ISSN 2278-3091
Volume 8, No.1.3, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse5081.32019.pdf
https://doi.org/10.30534/ijatcse/2019/5081.32019

# Enzyme Sub-Functional Class Prediction Based on Sequence-Structure Knowledge

**Sharon Kaur Guramand[1], Rohayanti Hassan[1,*], Dede Rohidin[2], Razib M. Othman[1], Asraful Syifaa' Ahmad[1], Shahreen Kasim[3]**
[1]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[2] School of Computing, Telkom University, 40257 Bandung, West Java, Indonesia
[3]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400, Batu Pahat, Johor, Malaysia
*Corresponding author E-mail: rohayanti@utm.my

## ABSTRACT

Enzyme sequences can be classified based on their structure similarity and/or common evolutionary origin called the sub-functional classes. Information on enzyme sub-functional class is readily available, easing the protein structure and enzyme function probing. ENZYME data-base and UniProt/Swiss-Prot are two prominent classification schemes used to assign the sub-functional class of enzymes. Both schemes determine the subclasses manually based on known main functional classes of enzyme. However, the quantity of known protein sequences is growing exponentially with respect to the quantity of known enzyme functional class. As pointed in the previous literature, it is estimated that only 3-4% of known protein sequences can be assigned to corresponding known sub-functional classes. The fact that this is a tedious and time consuming manually-determined method has further limited the enzyme sub-functional class assignment. In hybrid methods, the combination of sequence-structure knowledge in enzyme sub-functional class prediction allows for proper identification of true positives and true negatives for each query sequence. Besides, with the growing number of unannotated sequences, association of a new sequence to an enzyme of known structure can be a significant step towards the identification of its biological role in enzymatic function.

**Key words** : Enzyme Sub-Functional Classes; Conjoint Triad Feature; Sequence-Structure Knowledge; Support Vector Machine.

## 1. INTRODUCTION

This Due to the laborious manually-determined schemes, a repertoire of other computational methods has been investigated in order to produce high-throughput sequence-structure subclass assignment. These methods utilized the knowledge of known amino acid sequence contents and arrangements which is available in a larger quantity and is well-known for the use of sequence level information. At the early stage, the amino acid sequence classification method has been used to assign the sub-functional class for corresponding protein sequence. However no or lack sequence order and sequence length knowledge was used which in turn lead to arbitrated statistical inference. The insufficient knowledge of known sub-functional class also hindered the percentage of accuracy for amino acid based classification method. Only 51.9% attained for accuracy in prediction using amino acid composition based classification in ENZYME select dataset [2]. Using a similar method [5], in testing less 40% similarity dataset reported accuracy of 62.3%.

Currently, the enzyme sub-functional class is predicted using more sophisticated method which basically integrates two mechanisms: firstly, the amino acids of protein sequences are represented by features vector and secondly the features vector is then served into classification method to predict the corresponding sub-functional class. However, it is a challenging task to predict the sub-functional class for sequences that is characterized by low-identity to each other. Most related studies are primarily focused on complex features vector. Advanced representations such as merging the amino acid composition with its evolutionary and neighbourhood information, pseudo-amino acids that considered the effects of sequence order [4], [6] and multi composite features [7].. However, the knowledge of known sub-functional class from ENZYME and Uniprot/Swiss-Prot are frequently used as a standard of truth for classification method [8] even though both schemes show inconsistent sub-functional class assignments for some protein sequences. This study conducted a preliminary experiment onto two ENZYME dataset. As depicted in Table 1, $DS_I$ and $DS_{II}$ produced different sub-functional classes' assignment (EC) for EC.3 sequences of ENZYME dataset. $DS_I$ is referring to dataset with sequences have less than 40% sequence identity while $DS_{II}$ is referring to dataset with 25% sequence identity. These differing assignments could lead to wrong classes as well as overestimate error. Besides, it is equally important to test with the latest dataset ($DS_{II}$) available in current database as it provides additional and lengthy information on each subclass in every functional class. To date, there were no researches carried out using $DS_{II}$ in predicting the sub-functional classes of enzyme which corresponds to the ENZYME database.

This paper is organized as follows. In section 2, the materials and methods used are explained. The experimental results of comparative evaluation proposed in this paper is presented in section 3. Finally, our work of this paper is summarized in the last section.

