ENZYME SUB-FUNCTIONAL CLASS PREDICTION USING MULTI-BIOLOGICAL KNOWLEDGE FEATURE RESPRESENTATION 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.

ABSTRACT

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

Bidang pengkomputeran dalam struktur biologi telah berkembang pesat melalui kaedah pemprosesan berteknologi tinggi terutamanya dalam bentuk ramalan. Pengetahuan tentang kelas sub-fungsi enzim digunakan dalam pelbagai tugasan ramalan bagi menangani ciri-ciri struktur dan fungsi enzim. Walau bagaimanapun, isu tentang pengetahuan urutan struktur yang tidak mencukupi, kekurangan dalam kelas sub-fungsi enzim yang telah diramalkan, jujukan identiti rendah telah menyebabkan perwakilan ciri-ciri yang tidak tepat dan ketidakseimbangan kelas sub-fungsi enzim yang menyumbang kepada keputusan ramalan yang rendah. Oleh itu, kajian ini mencadangkan kaedah yang menggunakan ciri-ciri vektor terbitan melalui penyatuan komposisi asid amino; komposisi *dipeptide*; *hydrophobicity* dan *hydrophilicity* dinamakan APH yang berasaskan pengetahuan pelbagai sumber biologi. *Support Vector Machine* mewakil dan mengelaskan setiap urutan protein ke dalam vektor masing-masing. Proses ini akan meningkatkan pengetahuan urutan struktur dan mengatasi perwakilan ciri-ciri yang tidak tepat. Selain itu, *Twin Support Vector Machine* mengklasifikasikan kelas sub-fungsi enzim dan menyelesaikan ketidakseimbangan kelas sub-fungsi enzim. Dalam kajian ini, fungsi kernel bio-inspirasi telah diperkenalkan untuk meningkatkan ramalan kelas sub-fungsi enzim secara keseluruhan. Keputusan ramalan telah dinilai berdasarkan ketepatan, kepekaan, keperincian dan nilai pekali korelasi *Matthews*. Pengesahan statistik dan biologi menggunakan ujian-*t* dan Ontologi Gen menunjukkan pencapaian nilai ketepatan melebihi 98%. Hasil penemuan kaedah kajian ini dapat membantu dalam ramalan fungsi biologi, struktur dan fungsi enzim protein, kelas struktur protein serta memberi panduan dalam merekabentuk ubat-ubatan baharu untuk menyembuhkan pelbagai jenis penyakit.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAC	–	Amino Acid Composition
AFKNN	–	Adaptive Fuzzy K-Nearest Neighbor
APH	–	Hybrid of Amino Acid Composition, Dipeptide Composition, Hydrophobicity and Hydrophilicity
CD	–	Circular Dichroism
CDA	–	Covariant Discriminant Algorithm
CTF	–	Conjoint Triad Feature
DNA	–	Deoxyribonucleic Acid
DPC	–	Dipeptide Composition
DT	–	Decision Trees
EC	–	Enzyme Commission
HH	–	Hydrophobicity and Hydrophilicity
IUBMB	–	International Union of Biochemistry and Molecular Biology
KNN	–	K-Nearest Neighbor
NB	–	Naïve Bayesian
NMR	–	Nuclear Magnetic Resonance
NN	–	Neural Network
PC	–	Personal Computer
PDB	–	Protein Data Bank
PPI	–	Protein-protein Interaction
Pse-AAC	–	Pseudo Amino Acid Composition
RAM	–	Random Access Memory
RNA	–	Ribonucleic Acid
SVM	–	Support Vector Machine
TWSVM	–	Twin Support Vector Machine

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.

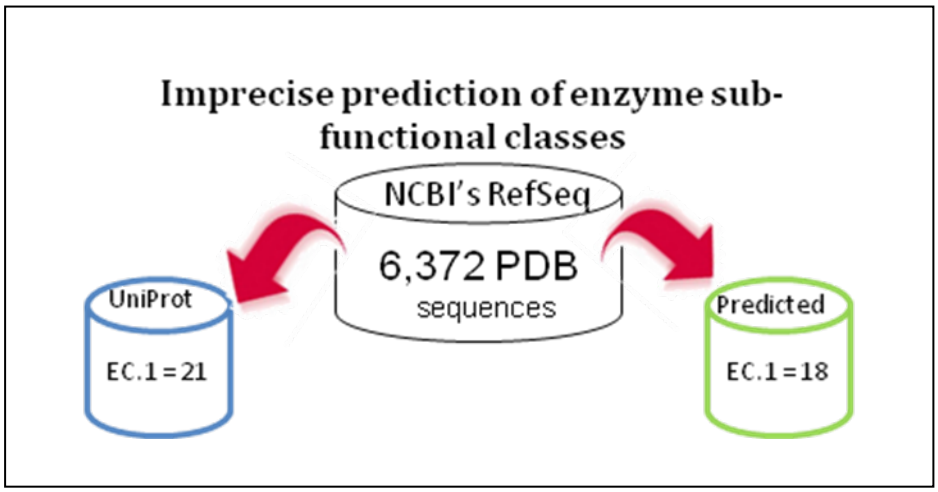

Figure 1.1: The differentiation in enzyme sub-functional class prediction

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 *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, *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

