International Journal on Advanced Science Engineering Information Technology

Vol.7 (2017) No. 4
ISSN: 2088-5334

Incorporating Multiple Biology based Knowledge to Amplify the Prophecy of Enzyme Sub-Functional Classes

Sharon K. Guramad S. [1], Rohayanti Hassan[1*], Razib M. Othman[1], Hishammuddin Asmuni[1], and Shahreen Kasim[2]

[1]*Universiti Teknologi Malaysia, 81310 UTM Skudai, Malaysia*

[2] *Soft Computing and Data Mining Centre, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn, Johor, Malaysia*

**E-mail: rohayanti@utm.my*

***Abstract*— Based on current in silico methods, enzyme sub-functional classes is distinguished from sequence level information, local order or sequence length and order knowledge. To date, no work has been done to predict the enzyme subclasses efficiently corresponding to the ENZYME database. In order to precisely predict the sub-functional classes of enzyme, we propose a derivative feature vector labelled as APH which unifies amino acid composition, dipeptide composition, hydrophobicity and hydrophilicity. Support Vector Machine is used for prediction and the performance is evaluated using accuracy obtained over 99% and Matthew's Correlation Coefficient *(MCC)* over 0.99 with the aid of biological validation from in vivo studies.**

***Keywords*— Enzyme sub-functional classes, amino acid composition, dipeptide composition, hydrophobicity and hydrophilicity, support vector machine.**

I. INTRODUCTION

An enzyme functional and sub-functional class plays a major role in protein evolution studies [1], structural class [2, 3], subcellular localization [4] as well as protein function prediction [5, 6]. According to the ENZYME database, enzymes can be categorized to six main functional classes based on their Enzyme Commission (EC) number accessible from (http://www.brenda-enzymes.org/) namely oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases abbreviated as EC.1, EC.2, EC.3, EC.4, EC.5, and EC.6 respectively. All of these main classes can be further classified into its sub-functional classes as illustrated in Figure 1. The primary knowledge of enzyme main and subclasses is significant as it embodies essential information that classes can be used to infer protein structures related in understanding the biological function of a protein used vastly as therapeutic strategy [7]. However, to date, no researches carried out had correctly predicted the sub-functional classes which correspond to the ENZYME database. This is due to the previous methods that focused merely on the use of sequence level information [8] with lack or no sequence order and sequence length knowledge [9, 10] to identify the subclasses.

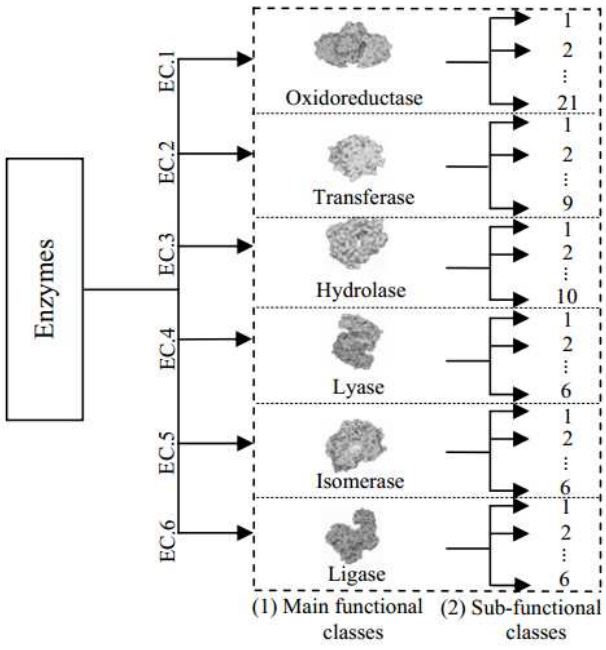

Fig. 1. The hierarchical structure of enzymes consists of the main and sub-functional classes

For instance, for EC.1, in Wang et al. [10] the number of subclasses correctly predicted is 12; Zhou et al. [9] and Huang et al. [11] predicted 16 subclasses and Shen and Chou [12] and Shi and Hu [13] predicted up to 18 subclasses whereas the ENZYME database has 21 subclasses all in all. Thus, by adopting the aforementioned rudiments, we can distinguish the uncategorized sequences in each class efficiently.

