ENZYME SUB-FUNCTIONAL CLASS PREDICTION USING MULTI-BIOLOGICAL KNOWLEDGE FEATURE REPRESENTATION AND TWIN SUPPORT VECTOR MACHINE

SHARON KAUR A/P GURAMAD SINGH

A thesis submitted in fulfillment of the requirements for the award of the degree of Master of Science (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

NOVEMBER 2013
To my beloved late father…
Loving and understanding mother, brothers and husband…
Thank you for your immense love, prayers and support...
ACKNOWLEDGEMENT

First and foremost, praises to God. With the strength, patience and determination given by Him, I finally completed my thesis for Masters Degree. I would like to express my greatest gratitude to my supervisor, Dr. Rohayanti binti Hassan for her time, continuous guidance and encouragement throughout this research also her patience, kindness, and for her healthier supports for the past a year and 5 months. Not forgetting my co-supervisor, Dr. Muhamad Razib Bin Othman who had allocated plenty of his time in reviewing the research conducted in ensuring efficiency and consistency. Their continued motivation had ensured the success of this research at all levels. A sincere appreciation to all of my fellow friends for being supportive and leading a helping hand.

A special thanks to my family members who had always been there to support and cherish me with love and prayers, my husband for his understanding and sacrifices. I appreciate the financial support by GATES IT Solution Sdn. Bhd. under the scheme of GATES Scholars Foundation (GSF), reference no. LTR/GSF/2011-01. Lastly, I would like to extend my appreciation to those who involved indirectly in ensuring the completion of this research.
The field of computational structural biology these days has become advanced especially in the continued development of new high-throughput methods for predicting enzyme sub-functional classes. Prior knowledge of enzyme sub-functional classes has been applied in numerous important predictive tasks that address structural and functional features of enzymes. However, issues on insufficient sequence-structure knowledge, lack of known enzyme sub-functional class, low-identity sequences have caused inaccurate feature representation and imbalance distribution of enzyme sub-functional class which has contributed to low prediction results. Thus, the research proposed a derivative features vector through the consolidation of amino acid composition; dipeptide composition; hydrophobicity and hydrophilicity known as APH which is based on multi-biological knowledge. The Support Vector Machine assigns and classifies every protein sequence into its respective vector. This process would enhance the sequence-structure knowledge and overcome inaccurate feature representation. Besides that, the Twin Support Vector Machine classifies the enzyme sub-functional class and solves the imbalance distribution of enzyme sub-functional class. In this study, bio-inspired kernel function was introduced to improve the overall enzyme sub-functional class prediction. The overall results were evaluated based on accuracy, sensitivity, specificity and Matthew’s Correlation Coefficient value. Statistical and biological validation using t-test and Gene Ontology showed that the experimental results achieved an accuracy of more than 98%. Findings from the research have shown that the proposed method could assist in the prediction of the enzyme biological function, protein structure and function, protein structural class and hence provide guidance in the designing of novel drugs to cure diseases.
ABSTRAK

# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DECLARATION</td>
<td>ii</td>
</tr>
<tr>
<td></td>
<td>DEDICATION</td>
<td>iii</td>
</tr>
<tr>
<td></td>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td></td>
<td>ABSTRACT</td>
<td>v</td>
</tr>
<tr>
<td></td>
<td>ABSTRAK</td>
<td>vi</td>
</tr>
<tr>
<td></td>
<td>TABLE OF CONTENTS</td>
<td>vii</td>
</tr>
<tr>
<td></td>
<td>LIST OF TABLES</td>
<td>xi</td>
</tr>
<tr>
<td></td>
<td>LIST OF FIGURES</td>
<td>xiii</td>
</tr>
<tr>
<td></td>
<td>LIST OF ABBREVIATIONS</td>
<td>xv</td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Overview</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Background</td>
<td>2</td>
</tr>
<tr>
<td>1.3</td>
<td>Challenges of Enzyme Sub-functional Class Prediction</td>
<td>5</td>
</tr>
<tr>
<td>1.4</td>
<td>Current Methods in Enzyme Sub-functional Class Prediction</td>
<td>6</td>
</tr>
<tr>
<td>1.5</td>
<td>Problem Statement</td>
<td>6</td>
</tr>
<tr>
<td>1.6</td>
<td>Objectives of the Study</td>
<td>8</td>
</tr>
<tr>
<td>1.7</td>
<td>Scope of the Study</td>
<td>8</td>
</tr>
<tr>
<td>1.8</td>
<td>Significance of the Study</td>
<td>9</td>
</tr>
<tr>
<td>1.9</td>
<td>Organization of the Thesis</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>LITERATURE REVIEW</td>
<td></td>
</tr>
</tbody>
</table>
## 2. Introduction