Aki, Y., Nagai, M., Nagai, Y., Imai, K., Aki, M., Sato, A., Kubo, M., Nagatomo, S., Kitagawa, T., 2010. Differences in Coordination States of Substituted Tyrosine Residues and Quaternary Structures among Hemoglobin M Probed by Resonance Raman Spectroscopy. *Journal of Biological Inorganic Chemistry*, 15(2), pp.147–158.

Anfinsen, C.B., 1973. Principles That Govern the Folding of Protein Chains. *Science*, 181(4096), pp.223–230.

Arjunan, S.P., Kumar, D.K., Naik, G.R., 2010. A Machine Learning Based Method For Classification Of Fractal Features Of Forearm Semg Using Twin Support Vector Machines. *Proceedings of the Annual International Conference on IEEE Engineering in Medicine and Biology Society*, (Buenos Aires, Argentina), pp.4821–4824.

Bairoch, A., 2000. The ENZYME Database in 2000. *Nucleic Acids Research*, 28(1), pp.304–305.

Barthel, D., Hirst, J.D., Blaewicz, J., Burke, E.K., Krasnogor, N., 2007. ProCKSI: A Decision Support System for Protein (Structure) Comparison, Knowledge, Similarity and Information. *BMC Bioinformatics*, 8(1), p.416.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. *Nucleic Acids Research*, 28(1), pp.235–242.

Billaut-Laden, I., Allorge, D., Crunelle-Thibaut, A., Rat, E., Cauffiez, C., Chevalier, D., Houdret, N., Lo-Guidice, J.M., Broly, F., 2006. Evidence for a Functional Genetic Polymorphism of The Human Thiosulfate Sulfurtransferase (Rhodanese), A Cyanide and H2S Detoxification Enzyme. *Toxicology*, 225(1), pp.1–11.

Bimczok, D., Verdonck, F., Hartig, R., Cox, E., Rothkotter, H.J., 2010. Primary Porcine CD11R1+ Antigen Presenting Cells Isolated from Small Intestinal Mucosa Mature but Lose Their T Cell Stimulatory Function in Response to Cholera Toxin Treatment. *Veterinary Immunology and Immunopathology*, 134(3-4), pp.239–248.

Borgwardt, K.M., Ong, C.S., Schonauer, S., Vishwanathan, S.V.N., Smola, A.J., Kriegel, H.P., 2005. Protein Function Prediction via Graph Kernels. *Bioinformatics*, 21(Suppl. 1), pp.i47–i56.

Borro, L.C., Oliveira, S.R., Yamagishi, M.E., Mancini, A.L., Jardine, J.G., Mazoni, I., Santos, E.H., Higa, R.H., Kuser, P.R., Neshich, G., 2006. Predicting Enzyme Class From Protein Structure Using Bayesian Classification. *Genetics and Molecular Research*, 5(1), pp.193–202.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bairoch, A., 2007. UniProtKB/Swiss-Prot. *Methods in Molecular Biology*, 406(1), pp.89–112.

Brown, D.P., Krishnamurthy, N., Sjolander, K., 2007. Automated Protein Subfamily Identification and Classification. *PLoS Computational Biology*, 3(8), pp.1526–1538.

Cai, Y.D., Chou, K.C., 2005. Predicting Enzyme Subclass by Functional Domain Composition and Pseudo Amino Acid Composition. *Journal of Proteome Research*, 4(3), pp.967–971.

Cai, Y.D., Chou, K.C., 2006. Predicting Membrane Protein Type by Functional Domain Composition and Pseudo-amino Acid Composition. *Journal of Theoretical Biology*, 238(2), pp.395–400.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002. Artificial Neural Network Method for Predicting Protein Secondary Structure Content. *Computers and Chemistry*, 26(4), pp.347–350.