Concerning to the prediction of enzyme sub-functional classes, many attempts [9-16] have been made using computational methods such as SVM [9, 11, 13, 15-16, 20]. It is believed that biological based knowledge has a strong relation to derive the enzyme sub-functional classes. This is because enzymes play an important role in regulating and initiating every biological reaction [8]. In regard to this, amino acid composition (AAC) is one of the initial successful methods used extensively due to its ingenuous characteristics [17]. Fundamentally, AAC is dependent on the proportion of amino acid residue occurrences quantified using statistical method found in the respective enzyme protein sequence [18]. In previous work, up to 90% [10, 12-13] accurate prediction of the enzyme sub-functional classes can be attained using the modified versions of AAC representation which are the pseudo amino acid composition (Pse-AAC) [10] and amphiphilic pseudo-amino acid composition (Am-Pse-AAC) [9]. The improved performances of Pse-AAC and AmPse-AAC was accredited to the incorporation of several encoding features and machine learning algorithm as in functional domain [12, 19], conjoint triad feature (CTF) [10, 20], and gene ontology (GO) [21] adopted using Support Vector Machines (SVM) [9, 13, 15-16, 20] and K-Nearest Neighbor (KNN) [11] classifier. Though, the discrete model of PseAAC is of use in statistical prediction but complex in terms of integrating the sequence-order information whereas the Am-Pse-AAC could lead to an infinite number of sample patterns due to different types of amphiphilic features. Thus, we believe that finding the optimal amphiphilic features for instance hydrophobicity [22] and hydrophilicity [23]; as input vector could lead to a better prediction. Previous work [13] done had proved that the use of these features increased the overall prediction accuracy by 3.4% compared to Shen and Chou's results [12]. Subsequently, the dipeptide composition yield upon the hydrolysis of two amino acids resorted by many researchers in expressing the enzyme sequence information efficiently such as the increment of diversity [24].

The aforementioned methods in predicting sub-functional classes of enzymes each have their own merits. As is well known, an enzyme sub-functional class is very dense that involves many physical and chemical properties. For this kind of complicated biological system, it would be particularly effective to treat it by assembling many individual predictors with each operated based on its own special feature. Hence, this paper focuses primarily on presenting a new alternative feature vector, abbreviated as APH. APH is the consolidation between (i) AAC, (ii) dipeptide composition, (iii) hydrophobicity and hydrophilicity properties of protein sequence in order to tackle three major elements: (1) sequence level information, (2) local order of protein sequences, and (3) sequence order and length respectively. In computational approach, SVM [20] is utilized in predictive tasks to deal with multi-class classification problems. By utilizing APH and SVM, we evaluate the prediction outcome based on accuracy (acc) and Matthew's Correlation Coefficient (MCC).

Additionally, biological validation is also done as supporting source. The rest of the paper is organized as follows. In Section 2, we illustrate the performance of computational methods and algorithms on the benchmark datasets. Section 3 presents the experimental results and discussion. Finally, the last section deduces the summary of this study.

II. Material And Methods

A. Dataset

The sub-functional classes were classified into each of the six main enzyme classes based on the accession numbers extracted from the ENZYME database at ftp://ftp.expasy.org/databases/enzyme/ (Release of 19-Oct-2011). The corresponding protein sequences represented by its EC number were taken from the databank of Uniprot/Swiss-Prot at http://www.ebi.ac.uk/swissprot/ (Release of 21-Sep- 2011). The dataset used in [10] and [13] were also applied in order to examine the effectiveness of the proposed method as compared to the previously used techniques. The benchmark dataset were curated based on the following screening procedures as illustrated in Figure 2: (1) a redundancy cutoff was set to avoid any homologous bias where no sequences had $\geqslant$ 25% sequence identity to any other, (2) enzymes with less than 50 amino acids were excluded to avoid fragment data, and (3) enzymes consist of multi-domain proteins with multiple enzymatic functions was removed. Thus, six datasets (SeqEC.1 – SeqEC.6) were constructed based on the main functional classes which can be formulated as:

$$Seq_k = \sum_{i=1}^{w} S_{k.i} + \cdots + S_{k.w} \quad (1)$$

i=1 where Seq_k = {$k \in$EC.1, EC.2, EC.3, EC.4, EC.5, EC.6} are defined as the main functional classes of enzyme, $S_{k.i}$ and $S_{k.w}$ are the sub-functional classes given that w = {21, 9, 10, 6, 6, 6} respectively which varies depending on the elements in k.

