IEC
Institute of Electronics and Computer

Journal of Artificial Intelligence and Systems, 2020, 2, 38-52
https://iecscience.org/journals/AIS
ISSN Online: 2642-2859

Set of Approaches Based on Position Specific Scoring Matrix and Amino Acid Sequence for Primary Category Enzyme Classification

L. Nanni[1,*], S. Brahnam[2]

[1] DEI, University of Padua, viale Gradenigo 6, Padua, Italy.
loris.nanni@unipd.it
[2] Information Technology and Cybersecurity, Missouri State University, 901 S. National, Springfield, MO 65804, USA.
sbrahnam@missouristate.edu
*Corresponding author

How to cite this paper: L. Nanni, S. Brahnam (2020). Set of Approaches Based on Position Specific Scoring Matrix and Amino Acid Sequence for Primary Category Enzyme Classification. Journal of Artificial Intelligence and Systems, 2, 38–52. https://doi.org/10.33969/AIS.2020.21004.

Received: December 29, 2019
Accepted: February 3, 2020
Published: February 5, 2020

Abstract

The last decade has witnessed an unprecedented accumulation of proteins in large online databases which has led to the need for automatic prediction of protein function essential for massive and timely annotations of the proteins in these datasets. Protein databases, combined with functional annotations and machine learning (ML) techniques, offer many potential benefits, including significantly facilitating rapid pharmacological target identification. The main objective of this study is to identify, for the problem of enzyme classification, the most powerful combinations of descriptors taken from different protein representations. To achieve this objective, four approaches for representing the Position-Specific Scoring Matrix (PSSM) combined with three methods for representing the Amino Acid Sequence (AAS) are evaluated with the aim of experimentally producing a powerful ensemble of descriptors for enzyme function prediction. Each protein descriptor is classified by a Support Vector Machine (SVM), with the set of SVMs finally combined by sum rule. Cross-validation experiments using these descriptors on single-functional enzymes (n=44,661) extracted from the PDB database demonstrate that the ensemble proposed here achieves superior classification rates compared to state-of-the-art ML techniques reported in the literature on the same dataset. Although the proposed ensemble strongly outperforms these other techniques, it is computationally much heavier, mainly because the PSSM extraction process is time consuming. However, there is a growing repository of proteins where PSSM has already been extracted, making the proposed method more practical and attractive. The MATLAB code and the dataset used in the experiments reported here are available at https://github.com/LorisNanni.

Keywords

Protein classification, Enzyme classification, Support Vector Machine, Position-Specific Scoring Matrix

1. Introduction

Environmental genomics has led to the discovery of many new protein families and the creation of massive protein databases [1]. Providing these proteins with experimentally verified functional annotations using current methods (e.g. NMR spectroscopy) is laborious, expensive, and inadequate to the task of annotating these growing collections of protein sequences. Fortunately, collections of proteins that have been annotated are sufficiently large enough now to train supervised machine learning (ML) approaches to perform functional annotations on protein sequences that have yet to be annotated [2].

Many machine learning (ML) approaches have been proposed for automatic protein annotation, many of which consider the enzymatic structures available in the Protein Data Bank (PDB) and the enzyme commission (EC) number as a comprehensive framework for annotation. Enzymes are proteins classified into six primary classes based on the chemical reactions the enzymes catalyze: i. oxidoreductases, ii. transferases, iii. hydrolases, iv. lyases, v. isomerases, and vi. ligases. These classes are labeled with the enzyme commission (EC) number (NC-IUBMB, 1992), a system that is based on experimental evidence [3, 4]. A given enzyme code starts with the letters "EC" followed by four numbers representing a hierarchical classification level, each separated by periods. The first level designates one of six high-level classes that represent a generic reaction type. The second and third levels mostly describe the bonds or functional groups in a reaction. The fourth level defines the substrate and specific enzyme-catalyzed chemical reaction.

Important in any machine learning task is representation and the extraction of powerful descriptors. Ruiz-Blanco, et al. [5] have identified four protein descriptor families based on protein representations: sequence-based, linear-topology-based (both 1D), pseudo-fold-topology-based (2D), and 3D–structure-based (3D). Early work in automatic enzyme annotation focused on extracting 1D features from sequence-based representations, mostly representations based on the 1D amino acid sequence (AAS), which is a string that represents the arrangement of the amino acids composing a protein. Features taken from the AAS have been trained by such classic classifiers as the Support Vector Machine (SVM) ([6-13]), k-Nearest Neighbor (kNN) ([14-16]), decision trees, random forest (([17-19]), and neural networks (NN) ([20]). Agüero-Chapin et. al [21] proposed using Topological Indices to BioPolymers (TI2BioP) for calculating topological indices extracted from DNA, RNA, and protein biopolymers. More recently, Ruiz-Blanco, et al. [5] successfully extracted Pseudo-fold-topology-based 2D descriptors where TI2BioP were applied for projecting biopolymeric sequences into bidimensional graphs. For two comprehensive reviews of the early literature on automatic enzyme function classification, see [22] and [23].