Calzada, J., Zamarro, M.T., Alcón, A., Santos, V.E., Díaz, E., García, J.L., Garcia-Ochoa, F., 2009. Analysis of Dibenzothiophene Desulfurization in a Recombinant Pseudomonas Putida Strain. *Applied and Environmental Microbiology*, 75(3), pp.875–877.

Cámara, B., Nikodem, P., Bielecki, P., Bobadilla, R., Junca, H., Pieper, D.H., 2009. Characterization of a Gene Cluster involved in 4-chlorocatechol Degradation

by Pseudomonas Reinekei MT1. *Journal of Bacteriology*, 191(15), pp.4905–4915.

Chang, C.C., Lin, C.J., 2011. LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27), pp.1–27.

Chen, C., Chen, L.X., Zou, X.Y., Cai, P.X., 2008. Predicting Enzyme Subclass based on Multi-Features Fusion. *Journal of Theoretical Biology*, 253(2), pp.388–392.

Chen, C., Zhou, X., Tian, Y., Zou, X., Cai. P., 2006. Predicting Enzyme Subclass with Pseudo-Amino Acid Composition and Support Vector Machine Fusion Network. *Analytical Biochemistry*, 357(1), pp.116–121.

Chen, W., Wang, J., Wang, E., Lu, Y., Lau, S.K., Lawrence, M., Huang, Q., 2010. Detection of Clonal Lymphoid Receptor Gene Rearrangements in Langerhans Cell Histiocytosis. *American Journal of Surgical Pathology*, 34(7), pp.1049–1057.

Chinnasamy, A., Sung, W.K. , Mittal, A., 2005. Protein Structure and Fold Prediction Using Tree-augmented Naive Bayesian Classifier. *Journal of Bioinformatics and Computational Biology*, 3(4), pp.803–819.

Chou, K.C., 1989. Low-frequency Resonance and Cooperativity of Hemoglobin. *Trends in Biochemical Sciences*, 14(6), pp.212–213.

Chou, K.C., 2005. Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *BMC Bioinformatics*, 21(1), pp.10–19.

Chou, K.C., Cai, Y.D., 2004. Using GO-PseAA Predictor to Predict Enzyme Sub-class. *Biochemical and Biophysical Research Communications*, 325(2), pp.506–509.

Chou, K.C., Elrod, D.W., 2003. Prediction of Enzyme Family Classes. *Journal of Proteome Research*, (2)2, pp.183–190.

Chou, W.Y., Pai, T.W., Jiang, T.Y., Chou, W.I., Tang, C.Y., Chang, M.D.T., 2011. Hydrophilic Aromatic Residue and In Silico Structure for Carbohydrate Binding Module. *PLoS ONE*, 6(9), p.e24814.

Christopher, J.C.B., 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2), pp.121–167.

Costantini, S., Costantini, M., Colonna, G., 2010. Frequencies of Specific Peptides in Intrinsic Disordered Protein Domains. *Protein and Peptide Letters*, 17(11), pp.1398–1402.

Crammer, K., Singer, Y., 2001. On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines. *Journal of Machine Learning Research*, 2, pp.265–292.

Dessailly, B.H., Adam, A.J., Yeats, C., Lees, J.G., Cuff, A., Orengo, C.A., 2009. The Evolution of Protein Functions and Networks: A Family-centric Approach. *Biochemical Society Transactions*, 37(Part 4), pp.745–750.

Ding, S., Yu, J., Qi, B., Huang, H., 2012. An Overview on Twin Support Vector Machines. *Artificial Intelligence Review*.

Ding, Y.S., Zhang, T.L., 2008. Using Chou's Pseudo Amino Acid Composition to Predict Subcellular Localization of Apoptosis Proteins: An Approach with Immune Genetic Algorithm-based Ensemble Classifier. *Pattern Recognition Letters*, 29(13), pp.1887–1892.

Dodsworth, J.A., Leigh, J.A., 2007. Nifi Inhibits Nitrogenase by Competing with Fe Protein for Binding to the MoFe Protein. *Biochemical and Biophysical Research Communications*, 364(2), pp.378–382.

Drotar, D., 2010. Treatment Adherence in Patients with Sickle Cell Anemia. *Journal of Pediatrics*, 156(3), pp.415–419.

Esmaeili, M., Mohabatkar, H., Mohsenzadeh, S., 2009. Using the Concept of Chou's Pseudo Amino Acid Composition for Risk Type Prediction of Human Papillomaviruses. *Journal of Theoretical Biology*, 263(2), pp.203–209.

Fetrow, J.S., 2006. Active Site Profiling to Identify Protein Functional Sites in Sequences and Structures Using the Deacon Active Site Profiler (DASP). *Current Protocols in Bioinformatics*, 14(8), p.10.1–8.10.16.

