Published in final edited form as:
J Proteomics Bioinform. 2011 August 23; 4: 147–152.

Computational Approaches for Automated Classification of Enzyme Sequences

Akram Mohammed[1] and **Chittibabu Guda**[1,2,*]

[1]Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, NE, USA

[2]Center for Bioinformatics and Systems Biology, University of Nebraska Medical Center, NE, USA

Abstract

Determining the functional role(s) of enzymes is very important to build the metabolic blueprint of an organism and to identify the potential roles enzymes may play in metabolic and disease pathways. With exponential growth in gene and protein sequence data, it is not feasible to experimentally characterize the function(s) of all enzymes. Alternatively, computational methods can be used to annotate the enormous amount of unannotated enzyme sequences. For function prediction and classification of enzymes, features based on amino acid composition, sequence and structural properties, domain composition and specific peptide information have been widely used by different computational approaches. Each feature space has its own merits and limitations on the overall prediction accuracy. Prediction accuracy improves when machine-learning methods are used to classify enzymes. Given the incomplete and unbalanced nature of annotations in biological databases, ensemble methods or methods that bank on a combination of orthogonal feature are more desirable for achieving higher accuracy and coverage in enzyme classification. In this review article, we systematically describe all the features and methods used thus far for enzyme class prediction. To the authors' knowledge, this review represents the most exhaustive description of methods used for computational prediction of enzyme classes.

Keywords

Enzyme classification; Amino acid composition; Sequence similarity; Structural information; Domain composition; Machine learning; Support vector machine; Nearest neighbor predictor; Ensemble method

Introduction

Identification and classification of enzymes is extremely beneficial in understanding their cellular functions and consequently in the design and development of drugs from a therapeutic perspective. Enzymes are very specific in their action and usually catalyze only one specific reaction [1,2]. Enzymes represent a significant fraction of a proteome [3] and catalyze a variety of reactions in the cellular systems. Hence functional identification of the entire enzyme complement of an organism provides a metabolic blue print for that species.

*Corresponding author: Chittibabu Guda, Ph.D, Director of Center for Bioinformatics and Systems Biology, Associate Professor, Dept of Genetics, Cell Biology & Anatomy College of Medicine, University of Nebraska Medical Center, Omaha, NE 68198-5145, Tel: (402) 559-5954; Fax: (402) 559-5942; babu.guda@unmc.edu.

Since the genomic data is increasing at an exponential pace, it is extremely tedious and expensive to experimentally determine the function(s) of all proteins. A task of such magnitude can be partly addressed by developing computational methods to determine whether a given new protein sequence is an enzyme or a non-enzyme; and if it is an enzyme, to which enzyme family, class and sub-class does it belong [4]? Such information will guide us to design experiments to further test their catalytic activities.

Each enzymatic activity has a recommended name, and the Enzyme Commission (EC) [5] organizes all enzymes into six major classes. These include (1) oxidoreductases - catalyzing oxidoreduction reactions; (2) transferases – catalyzing the transfer of a chemical group from a donor to an acceptor; (3) hydrolases - catalyzing the hydrolysis of various bonds; (4) lyases - enzymes cleaving bonds by means other than by hydrolysis; (5) isomerases - catalyzing geometrical or structural changes within one molecule; (6) ligases - catalyzing the joining of two molecules coupled with hydrolysis of a pyrophosphate bond in ATP or a similar triphosphate. The EC's hierarchical classification assigns unique four-field numbers (such as EC 1.2.1.1) to different enzymatic activities where, the first three digits of an EC number describe the overall type of enzymatic reaction and the last digit represents the substrate specificity of a reaction [6,7]. Given a dataset of labeled protein sequences belonging to different enzyme classes, class-specific features can be extracted to build models that can predict the enzyme class of an unknown protein sequence. This concept has been widely exploited by machine learning algorithms to develop automated methods for enzyme classification and function prediction. Machine learning methods also offer flexibility in handling very high dimensionality in their classifiers. These methods primarily vary by type, size of labeled data, feature space used and the computational approach employed to build models.