Comparatively little work has considered the structural model of a protein, such as the high-resolution 3-D structure of protein sequences. This can be accounted for in the main because the rate at which protein structures are being solved is increasing faster than advances in experimental knowledge [9]. Thus, there is an urgent need to add functional annotations to protein structures. The benefit of using structural attributes in the prediction of enzyme function has been demonstrated in [24] utilizing a Bayesian approach and in [13] where classification looked at the utility of a protein's structural information along with its combination with sequence-based descriptors. Excellent results were achieved in [25] when protein structure and chemical information were combined with sequential information into a graph model from which features were extracted and classified by SVM.

Relevant to this study are the fusion approaches proposed in [2] and [13] that are based on structural information, esp., the systems named SIfusion in [2] and SVMstructural in [13], where structural information trained on SVMs is coupled with kNNs trained on AAS. Finally, Ruiz-Blanco et al. [5] have demonstrated in their assessment of different descriptors that 3D structural features rank first class in information load though all classes were shown to encode significant information.

Although descriptors based on AAS are essential for training classification models, another representation in common use is the Position-Specific Scoring Matrix (PSSM), which is more informative than AAS in that PSSM incorporates evolutionary information. PSSM [26], is a scoring matrix representation for proteins calculated with the application PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool). PSI-BLAST is a sequence similarity search method that specifies the scores for observing particular amino acids or nucleotides at specific positions. In other words, this tool compares PSSM profiles to discover related, though sometimes remote, homologous proteins or DNA. Descriptors based on PSSM have been shown to improve the prediction performance of both the structural and functional properties of proteins across a range of bioinformatics problems [27], including the prediction of protein structural classes [28], protein fold recognition [29], protein-protein interactions [30], protein subcellular localization [31], RNA-binding sites [32] and, relevant here, protein functions [33-36]. In [35], for instance, 1D descriptors taken from PSSM were classified using probabilistic neural networks (PNN), kNN, decision tree, multi-layer perceptron, and SVM. In [34], multi-functional enzymes were identified from the ENZYME database via a combinational model of SVM and random forest.

In addition to training descriptors taken from different protein representations on SVM, NN, random forest, etc., more recent works have explored deep learning techniques, which have revolutionized ML, particularly machine vision, with the introduction of such deep learners as the Convolutional Neural Network (CNN) [37]. The primary advantage of using deep learning techniques is that they automatically derive or learn a set of features for a specific classification problem

during the training process, unlike more traditional approaches that require ML engineers to select and extract an appropriate set of handcrafted features upfront before the training phase begins. CNNs have been applied to many protein classification problems, specifically to protein secondary structure prediction in [38][39] and to protein function prediction in [40-43]. In [38], classification was based on PSSM and in [39] on one-dimensional convolution applied on features based on AAS. For protein function prediction, DeepGO, proposed in [41], was one of the first methods to employ a CNN to predict protein function from the protein AAS and cross-species protein-protein interaction networks. DeepGO was improved in [42] with DeepGOPlus, which used a CNN to scan a sequence for motifs predictive of a given protein's function and then by combining this approach with sequence similarity-based predictions; DeepGo was also enhanced in [43] with DeepText2GO, which integrated probabilistic sequence-based learners utilizing deep semantic representations of texts combined with domains, sequence homology, families, and motifs. In [40], shape features were extracted that represented protein structure as a local (for each amino acid) dispersion of angles and amino acid distances. The multi-channel image and feature maps served as the inputs to an ensemble of CNNs for function prediction (MultiChannel CNN). The outputs were then combined with a kNN or SVM.

The goal of this work is to develop an ensemble of powerful protein descriptors for the problem of enzyme classification. Many well-known protein descriptors are investigated, including those derived from the AAS model and variants of the PSSM matrix representations of a protein. Produced in this investigation is an effective and competitive ensemble of descriptors/features for enzyme classification, a set that obtains state-of-the-art performance on a large (n=44,661) dataset of single-functional enzymes that were extracted from the PDB database.

The main contributions of this study include:

- Using the Discrete Cosine Transform (DCT) to extract a set of features from the PSSM matrix; these DCT coefficients are used to train an SVM classifier.
- Combining PSSM and DCT, which produced the best performance among the stand-alone approaches based on PSSM, outperforming the widely used PSSM descriptors;
- Combining both aminoacidic sequence descriptors with the different PSSM descriptors, which obtained the best performance on the dataset used in this study;
- Successfully producing an ensemble that outperforms those reported in the literature.