Fogle, E.J., van der Donk, W.A., 2007. Pre-Steady-State Studies of Phosphite Dehydrogenase Demonstrate that Hydride Transfer is Fully Rate Limiting. *Biochemistry*, 46(45), pp.13101–13108.

Freeman, T.C., Wimley, W.C., 2010. A Highly Accurate Statistical Approach for the Prediction of Transmembrane β-barrels. *Bioinformatics*, 26(16). pp.1965–1974.

Gao, Q.B., Ye, X.F., Jin, Z.C., He, J., 2010. Improving Discrimination of Outer Membrane Proteins by Fusing Different Forms of Pseudo Amino Acid Composition. *Analytical Biochemistry*, 398(1), pp.52–59.

Garg, A., Raghava, G.P.S., 2008. A Machine Learning based Method for the Prediction of Secretory Proteins using Amino Acid Composition, their Order and Similarity-Search. *In Silico Biology*, 8(2), pp.129–140.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., Bairoch, A., 2003. ExPASy: The Proteomics Server for In-Depth Protein Knowledge and Analysis. *Nucleic Acids Research*, 31(13), pp.3784 – 3788.

Getreuer, P., 2011. Linear methods for image interpolation. *Image Processing On Line*.

Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A. Thornton, J.M., 2006. A Method for Localizing Ligand Binding Pockets in Protein Structures. *Proteins*, 62(2), pp.479–488.

Gronwald, W., Hohm, T., Hoffman, D., 2008. Evolutionary Pareto-optimization of Stably Folding Peptides. *BMC Bioinformatics*, 9(1), p.109.

Haykin, S.S., 1999. *Neural Networks: A Comprehensive Foundation*, (New Jersey, USA).

Hermes, M., Osswald, H., Riehle, R., Piesch, C., Kloor, D., 2008. S-Adenosylhomocysteine Hydrolase Overexpression in HEK-293 Cells: Effect on Intracellular Adenosine Levels, Cell Viability, and DNA Methylation. *Cellular Physiology and Biochemistry*, 22(1-4), pp.223–236.

Hu, X., Wang, T., 2011. Prediction of Enzyme Subclass by using Support Vector Machine based on Improved Parameters. *Proceedings of 7th International Conference in Natural Computation*, pp.593–598.

Huang, B., Schroeder, M., 2006. LIGSITEcsc: Predicting Ligand Binding Sites Using the Connolly Surface and Degree of Conservation. *BMC Structural Biology*, 6(1), p.9.

Huang, W.L., Chen, H.M., Hwang, S.F., Ho, S.Y., 2007. Accurate Prediction of Enzyme Subfamily Class Using an Adaptive Fuzzy K-Nearest Neighbor Method. *Biosystems*, 90(2), pp.405–413.

Huang, W.L., Chen, H.M., Hwang, S.F., Ho, S.Y., 2007. Accurate Prediction of Enzyme Subfamily Class using an Adaptive Fuzzy K-Nearest Neighbor Method. *BioSystems*, 90(2), pp.405– 413.

Inbar, Y., Benyamini, H., Nussinov, R., Wolfson, H. J., 2003. Protein Structure Prediction via Combinatorial Assembly of Sub-structural Unit. *Bioinformatics*, 19(1), pp.i158–i168.

Jayadeva, Khemchandni, R., Chandra, S., 2007. Twin Support Vector Machines for Pattern Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5), pp.905–910.

Joosten, H.J., Han, Y., Niu, W., Vervoort, J., Dunaway-Mariano, D., Schaap, P.J., 2008. Identification of Fungal Oxaloacetate Hydrolyase within the Isocitrate Lyase/PEP Mutase Enzyme Superfamily using a Sequence Marker-Based Method. *Proteins*, 70(1), pp.157–166.

Juncker, A.S., Jensen, L.J., Pierleoni, A., Bernsel, A., Tress, M.L., Bork, P., Heijne, G.V., Valencia, A., Ouzounis, C.A., Casadio, R., Brunak, S., 2009. Sequence-based Feature Prediction and Annotation of Proteins. *Genome Biology*, 10(2), p.206.

Karplus, M., Shakhnovich, E., 1992. Theoretical Studies of Thermodynamics and Dynamics. In Creighton, T.E. (Ed.) *Protein Folding*, pp.127–195, (New York, USA).

Kearns, M., Valiant, L.G., 1994. Cryptographic Limitations on Learning Boolean Formulae and Finite Automata. *Journal of the ACM*, 41(1), pp.67–95.