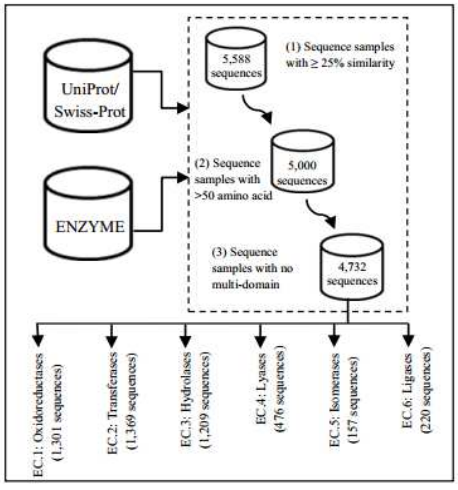

Fig. 2. Steps of dataset preprocessing

From Figure 2, each of the enzyme functional class carries its own enzymatic functionalities which contribute to the genesis of interesting problems in bioinformatics [8]. The six functional classes are classified as the following: (i) EC.1: take responsibility in catalyzing oxidation and reduction reactions, (ii) EC.2: used in a functional group transfer reaction from one molecule to another, (iii) EC.3: responsible for catalyzing the hydrolysis reaction and breakage of single bonds, (iv) EC.4: take responsibility in formation or removal of double bond with group transfer, (v) EC.5: functions in catalyzing the isomerization of functional groups within a single molecule, and (vi) EC.6: catalyzing the single bond formation by eliminating the elements of water from two functional groups to form a single bond.

B. Generation of AAC

AAC alone performs at its best with existing yet more complex features indicating the presence of sequence-level information that is predictive of interaction, but which is not necessarily restricted to domains. AAC is a fraction of each amino acid present in the protein sequence. Suppose a protein sequence with *L* amino acid residues:

$$R_1R_2R_3R_4R_5\ldots R_L \tag{2}$$

where *R* represents the amino acid residue and the subscript number represents the position of amino acid residue of length, *L* in a protein sequence (Seq_k).

If λ is the length of protein sequence and β_i is the frequency of occurrence of an amino acid *i*, then AAC_i is:

$$AAC_i = \frac{\beta_i}{\lambda} \tag{3}$$

where *i* is any of the 20 amino acids.

C. Composition of Dipeptide

In order to implement information about frequency as well as local order of residues in proteins, we also constructed dipeptide composition (DPC) based model. DPC is considered as better feature as compared to AAC as it encapsulates global as well as local information of the sequence. The DPC based model encompasses the information about AAC along local order of amino acid. It gives the fixed pattern length of a vector with 400 (20x20) dimensions. The fraction of each dipeptide, DPC_i was computed using the following equation:

$$DPC_i = \frac{\sigma_{ij}}{\omega} \tag{4}$$

where *i*, *j* are any of the 20 amino acid residues, σ_{ij} is the fraction of a pair of amino acids (*i*, *j* = 1, 2,...,20) and ω is the total number of all possible dipeptides.

D. Generation of Pse-AAC Features

The concept of Pse-AAC concerning on the use of hydrophobicity and hydrophilicity factors was proposed in order to avoid a complete lost in the sequence order information. In contrast with the conventional AAC that contains 20 components with each reflecting the occurrence frequency for one of the 20 native amino acids in a protein, the essence of Pse-AAC is that it includes information beyond AAC where the first 20 represent the components of its conventional AAC while the additional factors reflects the sequence order effect of a protein through a discrete model. Thus, according to the definition of Pse-AAC, a protein sequence can be expressed as a vector *P* which is formulated as follows:

$$P = \{P_1,..., P_{20}, P_{20+\lambda}\} \tag{5}$$

where the first 20 numbers in Eq. (5) represent the classic AAC, and the next λ discrete numbers describe sequence correlation factor which is the hydrophobicity and hydrophilicity values calculated based on [9] by the following equation:

$$h^1(i) = \frac{h_\varphi^1(i) - T_1}{\sqrt{\frac{\sum_{i=1}^{20}\left[h_\varphi^1(i) - T_1\right]^2}{20}}}, T_1 = \sum_{i=1}^{20}\left[h_\varphi^1(i)/20\right] \tag{6}$$

$$h^2(i) = \frac{h_\varphi^2(i) - T_2}{\sqrt{\frac{\sum_{i=1}^{20}\left[h_\varphi^2(i) - T_2\right]^2}{20}}}, T_2 = \sum_{i=1}^{20}\left[h_\varphi^2(i)/20\right] \tag{7}$$

where *i* is the indices of amino acid residue; $_{h\varphi}1$ and $h_\varphi{}^2$ are the original hydrophobic and hydrophilic values of the *i*th amino acid.

E. APH with SVM Classification

In the present study, a freely downloadable package of SVM which is the SVM^{light} (*http://svmlight.joachims.org/*) has been used to predict the sub-functional classes of enzyme. The define parameters used such as the linear inbuilt kernel function by adopting the 10-fold cross validation (CV) technique. Since the prediction of sub-functional classes is a multi-class classification problem, we constructed *N* SVMs for *N* class classification. Here, the class number was equal to six for enzyme functional classes. Hence, *i*th SVM was trained with all the samples in the *i*th class with positive label and negative label for the sequences of remaining sub-functional classes. This kind of SVM is known as one versus all (1-v-a) SVM. The training and testing dataset was partitioned in ratio of 0.8:0.2.