**Table 1:** Inconsistent enzyme sub-functional class assignment between two datasets for 10 sub-functional class protein sequences from EC.3.

| EC.3 | Number of protein sequences | Number of correct prediction ($DS_I$) | Number of correct prediction ($DS_{II}$) |
|---|---|---|---|
| EC.3.1 | 427 | 425 | 399 |
| EC.3.2 | 136 | 134 | 122 |
| EC.3.3 | 108 | 106 | 104 |
| EC.3.4 | 210 | 209 | 208 |
| EC.3.5 | 98 | 98 | 98 |
| EC.3.6 | 73 | 72 | 72 |
| EC.3.7 | 102 | unknown | 102 |
| EC.3.8 | 11 | unknown | 11 |
| EC.3.11 | 15 | unknown | 15 |
| EC.3.13 | 29 | unknown | 28 |
| **Overall** | **1,209** | **1,044** | **1,159** |

## 2. MATERIALS AND METHODS

Based on the aforementioned deficiencies in enzyme sub-functional class prediction method, this study repeats the method and algorithm from Wang [7] designated as SVM-CTF that aims to overcome the insufficient sequence-structure knowledge due to the lack of protein sequence identity also by only using a simpler feature vector. SVM-CTF is an abbreviation of SVM with Conjoint Triad Feature for enzyme sub-functional class prediction. Fig 1 demonstrates the analogy of sub-functional class prediction by the repeated method. Initially, ENZYME and UniProt/Swiss-Prot predicted the Nitrogenase Molybdenum-iron protein (UniProt ID: Q57118) as subclass EC.1.17 throughout the tedious manual experimental routine. However, in using the Chou's pseudo-amino acid composition (Pse-AAC) classification method [5] that bases on the whole protein sequences, the subclass was still unknown. This study has followed the use of physio-chemistry properties of protein sequences which were derived from the protein-protein interaction (PPI) of any sequence. In order to avoid the inconsistent standard of truth, the subclasses of enzymes were determined using the available ENZYME database (ftp://ftp.expasy.org/databases/enzyme/). Subsequently, SVM was implemented to predict the sub-functional class. To evaluate the performances of the repeated method, two measurement metrics were used: accuracy (*acc*) and Matthew's Correlation Coefficient (MCC).

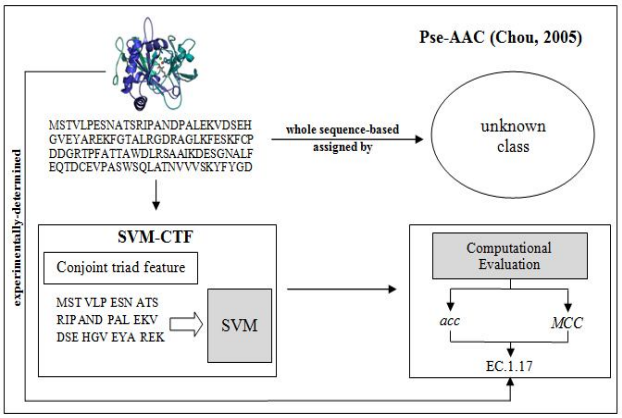


**Figure 1:** The sub-functional assignment/prediction for Nitrogenase Molybdenum-iron protein (UniProt ID: Q57118) using Chou's Pse-AAC classification method5 and repeated method termed as SVM-CTF

### 2.1 Dataset Preparation

The benchmark datasets used to validate the performance of repeated method is collected from ENZYME database. The sequences in these datasets have less than 40% sequence identity ($DS_I$) and 25% sequence identity ($DS_{II}$) to any other in a same functional class. The detail information of this datasets can be found in Shen and Chou (2007). In addition, for avoiding the extreme sub-functional class bias, those subclass which contain less than 50 proteins are excluded in our validation. Finally there are six main functional classes and thirty-four sub-functional classes for $DS_I$ (Release of 01–May–2007) and fifty-eight sub-functional classes for $DS_{II}$ (Release of 21–Sept–2011) in the benchmark datasets.