It should be noted that there is a major drawback to the approach proposed in this study. The PSSM extraction process is time-consuming, and ensemble methods

require considerable computational power. Although computation time is not a significant factor in the testing phase, it is a consideration in the training phase, esp. when evaluating large sets of descriptors across many large datasets. As far as PSSM extraction is concerned, however, there is a growing repository of proteins where PSSM has already been extracted. Moreover, these limitations are offset to some degree, given the power of PSSM descriptors.

2. Proposed approach

2.1 General machine learning approach for protein classification

It was noted in the introduction that the main goal of this study is to evaluate descriptors and features extracted from two protein representations. The method proposed here extracts different descriptors from both the ACC and PSSM protein representation. The descriptors are then trained on separate SVMs, with the scores of the SVMs summed for a final classification.

The two protein representations are discussed in section 2.2, and the different features/descriptors are described in section 2.3. In some cases, the extraction methods detailed in section 2.3 are repeated several times. This repetition is necessary so that each of the physicochemical properties are considered in the extraction process. The physiochemical properties were obtained from the amino acid index database [44] available at http://www.genome.jp/dbget/aaindex.html). An amino acid index is a set of twenty numerical values, each corresponding to a unique physicochemical property of the amino acid. Rejected are amino acids where all properties have a value of less than two. To date, the amino acid index database has 566 indices and 94 substitution matrices; though this number may seem small, a small set of properties is considered adequate for classifying most protein problems.

The ensembles of SVMs were implemented using the LibSVM toolbox (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). SVMs are trained with the training data using a grid search, thereby ensuring that the test is blind. All features are linearly normalized to [0, 1] based on the training data.

The ensemble approach employed in this work is based on the idea of combining all the information contained in sets of different descriptors. SVMs are fused using the weighted sum rule (which includes the standard sum rule since it is equivalent to the weighted sum rule when all the SVMs are assigned the same weight) [45].

2.2. Protein representations

In this study, the descriptors, or features, are extracted from two protein representations, AAS and PSSM, which are the focus of this section. How the different descriptors are extracted from these representations is detailed in section 2.3.

2.2.1 Amino acid sequence (AAS)

As noted in the introduction, sequential-based models of protein representations are based on AAS, a string that represents the arrangement of the 20 native amino acids composing a protein. AAS represents a protein as a linear sequence, defined as $P = (p_1, p_2, \dots, p_N)$, where $p_i \ \epsilon \mathcal{A} = [A, C, D, \dots, Y]$, and $\mathcal{A}$ represents the 20 native amino acids.

Research has demonstrated that combining AAS with information regarding the physicochemical properties of amino acids produces many robust protein descriptors, as discussed by Kawashima et al. [44], which show that this combination is sufficient for defining protein structure and functions.

2.2.2 Position Specific Scoring Matrix (PSSM)

PSSM [26] is a scoring matrix representation for proteins calculated with the application PSI-BLAST. In this study, PSI-BLAST is called from MATLAB. When calling PSI-BLAST, four parameters must be specified in the command-line flags: position, probe, profile, and consensus.

PSSM is an N×20 matrix:

$$PSSM = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,20} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ S_{N,1} & S_{N,2} & \cdots & S_{N,20} \end{bmatrix}, \quad (1)$$

where N is the protein length, and $S_{i,j}$ is the probability of amino acid j occurring at position i. Each row contains values representing the positions of the sequence, and the columns contain values representing the twenty amino acids. Each element of $PSSM(i,j)$ is calculated as $PSSM(i,j) = \sum_{k=1}^{20} w(i,k) \times Y(j,k)$ $(i = 1, \dots, N; j = 1, \dots, 20)$. The expression $w(i,k)$ is the ratio between the number of probes and the frequency at position i of the k^{th} amino acid. The expression $Y(j,k)$ is the value of the Dayhoff's mutation matrix and reflects the rate of change between the j^{th} and k^{th} amino acids.

PSSM scores are positive and negative integers, with smaller values indicating that an amino acid occurs more frequently in the alignment than expected and larger values signaling that the amino acid is occurring less often than expected. In this way, a given element in a PSSM profile approximates the occurrence probability of an amino acid at a given location.

2.3 Protein feature extraction approaches

The seven methods for extracting the features from the two protein representations (AAS and PSSM) are described in this section. Features extracted from AAS include Amino Acid Composition (AS), Quasi Residue Couple (QRC), and Autocovariance approach (AC). Descriptors extracted from PSSM include Pseudo PSSM (PP), Average Blocks (AB), Autocovariance matrix (AM), and Discrete Cosine Transform (DCT).

2.3.1 AS

AS is defined as $AS(i) = h(i)/N \quad i \in [1, \dots, 20]$, where $h(i)$ simply tabulates in a protein sequence of length N the number of instances of a given amino acid.