In this review, we provide a comprehensive discussion of various computational methods developed to date, and we will discuss these methods separately based on the feature space used and the computational approach employed for enzyme classification. Different feature-spaces used include amino acid composition, sequence-similarity, structural similarity, domain composition and specific peptides. Different computational approaches are based on machine learning algorithms such as nearest-neighbor method, association rule mining, self-organizing maps, Bayesian networks, support vector machines, decision trees and ensemble methods. To authors' knowledge, this review represents the most exhaustive description of methods used for computational prediction of enzyme classes.

Methods by Feature Space

Amino acid composition

Some of the earlier works [8,9] for the prediction of enzyme classes use amino acid composition (AAC) information because sequence information is readily available, and representation of the primary structure of a protein requires considerably less computational resources than those required for a three-dimensional (3-D) structure of a protein. Only oxidoreductases classes were used as the training and testing datasets in the classification of enzymes and comparing accuracies [8,9]. Based on amino acid composition, Chou et al. [8] has developed an enzyme classification method using covariant discriminant algorithm (CDA) and achieved an overall accuracy of about 64%. However, a protein sequence described only by AAC would lose its sequence-order effect and limits the prediction accuracy. Chou, [9] extended his original covariant discriminant algorithm by introducing new features derived from the protein sequence, which he calls amphiphilic pseudo-amino acid composition (AmPseAAC) that preserves the sequence order. The amino acid sequence is converted into different sequences of discrete values, consisting of hydrophobic and hydrophilic distribution patterns at each position of the sequence. This representation

considerably increased the feature-space derived from a protein sequence and results in improved accuracy (70.6%) to distinguish between enzyme and non-enzyme classes. Later the same group [10] reported an accuracy of 73% for the prediction of oxidoreductases subclasses using AmPseAAC.

Overall AAC-based methods are easy to implement and run faster; however, the accuracy of such methods is limited. Also, AmPseAAC has introduced some undetermined parameters to consider the physicochemical properties of amino acids [10]. Methods based on other criteria appear to fare well as described below.

Sequence and structural similarity

Homology-based tools such as BLAST [11], PSI-BLAST [12] and HMMER [13] are used to detect sequence homology between pairs of proteins or against protein family databases and infers functional similarity from homology. Some studies [2,14] suggest that homology-based tools are sufficient to determine the most probable EC number for the query sequence, but less coverage is achieved with these methods. However, simple pair-wise comparisons may be misleading due to the availability of redundant protein sequences in public databases [15]. Therefore, in addition to the sequence similarity, Tian et al. [6] have used the functional similarity from homology to predict enzyme classes; yet, this method only works well when two sequences are very similar. Similarly, by combining sequence similarity with other functional features such as interacting partners, Espadaler et al. [7] have shown that the protein sequences with sequence similarity are more likely to exhibit the same enzymatic activity if they share the same interacting partners. Otto et al. [16] and Galperin et al. [17] have developed methods for identification of analogous enzymes using sequence similarity by grouping proteins that share the same enzymatic activity (EC classes).

Sequence similarity methods rely heavily up on identifying similar proteins and transferring their annotations to a query sequence, therefore, fail when a similar protein is either not identified or lacked annotations in the target database. On the other hand, methods based on structural similarity [18–29] are relatively more tolerant to low-sequence similarity because structural properties are more conserved in evolution. In one of such methods, Dobson et al. [18] defined a protein by its residue fractions, surface properties, secondary structure information and ligands. The method intentionally incorporates structural similarity outside the functional class in order to maximize dataset size and better represent the full range of structures in each functional class. Similarly, by using dataset from Dobson et al. [18], Munteanu et al. [19] developed a method to identify enzymes and non-enzymes. They also showed that increasing the complexity of the data or method does not always improve the accuracy of enzyme/non-enzyme models. By combining the sequence information with structural information, Rottig et al. [20] proposed a novel method for enzyme classification. Similarly, by using 3D structural information, Concu et al. [21,22] developed a method by measuring the similarity or deviation for comparison of local structures with template structures. In another study, Izrailev et al. [24] have developed a method to predict enzyme function using protein-ligand interactions from BRENDA [26] database. Since this method is dependent on the protein-ligand interactions data, near complete and more accurate information on such interactions are necessary to achieve higher accuracy in function prediction.

