CHAPTER 1

INTRODUCTION

1.1 Bioinformatics

Bioinformatics is the integration of biology, informatics and mathematics and it employs computational tools and methods for managing, analyzing and manipulating sets of biological data. This is a multidisciplinary field which integrates computer science, mathematics, and engineering. Major research efforts in this field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, protein function prediction, the prediction of gene expression, protein-protein interactions, and the modeling of evolution processes. The prediction of protein function is at the heart of this work, since when sequencing a new genome, determining function and structure relationships are among the most important questions. To determine them, newly sequenced proteins can be compared to established databases or known training datasets via the use of similarity functions. Then their function and structure can either be inferred from the most similar, well-known protein sequences, or they can be predicted as a member of a known protein group by machine learning approaches like Artificial Neural Networks or Support Vector Machines.

1.2 The challenge in protein function prediction

One of the most well-known research topics in the field of bioinformatics is protein function prediction. Many researchers have proposed and developed various

computational methods for predicting different protein functions. Thus, it may be concluded that we cannot find a single computational method which can be used to predict all protein functions. Thus, we should select a suitable computational method to solve each protein function for obtaining optimized results.

Now, there are three major challenges in protein or sequence prediction. First, most of the classifiers, such as decision trees and neural networks, can only represent input data as a vector of features. However, currently, there are no explicit features in sequence data. Second, since, there are various feature selection methods; we can transform a sequence into a set of features. The dimensionality of the feature space for the sequence data can be very high and the computation can be costly. The last challenge in sequence classification is to interpret our selected classifier.

In this work, we focus on the prediction of different protein functions. We are interested in selecting computational methods that can be used to classify different datasets depending on their characteristic data. In this work, we are interested the three proteins as representative samples to show how to choose suitable computational-based methods to predict protein function.

1.3 Contributions

In the following we summarize the results by arranging them into three distinct thesis points

I. Human Leukocyte Antigen gene prediction [1]

We developed an approach to predict HLA genes using a combined method integrating a *k*-NN classifier with a spectrum kernel. We designed a series of combination parameters, which allowed us to determine relative contributions from a

k-NN classifier and spectrum kernel. Our approach was performed on 4,276 sequences using by ten-fold cross-validation. Moreover, well-known classifiers were used to compare to our proposed method. Our experimental results show that our simpler combining approach is comparable to other sophisticated and conventional methods for major class classification.

II. HIV-1 CRF01_AE coreceptor usage prediction [2]

In this work we were interested in predicting HIV-1 CRF01_AE coreceptor usage. We first used the LMT method to select the informative positions from the V3 amino acid sequence. This method found nine informative positions. The SVM classifier was then used to predict the HIV-1 CRF01_AE coreceptor usage based on the nine informative positions. Our approach was performed on 273 sequences (177 isolations showing R5 phenotype, and 96 showing X4 phenotype.), and measured by ten-fold cross-validation. Finally, our approach was compared with existing methods (genotypic or computational methods).

III. Protein crystallization prediction [3]

In this study, we utilize a newly-developed scoring card method (SCM) with a dipeptide composition feature to predict protein crystallization. The motivation of using SCM is to evaluate SCM as well as investigate protein crystallization based on the obtained propensity scores of dipeptides. The experimental results show that the SCM classifier has advantages of simplicity, high interpretability, and high accuracy in predicting protein crystallization. Moreover, the SCM method is compared with existing SVM-based ensemble classifiers.

1.4 Thesis outlines

Below is a summary of the rest of the chapters in the thesis:

Chapter 2. Computational and biological background: We review biological background, sequence classification and feature selection methods. In more details, we give a description of our selected computational-based methods, i.e. *k*-nearest neighbor (*k*-NN) classifier, support vector machine (SVM), logistic model tree (LMT), and the scoring card (SCM) method for analyzing the three protein functions.

Chapter 3. Human Leukocyte Antigen gene prediction: The goal of HLA gene prediction is to distinguish major and subclasses of the HLA gene. We examined the *k*-NN classifier based string kernel to predict Human Leukocyte Antigen (HLA) gene. We, in this work, intended to find the optimum parameter from these two methods to obtain high accuracy HLA gene prediction.

Chapter 4. HIV-1 CRF01_AE coreceptor usage prediction: We was interested in other protein function, i.e. Human Immunodeficiency Virus type 1 (HIV-I) (especially the CRF01_AE subtype). We used developed SVM classifier based the LMT method to predict HIV-1_ CRF01_AE coreceptor usage. Since exiting tools (almost) were mainly developed for B or C subtypes, these tools are not reliable to predict HIV-I CRF01_AE coreceptor usage. Thus our objective was to use a computation-based method to specifically predict the CRF01_AE subtype, and also applied the idea of feature selection (to improve performance.

Chapter 5. Protein crystallization prediction: We used the simple and efficient scoring card method to predict protein crystallization. In previous chapters we use complexity methods to predict two protein functions. Intuitively, the two methods are hard to interpret the knowledge and insights of the two protein functions. In contrast, the SCM method can overcome such problems and can be used to predict protein crystallization.

Chapter 6. Conclusions: We review the main contributions of the thesis and summarize their significance, applicability and limitations.