2.3.2 QRC

QRC [46], motivated by Chou's quasi-sequence-order model and Yuan's Markov chain model [47], extracts features from a protein sequence. Given the physicochemical property d, QRC is defined as

$$QRC_m^d(k) = \frac{1}{N-m}\sum_{n=1}^{N-m} H_{i,j}(n, n+m, d) + H_{j,i}(n+m, n, d), \quad (2)$$

where i and $j \in [1, \dots, 20]$ are the twenty different amino acids; $k = j + 20(i-1)$, N is the protein length, and $index(i,d)$ returns the value of the property d for the amino acid i. $H_{i,j}(a,b,d) = index(i,d)$, if $p_a = i$ and $p_b = j$; otherwise, $H_{i,j}(a,b,d) = 0$.

Here, QRC^d descriptors are extracted for m, range of 1-3, and concatenated. This produces a feature vector of length 1200. An ensemble of *QRC* descriptors is generated by randomly selecting twenty-five ties for d.

2.3.3 AC

As suggested by the name, AC [48] is based on autocovariance and is a sequence-based variant of Chou's pseudo amino acid composition (PseAAC) [49]. AC results in a set of PseAAC-based features that are produced by concatenating the twenty standard AAC values with the addition of max values that consider sequence order. The parameter max (set to 20 in this work) indicates the maximum distance between the two amino acids i and j.

Given a protein $P = (p_1, p_2, \dots, p_N)$ for a fixed physicochemical property d, the AC descriptor $(AC^d \in \Re^{20+max})$ is defined as

$$AC^d(i) = \begin{cases} h(i)/N & i \in [1, \dots, 20] \\ \sum_{k=1}^{N-i+20} \frac{(index(p_k,d)-\mu_d)\cdot(index(p_{k+i-20},d)-\mu_d)}{\sigma_d \cdot (N-i+20)} & i \in [21, \dots, 20+max] \end{cases} \quad (3)$$

where $index(i,d)$ is the value of property d for amino acid i, and $h(i)$ calculates the number of times a given amino acid occurs in a protein sequence. Both μ_d and σ_d are the mean and the variance, respectively, of d, normalized on the twenty amino acids:

$$\mu_d = \frac{1}{20}\sum_{i=1}^{20} index(i,d), \quad \sigma_d = \frac{1}{20}\sum_{i=1}^{20}(index(i,d) - \mu_d)^2 \quad (4)$$

As with QRC, the physicochemical properties are randomized twenty-five times to generate the ensemble of *AC* descriptors.

2.3.4 PP

PP is a popular PSSM descriptor [50, 51]. Like AC, it preserves information about the amino-acid sequence by taking into account the pseudo amino acid composition.

Given an input matrix $Mat \in \Re^{N\times 20}$, the PP descriptor is a vector $PP \in \Re^{320}$ defined as

$$PP(k) = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} E(i,j) & k = 1,\dots,20 \\ \frac{1}{N-lag}\sum_{i=1}^{N-lag}[E(i,j) - E(i+lag,j)]^2 & \begin{matrix} j = 1,\dots,20;\ lag = 1,\dots,15, \\ k = 20 + j + 20\cdot(\text{lag} - 1) \end{matrix} \end{cases} \quad (5)$$

where *k* is a linear index that scans the cells of *Mat*, *lag* is the distance between one residue and its neighbors, *N* is the length of the protein sequence, and $E \in \Re^{N\times 20}$ is the normalized version of *Mat*:

$$E(i,j) = \frac{Mat(i,j) - \frac{1}{20}\sum_{v=1}^{20} Mat(i,v)}{\sqrt{\frac{1}{20}\sum_{u=1}^{20}\left(Mat(i,u) - \frac{1}{20}\sum_{v=1}^{20} Mat(i,v)\right)^2}} \qquad i = 1,\dots,N; j = 1,\dots,20. \quad (6)$$

2.3.5 AB

AB [51] is a PSSM descriptor derived from the local average of the input matrix $Mat \in \Re^{N\times 20}$. AB is a fixed-length vector $AB \in \Re^{400}$ obtained as

$$AB(k) = \frac{20}{N}\sum_{z=1}^{N/20} Mat\left(z + (i-1) * \frac{N}{20}, j\right) \qquad i=1, \dots, 20; j=1, \dots, 20;$$

$$k=j+20\times(i-1), \quad (7)$$

where *k* is a linear index that scans the cells of *Mat.* As a result, AB is the average of *Mat* blocks.

2.3.6 AM

AM [52] is a PSSM descriptor that utilizes an autocovariance matrix (thus, the label AM) to preserve local sequence-order information. AM represents the average correlation between positions. To make each column of the matrix fixed length, we apply autocovariance variables to each column of the input matrix.

AM is calculated from the input matrix $Mat \in \Re^{N\times 20}$ as

$$AM(k) = \frac{1}{N-lag}\sum_{i=1}^{N-lag}\left(Mat(i,j) - \frac{1}{N}\sum_{i=1}^{N} Mat(i,j)\right) \times \left(Mat(i+lag,j) - \frac{1}{N}\sum_{i=1}^{N} Mat(i,j)\right) \quad (8)$$

with *j*=1, ..., 20; *lag*=1,…,15; *k*=*j*+20×(*lag*-1), and where *N* is the length of the protein sequence, *k* is an index that scans the cells of *Mat*, and *lag* is the distance between a given residue and its neighbors.