For the distantly related homologous enzymes and for the low sequence identity regions, where homology based annotations are least reliable, a new technique was developed to functionally annotate the enzymes using evolutionarily important residues [27]. Since enzyme reactions are dependent on the structural information of enzymes that catalyze them, comparing similarities among ligands [28] and computing the mechanistic similarity of enzyme reactions based on bond change information [29] has also been explored to classify

enzymes. Classification of enzyme classes based on structural properties is better than the sequence similarity approaches. Nevertheless, the coverage of structure-based methods is low due to sparse nature of structural data in the protein data bank (PDB). Below, we discuss, alternative approaches that are not solely based on sequence or structural similarity.

Functional domain composition

In order to enhance the accuracy of protein classification, it is essential to have an effective representation of protein, which includes as much information a protein has as possible. Reports [4,30–34] show that supplementing functional domain information such as sequence-order-related features, function-order-related features, domains and motifs have improved the prediction accuracy of enzyme classification. By capturing the core features from the Gene Ontology (GO) database such as biological processes and molecular functions of proteins and hybridizing with PseAAC, Chou et al. [30] and Cai et al. [31] developed an enzyme prediction method irrespective of sequence similarity. Similarly, Cai et al. [4,32] and Lu et al. [33] used functional domain composition from interPro and Pfam databases respectively, to predict enzyme subclasses.

In addition to evolutionary information, functional domain information of protein sequence was used to develop a method known as EzyPred [34] with an overall accuracy of 91%. EzyPred predicts whether a given protein sequence is an enzyme or a non-enzyme, and if an enzyme, it also predicts the main and sub-functional class. However in this study, functional classes were treated independently and the inter-class relationships were ignored.

Specific peptides

Sequence motifs are signatures of protein families that have been used as features in enzyme classification [35,36]. Properly chosen motifs expect to represent the key conserved regions of enzyme families and, therefore, reduce the noise that could otherwise result by considering the full-length sequences.

Motif Extraction (MEX) algorithm [37] extracts motifs from protein sequences using unsupervised learning. Based on MEX, Kunik et al. [38] have developed a method to identify and classify enzymes based on Specific Peptides (SPs). The SPs are strings of amino acids, derived from enzyme sequences using MEX and showed that the coverage of the SPs is better than that of PROSITE motifs in finding the function of enzyme families. Further, Weingart et al. [39] have demonstrated how SPs can be employed on Data Mining of Enzymes (DME) on any given set of protein sequences. They use a peptide length of greater than six for the protein sequences that carried the same EC assignment for better accuracy. In another study [40], reactive motifs derived from binding and catalytic sites were used to predict enzyme classes. These motifs combined with the knowledge on their physicochemical properties fared well with PROSITE-based motifs. The prediction accuracy of such methods can be improved as the quality and quantity of annotations on binding and catalytic sites get better.

Methods by Computational Approach

Nearest-neighbor (NN) method

NN predictor has been widely used for enzyme classification [4,31,41,42] and works best when the distributions of the samples are unknown. By coupling AmPseAAC with adaptive fuzzy k-nearest neighbor (AFK-NN) predictor, Huang et al. [41] reported 76.6% prediction accuracy and showed that the method is computationally intensive and hence is time-consuming. Nasibov et al. [42] showed that k-nearest neighbor and minimum distance-based classifiers can be used to classify enzymes according to their AAC by encoding each

enzyme sequence into a 20-Dimensional vector, where each entry represents the frequency of an amino acid. Since molecular functions can also be used to classify enzyme families, Cai et al. [31] represented a protein sequence in a 1930-dimensional vector where each dimension refers to a GO term. Similarly, a protein represented by a 7785-dimensional vector of known domains and motifs from interPro database was used with a nearest neighbor predictor [4]. This study resulted in an overall accuracy of 85% in identifying enzyme family classes.

SVM-based methods

Support vector machines (SVMs) are widely used for classification tasks in bioinformatics. It learns to classify data (protein sequences) by determining a hyperplane using the feature space that maximizes the margin required to separate two classes of data. By projecting a new sequence onto the hyperspace, SVMs could be used to determine whether it is an enzyme or a member of an enzyme class based on its location with respect to the hyperplane.