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

The development of highlight vectors for every datum overwhelms the learning ability of the SVM-based strategies. Since the CTF considers the synthesis of amino acids as well as succession request impact, so this proficient and straightforward encoding plan is under dreary thought here. The CTF is developed dependent on the dipoles and volumes of the side chains, the 20 amino acids can be grouped into seven classes: {A, G, V}, {I, L, F, P}, {Y, M, T, S}, {H, N, Q, W}, {R, K}, {D, E}, {C}. Subsequently, a (7x7x7) = 343-measurement vector is utilized to speak to a given protein grouping, where every component of this vector is the recurrence of the relating conjoint set of three showing up in the protein succession.

Here, we have demonstrated that SVM with the CTF beat other grouping based PPI forecast techniques. The CTF considers the properties of one amino corrosive as well as its vicinal amino acids and treats any three coterminous amino acids as a unit. That is, it contains the arrangement of amino acids as well as the succession request impact. It has likewise effectively been utilized in the forecast of DNA and RNA restricting proteins. Motivated by these, we used the CTF into our prescient model to speak to the info vector.

## 3. RESULTS AND DISCUSSION

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

By using sequence from ENZYME dataset, the effectiveness of the repeated method was tabulated (Table 2 and Table 3). Results indicated that the accuracy improved from 34.5% to 42.7% in both $DS_I$ and $DS_{II}$ compared to the earlier Pse-AAC based classification method5. Table 2 and Table 3 show that the highest accuracy based on the Pse-AAC classification method5 for all main functional class was below 53%. A low accuracy of Pse-AAC classification method [5] was caused by the irregular distribution of sequence length in predicting the sub-functional class. To make matters worse, the segregated nature of amino acid lead to the unsteady proportions of hydrophilic and hydrophobic values which was the most essential prediction criteria in Pse-AAC classification5. In both the datasets, DSII scored the highest acc due to the amount of positive prediction as well as the number of sequences used. The method utilized succeeded in improving the previous method for both the datasets tested by bridging the sequence-structure knowledge using the potential features which later SVM exploited to reveal the hindered sub-functional class. For instance, in EC.1 there is 22 subclasses, DSI scored 17 subclasses whereas DSII predicted 21 all in all.

### 3.2 Analysis of Single Feature Selection towards Prediction

Table 4 shows the results of the extended analysis on the effects of different features vector assignment methods using $DS_{II}$ on enzyme sub-functional class prediction. The results were yielded from SVM-CTF which was evaluated using acc and MCC. CTF in SVM-CTF category proved to be the best performer in all metrics with *acc* of 87.1% and MCC of 88.4% for EC.2. This is followed by SVM-DPC, also using the same dataset, succeeded with *acc* of 75.6% and MCC of 74.6% for EC.2. While SVM-AAC has been proven to excel in every enzyme main class prediction, it also showed a competent *acc* and MCC for sub-functional class prediction [3], [8], [9]. SVM-AAC achieved 70.6% in *acc* and 70.2% in MCC in EC.1. Overall, SVM-CTF has the highest average acc in all the classes.

The ultimate reason on the single feature selection is to determine the most significant features that can be applied to represent the raw dataset to predict its subclasses accurately. AAC is one of the underlying effective strategies utilized broadly because of its candid qualities [10]. Also, AAC is subject to the extent of amino corrosive buildup events evaluated utilizing measurable technique found in the particular protein grouping. DPC is the dipeptide piece yield upon the hydrolysis of two amino acids in communicating the compound succession data proficiently, for example, the augmentation of assorted variety [11].

**Table 2:** An increment of accuracy (%) presented by SVM-CTF compared to Pse-AAC classification method[5] using $DS_I$

| Method | Features | Enzyme main functional class (% of *acc*) | | | | | |
|---|---|---|---|---|---|---|---|
| | | **EC. 1** | **EC. 2** | **EC. 3** | **EC. 4** | **EC. 5** | **EC. 6** |
| **SVM-CTF** | PPI based-Conjoint Triad Feature | 78.0 | 61.2 | 64.4 | **87.3** | 79.0 | 80.8 |
| **Pse-$_5$AAC** | Hydrophilicity, Hydrophobicity | 51.3 | 49.7 | 52.8 | 40.5 | 39.0 | 33.6 |