2.3.7 DCT

DCT [53], a separable linear transformation for converting a signal into its elementary frequency components, is a PSSM descriptor. The value of DCT is its ability to collapse information into a small set of coefficients.

With $Mat \in \Re^{NxN}$ as the input matrix, DCT is defined as

$$DCT(i,j) = a_i a_j \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} Mat(m,n) cos\frac{\pi(2m+1)i}{2M} cos\frac{\pi(2n+1)i}{2N},$$

$$0 \le i \le M,\ 0 \le j \le N, \quad (9)$$

where

$$a_i = \begin{cases} \frac{1}{\sqrt{M}} & i = 0 \\ \sqrt{\frac{2}{M}} & 1 \le i \le M-1 \end{cases} \quad \text{and} \quad a_j = \begin{cases} \frac{1}{\sqrt{N}} & j = 0 \\ \sqrt{\frac{2}{N}} & 1 \le j \le N-1 \end{cases}.$$

Here the *DCT* descriptor retains the first 400 coefficients.

3. Results and Discussion

3.1 Dataset

The ensembles of descriptors described in section 2 are tested on a benchmark dataset [40] that contains 44,661 protein structures divided into the six top-level EC classes (EC Class [Sample Size]): EC1 [8,075]; EC2 [12,739]; EC3 [17,024]; EC4 [3,114]; EC5 [1,905]; and EC6 [1,804]. As can be seen, these classes are unbalanced. The protein structures were collected from PDB. All enzymes in PDB that occurred in more than one class were excluded.

A five-fold cross-validation testing protocol was used, four folds for training and one for testing. After the classifiers were trained on the training set, class probabilities were calculated for the testing samples.

3.2 Experimental results

Experimental results comparing each of the descriptors described in section 2.3 are presented in Table 1, alongside some of their fusions and state-of-the-art results from the literature. Before each fusion, the scores on the training data are normalized to mean 0 and standard deviation 1.

The fusions considered in Table 1 are the following:

- AllPSSM: fusion by sum rule of PSSM-AB, PSSM-AM, PSSM-PP, and PSSM- DCT;
- AllPSSM \AB: fusion by sum rule of PSSM-AM, PSSM-PP, and PSSM-DCT;
- (AllPSSM \AB) + QRC: fusion by sum rule of QRC, PSSM-AM, PSSM-PP, and PSSM- DCT.

Given the normalized scores, the sum rule sums the matching scores obtained by the different classifiers that are combined. In addition, we have reported the performance obtained using PSSM-DCT as the input to two other classifiers: Multilayer Feedforward Neural Networks (MFN) [54] and Random Decision Forests [55], labelled RF here. It will be observed that both MLP and RF obtain a performance that is lower than SVM.

Performance is measured as accuracy, i.e., as the ratio of correctly predicted observations to the total number of observations.

Table 1. Performance (accuracy) obtained by the methods proposed here and compared with the literature.

AC	SVM	72.49
QRC	SVM	93.61
PSSM -AB	SVM	92.98
PSSM- AM	SVM	93.23
PSSM- PP	SVM	94.44
PSSM- DCT	SVM	94.57
AllPSSM	SVM	96.23
AllPSSM \ AB	SVM	96.26
(AllPSSM \AB) + QRC	SVM	**96.63**
PSSM- DCT	MFN	93.12
PSSM- DCT	RF	92.95
MultiChannel CNN [40]		90.1
SIfusion [2]		83
SVMstructural [13]		73.5

Examining Table 1, we observe the following:

- PSSM-DCT obtains the best performances, outperforming the widely used PSSM descriptors;
- The fusion of each of the matrix representation descriptors provides additional information;
- The best performance is obtained by combining both aminoacidic sequence descriptors and the different PSSM descriptors;
- The proposed ensemble outperforms those reported in the literature.

4. Conclusion

In this paper, we evaluate the performance of some widely used protein representations and feature extraction methods. The goal is to identify the descriptors and combinations that prove most beneficial for enzyme classification. We show that an ensemble of representations based on PSSM significantly boosts performance. Comparing the best ensemble proposed here with the literature on the same benchmark dataset demonstrates the superiority of our approach.

Experiments that compare the performance of descriptors and their fusion, including descriptors that can be extracted from deep learners, as well as additional classification methods need to be conducted. In the future, we plan on examining ensembles made with AdaBoost and Rotation Forest. Ensemble methods require considerably more computational power than SVM, however. Although computation time is not a significant factor in the testing phase, it is a consideration in the training phase, esp. when evaluating large sets of descriptors across many large datasets.