Khemchandani, R., Jayadeva, Chandra, S., 2008. Optimal Kernel Selection in Twin Support Vector Machines. *Optimization Letters*, 3(1), pp.77–88.

Kim, J., Mrksich, M., 2010. Profiling the Selectivity of DNA Ligases in an Array Format with Mass Spectrometry. *Nucleic Acids Research*, 38(1), p.e2.

Kumar, C., Choudhary, A., 2012. A top-down approach to classify enzyme functional classes and sub-classes using random forest. *EURASIP Journal on Bioinformatics and Systems Biology*, 2012(1), pp.1–14.

Kurgan, L., Zhang, T., Zhang, H., Shen, S., Ruan, J., 2008b. Secondary Structure based Assignment of the Protein Structural Classes. *Amino Acids*, 35(3), pp.551–564.

Leadbeater, N.E., Schmink, J.R., 2008. Use of Raman Spectroscopy as a Tool for In Situ Monitoring of Microwave-promoted Reactions. *Natures Protocol*, 3(1), pp.1–7.

Lee, B.J., Lee, H.G., Lee, J.Y., Ryu, K.H., 2007. Classification of Enzyme Function from Protein Sequence based on Feature Representation. *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering*, (Boston, USA), pp.741–747.

Leslie, C., Eskin, E., Cohen, A., Weston, J., Noble, W.S., 2004. Mismatch String Kernels for Discriminative Protein Classification. *Bioinformatics*, 20(4), pp.467–476.

Leslie, C., Eskin, E., Noble, W.S., 2002. The Spectrum Kernel: A String Kernel for SVM Protein Classification. *Proceedings of the 7th Pacific Symposium on Biocomputing*, (Hawaii, USA), pp.564–575.

Li, B., Hu, J., Hirasawa, K., 2008. Support Vector Machine Classifier with WHM Offset for Unbalanced Data. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 12(1), pp.94–101.

Li, B., Hu, J., Hirasawa, K., Sun, P., Marko, K., 2006. Support Vector Machine with Fuzzy Decision-Making for Real-World Data Classification. *In IEEE World Congres on Computational Intelligence International Joint Conference on Neural Networks*, pp.587–592.

Li, X.C., Wang, L., Sung, E., 2008. AdaBoost with SVM-based Component Classifiers. *Applications of Artificial Intelligence*, 21(5), pp.785–795.

Liang, J., 2007. Computation of Protein Geometry and Its Applications: Packing and Function Prediction. In Xu, Y., Xu, D., Liang, J. (Eds.) *Physics and Astronomy*, pp.181–206, (Berlin, Germany).

Lin, M., Oliver, D.J., 2008. The Role of Acetyl-coenzyme a Synthetase in Arabidopsis. *Plant Physiol*, 147(4), pp.1822–1829.

Liras, P., Demain, A.L., 2009. Chapter 16: Enzymology Of Beta-Lactam Compounds with Cephem Structure Produced by Actinomycete. *Methods in Enzymology*, 458, pp.401–429.

Loening, A.M., Andrasi, A.D., Gambhir, S.S., 2010. A Red-shifted Renilla Luciferase for Transient Reporter-gene Expression. *Nature Methods*, 7(1), pp.5–6.

Lotan, I., Schwarzer, F., 2004. Approximation of Protein Structure for Fast Similarity Measures. *Journal of Computational Biology*, 11(3), pp.299–317.

Lu, J., Luo, L., Zhang, L., Chen, W., Zhang, Y., 2010. Increment of Diversity with Quadratic Discriminant Analysis - An Efficient Tool for Sequence Pattern Recognition in Bioinformatics. *Open Access Bioinformatics*, 2010(2), pp.89–96.

Lundqvist. J., Elmlund, H., Wulff, R.P., Berglund, L., Elmlund, D., Emanuelsson, C., Hebert, H., Willows, R.D., Hansson, M., Lindahl, M., Al-Karadaghi, S., 2010. ATP-induced Conformational Dynamics in the AAA+ Motor Unit of Magnesium Chelatase. *Structure*, 18(3), pp.354–365.

Mackay, D.H.J., Cross, A.J., Hagler, A.T., 1989. The Role of Energy Minimization in Simulation Strategies of Biomolecular Systems. In Fasman, G.D. (Ed.) Prediction of Protein Structure and the Principles of Protein Conformation, pp.317–4358, (New York, USA).

Malo, G.D., Wang, M., Wu, D., Stelling, A.L., Tonge, P.J., Wachter, R.M., 2008. Crystal Structure and Raman Studies of Dsfp483, a Cyan Fluorescent Protein from Discosoma Striata. *Journal of Molecular Biology*, 378(4), pp.871–886.

