CHAPTER I

INTRODUCTION

Proteins are very important and beneficial to biological and biochemical processes for living things. They are used to transport oxygen to cells and to convert chemical energy into mechanical energy and prevent harmful inflection. The first structure of a protein, myoglobin, was found in 1958 by using x-ray crystallography. This marked the beginning of a new era of biology. Simultaneously, biological research began to provide digital information into computational analysis.

A major goal of biology is to define rules of how cells work. All cells come from preexisting cells. DNA is the heritable material. Another rule comes from their forms that determine their function, which means that if we know the shape of a molecule, we should understand how they perform.

Although sequences of many proteins are known, amino acid sequence alone is insufficient to tell what proteins perform or interact. Many times, biologists need to find the shape or the three-dimensional structure of protein in order to understand its role. Progress of discovering the structure of proteins has been driven slowly in many individual laboratories around the world.

The proteins have to be purified in large amounts (about 1 mg.) and crystallized. Then either x-ray crystallography or nuclear magnetic resonance (NMR) is used in order to find information about the protein's structure at the atomic level, yielding the

location of every atom in the protein. A single three-dimensional structure of each protein can be determined one at a time by individual labs. The demand of structural proteomics methods has been increasing rapidly, but the progress is still slow.

The limitations of structural proteomics are the processes of preparation of samples and crystals. Protein structural information is collected by the Protein Data Bank (PDB). Although the three-dimensional structure information of proteins in the PDB has increased rapidly in recent years, such information still has not not met the demand for the structural proteomics by biologist.

Although, the prediction of protein structures has not been done successfully, grossly simplified solutions using the secondary structure prediction are still useful for theoreticians. Over the decades, many improvements have been made applying machine learning theories such as the Support Vector Machine. The improvement leads to more accurate in prediction.

The structural regularities of protein, α-helices and β-sheets, were found by the prediction of Pauling and Corey, and by Kendrew's work [1, 2]. There are local regularities of protein structures even though they are complex. These secondary structure elements lead to understanding of the overall structures of the protein. There are many levels of protein structures, which are secondary structure, super-secondary structure, tertiary structure, and quaternary structure. These levels must use the secondary structure as fundamental elements.

Currently, the number of known protein structures exceeds 20,000, much higher than those known 250 protein structures in 1988. The experimental determination of structure is still slow and costly. In addition, the gap between sequences and structure determinations is wider by the sequencing technology that becomes more advanced.

It is not only the protein structures that have to understand alone, but it is also fundamental molecular biology that has to be worked out. Understanding the structure of proteins leads to understand biological reactions and how they play within an organism.

1.1 Secondary Structure Prediction

One of the greatest challenges in sequence analysis is to find the accurate prediction of position of α-helices, β-strands and other secondary structures along the protein chain. Even today, the prediction still can not be done in high reliability. Structure prediction mostly begins with an analysis of a database of known structures, which are examined for possible relationships between sequence and structure. The ability to predict secondary structure also depends on identifying types, location and extent of each secondary structures element in known structures. The main types are α-helices, β-strands, which examined by their sequence variations. In order to simplify the prediction the other types of structures, including other types of helices, turns, and coils, are classified as Coils.

The three-dimensional structures in the PDB file sometime include secondary structures, which must be assigned to amino acids by examination of the structural coordinates of the atoms in the PDB file. The assumption on prediction methods are based on the correlation between amino acid sequence and secondary structure. From the assumption, a short stretch of sequence tends to form a particular kind of secondary structure (e.g. helix) over another kind (e.g. sheet). Therefore, many methods examine a sequence window of 13-17 residues as a pattern of conformation of amino acid.

Moreover, a further problem in prediction of structure is that tertiary structure of protein may not be the function of the amino acid sequence alone. There are other molecules that can cause the structure of protein as well.