Cai et al. [43] has developed a SVM method to classify remotely homologous enzymes of different functions using AAC. However, this method works only for known enzymes, and fails to distinguish between enzymes and non-enzymes. In another study, Han et al. [44] showed that SVMs could be used in predicting protein functional families directly from sequence parameters, irrespective of high sequence similarity. By representing each protein sequence as a feature vector assembled from its residue properties such as AAC and physicochemical properties, they achieved a prediction accuracy of 72% for enzymes that have no homologs of known function. However, this method does not demonstrate the capability of assigning distantly related or homologous protein sequences of different functions.

Dobson et al. [18] build binary classifiers using SVMs that discriminate between enzyme classes (two-class models), whereas, Huang et al. [41] observed that a large number of binary SVMs are not effective in dealing with classification of a large number of classes (multi-class models). In contrast to the above study, Lu et al. [33] have numerically represented a protein sequence as a 2657-dimensional feature vector from domain composition of Pfam database to identify and classify the enzyme classes. To improve the efficiency and reduce the input feature space, Cai et al. [46] used discrete wavelet transform (DWT) with SVMs and reported a prediction accuracy of 91.9% that is 9% higher than that of an earlier method [43].

By using information from AAC, low-frequency power spectral density and diversity values of enzymes, Shi et al. [47] developed an SVM method to predict enzyme subclasses. In order to preserve the sequence-order effect, neighbor relationships of the amino acids were used with AAC to develop an SVM-based method [48]. These SVMs are designed for unbalanced classification problems and performed better than standard SVMs in predicting enzyme subfamily classes. Using string kernels, a structured output prediction method [49], where both learning and prediction happens simultaneously is developed as opposed to predicting the membership in enzyme families one at a time [43].

Other machine learning methods

Other noteworthy machine learning approaches used for enzyme classification includes association rule mining and Bayesian classification. Using association rule mining technique, Chiu et al. [50] identified enzyme classes according to the rules associated with protein domain composition. Using physicochemical features with self-organizing neural networks, Sacher et al. [51] and Latino et al. [52] have classified enzymes by analyzing the

similarity of reactions. On the other hand, a Bayesian classifier assumes that each attribute value has an independent effect on each enzyme class [53,54]. Using protein structural properties with Bayesian algorithm, Borro et al. [53] were able to predict the first digit of EC number with 45% accuracy. Similarly, Levy et al. [54] used Bayesian methodologies with sequence similarity to predict protein function and validated the method against ENZYME database. Hung et al. [55] also developed a Bayesian framework for enzyme classification by considering the similarity scores of all relevant proteins, instead of relying on the sequence similarity of a single sequence.

Ensemble methods

The use of ensemble approaches is considered as advancement in the field of machine learning. Ensemble techniques works on a simple principle that a combination of diversified base models strengthens single-classifier based models [56–59]. These methods have been widely used in the area of supervised learning and aims at improving the predictive performance of a given statistical learning or model fitting technique. Instead of depending only on single classifier, Tian et al. [60], Arakaki et al. [61] developed enzyme prediction methods by combining predictions from independent components to infer the enzyme function. Another machine learning approach, decision trees were also used to predict enzyme function, where each family was modeled by a collection of decision trees [62]. Decision trees capture the features that help distinguish between families. Random forests, an ensemble of decision trees, were used in enzyme function classification [63], where the decision about the best predictor was taken by voting amongst the trees, thereby increasing the prediction accuracy of the model.

Another ensemble algorithm, AdaBoost [64] (adaptive boosting) that constructs a strong classifier as a linear combination of several weak classifiers, was used with RBFSVM (SVM with radial basis function kernel) to predict enzyme subfamily class function [65]. The method generates a set of component classifiers and combines them into a single prediction rule. It was also shown to perform better than standard SVM for the unbalanced classification datasets. Similarly, bagging [66] creates many similar training datasets and trains each of the datasets with a new model; the average of all the models' output is the final output for the prediction. Evaluating the prediction of a single model requires less computation than evaluating the prediction of ensemble methods, so ensembles may be thought of as a way to compensate for poor learning algorithms by performing extra computation.