In the future, we also plan on testing other feature transforms beyond DCT; moreover, tests will be performed using PSSM as an input to CNNs. However, this increases the issue addressed at the end of the Introduction regarding the need for increased computation power and time during the training phase. CNN requires far more computational power than other classifiers. Given the power of current GPUs, however, this limitation is becoming less of a concern.

Finally, as an alternative to Chou's recommended procedure of developing a web server, nearly all the MATLAB code used in this study, including the code for our best performing ensemble, is freely available. In our opinion, providing source code is of service to researchers in the field and enables any group to implement and set up any number of servers using the ensemble presented in this paper.

Conflicts of Interest

We have no conflict of interest to declare

References

[1] A. Godzik, "Metagenomics and the protein universe," Current Opinion in Structural Biology, vol. 21, no. 3, pp. 398-403, 2011/06/01/ 2011, doi: https://doi.org/10.1016/j.sbi.2011.03.010.

[2] S. Amidi, A. Amidi, D. Vlachakis, N. Paragios, and E. I. Zacharaki, "Automatic single- and multi-label enzymatic function prediction by machine learning," PeerJ, vol. 5, p. e3095, 2017/03/29 2017, doi: 10.7717/peerj.3095.

[3] E. C. Webb, "Enzyme nomenclature 1992," in Recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes. San Diego: Academic Press, 1992.

[4] S. Boyce and K. F. Tipton, A. F. Agrò, Ed. Nature encyclopedia of life sciences. London: Nature Publishing Group, 2001.

[5] Y. B. Ruiz-Blanco, G. Agüero-Chapin, E. García-Hernández, O. Álvarez, A. Antunes, and J. Green, "Exploring general-purpose protein features for distinguishing enzymes and non-enzymes within the twilight zone," (in eng), BMC bioinformatics, vol. 18, no. 1, pp. 349-349, 2017, doi: 10.1186/s12859-017-1758-x.

[6] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, "SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence," Nucleic Acids Research, vol. 31, no. 13, pp. 3692-3697, 2003, doi: 10.1093/nar/gkg600.

[7] L. Y. Han, C. Z. Cai, Z. L. Ji, Z. W. Cao, J. Cui, and Y. Z. Chen, "Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach," Nucleic Acids Research, vol. 32, no. 21, pp. 6437-6444, 2004, doi: 10.1093/nar/gkh984.

[8] C. Chen, Y.-X. Tian, X.-Y. Zou, P.-X. Cai, and J.-Y. Mo, "Using pseudo-amino acid composition and support vector machine to predict protein structural class," Journal of Theoretical Biology, vol. 243, no. 3, pp. 444-448, 2006/12/07/ 2006, doi: https://doi.org/10.1016/j.jtbi.2006.06.025.

[9] P. D. Dobson and A. J. Doig, "Predicting Enzyme Class From Protein Structure Without Alignments," Journal of Molecular Biology, vol. 345, no. 1, pp.

187-199, 2005/01/07/ 2005, doi: https://doi.org/10.1016/j.jmb.2004.10.024.

[10] X. B. Zhou, C. Chen, Z. C. Li, and X. Y. Zou, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes," Journal of Theoretical Biology, vol. 248, pp. 546-551, 2007, doi: DOI 10.1016/j.jtbi.2007.06.001.

[11] L. Lu, Z. Qian, Y.-D. Cai, and Y. Li, "ECS: An automatic enzyme classifier based on functional domain composition," Computational Biology and Chemistry, vol. 31, no. 3, pp. 226-232, 2007/06/01/ 2007, doi: https://doi.org/10.1016/j.compbiolchem.2007.03.008.

[12] Q. Jian-Ding, H. Jian-Hua, S. Shao-Ping, and L. Ru-Ping, "Using the Concept of Chous Pseudo Amino Acid Composition to Predict Enzyme Family Classes: An Approach with Support Vector Machine Based on Discrete Wavelet Transform," Protein & Peptide Letters, vol. 17, no. 6, pp. 715-722, 2010, doi: http://dx.doi.org/10.2174/092986610791190372.

[13] A. Amidi, S. Amidi, D. Vlachakis, N. Paragios, and E. I. Zacharaki, "A machine learning methodology for enzyme functional classification combining structural and protein sequence descriptors," in Bioinformatics and Biomedical Engineering. Cham: Springer, 2016, pp. 728-738.

[14] W.-L. Huang, H.-M. Chen, S.-F. Hwang, and S.-Y. Ho, "Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method," Biosystems, vol. 90, no. 2, pp. 405-413, 2007/09/01/ 2007, doi: https://doi.org/10.1016/j.biosystems.2006.10.004.

[15] H.-B. Shen and K.-C. Chou, "EzyPred: A top–down approach for predicting enzyme functional classes and subclasses," Biochemical and Biophysical Research Communications, vol. 364, no. 1, pp. 53-59, 2007/12/07/ 2007, doi: https://doi.org/10.1016/j.bbrc.2007.09.098.