We implemented the hybrid module which encapsulates the complete information of a protein such as AAC, DPC and hydrophobicity and hydrophilicity properties known as the APH features to all of the six datasets ($Seq_{EC.1} - Seq_{EC.6}$) used. SVM was provided with an input vector of 441 dimensions that consisted of 20 features for AAC, 400 features of DPC, and 21 features of the hydrophobicity and hydrophilicity factors.

F. Evaluation Measurement

In order to assess the prediction performances, *acc* and MCC_i were calculated as described by [19] using equations:

$$acc = \frac{\sum_{i=1}^{k}\rho(i)}{N} \tag{8}$$

$$MCC_i = \frac{\rho_i q_i - \tau_i S_i}{\sqrt{(\rho_i + S_i)(\rho_i + \tau_i)(q_i + S_i)(q_i + \tau_i)}} \tag{9}$$

where *i* is the index of a particular subclass, ℕ refers to the number of sequences predicted in subclass *i*, ρ_i represents the number of correctly predicted sequences of class *i*, q_i is

the number of correctly predicted sequences not of class i, τ_i represents the number of incorrectly predicted sequences of class i and s_i is the number of incorrectly predicted sequences not of class i.

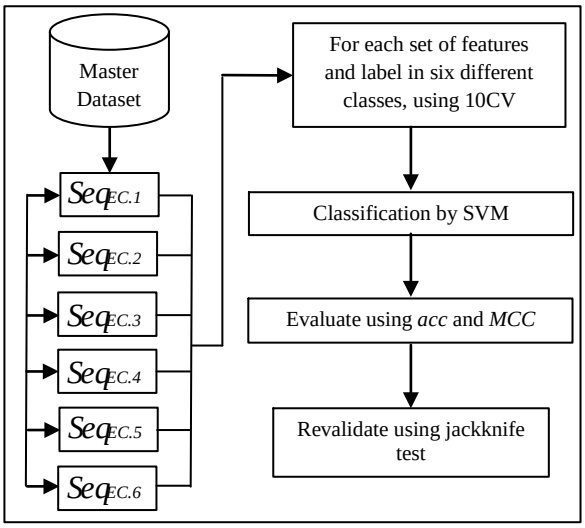

Fig. 3. Prediction of enzyme sub-functional classes.

III. Results And Discussion

A. Assessment of the most significant feature

The most significant feature is assessed using two measurements: *acc* is used to assess the degree of correctly predicted sub-functional classes with respect to the ground truth; *MCC* is used to access the degree of true and false positives and negatives and used even if the classes are of very different sizes.

Figure 4 and 5 presents the prediction performance that has been achieved using the discussed measurements.

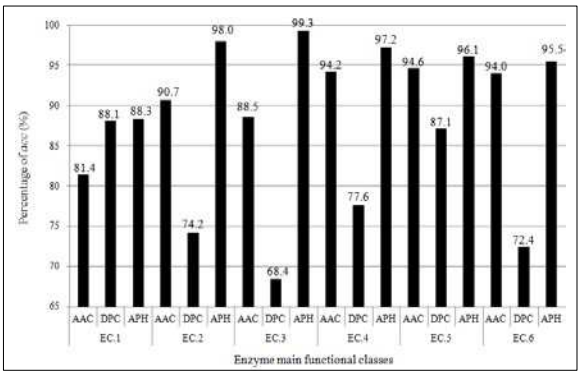

Fig. 4. Performance comparison across different method in terms of *acc*.

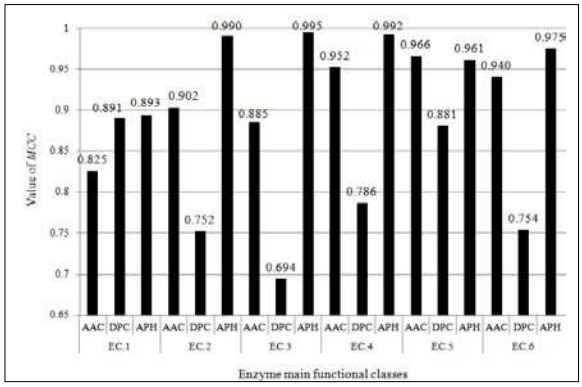

Fig. 5. Performance comparison across different method in terms of *MCC*.