Traditional machine learning methods expect data of fixed length and require protein sequences to be transformed into feature vectors. Unlike similarity-based methods, ML methods do not exploit the sequence data directly; instead, use derived features to obtain the feature space. Hence, there is some loss of information associated with this data transformation. Overall, using extraneous features from protein's functional annotation appears to improve the accuracy than using just the sequence information. A complete summary of methods published on enzyme classification is shown in Table 1 along with scoring features and computational methods used.

Conclusion

Knowledge about enzyme function(s) is extremely beneficial in understanding the holistic cellular function. Several prediction methods identify and classify the enzyme families; however, these methods suffer from many known limitations. Amino acid composition based methods lose sequence order effect, whereas sequence similarity-based methods fail to predict in case of weak or no similarity among protein sequences and needs extraneous annotations to identify their relationships. Methods based on structural features are more

tolerant with weaker sequence identities, but the coverage of enzymes with known structures is sparse in the protein data bank (PDB). Each feature-space and each computational approach work well for a certain set of enzyme families and certain sized datasets. Some machine learning methods works well only for certain feature combinations, because such feature space is dependent on the availability of functional annotations for enzymes in public databases. Given the strengths and limitations of existing methods and the unbalanced and incomplete nature of datasets on enzyme classes, new methods using multiple features and ensemble approaches are more favorable than single-feature-based individual classifiers for accurate prediction of enzyme classes. The growing information both on the functional annotation front and structure determination front will help develop methods for more accurate identification and classification of enzymes.

Acknowledgments

Funding

This work was supported by National Institutes of Health [1R01GM086533-01A1 to CG]; and startup funds to CG from University of Nebraska Medical Center.

References

1. Schmidt S, Sunyaev S, Bork P, Dandekar T. Metabolites: a helping hand for pathway evolution? Trends Biochem Sci. 2003; 28:336–341. [PubMed: 12826406]
2. Shah I, Hunter L. Predicting enzyme function from sequence: a systematic appraisal. Proc Int Conf Intell Syst Mol Biol. 1997; 5:276–283. [PubMed: 9322050]
3. Jeremy, MB.; John, LT.; Lubert, S. Biochemistry. New York: WH Freeman; 1988.
4. Cai YD, Chou KC. Using functional domain composition to predict enzyme family classes. J Proteome Res. 2005; 4:109–111. [PubMed: 15707365]
5. Webb, EC. Enzyme nomenclature. San Diego: Academic Press; 1992.
6. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 2003; 333:863–882. [PubMed: 14568541]
7. Espadaler J, Eswar N, Querol E, Aviles FX, Sali A, et al. Prediction of enzyme function by combining sequence similarity and protein interactions. BMC Bioinformatics. 2008; 9:249. [PubMed: 18505562]
8. Chou KC, Elrod DW. Prediction of enzyme family classes. J Proteome Res. 2003; 2:183–190. [PubMed: 12716132]
9. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics. 2005; 21:10–19. [PubMed: 15308540]
10. Chou KC, Cai YD. Using GO-PseAA predictor to predict enzyme subclass. Biochem Biophys Res Commun. 2004; 325:506–509. [PubMed: 15530421]
11. Altschul SF, Gish W. Local alignment statistics. Methods Enzymol. 1996; 266:460–480. [PubMed: 8743700]
12. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]
13. Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press; 1998.
14. Audit B, Levy ED, Gilks WR, Goldovsky L, Ouzounis CA. CORRIE: enzyme sequence annotation with confidence estimates. BMC Bioinformatics. 2007; 8:S3. [PubMed: 17570146]
15. Rost B. Enzyme function less conserved than anticipated. J Mol Biol. 2002; 318:595–608. [PubMed: 12051862]
16. Otto TD, Guimaraes AC, Degrave WM, de Miranda AB. AnEnPi: identification and annotation of analogous enzymes. BMC Bioinformatics. 2008; 9:544. [PubMed: 19091081]