### 2.1 Introduction

### 2.2 Protein Sequence

### 2.3 Enzyme Sub-Functional Class Prediction

### 2.4 Sequence-based Knowledge Representation in Enzyme Sub-functional Class Prediction

### 2.5 Significant Features Vector for Amino Acid Representation

### 2.6 Multiclass Classifier to Solve Imbalance Class Distribution Problem

### 2.7 Trends and Direction

### 2.8 Summary

## 3. Research Methodology

### 3.1 Introduction

### 3.2 Research Framework

### 3.3 Data Sources and Preparation

#### 3.3.1 Protein Sequences

#### 3.3.2 Features Vector Quantification

#### 3.3.3 Enzyme Functional and Sub-Functional Classes

### 3.4 Instrumentation and Results Analysis

#### 3.4.1 Hardware and Software Requirements

#### 3.4.2 Testing and Analysis

#### 3.4.3 Evaluation Metrics

### 3.5 Summary

## 4. Enzyme Sub-Functional Class Prediction Based on Single Feature Selection

### 4.1 Introduction

### 4.2 Materials and Methods

#### 4.2.1 Dataset Preparation

#### 4.2.2 Input Feature: The Conjoint Triad Feature (CTF)

#### 4.2.3 SVM to Solve the Classification Problem
4.3 Results and Discussion

4.3.1 Effect of Dataset in Improvement of Accuracy of Enzyme Sub-Functional Class Prediction

4.3.2 Analysis of Single Feature Selection Towards Prediction

4.3.3 Comparison to Other Classification Methods

4.3.4 Comparison to Other Related Works

4.4 Summary

5 MULTI-BIOLOGICAL BASED KNOWLEDGE FEATURES WITH SUPPORT VECTOR MACHINES FOR ENZYME SUB-FUNCTIONAL CLASS PREDICTION

5.1 Introduction

5.2 Materials and Methods

5.2.1 Dataset Preparation

5.2.2 Generation of AAC

5.2.3 Composition of Dipeptide

5.2.4 Generation of Pse-AAC Features

5.2.5 APH with SVM Classification

5.2.6 Evaluation Measurement

5.3 Results and Discussion

5.3.1 Assessment of the Most Significant Feature

5.3.2 Assessment on the Optimal Number of CV

5.3.3 Prediction on the Subclasses using the Best Classification Method

5.3.4 Prediction of Unidentified Enzyme Sub-functional Classes

5.3.5 Comparison to Other Related Works

5.4 Summary

6 INCORPORATING TWIN SUPPORT VECTOR MACHINE WITH LOW IDENTITY SEQUENCES FOR ENZYME SUB-FUNCTIONAL CLASS PREDICTION
6.1 Introduction

6.2 Materials and Methods
   6.2.1 Extraction of Amino Acid Sequences
   6.2.2 Quantification of Datasets with Various Sequence Identities (IDs)
   6.2.3 Generation of Input Features
   6.2.4 Prediction by TWSVM
   6.2.5 Kernel Selection
   6.2.6 Evaluation Measures

6.3 Results and Discussion
   6.3.1 Assessment on the Most Significant Feature using Different Rate of Sequence Similarities
   6.3.2 Assessment on the Effects of Classifiers by using Different Rate of Sequence Similarities
   6.3.3 Assessment on the Effect of Bio-inspired Kernels on TWSVM
   6.3.4 Validation on the Unclassified Enzyme Subclasses using Gene Ontology (GO)
   6.3.5 Comparison to Other Related Works