Based on the figures above, it shows APH outperformed the others in main classes for both *acc* and *MCC* with 99.3% and 0.995 in EC.2 and EC.3 respectively. This can be due to the property of APH which takes into account the knowledge of sequence order and length information. From these results, it suggests that the more properties of dataset are incorporated into the predictive model; an improved result can be obtained. Despite the promising results from all functional classes, the *MCC* value in EC.5 for APH is slightly lower than the AAC. It can be mainly due to the lesser number of sequences being analyzed as compared to the other classes. Otherwise, APH performs relatively well even if the protein sequences are of low homologous to each other.

B. Assessment on the optimal number of CV

From Supplementary Figure 1, it shows the trends of enzyme functional class elements with different number of CVs; represented as 5, 8, 10, 12, and 15. According to Chan and Lin [25], a notable increment trend for all class occurs within the 10-fold CV. Surprisingly, all classes also exhibit similar trends with an optimal value at peak with 10CV. This strongly owns to the definition that 10CV tends to provide less biased estimation of the *acc*. By employing the optimal 10CV using APH as feature, the results obtained outperforms with the highest *acc* of 96.1% for EC.2. This explains the feature representation criteria which considers the intrinsic information of the sequences used in this study.

C. Prediction on the subclasses using the best classification method

In this paper, we compared three classifiers namely SVM, KNN and Naïve Bayes (NB) for predicting sub-functional classes of enzyme for low homologies to known enzymes. From Supplementary Table 1, the SVM classifier outperformed the others for all classes. This reports that SVM is capable of solving the imbalance multi-class classification problem which occurs in predicting enzyme sub-functional classes by improving the predictive rule. SVM with APH as sequence-based feature showed a promising output with the highest *acc* of 95.7% for EC.2. This result simply implies that SVM works at its best with feature representation of higher degree where several

properties of problem that is needed to be solved had been taken into consideration.

D. Prediction of unidentified enzyme sub- functional classes

TABLE I
THE BIOLOGICAL VALIDATION OF ENZYME SUB-FUNCTIONAL CLASS PREDICTION.

Sub-functional Class	Enzyme Functional Class				
	ENZYME database	**This study**	**Wang *et al.* [10]**	**Shi and Hu [13]**	**Shen and Chou [12]**
EC.1.19 [37]	EC.1	EC.1	unknown	unknown	unknown
EC.1.20 [39]	EC.1	EC.1	unknown	unknown	unknown
EC.1.21 [38]	EC.1	EC.1	unknown	unknown	unknown
EC.2.9 [30]	EC.2	EC.2	unknown	unknown	unknown
EC.3.3 [32]	EC.3	EC.3	unknown	unknown	unknown
EC.3.7 [33]	EC.3	EC.3	unknown	unknown	unknown
EC.3.8 [34]	EC.3	EC.3	unknown	unknown	unknown
EC.3.11 [35]	EC.3	EC.3	unknown	unknown	unknown
EC.3.13 [36]	EC.3	EC.3	unknown	unknown	unknown
EC.2.8 [31]	EC.2	EC.2	unknown	EC.2	EC.2
EC.4.4 [29]	EC.4	EC.4	unknown	EC.4	EC.4
EC.4.99 [28]	EC.4	EC.4	unknown	EC.4	EC.4
EC.5.5 [40]	EC.5	EC.5	unknown	EC.5	EC.5
EC.5.99 [41]	EC.5	EC.5	unknown	EC.5	EC.5
EC.6.2 [42]	EC.6	EC.6	unknown	EC.6	EC.6
EC.6.4 [43]	EC.6	EC.6	unknown	EC.6	EC.6
EC.6.5 [45]	EC.6	EC.6	unknown	EC.6	EC.6
EC.6.6 [44]	EC.6	EC.6	unknown	EC.6	EC.6

Discriminating biochemical structure transformation patterns is an initial step toward reaction prediction. Thus, several studies have been carried out by researchers that have scientifically proven the existence of biochemical reaction in enzyme prediction via *in vivo* validation. The sub-functional class EC.2.9 currently has only one sub-subclass EC.2.9.1 which is selenotransferases, despite the very broad definition of enzymes transferring selenium-containing groups. This sub-functional class contains miscellaneous enzymes and includes several reactions for which the classification may have to be reviewed further by incorporating further knowledge-based context of functional groups.