[16] E. Nasibov and C. Kandemir-Cavas, "Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction," Computational Biology and Chemistry, vol. 33, no. 6, pp. 461-464, 2009/12/01/ 2009, doi: https://doi.org/10.1016/j.compbiolchem.2009.09.002.

[17] B. J. Lee, M. S. Shin, Y. J. Oh, H. S. Oh, and K. H. Ryu, "Identification of protein functions using a machine-learning approach based on sequence-derived properties," (in eng), Proteome science, vol. 7, pp. 27-27, 2009, doi: 10.1186/1477-5956-7-27.

[18] C. Kumar and A. Choudhary, "A top-down approach to classify enzyme functional classes and sub-classes using random forest," (in eng), EURASIP journal on bioinformatics & systems biology, vol. 2012, no. 1, p. 1, Feb 29 2012, doi: 10.1186/1687-4153-2012-1.

[19] C. Nagao, N. Nagano, and K. Mizuguchi, "Prediction of detailed enzyme functions and identification of specificity determining residues by random forests," PLoS ONE, vol. 9, no. 1, p. e84623, 2014, doi: https://doi.org/10.1371/journal.pone.0084623.

[20] V. Volpato, A. Adelfio, and G. Pollastri, "Accurate prediction of protein enzymatic class by N-to-1 Neural Networks," BMC Bioinformatics, journal article vol. 14, no. 1, p. S11, January 14 2013, doi: 10.1186/1471-2105-14-s1-s11.

[21] G. Agüero-Chapin, G. Pérez-Machado, R. Molina-Ruiz, Y. Morales-Helguera, V. Vasconcelos, and A. Antunes, "Ti2biop: topological indices to biopolymers. Its practical use to unravel cryptic bacteriocin-like domains," Amino Acids, vol. 40, no. 2, pp. 431-442, 2011, doi: https://doi.org/10.1007/s00726-010-0653-9.