6.4 Summary

7 CONCLUSION
   7.1 Concluding Remarks
   7.2 Research Contributions
   7.3 Future Works
   7.4 Closing Remarks

REFERENCES

LIST OF RELATED PUBLICATIONS
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Examples of publicly accessible protein databases</td>
<td>17</td>
</tr>
<tr>
<td>2.2</td>
<td>Enzyme sub-functional class prediction methods</td>
<td>27</td>
</tr>
<tr>
<td>3.1</td>
<td>Analysis with different testing parameter</td>
<td>41</td>
</tr>
<tr>
<td>4.1</td>
<td>Inconsistent enzyme sub-functional class assignment between two datasets for 10 sub-functional class protein sequences from EC.3</td>
<td>46</td>
</tr>
<tr>
<td>4.2</td>
<td>An increment of accuracy (%) presented by SVM-CTF compared to Pse-AAC classification method (Chou, 2005) using $DS_I$</td>
<td>49</td>
</tr>
<tr>
<td>4.3</td>
<td>An increment of accuracy (%) presented by SVM-CTF compared to Pse-AAC classification method (Chou, 2005) using $DS_{II}$</td>
<td>49</td>
</tr>
<tr>
<td>4.4</td>
<td>Performance of different feature representations using SVM for $DS_{II}$</td>
<td>50</td>
</tr>
<tr>
<td>4.5</td>
<td>Comparison amongst classification methods</td>
<td>52</td>
</tr>
<tr>
<td>4.6</td>
<td>Comparison with other related works</td>
<td>53</td>
</tr>
<tr>
<td>5.1</td>
<td>Evaluation results on enzyme classes using different feature vectors</td>
<td>66</td>
</tr>
<tr>
<td>5.2</td>
<td>Performance comparison using various computational approaches</td>
<td>66</td>
</tr>
<tr>
<td>5.3</td>
<td>The biological validation of enzyme sub-functional class prediction</td>
<td>67</td>
</tr>
<tr>
<td>5.4</td>
<td>Performance comparison with other related works</td>
<td>68</td>
</tr>
<tr>
<td>6.1</td>
<td>Datasets used in the study</td>
<td>72</td>
</tr>
</tbody>
</table>
6.2 Samples of sequence structure from EC.3.2 in different sequence similarities represented by three input features 81
6.3 Examples of classification between previous and this study based on Gene Ontology (GO) 87
6.4 Performance comparison with other methods in predicting enzyme sub-functional classes 88
<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The differentiation in enzyme sub-functional class prediction</td>
<td>4</td>
</tr>
<tr>
<td>2.1</td>
<td>The content structure of Chapter 2</td>
<td>11</td>
</tr>
<tr>
<td>2.2</td>
<td>The characterization of enzyme sub-functional class prediction</td>
<td>12</td>
</tr>
<tr>
<td>2.3</td>
<td>Hierarchical structure of enzymes sub-functional class</td>
<td>13</td>
</tr>
<tr>
<td>2.4</td>
<td>The hierarchical structure of enzymes consist of the main and sub-functional classes</td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>An overview of the research framework</td>
<td>34</td>
</tr>
<tr>
<td>3.2</td>
<td>The process of sequence extraction and features vector quantification</td>
<td>38</td>
</tr>
<tr>
<td>4.1</td>
<td>The sub-functional assignment/prediction for Nitrogenase Molybdenum-iron protein (UniProt ID: Q57118) using Chou’s Pse-AAC classification method (Chou, 2005) and repeated method termed as SVM-CTF</td>
<td>45</td>
</tr>
<tr>
<td>4.2</td>
<td>The intensity of features vector content for Putative Thiosulfate Sulfurtransferase protein (UniProt/Swiss-Prot ID: P91247) from EC.2.8 using $DS_I$</td>
<td>51</td>
</tr>
<tr>
<td>4.3</td>
<td>The intensity of features vector content for Putative Thiosulfate Sulfurtransferase protein (UniProt/Swiss-Prot ID: P91247) from EC.2.8 using $DS_{II}$</td>
<td>51</td>
</tr>
<tr>
<td>5.1</td>
<td>Steps of dataset preprocessing</td>
<td>57</td>
</tr>
<tr>
<td>5.2</td>
<td>Overview on prediction of enzyme sub-functional classes</td>