McCammon, J.A., Wong, C.F., Lybrand, T.P., 1989. Protein Stability and Function. In Fasman, G.D. (Ed.) Prediction of Protein Structure and the Principles of Protein Conformation, pp.149–4159, (New York, USA).

McDonald, A.G., Boyce, S., Moss, G.P., Dixon, H.B.F., Tipton, K.F., 2007. ExplorEnz: A MySQL Database of the IUBMB Enzyme Nomenclature. *BMC Biochemistry*, 8, p.14.

Mohammed, A., Guda, C., 2011. Computational Approaches for Automated Classification of Enzyme Sequences. *Journal of Proteomics and Bioinformatics*, 4(8), pp.147–152.

Naik, P.K., Mishra, V.S., Gupta, M., Jaiswal, K., 2007. Prediction of Enzymes and Non-enzymes from Protein Sequences based on Sequence Derived Features and PSSM Matrix using Artificial Neural Network. *Bioinformation*, 2(3), pp.107–112.

Nalivaeva, N.N., Fisk, L.R., Belyaev, N.D., Turner, A.J., 2008. Amyloid-degrading Enzymes as Therapeutic Targets in Alzheimer's Disease. *Current Alzheimer Research*, 5(2), pp.212–224.

Nigsch, F., Bender, A., Jenkins, J.L., Mitchell, J.B., 2008. Ligand-target Prediction Using Window and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *Journal of Chemical Information Model*, 48(12), pp.2313–2325.

Pál, C., Papp, B., Lercher, M.J., 2006. An Integrated View of Protein Evolution. *Nature Reviews*. *Genetics*, 7(5), pp.337–348.

Palioura, S., Sherrer, R.L., Steitz, T.A., Söll, D., Simonovic, M., 2009. The Human Sepsecs-Trnasec Complex Reveals the Mechanism of Selenocysteine Formation. *Science*, 325(5938), pp.321–325.

Pisal, D.S., Koloski, M.P., Balu-Iyer, S.V., 2010. Delivery on Therapeutic Proteins. *Journal of Pharmaceutical Sciences*, 99(6), pp.2557–2575.

Prudnikova, T., Mozga, T., Rezacova, P., Chaloupkova, R., Sato, Y., Nagata, Y., Brynda, J., Kuty, M., Damborsky, J., Smatanova, I.K., 2009. Crystallization and Preliminary X-ray Analysis of a Novel Haloalkane Dehalogenase DbeA from Bradyrhizobium Elkani USDA94. *Structural Biology and Crystallization Community*, 65(Part 4), pp.353–356.

Pruittt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI Reference Sequences (Refseq): A Curated Non-redundant Sequence Database of Genomes, Transcripts and Proteins. *Nucleic Acids Research*, 35(Database Issue), pp.D61–D65.

Pugalenthi, G., Tang, K., Suganthan, P.N., Chakrabarti, S., 2009. Identification of Structurally Conserved Residues of Proteins in Absence of Structural Homologs Using Neural Network Ensemble. *Bioinformatics*, 25(2), pp.204–210.

Punta, M., Ofran, Y., 2008. The Rough Guide to In Silico Function Prediction, or How to Use Sequence and Structure Information to Predict Protein Function. *PLoS Computational Biology*, 4(10), p.e1000160.

Qiu, J.D., Huang, J.H., Shi, S.P., Liang, R.P, 2010. Using the Concept of Chou's Pseudo Amino Acid Composition to Predict Enzyme Family Classes: An Approach with Support Vector Machine based on Discrete Wavelet Transform. *Protein and Peptide Letters*, 17(6), pp.715–722.

Rouhier, N., Unno, H., Bandyopadhyays, S., Masip, L., Kim, S.K., Hirasawa, M., Gualberto, J.M., Lattard, V., Kusunoki, M., Knaff, D.B, Georgious, G., Hase,

T, Johnsons, M.K, Jacquot, J.P., 2007. Functional, Structural and Spectroscopic Characterization of a Glutathione-Ligated [2Fe-2S] Cluster in Poplar Glutaredoxin C1. *The National Academy of Sciences*, 104(18), pp.7379–7384.

Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J., Schomburg, D., 2011. BRENDA, the Enzyme Information System in 2011. *Nucleic Acids Research*, 39(Database issue), pp.D670–D676.

Schoelkopf, B., Smola, A.J., 2002. Learning with Kernels. *MIT Press*, (Cambridge, USA).