17. Galperin MY, Walker DR, Koonin EV. Analogous enzymes: independent inventions in enzyme evolution. Genome Res. 1998; 8:779–790. [PubMed: 9724324]
18. Dobson PD, Doig AJ. Predicting enzyme class from protein structure without alignments. J Mol Biol. 2005; 345:187–199. [PubMed: 15567421]
19. Munteanu CR, Gonzalez-Diaz H, Magalhaes AL. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. J Theor Biol. 2008; 254:476–482. [PubMed: 18606172]
20. Rottig M, Rausch C, Kohlbacher O. Combining structure and sequence information allows automated prediction of substrate specificities within enzyme families. PLoS Comput Biol. 2010; 6:e1000636. [PubMed: 20072606]
21. Concu R, Dea-Ayuela MA, Perez-Montoto LG, Bolas-Fernandez F, Prado-Prado FJ, et al. Prediction of enzyme classes from 3D structure: a general model and examples of experimental-theoretic scoring of peptide mass fingerprints of Leishmania proteins. J Proteome Res. 2009; 8:4372–4382. [PubMed: 19603824]
22. Concu R, Dea-Ayuela MA, Perez-Montoto LG, Prado-Prado FJ, Uriarte E, et al. 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in Leishmania parasites. Biochim Biophys Acta. 2009; 1794:1784–1794. [PubMed: 19716935]
23. Kato T, Nagano N. Discriminative structural approaches for enzyme active-site prediction. BMC Bioinformatics. 2011; 12:S49. [PubMed: 21342581]
24. Izrailev S, Farnum MA. Enzyme classification by ligand binding. Proteins. 2004; 57:711–724. [PubMed: 15476211]
25. Bray T, Doig AJ, Warwicker J. Sequence and structural features of enzymes and their active sites by EC class. J Mol Biol. 2009; 386:1423–1436. [PubMed: 19100748]
26. Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res. 2002; 30:47–49. [PubMed: 11752250]
27. Kristensen DM, Ward RM, Lisewski AM, Erdin S, Chen BY, et al. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. BMC Bioinformatics. 2008; 9:17. [PubMed: 18190718]
28. Almonacid DE, Babbitt PC. Toward mechanistic classification of enzyme functions. Curr Opin Chem Biol. 2011; 15:435–442. [PubMed: 21489855]
29. Almonacid DE, Yera ER, Mitchell JB, Babbitt PC. Quantitative comparison of catalytic mechanisms and overall reactions in convergently evolved enzymes: implications for classification of enzyme function. PLoS Comput Biol. 2010; 6:e1000700. [PubMed: 20300652]
30. Chou KC, Cai YD. Predicting enzyme family class in a hybridization space. Protein Sci. 2004; 13:2857–2863. [PubMed: 15498934]
31. Cai YD, Zhou GP, Chou KC. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J Theor Biol. 2005; 234:145–149. [PubMed: 15721043]
32. Cai YD, Chou KC. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J Proteome Res. 2005; 4:967–971. [PubMed: 15952744]
33. Lu L, Qian Z, Cai YD, Li Y. ECS: an automatic enzyme classifier based on functional domain composition. Comput Biol Chem. 2007; 31:226–232. [PubMed: 17500036]
34. Shen HB, Chou KC. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun. 2007; 364:53–59. [PubMed: 17931599]
35. Bork P, Koonin EV. Protein sequence motifs. Curr Opin Struct Biol. 1996; 6:366–376. [PubMed: 8804823]
36. Bairoch A, Bucher P, Hofmann K. The PROSITE database, its status in 1997. Nucleic Acids Res. 1997; 25:217–221. [PubMed: 9016539]
37. Solan Z, Horn D, Ruppin E, Edelman S. Unsupervised learning of natural languages. Proc Natl Acad Sci U S A. 2005; 102:11629–11634. [PubMed: 16087885]
38. Kunik V, Meroz Y, Solan Z, Sandbank B, Weingart U, et al. Functional representation of enzymes by specific peptides. PLoS Comput Biol. 2007; 3:e167. [PubMed: 17722976]