<td>61</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Performance comparison across different method in terms of \textit{acc}</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>Performance comparison across different method in terms of \textit{MCC}</td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>The trends of feature representation for different number of CVs ranging from 5 to 15 across different main functional classes where (1) - (6) represents EC.1- EC.6 respectively</td>
<td></td>
</tr>
<tr>
<td>6.1</td>
<td>The Bio-TWSVM embodies steps of preparation of datasets and features (top), determination of the most significant feature (best feature and classifier) and the optimal sequence identities (bottom)</td>
<td></td>
</tr>
<tr>
<td>6.2</td>
<td>Quantification of datasets with different similarities using BLAST</td>
<td></td>
</tr>
<tr>
<td>6.3</td>
<td>Three different feature vector representation used in this study</td>
<td></td>
</tr>
<tr>
<td>6.4</td>
<td>Comparison in separation of hyperplanes using (i) TWSVM and (ii) SVM</td>
<td></td>
</tr>
<tr>
<td>6.5</td>
<td>Two nonlinear kernels generated based on Bio–TWSVM</td>
<td></td>
</tr>
<tr>
<td>6.6</td>
<td>Results for enzyme subclasses prediction using different rate of sequence similarities</td>
<td></td>
</tr>
<tr>
<td>6.7</td>
<td>Results based on classifiers in terms of \textit{acc} for various rate of sequence similarities</td>
<td></td>
</tr>
<tr>
<td>6.8</td>
<td>The difference in classification process using (i) ANN, (ii) KNN, (iii) SVM and (iv) TWSVM classifier based on sequence from EC.3.2</td>
<td></td>
</tr>
<tr>
<td>6.9</td>
<td>Performance comparison in terms of sensitivity and specificity for selected subclasses using (i) Fisher; (ii) Mismatch; and (iii) Spectrum kernels</td>
<td></td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>AAC</td>
<td>Amino Acid Composition</td>
<td></td>
</tr>
<tr>
<td>AFKNN</td>
<td>Adaptive Fuzzy K-Nearest Neighbor</td>
<td></td>
</tr>
<tr>
<td>APH</td>
<td>Hybrid of Amino Acid Composition, Dipeptide Composition, Hydrophobicity and Hydrophilicity</td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>Circular Dichroism</td>
<td></td>
</tr>
<tr>
<td>CDA</td>
<td>Covariant Discriminant Algorithm</td>
<td></td>
</tr>
<tr>
<td>CTF</td>
<td>Conjoint Triad Feature</td>
<td></td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
<td></td>
</tr>
<tr>
<td>DPC</td>
<td>Dipeptide Composition</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>Decision Trees</td>
<td></td>
</tr>
<tr>
<td>EC</td>
<td>Enzyme Commission</td>
<td></td>
</tr>
<tr>
<td>HH</td>
<td>Hydrophobicity and Hydrophilicity</td>
<td></td>
</tr>
<tr>
<td>IUBMB</td>
<td>International Union of Biochemistry and Molecular Biology</td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td>K-Nearest Neighbor</td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>Naïve Bayesian</td>
<td></td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear Magnetic Resonance</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td>Neural Network</td>
<td></td>
</tr>
<tr>
<td>PC</td>
<td>Personal Computer</td>
<td></td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
<td></td>
</tr>
<tr>
<td>PPI</td>
<td>Protein-protein Interaction</td>
<td></td>
</tr>
<tr>
<td>Pse-AAC</td>
<td>Pseudo Amino Acid Composition</td>
<td></td>
</tr>
<tr>
<td>RAM</td>
<td>Random Access Memory</td>
<td></td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
<td></td>
</tr>
<tr>
<td>TWSVM</td>
<td>Twin Support Vector Machine</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Overview