Shao, Z., Zhao, H., Zhao, H., 2009. DNA Assembler, an in vivo Genetic Method for Rapid Construction of Biochemical Pathways. *Nucleic Acids Research*, 37(2), p.e16.

Shen, H.B., Chou, K.C., 2007. EzyPred: A Top-down Approach for Predicting Enzyme Functional Classes and Subclasses. *Biochemical and Biophysical Research Communications*, 364(1), pp.53–59.

Shi, R., Hu, X., 2010. Predicting Enzyme Subclasses by using Support Vector Machine with Composite Vectors. *Protein and Peptide Letters*, 17(5), pp.599–604.

Shi, Z., Woody, R.W., Kallenbach, N.R., 2002. Is Polyproline II a Major Backbone Conformation in Unfolded Protein? *Advances in Protein Chemistry*, 62(1), pp.163–240.

Singh, N.P., Tyagi, V.P., Ratnam, B., 2010. Synthesis and Spectroscopic Studies of Tetradentate Schiff Base Complexes of Cu(II), Ni(II), Mn(II) and Co(II). *Journal of Chemical and Pharmaceutical Research*, 2(1), pp.473–477.

Söhngen C., Chang A., Schomburg D., 2011. Development of a Classification Scheme for Disease-Related Enzyme Information. *BMC Bioinformatics*, 12, p.329.

Sonnenburg, S., Zien, A., Philips, P., Ratsch, G., 2008. POIMs: Positional Oligomer Importance Matrices–understanding Support Vector Machine-based Signal Detectors. *Bioinformatics*, 24(13), pp.i6–i14.

Sreerama, N., Woody, R.W., 2003. Structural Composition Of $Beta_I$- and $Beta_{II}$-Proteins. *Protein Science*, 12(1), pp.384–388.

Sudhamsu, J., Kabir, M., Airola, M.V., Patel, B.A., Yeh, S.R., Rousseau, D.L., Crane, B.R., 2010. Co-Expression of Ferrochelatase Allows for Complete Heme Incorporation Into Recombinant Proteins Produced In E. Coli. *Protein Expression and Purification*, 73(1), pp.78–82.

Syed, U., Yona, G., 2009. Enzyme Function Prediction with Interpretable Models. *Methods in Molecular Biology*, 541, pp.373–420.

Szefczyk, B., 2008. Towards Understanding Phosphonoacetaldehyde Hydrolase: An Alternative Mechanism involving Proton Transfer that Triggers P-C Bond Cleavage. *Chemical Communications*, 35, pp.4162–4164.

Tian, W., Skolnick, J., 2003. How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity? *Journal of Molecular Biology*, 333(4), pp.863–882.

Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y., 2005. Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research*, 6(1), pp.1453–1484.

Tsuda, K., Akaho, S., Kawanabe, M., Muller, K.R., 2004. Asymptotic Properties of the Fisher Kernel. *Neural Computation*, 16(1), pp.115–137.

Tung, C.H., Huang, J.W., Yang, J.M., 2007. Kappa-alpha Plot Derived Structural Alphabet and BLOSUM-like Substitution Matrix for Rapid Search of Protein Structure Database. *Genome Biology*, 8(1), p.R31.

Tyagi, N., Givvimani, S., Qipshidze, N., Kundu, S., Kapoor, S., Vacek, J.C., Tyagi, S.C., 2010. Hydrogen Sulfide Mitigates Matrix Metalloproteinase-9 activity and Neurovascular Permeability in Hyperhomocysteinemic Mice. *Neurochemistry International*, 56(2), pp.301–307.

Voss, N.R., Gerstein, M., 2010. 3V: Cavity, Channel and Cleft Volume Calculator and Extractor. *Nucleic Acids Research*, 38(2), pp.W555–W562.

Walder, C., Kim, K.I., Schölkopf, B., 2008. Sparse Multiscale Gaussian Process Regression. *Proceedings of the 25th International Conference on Machine Learning*. (New York, USA), pp.1112–1119.

Walsh, I., Bau, D., Martin, A.J.M., Mooney, C., Vullo, A., Pollastri, G., 2009. Ab Initio and Template-based Prediction of Multi-class Distance Maps by Two-dimensional Recursive Neural Networks. *BMC Structural Biology*, 9(1), p.5.

Wang, P., Ownby, S., Zhang, Z., Yuan, W., Li, S., 2010. Cytotoxicity and Inhibition of DNA Topoisomerase I of Polyhydroxylated Triterpenoids and Triterpenoid Glycosides. *Bioorganic and Medical Chemistry Letters*, 20(9), pp.2790–2796.