There is evidence that the interaction within the primary amino acid chain influences the local secondary structure. The length up to 5 residues of the same amino acid sequence can be found in different secondary structures. Furthermore, "chameleon" which is an amino acid sequence with length of 11 residues, has been found to form an α-helices when inserted into a part of a primary protein sequence and form a β-sheet when inserted into another part of the sequence. By analysis of local regions, the longer distant interactions create the result of worse prediction in β-sheet than in the other secondary structures. However, the predicting methods using small window of sequence can perform more accurately than using larger windows covering more distant amino acids.

1.2 Problem Formulation and Proposed Solutions

There are many machine learning approaches developed for solving the problem of secondary structure prediction [8, 9]. The advanced techniques of machine learning methods such as Artificial Neural Network (ANN) and Nearest-Neighbor Classification have been used as the core of classifying algorithm for the structural class of protein. Recently, Support Vector Machine (SVM), the powerful classification method based on statistical learning theory, has been applied to the problem of secondary structure prediction. Due to the good results of high percent accuracy together with a good generalization performance, SVM becomes a major classification method at the present.

Performance of the learning model used for predicting of secondary structure class depends on two main factors, designing of classification network model and preprocessing of input data. The former factor has been considered by many researcher groups and some efficient models of network design have been proposed. The later factor, on another hand, is not well studied. The efficient feature extraction method based on biological sequence data has not been explored. The method of reducing the input dimensions while keeping the essential feature of protein pattern has been studied in this research. There are three main problems to be considered in this dissertation as shown in Table 1.1.

1.3 The Contributions of Dissertation

One approach to protein structure prediction is to first predict the secondary structure as a stepping stone toward the full structure. Since secondary structure prediction is a one-dimensional problem, it is considerably less complicated than a full prediction. Bioinformatists aim to predict which secondary structural element will be formed by each residue of the protein.

Since two-dimensional structure prediction is less complicated than the full structure prediction, two-dimensional structure information is often used to predict the three-dimensional structure [11, 12, 13]. Furthermore, it is especially useful in the prediction of regions of the protein likely to undergo structural change [14] and in the classification of proteins for genome analysis [15]. For example, in a matching attempt between a candidate sequence in a sequence database or a protein motif database, if there is a significant matching result on known secondary structure, the

candidate protein may share the three-dimensional structural features of the matched protein.

Table 1.1 Three formulated problems and the proposed solutions.

	Problem	**Solutions**
I	How can we find the representative set of input features for protein sequences data?	The new encoding scheme of amino acid sequence data based on Markov model will be introduced.
II	What is an appropriate learning model that can be efficiently used to classify protein secondary structures?	The multi-layer network structure of SVM classifiers is well studied and implemented. All free parameters of SVM classifiers as well as the model structure are experimented and adjusted appropriately.
III	How can we develop the powerful predicting algorithm and software system for protein secondary structure?	A web-based application for protein secondary prediction has been developed by using the new predicting algorithm and new encoding scheme.

1.4 Scope and Organization

This dissertation is organized into four chapters starting with Introduction chapter that summarizes the basic background of protein element and protein structure. This chapter also reviews a conventional method of finding the protein structure, limitation of laboratory experiment to obtain those structures, and importance of the structure prediction by machine learning approach. Finally, the problem formulation and proposed solutions of the dissertation are also defined in this chapter.

In Chapter 2, existing techniques of protein secondary structure prediction are reviewed and divided into three generations. Then, the theoretical mathematical model of machine learning technique, Support Vector Machine (SVM), and data preprocessing technique, Markov Model, which are the major computational tools used in our experiment are presented. In chapter 3, a new algorithm for the data preprocessing technique, the "multi-orders Markov model encoding scheme", is proposed. Then, the detail of method and algorithms that are used to implement and test our approach is explained. In Chapter 4, the improving results of our experiments are presented and compared to those of other approaches. Additionally, optimizing parameters as well as improving network design are discussed. Finally, the conclusion and guideline of future works will be presented.