Enzyme sub-functional class plays an important role in the foundation of enzyme structure information and hence leading to the determination of enzyme function in field of biomedicine. In finding the structure and function of an enzyme, a useful first step is predicting the functional class of enzymes and thereafter its sub-functional classes. Since enzymes are made up of proteins, the protein three dimensional structures can also be identified and utilized in identifying the details of interaction of protein with other biomolecules and finally providing guidelines to infer protein function. Chou and Elrod (2003) stated that the sequence-structure gap is widening rapidly due to the unavailability of protein sequences. Therefore, by predicting enzyme sub-functional class and assigning those into corresponding structures and functions may reduce this gap.

Prior to the time, several features vector and computational methods have been applied in prediction of the enzyme sub-functional class from their amino acid sequences. Enzyme sub-functional class prediction for supervised machine learning based method is gaining wide spread attention in the field of computational biology. Several in-depth review of computational methods used for predicting enzyme sub-functional class using different machine learning approach can be found. The predictions are performed using variety of classification algorithms in early research
includes Support Vector Machine (SVM: Wang et al., 2010; Shi and Hu, 2010, Zhou et al., 2007), Neural Network (NN: Huang et al., 2007; Shen and Chou, 2007; Naik et al., 2007) and Random Forest (Kumar and Choudhary, 2012). Hence, supervised machine learning technique gives a remarkable improvement of more than 80% in prediction quality as well as generalization capability in managing nonlinear classification.

The remainder of the chapter will provide a basic concepts regarding enzyme sub-functional class prediction using biological based knowledge. This is crucial as the thorough understanding on the fundamental information that is related to this research is needed. The following few sections will discuss the background and challenges as well as current respective solutions, toward achieving precise enzyme sub-functional class prediction. Research goal, objectives, scopes and significance ensue thereafter. The chapter ends with thesis organization.

1.2 Background

Enzymes are made up of proteins which are the fundamental components of all living cells. They are made up of a combination of varying amounts of the same 20 amino acids in sequence linked by peptide bonds. Enzymes cater most of the important functions, such as catalysis of biochemical reactions, transcription factors to guide the differentiation of the cell and its later responsiveness to signals, transport of materials in body fluids, receptors for hormones and other signaling molecules, and formation of tissues and muscular fiber. It is widely believed that the protein enzyme structures play key roles in determining its functions. However, it is extremely labor-expensive and sometimes even impossible to experimentally determine the structures for every protein sequence.

In pharmaceutical, the structure and function of enzymes are used to design drugs (Singh et al., 2010; Pisal et al., 2010). In addition to probe those structure and function, the knowledge of functional classes is essential. The primary knowledge of
enzyme main and sub-functional classes is significant as it exemplify essential information that can be used to infer enzyme structures related in understanding the biological function of an enzyme used vastly as therapeutic strategy. Other than that, the knowledge of sub-functional class of enzyme can be applied to identify a sickle protein (Drotar, 2010) in which the enzyme sub-functional class can be discerned using amino acids content. In early study, Chou and Elrod (2003) have catalogued the enzyme functional class into six common classes namely oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases abbreviated as EC.1, EC.2, EC.3, EC.4, EC.5, and EC.6 respectively.

In enzyme sub-functional class prediction, computational methods have been gaining widespread attention due to the laborious and time-consuming constraints in experimental wet lab or also known as in vivo methods. Enzyme sub-functional class is represented either based on knowledge-based method (Chou, 2005; Cai and Chou, 2005; Chou and Elrod, 2003; Shi and Hu, 2010) or chemical atomic-based potentials (Szefczyk, 2008; Calzada et al., 2009; Lin and Oliver, 2008). The former approach is highly complicated in which it needs to determine the enzyme sub-functional class by calculating the detailed amino acids coordinates that traversed a vast number of accessible polypeptide conformations. In contrast, knowledge-based method exploits the structures information of enzymes from in vivo analysis. However, both methods rely upon tedious visual inspection or statistical inference from the sequence.