Wang, Y.C., Wang, X.B., Yang, Z.X., Deng, N.Y., 2010. Prediction of Enzyme Subfamily Class via Pseudo Amino Acid Composition by Incorporating the Conjoint Triad Feature. *Protein and Peptide Letters*, 17(11), pp.1441–1439.

Wang, Y.C., Wang, Y., Yang, Z.X., Deng, N.Y., 2011. Support Vector Machine Prediction of Enzyme Function with Conjoint Triad Feature and Hierarchical Context. *BMC Systems Biology*, 5(Suppl. 1), p.S6.

Wang, Y., Hu, X., 2011. Predicting of Oxidoreductase and Lyase Subclasses by using Support Vector Machine. *Computer and Information Science 10th International Conference*, (Sanya, China), pp. 27–31.

Webb, E.C., 1992. Enzyme Nomenclature. *Academic Press*, (San Diego, USA). Available at: http://www.chem.qmul.ac.uk/iubmb/enzyme [Accessed January 07, 2012].

Webb, G.I., Janice, R., Wang, B.Z., 2005. Not So Naive Bayes: Aggregatin One-dependence Estimators. *Machine Learning*, 58(1), pp.5–24.

Wroblewska, L., Jagielska, A., Skolnick, J., 2008. Development of a Physics-based Force Field for the Scoring and Refinement of Protein Models. *Biophysical Journal*, 94(8), pp.3227–3240.

Wu, G., Chang, E.Y., 2005. KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), pp.786–795.

Wuthrich, K., 2002. NMR Studies of Structure and Function of Biological Macromolecules. *Journal of Biomolecular NMR*, 27(1), pp.13–39.

Xie, D., Li, A., Wang, M., Fan, Z., Feng, H., 2005. LOCSVMPSI: A Web Server for Subcellular Localization of Eukaryotic Proteins Using SVM and Profile of PSI-BLAST. *Nucleic Acids Research*, 33(2), pp.W105–W110.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., Madden, T., 2012. Primer-BLAST: A Tool to Design Target-specific Primers for Polymerase Chain Reaction. *BMC Bioinformatics*, 13(1), p.134.

Yu, C.S., Chen, Y.C., Lu, C.H., Hwang, J.K., 2006. Prediction of Protein Subcellular Localization. *Proteins*, 64(3), pp.643–651.

Zeczycki, T.N., St Maurice, M., Jitrapakdee, S., Wallace, J.C., Attwood, P.V., Cleland, W.W., 2009. Insight into the Carboxyl Transferase Domain Mechanism of Pyruvate Carboxylase from Rhizobium Etli. *Biochemistry*, 48(20), pp.4305–4313.

Zhang, T.L., Ding, Y.S., Chou, K.C., 2008. Prediction Protein Structural Classes with Pseudo-Amino Acid Composition: Approximate Entropy and Hydrophobicity Pattern. *Journal of Theoretical Biology*, 250(1), pp.186–193.

Zhang, L.R., Luo, L.F., 2003. Splice Site Prediction with Quadratic Discriminate analysis using Diversity Measure. *Nucleic Acids Research*, 31(21), pp.6214–6220

Zhou, X.B., Chen, C., Li, Z.C., Zou, X.Y., 2007. Using Chou's Amphiphilic Pseudo-Amino Acid Composition and Support Vector Machine for Prediction of Enzyme Subfamily Classes. *Journal of Theoretical Biology*, 248(3), pp.546–551.

LIST OF RELATED PUBLICATIONS

NO.	PUBLICATIONS	RELATED CHAPTERS
1.	**Journal (Author)**	
	Guramad S., Sharon K., Hassan, R., Othman, R. M., Asmuni, H. (2012). Incorporating Multiple Biology based Knowledge to Amplify the Prophecy of Enzyme Sub-functional Classes. *Protein and Peptide Letters*. Under Review. Impact Factor 2013: 1.942.	5
2.	Guramad S., Sharon K., Hassan, R., Othman, R. M., Asmuni, H. (2012). Twin Support Vector Machine With Bio-Inspired Optimization And Multi-Biological Based Knowledge For Enzyme Sub-Functional Classes Prediction. *Computers in Biology and Medicine*. Under Review. Impact Factor 2012: 1.089	6
	Proceeding (Author)	
1.	Guramad S., Sharon K., Hassan, R. (2012). Classification of Enzyme Subfamily Class based on Feature Representation. *Postgraduate Annual Research Seminar (PARS'12).* 27-29 Novermber, Johor, Malaysia: UTM.	5