39. Weingart U, Lavi Y, Horn D. Data mining of enzymes using specific peptides. BMC Bioinformatics. 2009; 10:446. [PubMed: 20034383]
40. Liewlom P, Rakthanmanon T, Waiyamai K. Prediction of enzyme class by using reactive motifs generated from binding and catalytic sites. ADMA. 2007; 4637:442–453.
41. Huang WL, Chen HM, Hwang SF, Ho SY. Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method. BioSystems. 2007; 90:405–413. [PubMed: 17140725]
42. Nasibov E, Kandemir-Cavas C. Efficiency analysis of KNN and minimum in enzyme family prediction. Comput Biol Chem. 2009:33. distance-based classifiers 461–464. [PubMed: 18799356]
43. Cai CZ, Han LY, Ji ZL, Chen YZ. Enzyme family classification by support vector machines. Proteins. 2004; 55:66–76. [PubMed: 14997540]
44. Han LY, Cai CZ, Ji ZL, Cao ZW, Cui J, et al. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. Nucleic Acids Res. 2004; 32:6437–6444. [PubMed: 15585667]
45. Zhou XB, Chen C, Li ZC, Zou XY. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol. 2007; 248:546–551. [PubMed: 17628605]
46. Qiu JD, Huang JH, Shi SP, Liang RP. Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. Protein Pept Lett. 2010; 17:715–722. [PubMed: 19961429]
47. Shi R, Hu X. Predicting enzyme subclasses by using support vector machine with composite vectors. Protein Peptide Lett. 2010; 17:599–604.
48. Wang YC, Wang Y, Yang ZX, Deng NY. Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical context. BMC Syst Biol. 2011; 5:S6. [PubMed: 21689481]
49. Astikainen K, Holm L, Pitkanen E, Szedmak S, Rousu J. Towards structured output prediction of enzyme function. BMC Proc. 2008; 2(Suppl 4):S2. [PubMed: 19091049]
50. Chiu SH, Chen CC, Yuan GF, Lin TH. Association algorithm to mine the rules that govern enzyme definition and to classify protein sequences. BMC Bioinformatics. 2006; 7:304. [PubMed: 16776838]
51. Sacher O, Reitz M, Gasteiger J. Investigations of enzyme-catalyzed reactions based on physicochemical descriptors applied to hydrolases. J Chem Inf Model. 2009; 49:1525–1534. [PubMed: 19445497]
52. Latino DA, Zhang QY, Aires-de-Sousa J. Genome-scale classification of metabolic reactions and assignment of EC numbers with self-organizing maps. Bioinformatics. 2008; 24:2236–2244. [PubMed: 18676416]
53. Borro LC, Oliveira SR, Yamagishi ME, Mancini AL, Jardine JG, et al. Predicting enzyme class from protein structure using Bayesian classification. Genet Mol Res. 2006; 5:193–202. [PubMed: 16755510]
54. Levy ED, Ouzounis CA, Gilks WR, Audit B. Probabilistic annotation of protein sequences based on functional classifications. BMC Bioinformatics. 2005; 6:302. [PubMed: 16354297]
55. Hung SS, Wasmuth J, Sanford C, Parkinson J. DETECT--a density estimation tool for enzyme classification and its application to Plasmodium falciparum. Bioinformatics. 2010; 26:1690–1698. [PubMed: 20513663]
56. Chou KC, Shen HB. Hum-PLoc: a novel ensemble classifierfor predicting human protein subcellular localization. Biochem Biophys Res Commun. 2006; 347:150–157. [PubMed: 16808903]
57. Shen HB, Chou KC. Ensemble classifier for protein fold pattern recognition. Bioinformatics. 2006; 22:1717–1722. [PubMed: 16672258]
58. Shen HB, Chou KC. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel. 2007; 20:39–46. [PubMed: 17244638]
59. Nanni L, Lumini A. Ensemblator: An ensemble of classifiers for reliable classification of biological data. Pattern Recog Lett. 2007; 28:622–630.

60. Tian W, Arakaki AK, Skolnick J. EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res. 2004; 32:6226–6239. [PubMed: 15576349]

61. Arakaki AK, Huang Y, Skolnick J. EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. BMC Bioinformatics. 2009; 10:107. [PubMed: 19361344]

62. Syed U, Yona G. Enzyme function prediction with interpretable models. Methods Mol Biol. 2009; 541:373–420. [PubMed: 19381539]

63. Kumar, C.; Li, G.; Choudhary, A. Enzyme function classification using protein sequence features and random forest. ICBBE; 2009.

64. Li X, Wang L, Sung E. AdaBoost with SVM-based component classifiers. Engineering Applications of Artificial Intelligence. 2008; 21:785–795.