In addition, the enzyme sub-functional class is known to yield relatively small number of proteins. In the most recent release, ExPasy database (Gasteiger et al., 2005) based on recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) includes 5026 active entries. Meanwhile, enzyme sub-functional based on known domains are listed in the recent ENZYME database (Bairoch, 2000), release of 19-Oct-2011 and UniProt/Swiss-Prot database (Boutet et al., 2007), release of 21-Sept-2011 which contains 538,010 sequence entries comprising 190,998,508 amino acids abstracted from 213,490 references. These databases illustrate a huge gap between known sequence and known enzyme sub-functional class in which only 1%-2% of the sequences can be assigned to the corresponding enzyme sub-functional class.
Inspired by the aforementioned challenges, this study is devoted to further investigate how enzyme sub-functional class prediction using computational methods with the application of the information of biological knowledge, can be more beneficial than the ones based on the information of protein sequences (Huang et al., 2007).

Several computational classification methods have been introduced. Support Vector Machine (SVM, NN, Bayesian classification (Green and Karp, 2004; Borro et al., 2006), Random Forest and Decision Trees (Syed and Yona, 2009; Syed and Yona, 2003) are amongst many classification methods, which are able to exploit the latent pattern within the identified structures of enzyme. However, there is still the challenge of representing the underlying pattern with significant features vector; from a simple features vector to represent the known enzyme structures such as amino acid composition (Chou and Elrod, 2003), pseudo amino acid composition (Chou, 2005; Chou and Cai, 2004; Cai and Chou, 2005), polypeptide composition (Shi and Hu, 2010), conjoint triad feature based on protein-protein interaction (PPI: Wang et al., 2010; Wang et al., 2011), to a more complex hybrid features vector that considers evolutionary information encoded in PSI-Blast profiles (Liu et al., 2010; Tung et al., 2007). In addition, some features vector exhibit inferior prediction performance when it lack in sequence identity.

![Figure 1.1: The differentiation in enzyme sub-functional class prediction](image)
1.3 Challenges of Enzyme Sub-functional Class Prediction

Although ENZYME and UniProt/Swiss-Prot are examples of well-established databases that contain more reliable information of enzyme sub-functional class, yet the lack of known sub-functional class of enzyme due to the laborious wet-lab experimental routine limits the high throughput enzyme sub-functional assignment. As a consequence, the assignment of enzyme sub-functional class by computational method suffers from the low prediction accuracy. In turn, the first challenge belongs to the unclassified enzyme sub-functional class prediction which limits the sequence-structure class assignment.

In order to produce an accurate sequence-structure assignment, the second challenge must be tackled, which is pertaining to the investigation of heterogeneous physiochemical characteristics of amino acids in specified sequence of protein. These physiochemical characteristics are transformed into numerical value and used to represent the input features vector for the enzyme sub-functional class prediction method. Unfortunately, the prediction performance is often poor because of the inaccurate features vector is used to signify the heterogeneous characteristics (Costantini et al., 2010; Chou, 2005). The situation is aggravated in the presence of low-identity sequences (Tian and Skolnick, 2003).

The third challenge stems from the nature of sequences length in every enzyme sub-functional class that exhibits the imbalance class distribution misleads the prediction of enzyme sub-functional class. Consequently, some researchers resorted to this issue using multi-class analysis (Wang et al., 2010; Jayadeva et al., 2009; Reshma et al., 2008). As a result, this will lead to over or underfitted prediction model for some particular classes if used without the aid of suitable kernel selection.
1.4 Current Methods in Enzyme Sub-functional Class Prediction

Generally, current methods for enzyme sub-functional class prediction can be categorized into two: experimental based and computational based (the details are presented in Chapter 2):

(i) Experimental based method predicts the enzyme sub-functional class of protein from physical characterization of the functional class when \textit{in vivo} analysis is employed. It can be identified either from the primary protein structure using X-ray crystallography (Palioura et al., 2009; Joosten et al., 2008) and nuclear magnetic resonance (NMR) spectroscopy (Sudhamsu et al., 2010; Liras and Demain, 2009; Cámara et al., 2009), or from the enzyme structure using circular dichroism (CD) spectroscopy (Dodsworth and Leigh, 2007; Kim and Mrksich, 2010; Shi et al., 2002) and Raman spectroscopy (Leadbeater and Schmink, 2008; Aki et al., 2010; Malo et al., 2008).