E. Comparison to other related works

According to Table 2, comparing the prediction performance with existing works is difficult due to the different specification of computational framework and datasets. However, it is obvious that previous works uses larger number of sequences with greater similarity than our method. To the best of our knowledge, the best acc_{all} was from Wang *et al*. [41] with 93.5%, using CTF as features. Notwithstanding of utilizing only 25% sequences ID, our proposed method surpasses others at 95.9%. We believe this was due to the efficiency of APH to precisely predict the enzyme sub-functional classes. This was directly accredited to the hybridization of different features to form APH, which cancelled out each other weaknesses. Moreover, many other works used fixed number of subclasses as previously done.

TABLE II
PERFORMANCE COMPARISON WITH OTHER RELATED WORKS.

Sequence Similarity/Features Vector/Method	**References**	acc_{all} **(%)**
<40ED/PseAAC+CTF/AMSVM	[41]	93.5
≤40ED/LFD+ID/SVM	[13]	94.7
≤40ED/Top-Down Approach/KNN	[12]	92.4
<40ED/Am-Pse-AAC/ AFK-NN	[11]	92.1
<40ED/Functional Domain+PseAAC/ISort	[19]	94.8

≤25ED/GO-PseAAC	[21]	86.5
<40ED/Am-PseAAC/ Augmented CDA	[26]	76.6
<40ED/AAC/CDA	[27])	63.6
25ED/APH/SVM	This study	**95.9**

*ED: Percentage of sequence ID; AM-SVM: Arithmetic mean (AM) offset SVM; LFD+ID: Low-frequency power spectral density and increment of diversity; AFK-NN: Adaptive fuzzy KNN; CDA: Covariant-discriminant algorithm.

IV. CONCLUSIONS

In this paper, we can conclude that enzyme sub-functional classes are an essential protein fold prior to the prediction of protein structure and function. In this study, we proposed APH in order to overcome the weaknesses of subclasses prediction by employing (i) the most significant features, (ii) the optimal classifier and (iii) the distinguishable nonidentified subclasses. We devised APH, which is hybridized from the different features in predicting low homologous sequence similarities. Based on the results of five different types of evaluation carried out; (i) assessment of the most significant feature, (ii) assessment on the optimal number of CV, (iii) prediction on the subclasses using the best classification method, (iv) prediction of enzyme sub-functional classes with biological studies done previously as supporting material, and (v) comparison to other related works. In near future, we plan to investigate larger amount of sequences and further exploring the sub-functional classes for more latent information that might open up to a whole new research direction. We shall also make effort in designing a web-server for the method presented in this paper.

ACKNOWLEDGMENT

The authors would like to thank GATES BIOTECH Solution Sdn. Bhd. (GBIT) for funding this work under Reference No. LTR/GSF/2011-01 in collaboration with Universiti Teknologi Malaysia.

REFERENCES

[1] Pál, C.; Papp, B.; Lercher, M.J. An Integrated View of Protein Evolution. *Nat. Rev. Genet.*, 2006, *7*, 337-346.

[2] Chen, C.; Zhou, X.; Tian, Y.; Zou, X.; Cai. P. Predicting Protein Structural Class with Pseudo-Amino Acid Composition and Support Vector Machine Fusion Network. *Anal. Biochem.*, 2006, *357*, 116-121.

[3] Chen, C.; Chen, L.X.; Zou, X.Y.; Cai, P.X. Predicting Protein Structural Class based on Multi-Features Fusion. *J. Theor. Biol.*, 2008, *253*, 388-392.

[4] Yu, C.S.; Chen, Y.C.; Lu, C.H.; Hwang, J.K. Prediction of Protein Subcellular Localization. *Proteins*, 2006, *64*, 643651.

[5] Punta, M.; Ofran, Y. The Rough Guide to In Silico Function Prediction, or How to Use Sequence and Structure Information to Predict Protein Function. *PLoS Comput. Biol.*, 2008, *4*, e1000160.

[6] Syed, U.; Yona, G. Enzyme Function Prediction with Interpretable Models. *Methods Mol. Biol.*, 2009, *541*, 373420.

[7] Nalivaeva, N.N.; Fisk, L.R.; Belyaev, N.D.; Turner, A.J. Amyloid-degrading Enzymes as Therapeutic Targets in Alzheimer's Disease. *Curr. Alzheimer Res.*, 2008, *5*, 212-224.

[8] Kumar, C.; Li, G.; Choudhary, A. Enzyme Function Classification using Protein Sequence Features and Random Forest. In: *Bioinformatics and Biomedical Engineering*, 3rd International Conference, Beijing, June 11-13, 2009; pp. 1-4.

[9] Zhou, X.B.; Chen, C.; Li, Z.C.; Zou, X.Y. Using Chou's Amphiphilic Pseudo-Amino Acid Composition and Support Vector Machine for Prediction of Enzyme Subfamily Classes. *J. Theor. Biol.*, 2007, *2248*, 546-551.

