to the protein structural similarity information. Protein function controls how each aspects of an organism work and to a great extent, how the organism looks and behaves. In general, several genes contribute to the characteristics of a human trait and thus make this process even more complicated. If a method can be designed to reliably group all proteins, which share the same, functional domains into families, newly discovered proteins can be classified more easily. In long term, it is hoped that the study of remote protein homology detection and classification will help in the fight to cure deadly diseases such as Acquired Immune Deficiency Syndrome (AIDS). Most of the genetic diseases are caused by the body creating proteins incorrectly or at an incorrect time. Therefore, if the deviation of these proteins from the normal is known, we will be able to design drugs accordingly (Payne, 2001). For instance, in drug discovery, if sequence *S* is obtained from some disease *X* and it is determined that *S* belongs to the family *F*, and then one may try a combination of the existing drugs for *F* to treat the disease *X*. Biological sequence similarity can tell us what is common and what is unique between different species at the genome level. One application is to identify unique, crucial proteins in pathogens to be used as targets for products that are both safe and effective. The functions of human genes and other Deoxyribonucleic acid (DNA) regions could be revealed by studying their counterparts in lower organisms.

Traditionally, understanding the biological function of the various macromolecules is by using biochemical techniques. However, traditional biochemical techniques are very time consuming. For example, the straightforward way to determine the three-Dimensional (3D) structure of a protein involves the use of physical methods like X-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy. Such methods can take months, even years, to produce the desired results (Logan *et al.*, 2001). Furthermore, due to physical constraints, these techniques are not applicable to all types of proteins. The inherent difficulty of applying physicochemical analysis methods in conjunction with the feasibility of predicting a protein function from primary sequence necessitates the development of new techniques for detecting homologous proteins. Although computer science is expected to play a key role in this new area, conventional computer science algorithms are unable to address this problem. This is due to the complexity of the biological systems and the lack of fundamental theory at the molecular level. Another reason is that, conventional computational tools are unable to handle the large and rapidly expanding amount of biological data. Consequently, remote protein homology detection is expected to perform the annotation of new protein sequence to its homogeneous family without a certain degree of confidence.

The interest and need to design effective and efficient remote protein homology detection method was established and intensified in the last few years. Many methods have been developed (Yang *et al*., 2008; Argawal *et al*., 2008; Zaki and Deris, 2007; Mohseni-Zadeh *et al*., 2004; Liao and Noble, 2003; Pipenbacher *et al*., 2002) in order to classify all the protein sequences into their family by referring to the protein structural similarity information. The scope of the study presented in this thesis is limited to producing a method which can extract the best quality protein feature vectors referring to low dimensional protein feature vectors extracted from high structural similarity information. Referring to this statement, an attractive computational framework that combines the sequence comparison model, generative model, and discriminative classifier model is introduced. The sequence comparison model will assists in obtaining high structural similarity information between the protein sequences. Meanwhile, the generative model will involve building low dimensional protein feature vectors caused by abandonment of noisy and redundant data in the protein feature vectors and protein patterns respectively. Whereas, the discriminative classifier model is used in order to classify all the protein sequences into homologue or non-homologue members. In order to achieve the scope, the sequence comparison model and generative model were studied deeply in order to extract best protein feature vectors with low dimensional and contain high structural similarity information. The protein structural similarity information contains same features of ancestor information shared by each protein sequences. This high structural similarity information is extracted from four steps: split technique, pairwise protein substring alignment, guide tree, and multiple protein sequence alignment. This protein structural similarity information then will be used as input to LSA (Bellegarda, 2000) in order to produce protein feature vectors. Three steps are included in LSA to produce protein feature vectors: (i) TEIRESIAS (Rigoutsos and Floratos, 1998) to extract protein pattern; (ii) chi-square (Yang and Pedersen, 1997) to remove redundant data in protein pattern; and (iii) Singular Value Decomposition (SVD:Alter *et al*., 2000) to reduce the high dimensional protein feature vectors by removing noisy data. Then, SVM will be used to train the extracted protein feature vectors from LSA to a set of a model. The performance of the model is assessed by assigning labels to each protein sequence. To test the proposed computational framework, the dataset of protein sequences from Structural Classification of Proteins (SCOP: Conte *et al.*, 2000) database, limited to version 1.53, is used. On the other hand, to measure the performance of the proposed computational framework, three measures are chosen, mean ROC, mean MRFP, and family by family comparison of ROC score.

1.7 Organization of the Thesis

A general content description of the subsequent chapters in this proposal is given as follows:

(i) Chapter 1 describes the challenges, current methods, problems, objectives, significance, and scope of the study.

(ii) In Chapter 2, the basic concepts of remote protein homology detection, the raised problems in remote protein homology detection, and the trend and tendencies are presented. Exhaustive review of the previous related work is also presented.

(iii) Chapter 3 begins with a brief review of the proposed computational framework followed by detailed description for all instruments involved such as hardware and software requirements, testing and analysis, and performance measurement.

(iv) Chapter 4 describes a new approach of remote protein homology detection framework known as SVM-LSA-SS1 algorithm which is developed to solve the high dimensional protein feature vectors caused by noisy and redundant data. The main focus is on LSA which consist three main components: (i) TEIRESIAS algorithm to extract protein pattern; (ii) chi-square algorithm to reduce redundant data in protein patterns, and; (iii) SVD algorithm to reduce noisy data and produce protein feature vectors.

(v) Chapter 5 describes another new approach of remote protein homology detection framework known as SVM-LSA-SS2 algorithm, extended from SVM-LSA-SS1 that is developed to solve the hard-to-align proteins. The main focus is on highly sensitive structural similarity algorithm which consist of four main components: (i) split technique to split all the protein sequences into short length of protein sequences; (ii) Substitution Matching Similarity (SMS) to extract the protein structural similarity and solve hard-to-align proteins problem; (ii) guide tree to group proteins sequences based on the protein structure similarity; and (iii) multiple protein sequence alignment to extract high protein structural similarity information according to the guide tree.

(vi) Chapter 6 contains the conclusion of the study and the achieved results. The contributions and future works of the study are also described.