(ii) Computational based method upon input of the protein sequence predicts the enzyme sub-functional class of enzyme by utilizing mathematical inference and/or computational algorithms. It can be broadened into two categories: knowledge-based method and chemical atomic-based potentials method. Consecutively, knowledge-based method is branched into four major categories: pseudo amino acid composition (Chou, 2005; Cai and Chou, 2005; Cai et al., 2005), amino acid composition (Chou and Elrod, 2003; Esmaeili et al., 2010), functional domain composition (Cai and Chou, 2004; Cai and Chou, 2005; Cai and Chou, 2006; Chou and Cai, 2004) and polypeptides/peptides composition (Costantini et al., 2010; Shi and Hu, 2010; Ding and Zhang, 2008; Zhang and Luo, 2003).

1.5 Problem Statement

To date, classification of enzyme sub-functional class using sequence-structure knowledge instead of the sequence information is still a hot research field
and has been gaining various attentions. The enzyme sequence-structure gap is growing tremendously at a rapid pace. Generally, due to the numerous active genomes and sequencing projects, there exist more protein sequences to be classified within a certain period of time as compared to solving enzyme structures and functions. Hence, to reduce the distance of the gap, efficient computational approach has been introduced to predict enzyme sub-functional class. Based on the above mentioned challenges (Section 1.2), some factors will need to be addressed by the possible solution.

The first factor is related to the insufficient knowledge of known enzyme sub-functional captured during in vivo. It is observed that the quantities of known sequences are growing exponentially with respect to the quantity of known enzyme sub-functional (Chou and Elrod, 2003). The wide sequence-structure gap has a direct effect on the enzyme sub-functional class prediction. Thus, this study aims to provide an enzyme sub-functional class prediction method that can acquire the biological based knowledge, derived from known excessive protein sequences, in order to produce high-throughput sequence-structure class assignment instead of the laborious experimental based method.

The second factor is pertaining to the inaccurate feature representation of the protein sequences. Recently, large quantity and high-identity of sequences hold the key to achieve higher accuracy in enzyme sub-functional class prediction. In contrast, this study aims to generate alternative features vector that is more robust without degrading the prediction performance. In this study, the biological based features carried on sequence level were introduced in the predictive of enzyme sub-functional classification. The additional sequence order and sequence length knowledge is expected to avoid the inconsistency of enzyme sub-functional class prediction.

The third factor is related to the imbalance class distribution of enzyme sub-functional class due to the amount of sequences in every class is irregular. Consequently, classification rules become too restrictive due to the unsteady amount of protein sequences acquired during in vivo. More specifically, it suffers from the
tightly bounded maximum or minimum margins when classifying the enzyme sub-functional class using the conventional SVM classifier. Hence, an optimized Twin SVM method is proposed to rectify such inadequacy.

1.6 Objectives of the Study

The goal of this study is to predict the enzyme sub-functional class from the protein sequences using the multi-biological features and multi-class classifier as computational method. This can be objectified into:

(i) To construct the optimum single feature with the incorporation of SVM algorithm in order to bridge the sequence-structure knowledge.
(ii) To develop the multi-biological knowledge based feature representation in order to improve the accuracy of the enzyme sub-functional class prediction.
(iii) To optimize the multi-class classifier algorithm by exploiting the newly designed features vector in (ii) to resolve the imbalance classification issue in enzyme sub-functional class prediction.