[10] Wang, Y.C.; Wang, X.B.; Yang, Z.X.; Deng, N.Y. Prediction of Enzyme Subfamily Class via Pseudo Amino Acid Composition by Incorporating the Conjoint Triad Feature. *Protein Peptide Lett.*, 2010, *17*, 1441-1439.

[11] Huang, W.L.; Chen, H.M.; Hwang, S.F.; Ho, S.Y. Accurate Prediction of Enzyme Subfamily Class using an Adaptive Fuzzy *K*-Nearest Neighbor Method. *BioSystems*, 2007, *90*, 405-413.

[12] Shen, H.B.; Chou, K.C. EzyPred: A Top-down Approach for Predicting Enzyme Functional Classes and Subclasses. *Biochem. Biophys. Res. Commun*., 2007, *364*, 53-59.

[13] Shi, R.; Hu, X. Predicting Enzyme Subclasses by using Support Vector Machine with Composite Vectors. *Protein Peptide Lett.*, 2010, *17*, 599-604.

[14] Brown, D.P.; Krishnamurthy, N.; Sjolander, K. Automated Protein Subfamily Identification and Classification. *PLoS Comput. Biol*., 2007, *3*, 1526-1538.

[15] Hu, X.; Wang, T. Prediction of Enzyme Subclass by using Support Vector Machine based on Improved Parameters. In: *Natural Computation (ICNC)*, 7th International Conference, Shanghai, China, July 26-28, 2011; pp. 593-598.

[16] Wang, Y.; Hu, X. Predicting of Oxidoreductase and Lyase Subclasses by using Support Vector Machine. In: *Computer and Information Science (ICIS)*, 10th International Conference, Sanya, China, May 16-18, 2011; pp. 27-31.

[17] Garg, A.; Raghava, G.P.S. A Machine Learning based Method for the Prediction of Secretory Proteins using Amino Acid Composition, their Order and Similarity-Search. *In Silico Biol*., 2008, *8*, 129-140.

[18] Mohammed, A.; Guda, C. Computational Approaches for Automated Classification of Enzyme Sequences. *J. Proteomics & Bioinformatics*, 2011, *4*, 147-152.

[19] Cai, Y.D.; Chou, K.C. Predicting Enzyme Subclass by Functional Domain Composition and Pseudo Amino Acid Composition. *J. Proteome Res.*, 2005, *4*, 967-971.

[20] Wang, Y.C.; Wang, Y.; Yang, Z.X.; Deng, N.Y. Support Vector Machine Prediction of Enzyme Function with Conjoint Triad Feature and Hierarchical Context. *BMC Syst. Biol*., 2011, *5*, S6.

[21] Chou, K.C.; Cai, Y.D. Using GO-PseAA Predictor to Predict Enzyme Sub-class. *Biochem. Biophys. Res. Commun*., 2004, *325*, 506-509.

[22] Zhang, T.L.; Ding, Y.S.; Chou, K.C. Prediction Protein Structural Classes with Pseudo-Amino Acid Composition: Approximate Entropy and Hydrophobicity Pattern. *J. Theor. Biol.*, 2008, *250*, 186-193.

[23] Chou, W.Y.; Pai, T.W.; Jiang, T.Y.; Chou, W.I.; Tang, C.Y.; Chang, M.D.T. Hydrophilic Aromatic Residue and In Silico Structure for Carbohydrate Binding Module. *PLoS ONE*, 2011, *6*, e24814.

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

[25] Chang, C.C.; Lin, C.J. LIBSVM : A Library for Support Vector Machines. ACM Transactions on *Intelligent Systems and Technology*, 2011, *2*, 1-27.

[26] Chou, K.C. Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *BMC Bioinformatics*, 2005, *21*, 10-19.

[27] Chou, K.C.; Elrod, D.W. Prediction of Enzyme Family Classes. *J. Proteome Res.*, 2003, *2*, 183-190.

[28] Sudhamsu, J.; Kabir, M.; Airola, M.V.; Patel, B.A.; Yeh, S.R.; Rousseau, D.L.; Crane, B.R. Co-expression of ferrochelatase allows for complete heme incorporation into recombinant proteins produced in E. coli. *Protein Expr. Purif*., 2010, *73*, 78-82.

[29] Tyagi, N.; Givvimani, S.; Qipshidze, N.; Kundu, S.; Kapoor, S.; Vacek, J.C.; Tyagi, S.C. Hydrogen sulfide mitigates matrix metalloproteinase-9 activity and neurovascular permeability in hyperhomocysteinemic mice. *Neurochem. Int*., 2010, *56*, 301-307.