**Table 3:** An increment of accuracy (%) presented by SVM-CTF compared to Pse-AAC classification method[5] using $DS_{II}$

| Method | Features | Enzyme main functional class (% of *acc*) | | | | | |
|---|---|---|---|---|---|---|---|
| | | EC.1 | EC. 2 | EC. 3 | EC. 4 | EC. 5 | EC. 6 |
| **SVM-CTF** | PPI based-Conjoint Triad Feature | 80.0 | 71.8 | 63.4 | **92.7** | 90.0 | 90.9 |
| **Pse-$_5$AAC** | Hydrophilicity, Hydrophobicity | 49.1 | 50.0 | 48.2 | 34.5 | 49.1 | 44.5 |

### 3.3. Comparison to Other Related Works

As depicted in Table 5, the comparison is split into three: with SVM and Random Forest classifier also with no classifier. Without using any classifier, Chou [5] argued that a large and homologues dataset resulted in a better performance; however the accuracy was still degraded at 57.4%. Chou and Elrod [2] then extended the work by integrating tri-peptides frequency features to discern the sub-functional class. Unfortunately, they were only able to gain not more than 3% compared to their previous works [6]. In similar work, Wang [9] concluded that the knowledge of amino acid alone with no classifier limits the accuracy to just 66.7%.

**Table 4:** Performance of different feature representations using SVM for $DS_{II}$

| Feature | Metric | Enzyme main functional class (% of *acc*) | | | | | |
|---|---|---|---|---|---|---|---|
| | | EC1 | EC2 | EC3 | EC4 | EC5 | EC6 |
| AAC | *acc* | 70.6 | 66.1 | 68.2 | 69.0 | 66.9 | 67.4 |
| | *MCC* | 70.2 | 67.3 | 68.2 | 67.5 | 63.3 | 65.0 |
| DPC | *acc* | 73.2 | 75.6 | 75.1 | 72.8 | 73.9 | 75.5 |
| | *MCC* | 73.0 | 74.6 | 75.1 | 71.3 | 71.4 | 72.7 |
| CTF | *acc* | 86.3 | **87.7** | 85.2 | 86.4 | 88.6 | 85.5 |
| | *MCC* | 86.0 | **88.4** | 84.9 | 86.5 | 82.9 | 83.6 |

**Table 5:** Comparison with other related works

| Classification Method | Reference | Features vector | Overall accuracy, $acc_{all}$ (%) |
|---|---|---|---|
| None | Chou [5] | Pse-AAC | 57.4 |
| | Wang [9] | Tri-peptide | 59.9 |
| | Chou and Elrod [2] | Propensity score | 66.7 |
| Random Forest | Kumar and Choudhary[12] | Top-down approach | 84.3 |
| SVM | Chou and Cai [13] | Pse-AAC | 74.3 |
| | Chen [6] | Trio amino acids | 85.3 |
| | Zhou [4] | Amp Pse-AAC | 80.9 |
| | Shi and Hu [8] | Amp-AAC | 80.1 |
| | **This study** | **CTF** | **92.7** |

Hence, the incorporation between various features and SVM in the repeated method has been proven to be able to reveal the hindered sequence-structure knowledge with the best accuracy of 92.7%. By using simpler features vector, the repeated method yielded 7.4% higher than Chen [6], 11.8% higher than Zhou [4], 12.6% higher than Shi and Hu [8] and 18.4% higher than Chou and Cai [4]. Besides, the Random Forest classifier introduced by Kumar and Choudhary [12] using a top-down three layer model where the top layer classifies a query protein sequence as an enzyme or non-enzyme, the second layer predicts the main function class and bottom layer further predicts the sub-function class scored the overall *acc* above 80%. Hence, a need for a computing method is felt that can distinguish protein enzyme sequences from those of non-enzymes and reliably predict the function of the former.

## 4. CONCLUSION

SVM is a powerful classifier, while the input features vector bases are able to enrich the knowledge between known protein sequences and known sub-functional classes. In this study, the advantages of both elements have been integrated to precisely predict the enzyme sub-functional class. The integration is known as SVM-CTF that has been developed to solve the problems of insufficient known sequence-structure knowledge as well as low prediction accuracy which are posed by the former Pse-AAC classification method. In this chapter, we have analyzed on the effect of different datasets on the predictive result. From the accuracy, we can conclude that the latest and newer dataset ($DS_{II}$) extracted contains additional sequence thus able to fill in the gap of unclassified subclasses. Besides, in the final section of analysis, a depth comparison with related works was done which shows that SVM-CTF succeeded with higher accuracy. There was also a vast difference in prediction result with and without classifier. Based on a higher similarity rate to ENZYME database, the repeated method might facilitate as an automatic sub-functional class prediction method specifically for low-identity sequences. It is anticipated that more influenced features vector can be adopted in the future works.