1.7 Scope of the Study

(i) This research uses dataset obtained from ENZYME and UniProt/Swiss-Prot database (Borgwardt et al., 2005).
(ii) International Commission on Enzymes to annotate the function of enzymes by the Enzyme Commission (EC) number (Bairoch, 2000).
(iii) The use of amino acid composition (AAC), dipeptide composition and hydrophilic and hydrophobic properties to attain the sequence order and sequence level information.
(iv) The introduction of APH feature which is the consolidation between AAC, dipeptide composition, hydrophobicity and hydrophilicity
properties as an efficient sequence encoding methods for representing given protein sequence.

(v) Twin SVM by incorporating the bio-inspired kernel function machine learning technique is used in order to solve the multiclass classification problem.

(vi) The prediction performances are assessed: (i) computationally: in terms of accuracy, sensitivity as well as specificity; and (ii) biologically: by cross-checking against ENZYME database and Gene Ontology. Finally, \textit{t-test} is employed for statistical validation.

1.8 Significance of the Study

The significance of this study can be branched into two main categories: computational and biological aspects. From computational aspect, the proposed method is intended to precisely predict the enzyme sub-functional class from protein sequences with low quantity and identity. It serves as an alternative for laborious and time consuming task of experimental prediction. From biological aspect, enzyme sub-functional class embodies structural information that provides detail insight into protein functionalities such as prediction of the outer membrane protein (Gao et al., 2010), prediction of the structural class (Kurgan et al., 2008) and prediction of the subcellular localization of protein (Xie et al., 2005). In molecular medicine, enzymes are used to design highly specialized drugs for treating diseases. For example, in the treatment of Type I diabetes, human insulin is given fast and slow reaction forms of damage β structure cell of the islet Langerhans (Chen et al., 2010). In the investigation of sickle protein, enzyme sub-functional knowledge is considered a milestone. For example, the sickle protein in anaemia cell arose from the substitution of glutamate by valine at the sixth position of the β subunit structure of haemoglobin (Drotar, 2010). Furthermore, enzyme sub-functional knowledge can be adopted as a therapeutic strategy in which it inhibits the function of viral diseases. For example, in cholera treatment, some structural routes have been devised to minimize the viral infection (Bimczok et al., 2010).
1.9 Organization of the Thesis

This thesis is organized into seven chapters. A brief description on the content of each chapter is given below:

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

(ii) Chapter 2 reviews the main subjects of interest, which are enzyme sub-functional class prediction, computational based method for enzyme sub-functional class prediction, imbalance classification rules, biological based knowledge structure and significant features vector.

(iii) Chapter 3 presents the research methodology of the computational method that supports the objectives of the study. This includes data sources, instrumentations and analyses.

(iv) Chapter 4 lays out the development of the SVM-CTF that is resilient towards insufficient sequence-structure knowledge of known enzyme sub-functional class. The prediction result is validated and compared against experimentally-determined enzyme sub-functional class from Wang et al. (2010). SVM-CTF is an abbreviation of SVM with Conjoint Triad Feature for enzyme sub-functional class prediction.

(v) Chapter 5 describes the APH feature that addresses the problem of heterogeneous characteristics of amino acids as well as low-identity sequences and uncertain feature representation by integrating significant features vector using the biological based knowledge. APH is an abbreviation of consolidation between (a) amino acids, (b) dipeptide composition, (c) hydrophobicity and hydrophilicity properties of protein sequence.

(vi) Chapter 6 proposes an extension to the baseline method, namely the Bio-TWSVM introduces an additional bio-inspired kernel component represented by Twin SVM classification, so as to overcome the imbalance class distribution in enzyme sub-functional class of particular sequence.

(vii) Chapter 7 draws general conclusions of the accomplished results and presents the contributions of the study as well as recommends the potential enhancements for future study.
REFERENCES


# LIST OF RELATED PUBLICATIONS

<table>
<thead>
<tr>
<th>NO.</th>
<th>PUBLICATIONS</th>
<th>RELATED CHAPTERS</th>
</tr>
</thead>
</table>