[30] Palioura, S.; Sherrer, R.L.; Steitz, T.A.; Söll, D.; Simonovic, M. The human SepSecS-tRNASec complex reveals the mechanism of selenocysteine formation. *Science*, 2009, *325*, 321-325.

[31] Billaut-Laden, I.; Allorge, D.; Crunelle-Thibaut, A.; Rat, E.; Cauffiez, C.; Chevalier, D.; Houdret, N.; Lo-Guidice, J.M.; Broly, F. Evidence for a functional genetic polymorphism of the human thiosulfate

sulfurtransferase (Rhodanese), a cyanide and H2S detoxification enzyme. *Toxicology*, 2006, *225*, 1-11.

[32] Hermes, M.; Osswald, H.; Riehle, R.; Piesch, C.; Kloor, D. S-Adenosylhomocysteine hydrolase overexpression in HEK293 cells: effect on intracellular adenosine levels, cell viability, and DNA methylation. *Cell Physiol. Biochem*., 2008, *22*, 223-236.

[33] Joosten, H.J.; Han, Y.; Niu, W.; Vervoort, J.; DunawayMariano, D.; Schaap, P.J. Identification of fungal oxaloacetate hydrolyase within the isocitrate lyase/PEP mutase enzyme superfamily using a sequence marker-based method. *Proteins*, 2008, *70*, 157-166.

[34] Prudnikova, T.; Mozga, T.; Rezacova, P.; Chaloupkova, R.; Sato, Y.; Nagata, Y.; Brynda, J.; Kuty, M.; Damborsky, J.; Smatanova, I.K. Crystallization and preliminary X-ray analysis of a novel haloalkane dehalogenase DbeA from Bradyrhizobium elkani USDA94. *Struct. Biol. Cryst. Commun*., 2009, *65*, 353-356.

[35] Szefczyk, B. Towards understanding phosphonoacetaldehyde hydrolase: an alternative mechanism involving proton transfer that triggers P-C bond cleavage. *Chem. Commun*., 2008, 4162-4164.

[36] Calzada, J.; Zamarro, M.T.; Alcón, A.; Santos, V.E.; Díaz, E.; García, J.L.; Garcia-Ochoa, F. Analysis of dibenzothiophene desulfurization in a recombinant Pseudomonas putida strain. *Appl. Environ. Microbiol*., 2009, *75*, 875-877.

[37] Dodsworth, J.A.; Leigh, J.A. NifI inhibits nitrogenase by competing with Fe protein for binding to the MoFe protein. *Biochem. Biophys. Res. Commun*., 2007, *364*, 378-382.

[38] Liras, P.; Demain, A.L. Chapter 16. Enzymology of betalactam compounds with cephem structure produced by actinomycete. *Methods Enzymol*., 2009, *458*, 401-429.

[39] Fogle, E.J.; van der Donk, W.A. Pre-steady-state studies of phosphite dehydrogenase demonstrate that hydride transfer is fully rate limiting. *Biochemistry*, 2007, *46*, 13101-13108.

[40] Cámara, B.; Nikodem, P.; Bielecki, P.; Bobadilla, R.; Junca, H.; Pieper, D.H. Characterization of a gene cluster involved in 4-chlorocatechol degradation by Pseudomonas reinekei MT1. *J. Bacteriol*., 2009, *191*, 4905-4915.

[41] Wang, P.; Ownby, S.; Zhang, Z.; Yuan, W.; Li, S. Cytotoxicity and inhibition of DNA topoisomerase I of polyhydroxylated triterpenoids and triterpenoid glycosides. *Bioorg Med Chem Lett*., 2010, *20*, 2790-2796.

[42] Lin, M.; Oliver, D.J. The role of acetyl-coenzyme a synthetase in Arabidopsis. *Plant Physiol*., 2008, *147*, 18221829.

[43] Zeczycki, T.N.; St Maurice, M.; Jitrapakdee, S.; Wallace, J.C.; Attwood, P.V.; Cleland, W.W. Insight into the carboxyl transferase domain mechanism of pyruvate carboxylase from Rhizobium etli. *Biochemistry*, 2009, *48*, 4305-4313.

[44] 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. ATP-induced conformational dynamics in the AAA+ motor unit of magnesium chelatase. *Structure*, 2010, *18*, 354-365.

[45] Kim, J.; Mrksich, M. Profiling the selectivity of DNA ligases in an array format with mass spectrometry. *Nucleic Acids Res*., 2010, *38*, e2