65. Wang YC, Wang XB, Yang ZX, Deng NY. Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. Protein Pept Lett. 2010; 17:1441–1449. [PubMed: 20666729]

66. Breiman L. Bagging predictors. Machine Learning. 1996; 24:123–140.

Table 1

Enzyme classification work grouped by methods used.

Reference	Methods (URL)	Features
Chou et al. [8]	CDA	AAC
Chou [9]	CDA	AmPseAAC
Shah et al. [2]	BLAST, FASTA	Sequence information
Audit et al. [14]	BLAST	Sequence information
Tian et al. [6]	PSI-BLAST	Sequence and function information
Espadaler et al. [7]	PSI-BLAST and BLAST	Sequence information and protein interactions
Otto et al. [16]	BLASTp and HMMer	Sequence information
Galperin et al. [17]	PSI-BLAST	Sequence information
Chou et al. [10]	Statistical analysis	PseAAC
Bray et al. [25]	Statistical analysis	Structural and sequence properties
Munteanu et al. [19]	Statistical analysis	Structural properties and AAC
Cai et al. [31]	NN	Domain composition (GO) and PseAAC
Cai et al. [4]	NN	Domain composition (interPro)
EzyPred [34]	NN(http://www.csbio.sjtu.edu.cn/bioinf/EzyPred/)	Domain composition (Pfam) and Evolutionary information
Nasibov et al. [42]	K-NN	AAC
Huang et al. [41]	AFK-NN	AmPseAAC
Chou et al. [30]	ISort	Domain composition (GO) and AmPseAAC
Cai et al. [32]	ISort	Domain composition (interPro) and PseAAC
Borro et al. [53]	Bayesian	Structural information
Levy et al. [54]	Bayesian	Sequence information
Detect [55]	Bayesian	Sequence information
Latino et al. [52]	SOM	Physicochemical and topology descriptors
SVMProt [43,44]	SVM (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi)	AAC
ECS [33]	SVM (http://pcal.biosino.org/enzyme_classification.html)	Domain composition (Pfam)
Dobson et al. [18]	SVM	Structural properties
ASC [20]	SVM (http://asc.informatik.uni-tuebingen.de)	Structural and sequence properties
SPSearch [38]	SVM (http://adios.tau.ac.il/SPSearch/)	Specific peptides (MEX[38])
Qiu et al. [46]	SVM	PseAAC
Wang et al. [48]	SVM	AAC and Neighbor relationships
Shi et al. [47]	SVM	PseAAC
Almonacid et al. [28]	Tanimoto coefficients and similarity search	Structural information

Reference	Methods (URL)	Features
Almonacid et al. [29]	Tanimoto coefficients and similarity search	Structural information
Kristensen et al. [27]	Structure template matching	Structural information (Evolutionary)
Kato et al. [23]	Structure template matching	Structural properties
Concu et al. [21]	LDA and ANN	Structural information
Concu et al. [22]	LDA and ANN (http://miaja.tic.udc.es/Bio-AIMS/EnzClassPred.php)	Structural information
Astikainen et al. [49]	HM3 algorithm	String kernels
DME [39]	MEX [38] (http://adios.tau.ac.il/DME/)	Specific peptides (MEX[38])
Liewlom et al. [40]	Mutation control	Reactive motifs
Izrailev et al. [24]	Nearest neighbor distance	Ligand interactions
EFICAz [60]	Ensemble	Sequence similarity, Pfam and Prosite patterns
EFICAz^2 [61]	Ensemble (http:/cssb.biology.gatech.edu/skolnick/webservice/EFICAz2/index.html)	Sequence similarity, Pfam and Prosite patterns
Umar et al. [62]	Ensemble	Sequence and structure information
Kumar et al. [63]	Ensemble	Sequence information
Wang et al. [65]	Ensemble	AAC and Neighbor relationships

AAC: Amino Acid Composition

AFK-NN: Adaptive fuzzy k-NN

AmPseAAC: Amphiphilic Pseudo Amino Acid Composition

ANN: Artificial Neural Network

CDA: Covariant Discriminant Algorithm

Hierarchical Max-Margin Markov algorithm

ISort: Intimate Sort predictor

LDA: Linear Discriminant Analysis

NN: Nearest Neighbor predictor

MEX [38]: Motif Extraction algorithm

SVM: Support Vector Machine