## ACKNOWLEDGEMENT

This work is supported by MyMaster Scholarship of the Ministry of Education Malaysia, RMC UTM, G-Heart scheme under the Gates Scholars Foundation and GUP grant, with Vot No: 16H73. We also would like to thank Universiti Tun Hussein Onn Malaysia for supporting this research under the Fundamental Research Grants Scheme (FRGS Vot number: 1559), also, thanks to Gates IT Solution Sdn Bhd for the whole support.

## REFERENCES

[1] Cai YD and Chou KC, "**Predicting enzyme subclass by functional domain composition and pseudo amino acid composition**," *Journal of Proteome Research*, vol. 4, no. 3, pp. 967–971, 2005. https://doi.org/10.1021/pr0500399

[2] Chou KC and Elrod DW, "**Prediction of enzyme family classes**," *Journal of Proteome Research*, vol. 2, no. 2, pp. 183–190, 2003. https://doi.org/10.1021/pr0255710

[3] Shen HB and Chou KC, "**EzyPred: A top-down approach for predicting enzyme functional classes and subclasses**," *Biochemical and Biophysical Research Communications*, vol. 364, no. 1, pp. 53–59, Dec. 2007. https://doi.org/10.1016/j.bbrc.2007.09.098

[4] Zhou XB, Chen C, Li ZC, and Zou XY, "**Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes**," *Journal of Theoretical Biology*, vol. 248, no. 3, pp. 546–551, Oct. 2007. https://doi.org/10.1016/j.jtbi.2007.06.001

[5] Chou KC, "**Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes**," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, Jan. 2005. https://doi.org/10.1093/bioinformatics/bth466

[6] Chen C, Zhou X, Tian Y, Zou X, and Cai P, "**Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network**," *Analytical Biochemistry*, vol. 357, no. 1, pp. 116–121, Oct. 2006. https://doi.org/10.1016/j.ab.2006.07.022

[7] Wang YC, Wang XB, Yang ZX, and Deng NY, "**Prediction of Enzyme Subfamily Class via Pseudo Amino Acid Composition by Incorporating the Conjoint Triad Feature**," *Protein & Peptide Letters*, vol. 17, no. 11, pp. 1441–1449, 2010. https://doi.org/10.2174/092986611009011441

[8] Shi R and Hu X, "**Predicting Enzyme Subclasses by Using Support Vector Machine with Composite Vectors**," *Protein & Peptide Letters*, vol. 17, no. 5, pp. 599–604, May 2010. https://doi.org/10.2174/092986610791112710

[9] Wang P, Ownby S, Zhang Z, Yuan W, and Li S, "**Cytotoxicity and inhibition of DNA topoisomerase I of polyhydroxylated triterpenoids and triterpenoid glycosides**," *Bioorganic and Medicinal Chemistry Letters*, vol. 20, no. 9, pp. 2790–2796, May 2010.

https://doi.org/10.1016/j.bmcl.2010.03.063

[10] Mohammed A, "**Computational Approaches for Automated Classification of Enzyme Sequences**," *Journal of Proteomics & Bioinformatics*, vol. 4, no. 8, pp. 147–152, Aug. 2011.

[11] Lu, "**Increment of diversity with quadratic discriminant analysis – an efficient tool for sequence pattern recognition in bioinformatics**," *Open Access Bioinformatics*, vol. 2, p. 89, 2010. https://doi.org/10.2147/OAB.S10782

[12] Kumar C and Choudhary A, "**A top-down approach to classify enzyme functional classes and sub-classes using random forest**," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2012, no. 1, p. 1, Dec. 2012. https://doi.org/10.1186/1687-4153-2012-1

[13] Chou KC and Cai YD, "**Using GO-PseAA predictor to predict enzyme sub-class**," *Biochemical and Biophysical Research Communications*, vol. 325, no. 2, pp. 506–509, Dec. 2004. https://doi.org/10.1016/j.bbrc.2004.10.058