[22] S. K. Yadav and A. K. Tiwari, "Classification of enzymes using machine learning based approaches: a review," Machine Learning and Applications, vol. 2, no. 3/4, pp. 30-49, 2015.
[23] M. Sharma and P. Garg, "Computational Approaches for Enzyme Functional Class Prediction: A Review," Current Proteomics, vol. 11, no. 1, pp. 17-22, 2014, doi: http://dx.doi.org/10.2174/15701646116661404152250 13.
[24] L. C. Borro et al., "Predicting enzyme class from protein structure using Bayesian classification," Genetics and Molecular Research, vol. 5, no. 1, pp. 193-202, 2006.
[25] K. M. Borgwardt, C. S. Ong, S. Schönauer, S. V. N. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," Bioinformatics, vol. 21, no. suppl_1, pp. i47-i56, 2005, doi: 10.1093/bioinformatics/bti1007.
[26] M. Gribskov, A. D. McLachlan, and D. Eisenberg, "Profile analysis: Detection of distantly related proteins," presented at the Proceedings of the National Academy of Sciences (PNAS), 1987.
[27] J. Wang et al., "POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles," Bioinformatics, vol. 33, pp. 2756–2758, 2017.
[28] T. Liu, X. Zheng, and J. Wang, "Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile," Biochimie, vol. 92, no. 10, pp. 1330-1334, 2010/10/01/ 2010, doi: https://doi.org/10.1016/j.biochi.2010.06.013.
[29] A. Lobley, M. I. Sadowski, and D. T. Jones, "pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination," Bioinformatics, vol. 25, no. 14, pp. 1761-1767, 2009, doi: 10.1093/bioinformatics/btp302.
[30] J. Zahiri, O. Yaghoubi, M. Mohammad-Noori, R. Ebrahimpour, and A. Masoudi-Nejad, "PPIevo: Protein–protein interaction prediction from PSSM based evolutionary information," Genomics, vol. 102, no. 4, pp. 237-242, 2013/10/01/ 2013, doi: https://doi.org/10.1016/j.ygeno.2013.05.006.
[31] D. Xie, A. Li, M. Wang, Z. Fan, and H. Feng, "LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST," Nucleic Acids Research, vol. 33, no. suppl_2, pp. W105-W110, 2005, doi: 10.1093/nar/gki359.
[32] C.-W. Cheng, E. C.-Y. Su, J.-K. Hwang, T.-Y. Sung, and W.-L. Hsu, "Predicting RNA-binding sites of proteins using support vector machines and evolutionary information," (in eng), BMC bioinformatics, vol. 9 Suppl 12, no. Suppl 12, pp. S6-S6, 2008, doi: 10.1186/1471-2105-9-S12-S6.
[33] P. Radivojac et al., "A large-scale evaluation of computational protein function prediction," (in eng), Nature methods, vol. 10, no. 3, pp. 221-227, 2013, doi: 10.1038/nmeth.2340.
[34] X.-Y. Cheng et al., "A global characterization and identification of multifunctional enzymes," (in eng), PloS one, vol. 7, no. 6, pp. e38979-e38979, 2012, doi: 10.1371/journal.pone.0038979.
[35] Z. U. Khan, M. Hayat, and M. A. Khan, "Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model," Journal of Theoretical Biology, vol. 365, pp. 197-203, 2015/01/21/ 2015, doi: https://doi.org/10.1016/j.jtbi.2014.10.014.
[36] C. Fernandez-Lozano et al., "Improving enzyme regulatory protein classification by means of SVM-RFE feature selection," Molecular BioSystems, 10.1039/C3MB70489K vol. 10, no. 5, pp. 1063-1071, 2014, doi: 10.1039/C3MB70489K.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances In Neural Information Processing Systems, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger Eds. Red Hook, NY: Curran Associates, Inc., 2012, pp. 1097-1105.
[38] M. Spencer, J. Eickholt, and J. Cheng, "A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12, no. 1, pp. 103-112, 2015, doi: 10.1109/TCBB.2014.2343960.
[39] Y. Li and T. Shibuya, "Malphite: A convolutional neural network and ensemble learning based protein secondary structure predictor," in 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 9-12 Nov. 2015 2015, pp. 1260-1266, doi: 10.1109/BIBM.2015.7359861.
[40] E. I. Zacharaki, "Prediction of protein function using a deep convolutional neural network ensemble," PeerJ Computer Science, vol. 3, p. e123, 2017. [Online]. Available: https://peerj.com/articles/cs-124/.
[41] M. Kulmanov, M. A. Khan, and R. Hoehndorf, "DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier," Bioinformatics, vol. 34, no. 4, pp. 660-668, 2017, doi: 10.1093/bioinformatics/btx624.
[42] M. Kulmanov and R. Hoehndorf, "DeepGOPlus: improved protein function prediction from sequence," Bioinformatics, 2019, doi: 10.1093/bioinformatics/btz595.
[43] R. You, X. Huang, and S. Zhu, "DeepText2GO: Improving large-scale protein function prediction with deep semantic text representation," Methods, vol. 145, pp. 82-90, 2018/08/01/ 2018, doi: https://doi.org/10.1016/j.ymeth.2018.05.026.
[44] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," Nucleic Acids Research, vol. 27, no. 1, pp. 368-369, 374 1999. [Online]. Available: https://pdfs.semanticscholar.org/0e92/23abb1f973eff54d20486f0dab90c7dde9e0.pdf.
[45] G. Fumera and F. Roli, "Performance analysis and comparison of linear combiners for classifier fusion," presented at the Structural, Syntactic and Statistical Pattern Recognition and IAPR International Workshops, Ontario, Canada, 2002.
[46] L. Nanni, S. Brahnam, and A. Lumini, "High performance set of PseAAC descriptors extracted from the amino acid sequence for protein classification," Journal of Theoretical Biology, vol. 266, no. 1, pp. 1-10, 2010.
[47] J. Guo, Y. Lin, and Z. Sun, "A novel method for protein subcellular localization: Combining residue-couple model and SVM," presented at the Proceedings of 3rd Asia-Pacific Bioinformatics Conference, Singapore, 2005.
[48] Y. H. Zeng, Y. Z. Guo, R. Q. Xiao, L. Yang, L. Z. Yu, and M. L. Li, "Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach," Journal of Theoretical Biology, vol. 259, no. 2, pp. 366-72, 2009, doi: doi:10.1016/j.jtbi.2009.03.028.
[49] K.-C. Chou, "Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology," Current Proteomics, vol. 6, pp. 262-274, 2009.
[50] G.-L. Fan and Q.-Z. Li, "Predicting protein submitochondrion locations by combining different descriptors into the general form of Chou's pseudo amino acid composition," Amino Acids, vol. 20, no. Nov, pp. 1-11, 2011.

[51] J. C. Jeong, X. Lin, and X.-W. Chen, "On position-specific scoring matrix for protein function prediction," IEEE/ACM transactions on computational biology and bioinformatics, vol. 8, no. 2, pp. 308-315, 2011.
[52] L. Yang et al., "Using auto covariance method for functional discrimination of membrane proteins based on evolution information," Amino Acids, vol. 38, pp. 1497-1503, 2010.
[53] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," IEEE Trans Comput, vol. C-23, no. 1, pp. 90-93, 1974.
[54] P. Auer, H. Burgsteiner, and W. Maass, "A learning rule for very simple universal approximators consisting of a single layer of perceptrons," Neural Networks, vol. 21, no. 5, pp. 786-795, 2008/06/01/ 2008, doi: https://doi.org/10.1016/j.neunet.2007.12.036.
[55] T. K. Ho, "Random decision forests," presented at the ICDAR95 Third International Conference on Document Analysis and Recognition, Montreal, QC, 1995.